PALOMA HELENA DA SILVA LIBORIO

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA A FIXAÇÃO BIOLÓGICA DE NITROGÊNIO E TEOR DE PROTEÍNA EM SOJA (*Glycine max*)

Londrina
2023

PALOMA HELENA DA SILVA LIBORIO

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA A FIXAÇÃO BIOLÓGICA DE NITROGÊNIO E TEOR DE PROTEÍNA EM SOJA (*Glycine max*)

Tese apresentada à Universidade Estadual de Londrina - UEL, como requisito parcial para a obtenção do título de Doutora em Agronomia.

Orientador: Prof. Dr. Marco Antonio Nogueira

Londrina
2023

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Da Silva Liborio, Paloma Helena.
Estudo de associação genômica ampla (GWAS) para a fixação biológica de nitrogênio e teor de proteína em soja (Glycine max) / Paloma Helena Da Silva Liborio. - Londrina, 2023.
72 f. : il.

Orientador: Marco Antonio Nogueira.
Tese (Doutorado em Agronomia) - Universidade Estadual de Londrina, Centro de Ciências Agrárias, Programa de Pós-Graduação em Agronomia, 2023.
Inclui bibliografia.

1. SNP - Tese. 2. genes candidatos - Tese. 3. coinoculação - Tese. 4. qualidade de grãos - Tese. I. Nogueira, Marco Antonio. II. Universidade Estadual de Londrina. Centro de Ciências Agrárias. Programa de Pós-Graduação em Agronomia. III. Título.

CDU 63

PALOMA HELENA DA SILVA LIBORIO

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA A FIXAÇÃO BIOLÓGICA DE NITROGÊNIO E TEOR DE PROTEÍNA EM SOJA (*Glycine max*)

Tese apresentada à Universidade Estadual de Londrina - UEL, como requisito parcial para a obtenção do título de Doutora em Agronomia.

Banca Examinadora

Dr. Juliano Tadeu Vilela de Resende
Universidade Estadual de Londrina - UEL

Dra. Paula Cerezini
Syngenta

Dra. Francismar Marcelino-Guimarães
Embrapa Soja

Dr. Antonio Eduardo Pipolo
Embrapa Soja

Prof. Dr. Marco Antonio Nogueira
Embrapa Soja

Londrina, 30 de agosto de 2023.

AGRADECIMENTOS

A Deus pela vida e pelos anjos que cruzaram meus caminhos em diferentes situações.

À minha família, por ser meu porto seguro.

Ao Pedro pela resiliência, companheirismo e incentivo.

Ao Dr. Marco Antonio Nogueira, pelo ser humano incrível que é! Foi um privilégio receber sua orientação e conhecimentos, sem falar da parceria nas inúmeras colheitas de soja.

À Dra. Mariangela Hungria por ser fonte de inspiração, exemplo de força feminina e superação.

À Dra. Francismar Marcelino-Guimarães pelos ensinamentos, conversas, paciência e toda ajuda com a genotipagem.

Ao Dr. Antonio Eduardo Pípolo pelos ensinamentos, conversas e dicas valiosas.

À Dra. Ivani de Oliveira Negrão Lopes pela ajuda nas análises estatísticas, compreensão e carinho.

Aos funcionários do Laboratório de Biotecnologia do Solo: Ligia, Eduara, Renan, Natalice e Rinaldo, por serem simplesmente os melhores!

À toda equipe técnica da Embrapa Soja que me auxiliou nos experimentos.

Ao Dr. François Belzile por me acolher em seu laboratório e me possibilitar navegar pelo universo da bioinformática sob sua orientação e cuidados.

Aos amigos do laboratório de Biotecnologia do Solo e da UEL que fiz durante essa caminhada, pela amizade e por vezes me ajudarem na condução do trabalho.

Aos amigos e funcionários do *Département de Phytologie* – IBIS da Université Laval que estiveram ao meu lado no estágio no Canadá, especialmente ao Vicent-Thomas por ser meu mentor nas análises de bioinformática.

À UEL e Embrapa Soja pela estrutura que possibilitou minha formação e a condução da minha pesquisa.

À Capes pela concessão da bolsa de estudos no Brasil.

Ao Mitacs, Université Laval e Prof. François Belzile pela concessão da bolsa de estudos no Canadá.

LIBORIO, Paloma Helena da Silva. **Estudo de associação genômica ampla (GWAS) para a fixação biológica de nitrogênio e teor de proteína em soja (*Glycine max*)**. 2023. 72 folhas. Tese (Programa de Pós-graduação em Agronomia) – Universidade Estadual de Londrina, Londrina, 2023.

RESUMO

A compreensão dos mecanismos genéticos envolvidos na fixação biológica de nitrogênio viabiliza o desenvolvimento de cultivares de soja com maior eficiência simbiótica e capacidade de fixação biológica de nitrogênio. Além disso, a identificação de genes associados ao teor de proteína possibilita o desenvolvimento de cultivares com maior concentração de proteína nos grãos, atendendo à demanda da indústria de ração animal, principal destino da produção mundial de soja. Este estudo de associação genômica ampla objetivou investigar a FBN e o teor de proteína em soja por meio de uma análise abrangente do genoma de 100 acessos de soja. Os resultados destacaram a presença de SNPs significativos e genes candidatos relacionados à fixação biológica de nitrogênio e ao teor de proteína nos grãos. As regiões genômicas têm implicações práticas para a melhoria da performance simbiótica com incremento da qualidade dos grãos de soja. Espera-se que essas descobertas contribuam para o desenvolvimento de cultivares de soja mais produtivas, sustentáveis e nutritivas, fortalecendo a agricultura e promovendo a segurança alimentar.

Palavras-chave: coinoculação, genes candidatos, SNP, simbiose, qualidade de grãos.

LIBORIO, Paloma Helena Da Silva. **Genome-wide association study (GWAS) for biological nitrogen fixation and protein content in soybean (*Glycine max*).** 2023. 72 sheets. Thesis (Post Graduation in Agronomy) – State University of Londrina, Londrina, 2023.

ABSTRACT

Understanding the genetic mechanisms involved in biological nitrogen fixation enables the development of soybean cultivars with greater symbiotic efficiency and biological nitrogen fixation capacity. Furthermore, identifying genes associated with protein content enables the development of cultivars with a higher protein concentration in the grains, meeting the demand of the animal feed industry, the main destination for global soybean production. This genome-wide association study aimed to investigate BNF and protein content in soybeans through a comprehensive genome analysis of 100 soybean accessions. The results highlighted the presence of significant SNPs and candidate genes related to biological nitrogen fixation and protein content in grains. Genomic regions have practical implications for improving symbiotic performance and increasing the quality of soybeans. These discoveries are expected to contribute to the development of more productive, sustainable, and nutritious soybean cultivars, strengthening agriculture and promoting food security.

Key-words: co-inoculation, candidate genes, SNP, symbiosis, grain quality.

SUMÁRIO

INTRODUÇÃO GERAL

A soja é uma excelente fonte de proteína vegetal, utilizada na alimentação humana e animal. Os altos teores de proteína e a excelente combinação de aminoácidos, custo de produção baixo e disponibilidade, fazem da soja a principal fonte proteica na indústria de ração animal.

Os teores de proteína da soja apresentam tendência de queda, devido principalmente a maior preocupação com a produtividade de grãos, que apresenta correlação negativa com teor de proteína. A redução nos níveis de proteína dos grãos tem impactado a indústria de farelo de soja, reduzindo o lucro e a competitividade da soja.

Os estudos de Associação Genômica Ampla (GWAS) têm se mostrado ferramentas poderosas para a identificação de regiões genômicas envolvidas em características complexas, como a fixação biológica de nitrogênio (FBN) e o teor de proteína nos grãos de soja. Essas características desempenham um papel crucial tanto no aumento da produtividade, como na competitividade e qualidade da soja.

A FBN é um processo fundamental para a soja, permitindo que a planta converta o nitrogênio atmosférico em compostos nitrogenados essenciais para o desenvolvimento, sem a necessidade de adição de fertilizantes minerais nitrogenados. A simbiose eficiente com bactérias fixadoras de nitrogênio do gênero *Bradyrhizobium* é altamente desejável em soja, impactando positivamente a redução de custos de produção e a promoção da sustentabilidade agrícola.

Desta forma, compreender a base genética da FBN e do teor de proteína é essencial para o melhoramento genético da cultura visando maior eficiência produtiva e qualidade. A identificação de regiões genômicas associadas a essas características por meio de estudos de GWAS pode fornecer informações valiosas para o desenvolvimento de cultivares de soja com maior eficiência na FBN e maior teor de proteína.

Além disso, estudos de GWAS podem revelar informações sobre os mecanismos moleculares relacionados à FBN e à síntese proteica nos grãos por meio da identificação de genes candidatos. Essas informações podem orientar o desenvolvimento de estratégias de melhoramento genético mais precisas, como a seleção assistida por marcadores (SAM), que permite a rápida identificação de

genótipos com as características desejadas, conferindo agilidade aos programas de melhoramento genético.

Nessa perspectiva, os estudos de GWAS são estratégicos para ampliar o entendimento dessas características, fornecendo recursos para lidar com os desafios da produção de soja para os próximos anos. Tais desafios devem focar na obtenção de cultivares que sejam mais resilientes às mudanças climáticas, produtivas, de elevado valor nutricional e alinhadas com as práticas de sustentabilidade.

Capítulo 1

Title

Current status and future perspectives of genetic mapping of QTLs and genes involved in BNF and protein content in soybean

Journal

Current Biotechnology (B1)

Authors

Paloma Helena da Silva Liborio[a], Mariangela Hungria[b], Antonio Eduardo Pípolo[b], Francismar Marcelino-Guimarães[b], François Belzile[c], Marco Antonio Nogueira[b*]

Affiliations

[a]Universidade Estadual de Londrina, PO Box 10.011, 86057-970, Londrina, PR, Brazil.
[b]Embrapa Soja, PO Box 231, 86001-970, Londrina, PR, Brazil.
[c]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada

E-mail addresses

Liborio, P.H.S. paloma_liborio@hotmail.com
Hungria M. mariangela.hungria@embrapa.br
Pípolo, A.E. antonio.pipolo@embrapa.br
Marcelino-Guimarães, Francismar
Belzile, François francois.belzile@fsaa.ulaval.ca
Nogueira, M.A. marco.nogueira@embrapa.br

Correspondence

*Marco Antonio Nogueira
Embrapa Soja, PO Box 4006, 86085-981, Londrina, PR, Brazil. Phone: +55 43 3371-6215.
E-mail: marco.nogueira@embrapa.br

Current status and future perspectives of genetic mapping of QTLs and genes involved in BNF and protein content in soybean

Abstract
Due to the high protein content in their grains, soybean [*Glycine max* (L.) Merr.] has a high demand for nitrogen (N). However, plant responses to mineral-N supply are inconsistent, besides environmentally and economically costly. Therefore, the most rational way of providing N to soybeans is via inoculation with N_2-fixing bacteria. Over decades of breeding, grain yield has significantly increased but, conversely, the content of protein in grains has decreased. It has been hypothesized that more efficient N_2-fixing soybean genotypes may produce grains with higher protein content. Regarding this, we focused on research papers dealing with mapping and associative mapping studies on biological N_2 fixation and traits related to protein content in grains. Furthermore, we bring insights into factors affecting these traits and the relationship between protein content and biological N_2 fixation. Some studies report that these traits appear on the same chromosomes, however, new studies must be conducted to verify the interactions between them. Despite the challenges inherent to complex traits, the integration of technologies could smarten data collection for mapping studies. Finally, this review may stimulate reflections on pending questions regarding the genetic breeding for these important traits involved in soybean yield and the nutritional quality of their grains.

Key-words: BNF. Grain quality. GWAS. Relevance traits. Symbiosis. QTL.

Introduction

Soybean [*Glycine max* (L.) Merr.] has a prominent position in the global economy, with Brazil and the United States as the first and second largest producers, respectively. In the 2022/23 crop season, Brazil produced 152 million tons, with an average yield of 3,527 kg ha^{-1} (CONAB, 2023), whereas the United States produced 116 million tons, with an average yield of 3,480 kg ha^{-1} (USDA, 2023).

Due to the high protein content in their grains, soybeans have a high demand for nitrogen (N), which in Brazil is mostly supplied via biological nitrogen fixation (BNF) by inoculation with highly efficient selected strains (Santos et al., 2019), which makes the Brazilian soybeans production independent of nitrogen fertilizers (Hungria; Nogueira; Araújo, 2013), with economic savings above 15 US$ billion and avoided emissions over 180 million tons of eq.-CO_2 in the 2019-2020 crop season (Telles et al., 2023).

Soybean provides approximately 70% of the protein and 30% of the edible oil worldwide (Lam et al., 2010). Soybean meal is the main by-product derived from grain crushing for oil extraction, but its contents of protein must comply with contractual specifications to meet minimal requirements. However, the industry has found difficulties to meet the expected levels of protein due to the drop in protein levels in the raw material in the recent years (Pípolo et al., 2015; Lorini, 2018). Therefore, soybean must keep or increase their levels of protein in grains to remain a competitive protein source for the feed industry, without yield penalties.

Narrow genetic diversity, in addition to global climate changes, may partly explain the decrease in protein and increase in oil contents in soybean grains (Patil et al., 2017; Zhang et al., 2018; Wang et al., 2021). In the 1970s, cultivars produced in Brazil had protein contents in grains around 40% (Costa et al., 1981), but currently the average content dropped to 37.6% (Lorini, 2018).

Protein in grains is considered the most valuable component of soybean. However, scientific knowledge needs to advance on ways to increase the concentration of this important trait in new genotypes (Patil et al., 2017). Moreover, traits related to soybean grain composition involve major challenges due to polygenic inheritance, high environmental influence, and complex metabolic interactions during the plant development (Gupta et al., 2022) that may affect the final content of protein in grains.

Quantitative Trait Loci (QTLs) related to protein content in grains were found physically close to QTLs

associated with BNF traits (Torres et al., 2015), which increases the chances of these haplotypes being carried together due to the linkage disequilibrium. However, despite the environmental and economic importance of BNF for soybean, there are few studies on QTLs related to this trait, when compared with other traits of agronomic importance, such as grain yield and resistance to diseases.

BNF contributions to soybean and relationship with grain protein content

Some legumes have developed adaptive strategies along the evolutive process to acquire N for growth and development. Soybean is the most successful case of N acquisition based mainly on BNF, which consists of reducing the atmospheric N_2 into ammonia/ammonium that is assimilated by plants. The establishment of the symbioses involves initial attraction of the bacteria by isoflavonoids and betaines released in root exudates produced by the host plant (Gerahty et al., 1992). The infection process begins with the exchange of molecular signals released by the surrounding rhizobia that activate the nodulation genes in the host plant (Vargas; Hungria, 1997). However, BNF is a complex process that involves the interaction between microorganisms, plant genotypes, and environment.

The capacity to acquire N from BNF makes N-fixing legumes important components of production systems, promoting the sustainability of grain yield and improving the soil fertility (Kakraliya et al., 2018), since part of the fixed N remains in the plant residues and can be used for the next crop in rotation or succession systems, especially cereals, like in Latin America countries in which maize (*Zea mays* L.) or wheat (*Triticum aestivum*) is grown thereafter. However, several factors that are also limiting to the plant development, such as water stress (Cerezini et al. 2020), soil compaction (Siczek; Lipiec, 2011), deficiency of P and the micronutrients Co and Mo (Hungria; Campo; Mendes, 2001) may also impair the soybean capacity to perform BNF.

After photosynthesis, BNF is considered the most important biological process for plants (Graham; Vance, 2003) and contributes annually for 15% of the world's N demand (FAOSTAT, 2021). A soybean yield of 4,000 kg ha^{-1} with protein content of 38% requires 243.2 kg ha^{-1} of N for grain formation. If this N had to be supplied via N-fertilizers, the economic and environmental costs would be huge, considering that the plant responses to mineral N are inconsistent, becoming the process inefficient (Hungria et al., 2007; Saturno et al., 2017; Telles et al., 2023). However, under soybean conditions in Brazil the application of fertilizer N is not necessary, with a maximum of 20 kg ha^{-1} of N being tolerable if the P and K sources of the formulated fertilizers bring some mineral N content. Higher doses of N result in decreases in the nodulation and, consequently, on the BNF process, without any benefit on grain yield (Hungria et al., 1997).

Thus, the non-use of mineral N-fertilizers promote a sustainable and easy-access technology, given the low cost of bacterial inoculants, contributing for the success of soybean crop in Brazil, that stands out worldwide in the research, production and use of high-quality inoculants (Santos et al., 2019). The country has an annual adoption rate of approximately 80% of the soybean area (ANPII 2018). Recently, the combination of inoculants containing elite strains of *Bradyrhizobium* spp. and *Azospirillum brasilense* in co-inoculation have resulted in grain yield increase by 16% (Hungria; Nogueira; Araújo, 2013), with positive effects on several aspects like increased tolerance to drought (Cerezini et al., 2016) and production of seeds with higher physiological quality (Liborio et al., 2020); this successful technique has been applied in about 30% of the soybean area in the country (ANPII 2018).

At the initial developmental stages, before the establishment of the symbiosis, the soybean uses the N

from cotyledonal reserves and mineral N from the mineralization of soil organic matter and organic residues of previous crops. After the establishment of the symbiosis with *Bradyrhizobium*, the BNF process is the main source of N supply (Fabre; Planchon; 2000; Hungria et al., 2007), which may reach over 80% of the soybean required N (Herridge et al., 2008). Increasing yields and at the same time keeping or increasing the protein levels in soybean grains are challenges for breeding programs (Joaquim et al., 2022). A possible way can be by N supply via BNF, which N is more effectively converted into proteins in grains than the N coming from mineral fertilizers (Israel et al., 1985). N concentration was estimated across cultivars, treatments, and sampling dates as the ratio between total N uptake (kg ha^{-1}) and shoot biomass (kg ha^{-1}). The dilution curves of mineral N obtained from the soil showed a typical dilution with the increase of shoot biomass, while the N derived from BNF followed even concentration during the soybean development. This behavior suggests a continuous flow of N from the BNF along the plant growth cycle (Santachiara et al., 2018). Ferreira et al. (2016) verified that the supply of mineral N did not increase yield, protein, or oil concentrations in soybean grains. Unlike, BNF provided N more efficiently to the reproductive tissues, compared with plants supplied with mineral N (Warembourg; Fernandes, 1985; Zapata et al., 1987). Fabre and Planchon (2000) evaluated inbred soybean lines for BNF and capacity of nitrate assimilation from the soil. The authors concluded that the efficiency of BNF during the stages R2-R6 + 10 days is related to higher protein content in grains. Warembourg and Fernandes (1985) demonstrated that biologically fixed N was more mobile in the plant than mineral N. Increase in biologically fixed N provides a better supply of N to the grains, independently of the mineral N available in the soil. Mobilization of N from vegetative tissues to the reproductive ones was more effective in N_2 fixing than in mineral N-fertilized plants (Israel et al., 1985).

The maximum nitrogenase activity and BNF rates have been generally reported to occur at the initial reproductive stages and decreases as the plant advance in its cycle due to competition for photosynthates with the forming reproductive parts (Lawn; Brun, 1974). During the pod formation, there is a reduction in the N concentration of the aerial part due to the remobilization of N to grains (Imsande; Schmidt, 1998). Around 50% of the leaf N is in the form of the enzyme ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco), which has a strong relationship between N per unit of leaf area and photosynthetic rate (Sinclair, 2004). Under normal conditions of N supply, this remobilization of N from rubisco is not harmful to the photosynthetic process, since it is estimated that only about 10% of the photosynthetic potential of the plant is used (Sinclair; Rufty; Lewis, 2019). Under limiting conditions of N-supply, soybean will remobilize N from leaves to grains, which may reduce the photosynthetic and further compromise the yield (Salvagiotti et al., 2008). However, there is a self-regulation mechanism between the demand for carbohydrates and the photosynthetic rate, so increasing the drain force towards grains increases the photosynthetic rate, which is stimulated by BNF (Kaschuck et al., 2010). Until the end of the soybean cycle, the grains are the main sink of carbohydrates and the reduction of the drain of nutrients in this phase reduces the photosynthetic rate and accelerates leaf senescence (Haq; Mallarino, 2000).

The main product of BNF in soybean is ammonia (NH_3), which rapidly reacts with protons in the cytosol and forms ammonium (NH_4^+). Ureides (allantoin and allantoic acid) are efficient forms of N transport after their assimilation into glutamine and conversion to ureides. Thus, ureides are the primary forms of N exported from nodules to shoots via xylem (Mcclure; Israel, 1979; Atkins; Smith, 2007) and require the function of ureide permeases (UPS1) for transport them out of the N-fixing cells to the xylem (Pélissier et al. al., 2004; Collier; Tegeder, 2012). Glutamine synthetase (GS) activity is low in bacteroides, and in general, the infected plant cell performs the assimilation. GS is an enzyme that plays an essential role in N metabolism by catalyzing the

condensation of glutamate and ammonia to form glutamine (Hungria; Nogueira, 2022).

A higher rate of ureide exportation from nodules increases the BNF efficiency, increasing plant nutrition and grain yield. Using a model based on the overexpression of the UPS1 transporter, several steps involved in the regulation of metabolic events and N transport were verified, whose understanding can support new strategies for optimizing the BNF process (Carter; Tegeder, 2016; Joaquim et al., 2022).

Ureide metabolism, especially allantoin, may play a central role as regulator that promotes whole-plant adjustments for C and N acquisition, assimilation, and transport. This complex mechanism could lead, at least in UPS1-overexpressin plants, to increased growth and grain yield (Thu et al., 2020).The overexpression of the VfAAP1 gene in peas (*Pisum sativum*) and lima beans (*Vicia faba*) increased the assimilation of amino acids, providing higher protein content in larger grains (Rolletschek et al., 2005).

The bacteria × legume symbiosis is a highly specific interaction, so that certain genotypes match an efficient symbiosis only with a specific set of strains (Wang et al., 2017). The genetic control of symbiotic specificity is complex, involving a wide range of host plant and bacterial genes with diverse ways of action. The success of the BNF process depends on the host plant ability to selectively interact with the most mutualistic bacterial partners. In this way, it is of great importance to improve the capacity of the host plant to interact with beneficial microorganisms like effective N-fixing strains, so that it can choose and cooperate with the best partners.

The same bacterial strain can establish highly effective symbiosis with a given host and a poor interaction with another (Liu et al., 2020). Low efficiency is often related to genetic incompatibility due to a lack of co-adaptation between the symbiotic partners, which negatively affects the bacterial differentiation and survival within the nodule cells. It is worth reflecting on how great progress has been made in the selection of superior strains of bacteria, however, this has not occurred in the same proportion for selecting soybean genotypes for a more effective symbiosis (Qin et al., 2012). Research to obtain more effective symbiosis for N supply is an economic, environmentally sustainable, and promising strategy.

In studies comparing wild species and cultivars of soybean, pea and alfalfa, older cultivars performed better than new cultivars in terms of BNF rate, suggesting that the natural ability to distinguish the best symbiotic partners may have been impaired during bottlenecks in the domestication and/or breeding process (Kiers et al., 2007). Conversely, superior performance in new soybeans cultivars compared with wild soybeans for total nodule fresh mass, number of nodules, total nitrogen, and total ureide accumulation was verified. Thus, the authors suggesting that these traits were specially selected during the domestication and breeding process. (Muñoz et al., 2016).

Composition of soybean grains

The composition of grains has a decisive effect on the quality and uses of soybean-based products (Zhang et al., 2018). In recent years, soybean has achieved more space in the creation of products for several uses as it is rich in proteins, lipids, polyunsaturated fatty acids, and carbohydrates with prebiotic activity, in addition to soluble and insoluble fibers (Lorini, 2018).

Soybean grains contains the 10 essential amino acids, although the sulfurated amino acids (methionine and cysteine) are found in low levels (Canto; Turatti, 1989). The oil is rich in unsaturated fatty acids, consisting of palmitic acid (12%), stearic acid (4%), oleic acid (23%), linoleic acid (53%), and linolenic acid (8%) (Zhang et al., 2018). The carbohydrate composition of soybeans mainly consists of fiber, whereas the main non-fiber

carbohydrates are oligosaccharides like sucrose and glucose which are essential for cell functions (Zheng, 2009), and raffinoses (O'keefe; Bianchi; Sharman, 2015). Proteins can be classified as metabolic and storage, where storage proteins act in the supply of N and C for initial development. The metabolic proteins include structural proteins and enzymes involved in the synthesis of oil and proteins during the grain development. Based on their solubility, globulins and albumins are the main components of storage proteins, with globulin accounting for 70% of the total (Mandal; Mandal, 2000). Protein and oil concentrations in soybean grains are quantitative traits determined by the interaction among several genes with small or moderate genetic effects, as well as their interaction with the environment (Hwang et al., 2014).

During the formation of grains, storage proteins remain as protein bodies and are synthesized in the ribosomes (Smith, 1984). At this stage, C and N compounds are broken down into fractions that make up proteins, oils, and carbohydrates. In the first two weeks after pollination, an intense cell division takes place in the cotyledons and the cell size increases due to accumulation of protein and oil, which have the function of reserve stocks. Increasing the storage protein content by improving the ratio of glycinin to β-conglycinin is of great importance for soybean breeding (Panthee et al., 2006). Glycinins influence the amino acid profile, enhancing soybean value as a protein source. Meanwhile, b-conglycinins affecting properties like water absorption. The synthesis of protein is energetically more costly for the plant than is commonly assumed (Chung et al., 2003). The increase in soybean protein content decreases protein quality due to the reduction in the proportion of sulfurated amino acids, which content is already low in soybean seeds (Paek et al. 1997). Thus, understanding the genetic variability of these proteins is crucial because they are key storage proteins in soybean seeds.

Decrease of protein content in soybean grains

Monogastric animals cannot synthesize essential amino acids and need to obtain them from their diet (Patil et al., 2017). Formulations used in animal feed have been demanded for higher concentration of protein, to meet the feed quality to attend the demands of a more intensified livestock, with shorter life cycle of animals and higher quality of their products.

In this sense, soybean meal is one of the main, abundant, and cheaper, protein sources for cattle, pigs, fishes, and poultry around the world. When choosing the protein source, the feed protein industry has taken consider about two factors: cost and protein content of the raw material. Soybean meal is usually the most competitive protein source for feed in Brazil and in the external market. However, the reduction of protein content in grains of new, more yielding cultivars, is concerning and may be related to the widespread use of few parental lines to develop new varieties (Bonato et al., 2000; Zhang et al., 2018). Due to a negative correlation between protein and oil content and in some cases protein and yield, obtaining cultivars with higher protein content via conventional breeding does not result in expressive genetic gains and due to the speed, that current agriculture demands (Wilcox; Cavins, 1995; Panthee et al., 2006; Kambhampati et al., 2020; Joaquim et al., 2022).

Several factors may influence the protein levels in the soybean grains, such as the genotype, environment, the symbiosis with N_2-fixing bacteria, and their interactions (Vollmann et al., 2000; Fehr et al., 2003; Bueno et al., 2013). Soybean cultivars have relatively narrow genetic basis, which limits the genetic breeding processes (Hyten et al., 2006). Changes in the supply of C and N affects the chemical composition of the soybean grain and may be one mechanism involved in the variation of protein and oil contents (Hayati et al., 1996). Albrecht et al. (2008) evaluated variations in oil and protein contents in grains of soybean sown on different dates and observed that the

anticipation of the sowing date caused reduction in the protein contents in grains. Genetic variability has also been observed among soybean genotypes regarding protein and oil contents in grains, but different levels of inputs did not affect these variables (Carrão-Panizzi et al., 2021).

Soybean grown in the southern United States had higher protein content in grains, which was attributed to higher temperatures during the grain filling stage, in addition to greater genetic potential of those cultivars (Piper; Boote, 1999). In addition to climatic factors, the genetic basis of the breeding lines used in the development of recent cultivars can be also a cause of lower protein contents in grains, given that priority was given to the selection of more yielding parents in detriment of higher protein content in grains.

Mahmoud et al. (2006) comparing protein profiles of the soybean cultivars indicated that relative expression of most of the seed storage proteins had not varied substantially from the ancestral lines to commercial cultivars. These results suggests that the process of selecting and breeding for higher yield did not significantly impact the protein composition. This phenomenon appears to be attributed to the constrained genetic diversity among the initial parental lines.

Most of the Brazilian soybean cultivars derive from only 60 common ancestors, mainly CNS, S-100, Nanking, and Tokyo genotypes, which contribute for 55.26% of the genetic base of a set of 444 Brazilian cultivars registered up to 2009 (Wysmierski, 2013). The soybean cultivars Provar and Protana, launched in the United States in the 1960s, had an average protein content 3.5% higher than Amsoy and Corsoy, which became market leaders due to their higher yield potential (Wilcox, 1989). Bonato et al. (2000) reported that soybean cultivars released in the southern Brazil in the late 1990s already had lower protein content in their grains compared with predecessor cultivars released before 1980.

The process of obtaining more yielding soybean cultivars with superior grain composition is hampered by the inverse relationship between yield and protein content (Thorne; Fehr, 1970; Cober; Voldeng, 2000), and protein content and oil content (Diers et al., 1992; Sediyama et al.,1993; Chung et al., 2003; Patil et al., 2017; Kumar et al., 2021). A negative correlation between protein and oil contents indicates that the biosynthetic pathways for both traits share similar features, in which an increase in one lead to a decrease in the other (Kumar et al., 2021). This correlation also indicates that a decrease of 1% in oil content results in a 2% of protein increase in grains (Hymowitz et al., 1972).

Higher grain yields with satisfactory protein and oil contents are one of the major challenges in soybean breeding due to pleiotropy (Joaquim et al., 2022, Diers et al., 1992; Sebolt et al., 2000; Patil et al., 2017). Pleiotropic genes are those that influence multiple traits simultaneously, whereas non-pleiotropic genes have more specific effects on individual traits. This duality in genetic regulation plays a crucial role in shaping the complex phenotype of soybean. Nonetheless, the presence of non-pleiotropic genes should not be overlooked. These genes often have more focused functions, affecting protein content without causing extensive changes in other traits.

On the other hand, disequilibrium linkage (LD) pertains to the non-random association of alleles at different loci within a population. Genes that are physically close on the same chromosome are more likely to be inherited together, which can lead to the preservation of specific combinations of alleles across generations. Therefore, any factor that alters allele frequencies interferes LD (Gupta et al., 2005). In soybean, this phenomenon can influence the co-inheritance of alleles responsible for protein synthesis with those affecting other traits, resulting in consistent genetic associations.

Therefore, pleiotropy and LD are two complementary genetic mechanisms influencing soybean protein

genes. Pleiotropy can result in correlated changes in protein content and other traits, while disequilibrium linkage contributes to the co-inheritance of alleles responsible for protein synthesis with those affecting other traits. This understanding holds considerable significance for soybean breeding efforts aimed at optimizing protein content.

Several genomic tools are available nowadays to identify traits related to soybean grain composition (Bandillo et al. 2015; Phansak et al. 2016, Gupta and Manjaya, 2022). Protein and oil contents in 20,395 accessions from GRIN USDA show a range from 31.7% to 57.9% for protein and 6.5% to 25.6% for oil (https://www.soybase.org/grindata), evidencing the existence of genetic variability for use in breeding programs aiming at increasing protein and oil in soybean grains. The main QTL for protein is located on chromosome 20 and shows distinct haplotypes between populations that could assist breeders (Phansak et al. 2016). Introgression of wild soybean genes like *Glycine soja* Siebold & Zucc. has also been used to incorporate genetic diversity into new soybean varieties (Torkamaneh et al., 2020).

Zapata et al. (1987), evaluating BNF at different stages of development of soybean, concluded that the N derived from fixation assimilated between R3 and R7 stages was the predominant source of N for pod filling. In a study carried out by Torres et al. (2015), differences were found among soybean cultivars in terms of nodulation capacity and protein content in grains. Zimmer et al. (2016) verified the interaction between inoculation with *Bradyrhizobium* and soybean cultivars on protein content in grains. Co-inoculation (*Bradyrhizobium* spp. + *Azospirillum brasilense*) of 23 Brazilian cultivars, showed that co-inoculation increased the average protein content in grains by 5.6% (Liborio et al., 2020).

Genetic mapping for BNF and protein content in soybean grains

Genetic mapping or linkage mapping consists in determining the relative position and distances between markers along chromosomes. Genetic map distances between two markers are defined as the average number of recombination events (Semagn et al., 2006).

Genetic maps require the development of a mapping population, choice of the sample size, and the type of molecular marker for genotyping. Genetic maps with high levels of genome coverage are the first step toward locating genes or QTL that are associated with traits of interest. Regions in genomes that contain genes associated with quantitative traits are known as QTL. Molecular markers are highly used in QTL studies because identification based only on conventional phenotyping is not possible (Collard, 2005).

Detection of genes or QTLs depends on the level of LD between a causal mutation and the physically linked markers. Therefore, the higher degree of association between marker alleles and the variant phenotypes the more probability that the phenotype and causative mutation could be physically attached to the marker (Hwang et al., 2014).

QTL mapping can be performed by linkage mapping studies or biparental mapping (Panthee et al., 2005; Santos et al., 2006; Rodrigues et al., 2010; Wang et al., 2021). One of the major difficulties in linkage mapping is that the confidence intervals for the identified QTLs are usually quite wide, greater than 20 cM. The extensive regions of LD present in populations of recombinant inbred lines (RILs) (Borevitz et al., 2003) may contain many underlying genes, making difficult to identify candidate genes.

In recent years, genome-wide association studies (GWAS) also known as associative mapping has contributed significantly to discover new QTLs and to confirm QTLs obtained by linkage mapping studies. Understanding the genetic basis of variation in protein content in soybeans grains is essential for marker-assisted

selection (MAS). Co-location of a given QTL in the physical map of the soybean genome is important for location of associated markers (Seo et al., 2019; Yao et al., 2020). However, comparisons of QTLs across populations can be complex due to the high number of common markers shared across populations (Van; Mchale, 2017). Therefore, stable and consistent QTLs with a large additive effect are desirable for use in soybean breeding programs (Kadam et al. 2016).

Molecular components of BNF-related pathways and genes were initially identified in studies with mutants, where genes Rj 1, Rj 2, Rj 3, Rj 4, Rj 5, Rj 6, Rj 7, and Rj 8 were identified (Penmetsa; Cook 1997; Williams and Lynch 1954; Caldwell 1966; Vest 1970; Vest and Caldwell 1972; Harper and Nickell 1995; Vuong et al., 1996). The main QTLs for protein content in soybean grains were detected and mapped on chromosomes 20 (LG-I) and 15 (LG-E) (Diers et al., 1992; Sebolt et al., 2000; Chung et al., 2003; Rodrigues et al., 2010). These QTLs illustrate well what is expected with large additive effects and consistency in different populations. However, a few studies were performed to identify QTLs for BNF traits. A wide understanding of this complex biological process is essential to face the challenges of food security, environmental degradation, and climate change (Torkamaneh et al., 2020).

Torres et al. (2015) have reported that most of the QTLs related to BNF coincide with QTLs for protein content in grains. Nodulation-related QTLs close to QTLs associated with grain yield have also been reported (Orf et al. 1999; Hyten et al. 2004; Reinprecht et al., 2006; Yang et al., 2017), which suggests that plants capable of performing more promising symbiosis and more effective BNF process can achieve higher protein content and higher grain yield.

QTLs for BNF traits were identified at different stages of soybean development under greenhouse and field conditions. These QTLs are distributed in distinct linkage groups (D1b, A1, C2, O, B1, H, B2, E, J and I) (Tanya et al., 2005; Nicolás et al., 2006; Santos et al., 2013). Yang et al. (2017) concluded that in a relatively stable field environment, the heritability (h^2) for nodulation-related traits can be greater than 0.8.

Currently, 255 and 80 QTLs associated with protein content and BNF, respectively, are registered in Soybase (https://soybase.org). Most of these QTLs were discovered using linkage mapping studies, as GWAS studies boosted only in recent years with the advancement and popularization of large-scale genotyping technologies (Figure 1).

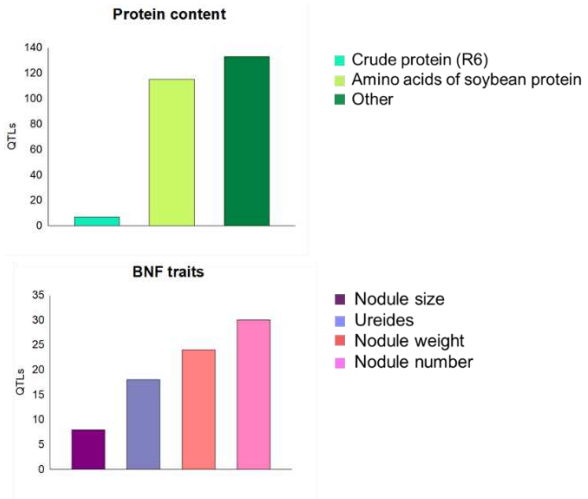

Figure 1. Number of QTLs identified for protein content in soybean grains and BNF traits through linkage

mapping and associative mapping (GWAS) studies (Soybase, 2023).

Due to the large number of inbred lines, breeding programs have had difficulties in evaluating BNF-related traits. The optimization of evaluations based on technological tools is essential for screening, genomic selection, and discovery of new genes (Torkamaneh et al., 2020). The number of nodules, nodule dry weight, ratio between shoot dry weight and nodules number, nodule size, shoot dry weight, root dry weight, N content, N accumulation, and ureides in tissues have been the most frequently evaluated traits for BNF (Ryle et al., 1981; Gan et al., 2003; Hwang et al., 2014; Torres et al., 2015; Grunvald et al., 2018; Torkamaneh et al., 2020). Most QTLs related to BNF have been obtained from a limited number of evaluations and genotypes, due to technical and labor limitations. Regard this, fine mapping of QTLs using robust evaluations in field trials, as the interaction with protein content in grains requires greater attention, mainly in phenotyping accuracy. A total of 12 QTLs explained from 12.6 to 59.0% of the phenotypic variation in a soybean population composed of 168 RILs for BNF evaluation. Among these QTLs, two new QTLs were revealed (qBNF-16 for number of nodules and qBNF-17 for size of nodules) (Yang et al., 2019).

Linkage mapping studies confirm the difficulty of simultaneously breeding for protein and oil traits. In most studies, these traits have shown negative correlation each other, which is a major bottleneck in soybean breeding for increasing them (Piper; Boote, 1999; Liang et al., 2010; Zhang et al., 2020). Diers et al. (1992) were pioneers in the application of Restriction Fragment Length Polymorphisms (RFLP) markers for protein and oil contents in soybean grains. The studied population was an offspring originated from a cross between *G. max* (lineage A81-356022) and *G. soja* (PI 468916). The segregation of the marker explained up to 43% of the total variation for the traits of interest. Cregan et al. (1999) reported that the Rj1 allele, related to high nodulation ability, was present in the D1b+W linkage group, while the Rj2 allele, related to low nodulation, was found in the J linkage group in three different mapping populations (59 F2 offspring from *G. max* × *G. soja*, 240 RILs from Minsoy × Noir, and 57 F2 offspring from Clark × Harosoy).

Based on the analysis of QTLs in the linkage group I for wild soybean (*Glycine soja*), alleles associated with this linkage group were related to higher content of protein in grains, lower content of oil, lower grain yield, smaller seeds, taller and earlier plants compared with QTLs found in populations of *Glycine max* (Sebolt et al., 2000). Panthee et al. (2006) verified microsatellite markers (Satt274, Satt420, and Satt479) in the D1b and O binding groups for protein content, oil, and seed size of soybean. Tanya et al. (2005) located an important QTL associated with BNF traits in the O linkage group. Minor effect QTLs were found in the A1, D1b + W, I, J, and K linkage groups.

Associations of 20 QTLs of minor effects in four linkage groups (B1, C2, D1b, and H), three for shoot dry weight, four for nodules number, two for nodule dry weight, and three for average nodule dry weight were verified in two populations of soybean with contrasting capacities of BNF, Bossier (high) and Embrapa 20 (medium) (Santos et al., 2006). Reinprecht et al. (2006), evaluating linoleic acid content in grains, yield, grain weight, protein content, and plant height, found three to eight QTLs per trait, which represented up to 78% of the total variation. Four QTLs associated with protein content were identified in the linkage groups D1a, G, A1, and I, whereas three QTLs associated with oil content were found in groups A1, I, and O. The phenotypic variation explained by the QTLs ranged from 6.24 to 18.94%, and 17.26 to 25.93%, for protein and oil contents, respectively (Rodrigues et al., 2010).

Using the composite interval mapping method for multiple traits (mCIM) Santos et al., 2013 identified two QTLs for SDW (located on LGs E and L), three QTLs for NN (located on LGs B1, E, and I), and one QTL for NDW/NN (located on LG I). These QTLs exhibited small effects (with R^2 values ranging from 1.7% to 10.0%) and accounted for 15.4%, 13.8%, and 6.5% of the total variation in these three traits, respectively. Hwang et al. (2014), provide the initial QTL insights into nodule traits in soybean derived from field experiments. These findings used Composite interval mapping (CIM) and achieved a total of five QTLs associated with nodule weight, one QTL was found close to a QTL detected through the Multiple interval mapping (MIM).

To determine the population structure and diversity related to BNF and protein content in soybean grains, Torres et al. (2015) identified 22 microsatellite markers distributed in 13 linkage groups (A1, A2, B1, B2, C1, C2, D1a, H, I, J, M, N, and O). Grunvald et al. (2018) identified two QTLs, one for nodule dry weight on chromosome 13 and the other for shoot dry weight on chromosome 19. The QTL located on chromosome 19 had previously been located by Santos et al. (2013), but in a different position, probably due to the different methods used for QTL analysis, based on inclusive composite interval mapping (ICIM) and composite interval method (CIM), respectively.

In a mapping study for grain yield, protein content in grains, plant height, and weight of grains, six, five, seven, and eight QTLs were identified, respectively (Zhu et al., 2021). Wang et al. (2021) found 50 QTLs for protein content in soybean grains distributed in 14 chromosomes. Among these QTLs, the main qSPC_20-1 and qSPC_20-2 located on chromosome 20 were repeatedly consistent in six environments where plants were grown.

Table 1 contains a summary of research on conventional mapping studies for traits related to BNF and protein content in soybeans. Only one study (Torres et al., 2015) simultaneously addressed traits related to BNF and protein content in soybean grains. Although many QTLs have been identified so far, the vast majority are yet to be functionally characterized.

Table 1. Linkage mapping studies for traits related to BNF and protein content in soybean grains.

Trait	Marker	Species	Panel*	Reference
Protein	RFLP	*G. max/G. soja*	60	Diers et al. (1992)
Protein	RAPD e SSR	*G. max*	77	Chung et al. (2003)
Protein	SSR	*G.max*	101	Panthee et al. (2005)
BNF	SSR	*G.max*	136	Tanya et al. (2005)
BNF	SSR	*G.max*	157	Santos et al. (2006)
Protein	RAPD, CAPS, STS e SSR	*G.max*	169	Reinprecht et al. (2006)
Protein	SSR	*G.max*	207	Rodrigues et al. (2010)
BNF	SSR	*G.max*	157	Santos et al. (2013)
BNF	SSR/SNP	*G.max*	80	Hwang et al. (2014a)
BNF and Protein	SSR	*G.max*	191	Torres et al. (2015)
BNF	SSR	*G.max*	175	Yang et al. (2017)
BNF	SNP	*G.max*	113	Grunvald et al. (2018)
BNF	SNP	*G.max*	168	Yang et al. (2019)
Protein	SNP	*G.max*	994	Zhu et al. (2021)
Protein	SNP	*G. max*	178	Wang et al. (2021)

*Number of RILs - Recombinant inbred lines. RFLP= Restriction Fragment Length Polymorphism; RAPD = Randomly amplified polymorphic DNA; CAPS = cleaved amplified polymorphic sequence; STS=sequence-tagged sites; SSR= Simple Sequence Repeat; SNP = Single Nucleotide Polymorphism.

Genome-wide association studies (GWAS) for BNF and protein content in soybean grains

In the conventional QTL mapping involving bi-parental populations, it is only possible to examine the allelic diversity segregated between the two population groups. This limits the genetic resolution, as few recombination events can be captured between these populations (Korte; Farlow, 2013). However, excellent results have been obtained in GWAS studies due to the allelic diversity found in a diversified panel of accessions that present lower LD and a higher resolution.

QTLs derived from segregating families such as RIL and F2 populations are generally restricted to those that segregate in the crossbreed (Zhang et al., 2015). Furthermore, one of the main advantages of GWAS is the possibility of using non-structured populations, which allows for the exploration of the genetic variability contained in panels obtained from Germplasm Banks.

Next Generation Sequencing (NGS) technologies use platforms that can generate information of millions of base pairs in a short time. These methodologies are based on the reduction of the genome complexity, as only a small portion of it is sampled in a single sequencing step that occurs in parallel with the construction of genomic libraries, allowing for detection of polymorphic positions in the sampled sequences quickly and efficiently (Shendure; Ji, 2008). Consequently, the identification of polymorphisms and genotyping in soybean have been focused on automation techniques and large-scale data collection. Over the years, molecular techniques, more efficient and low-cost, have been developed. The use of Single Nucleotide Polymorphism (SNP) can be highlighted, which is widely recognized for its robustness and accuracy. A wide range of markers is a great advantage for mapping studies, in addition to increasing the chances of correct association of the polymorphisms responsible for phenotypic variations of interest.

Genotyping-by-Sequencing (GBS) involves the use of restriction enzymes, adapters, and barcodes, and can be used in several species with a low cost per sample. Therefore, population studies, germplasm characterization, and mapping of traits of interest can be easily performed (Elshire et al., 2011). The choice of the ideal restriction enzyme depends on the coverage needed and the repetitive sequences present in the genome of the species under study. The ends of the cleaved fragments are linked to barcodes (previously known specific DNA sequences) that identify each genotype, in addition to adapter sequences. After connecting the adapters, a PCR is performed with all samples (multiplex). In the PCR multiplex, several loci (genomic regions) of interest are amplified in a single PCR reaction. This is achieved by using specific primers for each locus, where these primers are designed to bind to the flanking sequences of each locus. Simultaneous amplification of multiple loci in a single PCR reaction saves time, resources, and reagents.

After sequencing, bioinformatic analyzes for detection of SNPs takes place. Prospecting for markers in the raw sequences generated by sequencing (fastq) involves different bioinformatic analyses with the support of specific software for each step. The fragments are grouped by barcodes to be aligned according to the crop reference genome (Torkamaneh, 2017). In soybean, this search can be performed on SoyBase (https://www.soybase.org/). After the detection of SNPs, they are filtered considering missing data, minimum allele frequency (MAF) parameters, and sequencing coverage. After filtering, a file is generated with the call of SNPs (SNP calling) in hapmap and/or VCF formats (Glaubitz et al., 2014). However, the application of these parameters greatly reduces the number of initial SNPs. The reduction in the number of SNPs after filtering occurs due to factors such as data quality issues, including low sequencing coverage and poor quality, as well as the removal of rare variants with low allele frequencies. Filtering also eliminates SNPs stemming from contamination

or sequencing errors. Strong linkage disequilibrium with other SNPs, non-compliance with expected genotypes, and the focus on specific genomic loci are additional reasons for SNP exclusion. Moreover, statistical requirements, like a minimum threshold of complete genotypic data, may lead to the exclusion of SNPs with extensive missing data. These filtering processes are aimed at enhancing data accuracy, reliability, and relevance in genomic analyses.

Research efforts have been made to identify candidate genes and consider their possible roles in plants (Li et al., 2019; Trněný et al., 2019; Wang et al., 2021). GWAS not only accurately identifies most previously reported QTLs, but it also results in genomic regions that are closer. Regions in the genome identified by GWAS allow for more accurate marker-assisted selection (Hwang et al., 2014). Integration of GBS and GWAS can be used as a powerful approach for linkage mapping of complex traits in soybean (Sonah et al., 2015). This approach also allows the use of phenotypic and genotypic data from previous studies stored in databases. Associative mapping studies represented a paradigm shift, as classical QTL mappings prioritized the existence of high LD. Based on associative mapping, populations with low LD can be evaluated, which greatly increases the detection power of genes related to traits of interest.

False-positive associations can occur in GWAS due to the population structure (Mackay; Powell, 2007). These include scenarios such as genetically heterogeneous populations that can introduce spurious associations between genetic variants and phenotypic traits, potentially confounding the interpretation of GWAS results. Addressing population structure through appropriate statistical adjustments, such as principal component analysis (PCA) and kinship analysis, is crucial to ensure accurate and reliable GWAS outcomes by mitigating the risk of false positives. The mixed linear model is also widely used in GWAS, as it allows the inclusion of population structure and genetic correlations due to kinship to reduce spurious associations (Yu; Buckler, 2006).

The probability of detecting a QTL in the population, also called test power, depends on different factors such as the coefficient of determination (r^2) between the marker and the QTL, the extent of the effect of the QTL on the phenotypic variation, sample size, and significance level. To correct errors, there are several methodologies, the most used being Bonferroni and the False discovery rate (FDR). In a GWAS for BNF traits, Torkamaneh et al. (2020) identified 25 QTL regions encompassing 40 putative candidate genes for BNF traits including 20 genes with no prior known role in BNF. Ray et al., 2015, identified 53 putative loci on 18 chromosomes associated with ureide concentration. Two of the putative loci were located near previously reported QTL associated with ureide concentration and 30 loci were located near genes associated with ureide metabolism.

Based on the evaluation of 11 traits related to BNF, X SNPs were identified in eight soybean chromosomes. Three loci located on chromosome 17 were associated with N contents in grains in two environments. Two other loci located on chromosome 17 were associated with nodule number and nodule dry weight. Furthermore, the alleles were selected based on the phenotypes of different groups, which further demonstrated the reliability of these markers for traits of interest. Some of these candidate genes had different levels of expression in soybean during the development of nodules, two of which were later verified by qRT-PCR (Huo et al., 2019). Forty SNPs were identified in 17 genomic regions associated with protein content in grains. Of these, the five SNPs with the highest associations and seven adjacent SNPs were located in the 27.6-30.0 Mbp region on chromosome 20 (Hwang et al., 2014).

Seven QTLs associated with protein content in soybean grains were found in six chromosomes (Chr 2, 6, 11, 12, 13 and 16), explaining 60.9% of the variation for this trait. For oil content, eight QTLs were identified

on six chromosomes (Chr 1, 4, 5, 6, 17 and 19), explaining 78.3% of the variation in this trait. The correlation between the number of loci containing favorable alleles and the traits of interest was 0.49 for protein and 0.60 for oil contents, respectively. The identified molecular markers were mapped in genomic regions containing previously mapped QTLs for both traits, which increases the association between these regions and the genetic control of oil and protein contents in soybean grains (Dias et al., 2017).

Leamy et al. (2017), used associative mapping in a panel composed of 570 accessions of wild soybean (*Glycine soja*) and identified 29 SNPs associated with protein, oil, and fatty acid contents in grains, located in 10 chromosomes. Eight SNPs were co-localized, as previously described for (*Glycine max*). Some candidate genes associated with the metabolism and regulation of fatty acids were also identified. In the average of the wild soybean accessions, the protein content in grains was 48%, and the oil content was 11%. These results demonstrate the importance of exploring the existing variability in wild soybean accessions, as a promising source of alleles of interest for protein and oil contents. A QTL on chromosome 5 increased oil content with no effect on protein content, and a QTL on chromosome 10 increased protein content with little effect on oil content. The frequencies of positive effect haplotypes varied among maturation groups and geographic regions, providing guidance on which alleles have potential to contribute to soybean breeding in specific regions (Lee et al., 2019).

Li et al. (2019) obtained 31 SNPs associated with protein and oil contents in 12 soybean chromosomes. Genes rs12328685, rs49097495, rs29457452, and rs29979450 were pleiotropic for protein and oil content. Gaps on chromosomes 1, 15 and 20 were correlated with protein and oil contents in soybean grains. Overall, 25 candidate genes involved in protein and/or oil metabolism in two regions (qPro15-1, qPro20-1) were identified and eight of these genes had differential expression in parental lines during the final stages of reproductive development (Zhang et al., 2019).

Zhang et al. (2021a) identified three, four, and five QTLs related to protein, oil, and water-soluble protein contents, respectively. Furthermore, five QTLs (qPC-15-1, qOC-8-1, qOC-12-1, qOC-20-1, and qWSPC-8-1) were detected in different environments. Analysis of favorable alleles for oil and water-soluble protein content in soybean grains showed that qOC-8-1 (qWSPC-8-1) had opposite effects in the synthesis of soluble protein. Relative expression analysis suggested that Glyma.15G049200 in qPC-15-1 affects the protein synthesis and Glyma.08G107800 in qOC-8-1 and qWSPC-8-1 may be involved in the synthesis of water- and oil-soluble proteins, producing opposite effects.

The demand for soybean cultivars with higher protein and oil contents in grains to meet the market requirements is certainly a major challenge for breeders (Kumar et al., 2021). A linkage mapping and GWAS study for glycinin and β-conglycinin in soybean identified 67 QTLs and 11 genomic regions highly associated with glycinin-related loci (11S), β-conglycinin (7S), the sum of glycinin and β-conglycinin (SGC), and the ratio between glycinin and β-conglycinin (RGC). Among these loci, 19 QTLs were identified for the first time. Furthermore, eight genes in 11 genomic regions may be related to protein in grains. The Gy1 candidate gene promoter was conserved and the polymorphism in the Gy1 promoter sequence was significantly associated with 11S content (the main protein fractions found in soybean seeds). Association analysis revealed SNPs that are associated with these QTLs and these SNPs exhibited a low level of diversity in cultivated soybean varieties, typical for traits related to domestication (Zhang et al., 2021b).

For content of ureides, candidate genes were found on chromosomes 1, 2, 3, 5, 6, and 7 of the soybean genome (Ray et al., 2015). Hwang et al. (2014) identified 40 SNPs in 17 chromosomal regions for protein content

in grains. Five of these 40 SNPs had higher associations, and seven adjacent SNPs were in the 27.6-30.0 Mbp region of chromosome 20. Most of these loci found via GWAS were located within genomic regions that coincided with regions identified by previous linkage mapping studies. However, using GWAS, regions closer to genes of interest were found.

Zhang et al. (2018), in an associative mapping study for protein, oil, fatty acids, and amino acids in soybean grains found 138 candidate genes. These results revealed different genetic bases between amino acid content, grain weight, and protein content. On chromosome 15, two candidate genes for protein were found: Glyma 15g049700 and Glyma15g049900. Sonah et al. (2015), using GBS with about 47,000 SNPs for a panel of 139 soybean lines, identified

21 SNPs associated to protein content on Gm08 from 45.5 to 46.9 Mb and the lowest number (one single significant SNP each) on Gm19 (50.4 Mb) and Gm20 (10.0 Mb).

Huo et al. (2019) identified SNPs in eight soybean chromosomes, where three loci on chromosome 17 were associated with shoot N concentrations in two environments. Two other loci located on the same chromosome were associated with a higher nodule number and nodule dry weight per plant.

Table 2 contains a summary of research works based on GWAS studies for traits related to BNF and protein content in soybean grains. Again, GWAS studies involving both traits have not been located so far.

Table 2. GWAS studies for traits related to BNF and protein content in soybean grains. Note that no study has simultaneously studied both traits so far.

Trait	Platform/Technique	Specie	Panel*	Tested loci	Reference
Protein	Illumina Infinium and GoldenGate	*G. max*	298	31.954	Hwang et al. (2014b)
BNF	SoySNP50K iSelect SNP BeadChip	*G. max*	374	33.957	Ray et al. (2015)
Protein	Illumina Infinium BeadChip	*G.max*	169	6.000	Dias et al. (2017)
Protein	SoySNP50K iSelect SNP BeadChip	*G. soja*	570	52.041	Leamy et al. (2017)
BNF	SoySNP50K Illumina BeadChip	*G. max*	267	5.403	Huo et al. (2019)
Protein	SoySNP50K SNPs	*G. max*	621	34.014	Lee et al. (2019)
Protein	SLAF-seq e Illumina Genome Analyzer II	*G. max*	185	12.072	Li et al. (2019a)
Protein	Hi-Seq 2000	*G. max*	200	94.462	Zhang et al. (2019)
Protein	355K SoySNP	*G. max*	211	207.608	Zhang et al. (2021)
Protein	180K Axiom® Soya SNP array	*G. max*	203	96,432	Kim et al. (2023)

*Accessions number

Perspectives and other tools

The collection of soybean phenotypic data with high resolution and accuracy represents a challenge in research and has been a limiting factor for the effective use of genomic data in genetic breeding. At the initial stages of breeding programs, the number of accessions for phenotyping is quite high, and failures at this stage can compromise genetic gains. Therefore, new high-throughput phenotyping (HTP) platforms based on field evaluations can quickly evaluate thousands of accessions with high spatial and temporal resolution (Bai et al., 2016; Moreira et al., 2019; Baek et al., 2016; Moreira et al., 2019; Baek et al. al., 2020). In consequence, HTP methodologies for plant breeding can revolutionize the phenotyping of challenging characteristics such as BNF

traits.

A considerable number of QTLs related to protein content in soybean grains have already been identified. However, the use of marker-assisted selection has not been well explored yet. Although many molecular components that control the BNF have already been discovered, little information is available on the genetic diversity of cultivated germplasms for an economically relevant species such as soybean (Torkamaneh et al., 2020). Efforts should focus on inserting phenotypic and genotypic information related to BNF in soybean databases. These traits, when compared with others of equal agronomic relevance, are not grouped and available for many genotypes.

The popularization of soybean Whole Genome Resequencing (WGRS) brings many possibilities to be explored, such identification of domestication loci, selective scans, genetic diversity, population structure, linkage disequilibrium, and exploration of alleles of interest (Pawlowski et al., 2020; Kajiya-kanegae et al., 2021). Meta-QTL and meta-GWAS studies have also been successful in narrowing the genomic region to some loci by overlapping mapping results from multiple studies (Shook et al., 2021; Izquierdo et al., 2023; Chen et al., 2021). However, unlike meta-QTL studies, information for meta-GWAS studies has not yet been fully compiled, particularly for traits such as BNF.

An increase in plant genotyping capacity and the comparison among several related genotypes rises interest in Pan-Genomes, which aims to expand the linear reference genome coordinates system to accommodate more regions of genetic diversity (Bayer et al., 2020; Della Coletta et al., 2021; Torkamaneh et al., 2021). Pan-genomes are already available for cultivated soybean (Torkamaneh et al., 2021) and wild soybean (Liu et al., 2020).

Genetic breeding programs need to prioritize the insertion of markers related to BNF and grain protein composition in marker-assisted selection. The importance of integrating different approaches to obtain reliable results that have practical applicability has become even more evidenced.

Gene editing tools, such as CRISPR-Cas9 suggest that further legume varieties will be more resistant to climate changes and should be able to perform more promising symbioses, based on the selection of more symbiotically efficient partners (bacteria and plant genotype), in addition to adjustments in molecular circuits to deal with adverse climate conditions. Moreover, artificial intelligence (AI) such as machine learning (ML) in plant breeding will support plant breeders with efficient and effective tools to accelerate the development of new soybean cultivars harboring improved traits. Thus, breeding programs need to benefit from phenotyping and genotyping technologies to reduce the difficulty and delay in obtaining superior soybean genotypes for BNF traits and protein content in grains (Jain, Jones, and Roy, 2023).

Technological advances will make it possible to accurately identify causal loci and predict reproductive values (Kim et al. 2020). Mapping studies increasingly rely on high-throughput phenotyping tools, ML for statistical analyses of big data, cost-effective NGS, as well as numerous free bioinformatics tools to efficiently analyze complex genetic data. The integration of technologies can revolutionize studies of laborious trait mapping for phenotyping of high genetic complexity. In addition, genomic regions mapped with robustness and agility, enable advances in functional validation studies of candidate genes identified in mapping studies.

Conclusions

In the current scenario, the use of only conventional breeding is not enough to deal with the increased demand for soybean protein in the context of climate change. Mapping studies are useful tools for identifying QTLs and candidate genes associated with BNF and protein content traits. Soybean cultivars need to become more capable to provide grains with higher content of proteins based on an effective BNF resulting from an intelligent interaction between both symbiotic partners.

The appeal for ever more yielding soybean cultivars distanced genotypes with higher protein content in grains from the fields. In the coming years, research for new cultivars from a different perspective can support growers to better capitalize soybean based on protein content and increase their incomes. We hope that this review has shed light on the importance of developing soybean cultivars with higher protein content in grains and at the same time with greater efficiency in the use of biological N in the context of the sustainability provided by the BNF process.

Conflict of interest

The authors declare that they have no conflict of interest.

References

AGROSTAT - Estatísticas de Comercio Exterior do Agronegócio Brasileiro (2020). Disponível em: <https://sistemasweb.agricultura.gov.br/pages/AGROSTAT.html>. Acesso em: 20/08/2021.

Albrecht LP, Braccini ADL, Ávila MR, Suzuki LS, Scapim CA, Barbosa MC. Teores de óleo, proteínas e produtividade de soja em função da antecipação da semeadura na região oeste do Paraná. Bragantia. 2008;67:865-873.

ANPII—Associação Nacional dos Produtores e Importadores de Inoculantes (2018). Levantamento do uso de Inoculantes no Brasil. VIII Congresso Brasileiro de Soja, Goiânia, 2018.

Atkins CA, Smith PMC. Translocation in legumes: assimilates, nutrients, and signaling molecules. Plant Physiology. 2007;144(2):550-561.

Baek J, Lee E, Kim N, Kim SL, Choi I, Ji H, Chung YS, Choi MS, Moon JK, Kim KH. High throughput phenotyping for various traits on soybean seeds using image analysis. Sensors. 2020;20(1):248.

Bai G, Ge Y, Hussain W, Baenziger PS, Graef G. A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding. Computers and Electronics in Agriculture. 2016;128:181-192.

Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, Specht J, Lorenz A. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. The Plant Genome. 2015;8(3):1-14.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nature plants. 2020;6(8):914-920.

Bonato ER, Bertagnolli PF, Lange CE, Rubin SDAL. Teor de óleo e de proteína em genótipos de soja desenvolvidos após 1990. Pesquisa Agropecuária Brasileira. 2000;35:2391-2398.

Bueno RD, Borges LL, Arruda KMA, Bhering LL, Barros ED, Moreira MA. Genetic parameters and genotype x environment interaction for productivity, oil and protein content in soybean. Afr. J. Agric. Res. 2013;8(38):4853-4859.

Caldwell BE. Inheritance of a strain-specific ineffective nodulation in soybeans 1. Crop Science. 1966;6(5):427-428.

Canto WL, Turatti JM. Produção e mercado de produtos intermediários protéicos de soja no Brasil. Boletim do Centro de Pesquisa de Processamento de Alimentos. 1989.

Carrão-Panizzi MC et al. (2021). Teores de óleo e proteína em genótipos de soja em diferentes situações de manejo.

Circular Técnica 60. Embrapa.

Carter AM, Tegeder M. Increasing nitrogen fixation and seed development in soybean requires complex adjustments of nodule nitrogen metabolism and partitioning processes. Current Biology. 2016;26(15):2044-2051.

Cerezini P et al. Strategies to promote early nodulation in soybean under drought. Field Crops Research. 2016;196:160–167.

Cerezini P et al. Soybean tolerance to drought depends on the associated Bradyrhizobium strain. Brazilian Journal of Microbiology. 2020;51(4):1977-1986.

Chen H et al. Novel QTL and Meta-QTL mapping for major quality traits in soybean. Frontiers in plant science. 2021;12:774270.

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

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

Collard BCY, Grams RA, Bovill WD, Percy CD, Jolley R, Lehmensiek A, ... Sutherland MW. Development of molecular markers for crown rot resistance in wheat: mapping of QTLs for seedling resistance in a '2-49'×'Janz'population. Plant Breeding. 2005;124(6):532-537.

Collier R, Tegeder M. Soybean ureide transporters play a critical role in nodule development, function and nitrogen export. The Plant Journal. 2012;72(3):355-367.

Companhia Nacional de Abastecimento - CONAB (2023). Décimo Levantamento. Safra 2022/23. Disponível em: < Conab - Safra Brasileira de Grãos>. Acesso em: 25/04/23.

Costa SI da, Mori EEM, Fujita JT. Características químicas, organolépticas e nutricionais de algumas cultivares de soja. A soja no Brasil. Campinas: Instituto de Tecnologia de Alimentos. 1981. p. 823-827.

Cregan PB et al. An integrated genetic linkage map of the soybean genome. Agronomy-Faculty Publications. 1999;20.

Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN. How the pan-genome is changing crop genomics and improvement. Genome biology. 2021;22(1):1-9.

Dias DA, Polo LR, Lazzari F, Silva GJ, Schuster I. Genome-wide association for mapping QTLs linked to protein and oil contents in soybean. Pesquisa Agropecuária Brasileira. 2017;52:896-904.

Diers BW, Keim P, Fehr WR, Shoemaker RC. RFLP analysis of soybean seed protein and oil content. Theoretical and Applied Genetics. 1992;83:608-612.

Elshire RJ et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 2011;6(5):e19379.

Fabre F, Planchon C. Nitrogen nutrition, yield and protein content in soybean. Plant Science. 2000;152(1):51-58.

Fehr WR, Hoeck JA, Johnson SL, Murphy PA, Nott JD, Padilla GI, Welke GA. Genotype and environment influence on protein components of soybean. Crop Science. 2003;43(2):511-514.

Ferreira M, Fulaneti FS, Carvalho PD, Menezes HM, Beutler AN. Eficiência do inoculante e necessidade de aplicação de uréia em soja em solos de várzea. Anais do Salão Internacional de Ensino, Pesquisa e Extensão. 2016;8(2).

Food and Agriculture Organization of the United Nations – FAOSTAT (2021). Disponível em: <http://www.fao.org/faostat/en>. Acesso em 20/08/2021.

Gan Y, Stulen I, van Keulen H, Kuiper PJ. Effect of N fertilizer top-dressing at various reproductive stages on growth, N2 fixation and yield of three soybean (Glycine max (L.) Merr.) genotypes. Field Crops Research. 2003;80(2):147-155.

Gerahty N et al. Anatomical analysis of nodule development in soybean reveals an additional autoregulatory control point. Plant Science. 1992;85(1):1-7.

Glaubitz JC et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS one. 2014;9(2):e90346.

Graham PH, Vance CP. Legumes: importance and constraints to greater use. Plant physiology. 2003;131(3):872-

877.

Grunvald AK et al. Identification of QTLs Associated with Biological Nitrogen Fixation Traits in Soybean Using a Genotyping-by-Sequencing Approach. Crop Science. 2018;58(6):2523-2532.

Gupta SK, Manjaya JG. Advances in improvement of soybean seed composition traits using genetic, genomic and biotechnological approaches. Euphytica. 2022;218(7):99.

Haq MU, Mallarino AP. Soybean yield and nutrient composition as affected by early season foliar fertilization. Agronomy Journal. 2000;92(1):16-24.

Harper JE, Nickell CD. Genetic analysis of nonnodulating soybean mutants in a hypernodulated background. Soybean Genet. Newsl. 1995;22:185-190.

Hayati R, Egli DB, Crafts-Brandner SJ. Independence of nitrogen supply and seed growth in soybean: studies using an in vitro culture system. J Exp Bot. 1996;47(1):33-40.

Herridge DF, Peoples MB, Boddey RM. Global inputs of biological nitrogen fixation in agricultural systems. Plant Soil. 2008;311:1-18.

Hungria M et al. Importância do sistema de semeadura direta na população microbiana do solo. Embrapa Soja-Comunicado Técnico (INFOTECA-E). 1997.

Hungria M, Campo RJ, Mendes IC. Fixação biológica do nitrogênio na cultura da soja. Embrapa Soja-Circular Técnica (INFOTECA-E). 2001.

Hungria M, Campo RJ, Mendes IC. A importância do processo de fixação biológica do nitrogênio para a cultura da soja: componente essencial para a competitividade do produto brasileiro. Embrapa Soja-Documentos (INFOTECA-E). 2007.

Hungria M, Nogueira MA. Nitrogen fixation. In: Rengel Z, Cakmak I, White PJ, editors. Marschner´s Mineral Nutrition of Plants. 4th ed. Elsevier, London: Academic Press. 2022. p. 615-650.

Hungria M, Nogueira MA, Araujo RS. Co-inoculation of soybeans and common beans with rhizobia and azospirilla: strategies to improve sustainability. Biol Fertil Soils. 2013;49(7):791-801.

Huo X et al. Genetic loci and candidate genes of symbiotic nitrogen fixation–related characteristics revealed by a genome-wide association study in soybean. Mol Breeding. 2019;39:1-6.

Hwang S et al. Genetics and mapping of quantitative traits for nodule number, weight, and size in soybean (Glycine max L.[Merr.]). Euphytica. 2014;195(3):419-434.

Hymowitz T, Collins FI, Panczner J, Walker WM. Relationship between the content of oil, protein, and sugar in soybean seed 1. Agron J. 1972;64(5):613-616.

Hyten DL et al. Seed quality QTL in a prominent soybean population. Theor Appl Genet. 2004;109:552-561.

Hyten DL et al. Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci. 2006;103(45):16666-16671.

Imsande J, Schmidt JM. Effect of N source during soybean pod filling on nitrogen and sulfur assimilation and remobilization. Plant Soil. 1998;202(1):41-47.

Israel DW, Burton JW, Wilson RF. Studies on Genetic Male-Sterile Soybeans: IV. Effect of Male Sterility and Source of Nitrogen Nutrition on Accumulation, Partitioning, and Transport of Nitrogen. Plant Physiol. 1985;78(4):762-767.

Izquierdo P et al. Combination of meta-analysis of QTL and GWAS to uncover the genetic architecture of seed yield and seed yield components in common bean. Plant Genome. 2023;e20328.

Jain D, Jones L, Roy S. Gene editing to improve legume-rhizobia symbiosis in a changing climate. Curr Opin Plant Biol. 2023;71:102324.

Joaquim PI et al. Nitrogen compounds transporters: candidates to increase the protein content in soybean seeds. J Plant Interact. 2022;17(1):309-318.

Kadam S, Vuong TD, Qiu D, Meinhardt CG, Song L, Deshmukh R, ... & Nguyen HT. Genomic-assisted phylogenetic analysis and marker development for next generation soybean cyst nematode resistance breeding. Plant Science. 2016;242:342-350.

Kajiya-Kanegae H, Nagasaki H, Kaga A, Hirano K, Ogiso-Tanaka E, Matsuoka M, Ishimori M, Ishimoto M,

Hashiguchi M, Tanaka H, Akashi R. Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. DNA Research. 2021;28(1):dsaa032.

Kakraliya SK, Singh U, Bohra A, Choudhary KK, Kumar S, Meena RS, & Jat ML. Nitrogen and legumes: a meta-analysis. In: Legumes for soil health and sustainable management. 2018;277-314.

Kambhampati S, 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. 2020;10(1):18.

Kaschuk G, Alberton O, Hungria M. Three decades of soil microbial biomass studies in Brazilian ecosystems: lessons learned about soil quality and indications for improving sustainability. Soil Biology and Biochemistry. 2010;42(1):1-13.

Kiers ET, Hutton MG, Denison RF. Human selection and the relaxation of legume defenses against ineffective rhizobia. Proceedings of the Royal Society B: Biological Sciences. 2007;274(1629):3119-3126.

Kim JM, Kim KH, Jung J, Kang BK, Lee J, Ha BK, Kang S. Validation of marker-assisted selection in soybean breeding program for pod shattering resistance. Euphytica. 2020;216:1-9.

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

Kumar V, Vats S, Kumawat S, Bisht A, Bhatt V, Shivaraj SM, ... & Sonah H. Omics advances and integrative approaches for the simultaneous improvement of seed oil and protein content in soybean (Glycine max L.). Critical Reviews in Plant Sciences. 2021;40(5):398-421.

Lam HM, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature genetics. 2010;42(12):1053-1059.

Lawn RJ, Brun WA. Symbiotic nitrogen fixation in soybeans. I. Effect of photosynthetic source-sink manipulations. Crop Science. 1974;14(1):11-16.

Leamy LJ, Zhang H, Li C, Chen CY, Song BH. A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC genomics. 2017;18:1-5.

Lee S, Van K, Sung M, Nelson R, LaMantia J, McHale LK, Mian MR. Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. Theoretical and Applied Genetics. 2019;132:1639-59.

Li D, Zhao X, Han Y, Li W, Xie F. Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. Genomics. 2019;111(1):90-5.

Li X, Shao Z, Tian R, Zhang H, Du H, Kong Y, Li W, Zhang C. Mining QTLs and candidate genes for seed protein and oil contents across multiple environments and backgrounds in soybean. Molecular Breeding. 2019;39:1-6.

Liang HZ, YU YL, Wang SF, Yun LIAN, Wang TF, Wei YL, ... & ZHANG MC. QTL mapping of isoflavone, oil and protein contents in soybean (Glycine max L. Merr.). Agricultural Sciences in China. 2010;9(8):1108-1116.

Liborio PH, et al. Co-inoculation of Bradyrhizobium japonicum and Azospirillum brasilense on the physiological quality of soybean seeds. Semina: Ciências Agrárias. 2020;41(6):2937-2950.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X. Pan-genome of wild and cultivated soybeans. Cell. 2020;182.

Lorini I. Qualidade de sementes e grãos comerciais de soja no Brasil-safra 2016/17. Embrapa Soja-Documentos (INFOTECA-E). 2018.

Mackay I, Powell W. Methods for linkage disequilibrium mapping in crops. Trends in plant science. 2007;12(2):57-63.

Mahmoud AA, Natarajan SS, Bennett JO, Mawhinney TP, Wiebold WJ, Krishnan HB. Effect of six decades of selective breeding on soybean protein composition and quality: a biochemical and molecular analysis. Journal of Agricultural and Food Chemistry. 2006;54(11):3916-3922.

Mcclure PR, ISRAEL DW. Transport of nitrogen in the xylem of soybean plants. Plant Physiology. 1979;64(3):411-416.

Moreira FF, Hearst AA, Cherkauer KA, Rainey KM. Improving the efficiency of soybean breeding with high-throughput canopy phenotyping. Plant Methods. 2019 Dec;15:1-9.

Muñoz N, Qi X, Li MW, Xie M, Gao Y, Cheung MY, ... & Lam HM. Improvement in nitrogen fixation capacity could be part of the domestication process in soybean. Heredity. 2016;117(2):84-93.

Nicolás MF, Hungria M, Arias CA. Identification of quantitative trait loci controlling nodulation and shoot mass in progenies from two Brazilian soybean cultivars. Field Crops Research. 2006;95(2-3):355-366.

O'keefe SF, BIANCHI L, SHARMAN J. Soybean nutrition. 2015.

Orf JH, Chase K, Adler FR, Mansur LM, Lark KG. Genetics of soybean agronomic traits: II. Interactions between yield quantitative trait loci in soybean. Crop Science. 1999;39(6):1652-1657.

Paek NC, Imsande J, Shoemaker RC, Shibles R. Nutritional control of soybean seed storage protein. Crop Science. 1997;37(2):498-503.

Panthee DR, Pantalone VR, Saxton AM, West DR, Sams CE. Genomic regions associated with amino acid composition in soybean. Molecular Breeding. 2006;17:79-89.

Patil G, et al. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theoretical and Applied Genetics. 2017;130(10):1975-1991.

Pawlowski ML, Vuong TD, Valliyodan B, Nguyen HT, Hartman GL. Whole-genome resequencing identifies quantitative trait loci associated with mycorrhizal colonization of soybean. Theoretical and Applied Genetics. 2020 Feb;133:409-17.

Pélissier HC, et al. Pvups1, an allantoin transporter in nodulated roots of French bean. Plant Physiology. 2004;134(2):664-675.

Penmetsa RV, Cook DR. A legume ethylene-insensitive mutant hyperinfected by its rhizobial symbiont. Science. 1997;275(5299):527-530.

Phansak P, Soonsuwon W, Hyten DL, Song Q, Cregan PB, Graef GL, Specht JE. Multi-population selective genotyping to identify soybean [Glycine max (L.) Merr.] seed protein and oil QTLs. G3: Genes, Genomes, Genetics. 2016;6(6):1635-1648.

Piper EL, Boote KI. Temperature and cultivar effects on soybean seed oil and protein concentrations. Journal of the American Oil Chemists' Society. 1999;76(10):1233-1241.

Pípolo AE, et al. Teores de óleo e proteína em soja: fatores envolvidos e qualidade para a indústria. Embrapa Soja-Comunicado Técnico (INFOTECA-E). 2015.

Qin L, et al. The high-affinity phosphate transporter gmpt5 regulates phosphate transport to nodules and nodulation in soybean. Plant physiology. 2012;159(4):1634-1643.

Ramalho MAP. Genética na agropecuária. FAEPE. 1990.

Ray JD, et al. Genome-wide association study of ureide concentration in diverse maturity group IV soybean [Glycine max (L.) Merr.] accessions. G3: Genes, Genomes, Genetics. 2015 Nov 1;5(11):2391-403.

Reinprecht Y, et al. Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (Glycine max (L.) Merrill) germplasm. Genome. 2006;49(12):1510-1527.

Rodrigues JIS, et al. Mapping QTL for protein and oil content in soybean. Pesquisa Agropecuária Brasileira. 2010;45(5):472-480.

Rolletschek H, et al. Ectopic expression of an amino acid transporter (VfAAP1) in seeds of Vicia narbonensis and pea increases storage proteins. Plant Physiology. 2005;137(4):1236-1249.

Ryle GJA, Arnott RA, Powell CE. Distribution of dry weight between root and shoot in white clover dependent on N2 fixation or utilizing abundant nitrate nitrogen. Plant and Soil. 1981;60:29-39.

Salvagiotti F, et al. Nitrogen uptake, fixation and response to fertilizer N in soybeans: A review. Field Crops Research. 2008;108(1):1-13.

Santachiara G, et al. Does biological nitrogen fixation modify soybean nitrogen dilution curves?. Field Crops Research. 2018;223:171-178.

Santos MA, et al. Mapping of qtls associated with biological nitrogen fixation traits in soybean. Hereditas. 2013;150(2-3):17-25.

Santos MA, Nicolás MF, Hungria M. Identificação de QTL associados à simbiose entre Bradyrhizobium

japonicum, B. Elkanii e soja. Pesquisa Agropecuária Brasileira. 2006;41:67-75.

Santos MS, Nogueira MA, Hungria M. Microbial inoculants: reviewing the past, discussing the present and previewing an outstanding future for the use of beneficial bacteria in agriculture. AMB Express. 2019;9(1):1-22.

Saturno DF, et al. Mineral nitrogen impairs the biological nitrogen fixation in soybean of determinate and indeterminate growth types. Journal of Plant Nutrition. 2017;40(12):1690-1701.

Sebolt AM, Shoemaker RC, Diers BW. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Science. 2000;40(5):1438-1444.

Sediyama T, Pereira MG, Sediyama CS, Gomes JLL. Cultura da Soja, Parte I. UFV, Minas Gerais, 97p. 1993.

Semagn K, Bjørnstad Å, Ndjiondjop MN. Principles, requirements and prospects of genetic mapping in plants. African Journal of Biotechnology. 2006;5(25).

Seo JH, Kim KS, Ko JM, Choi MS, Kang BK, Kwon SW, Jun TH. Quantitative trait locus analysis for soybean (Glycine max) seed protein and oil concentrations using selected breeding populations. Plant Breeding. 2019;138(1):95-104.

Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135-1145.

Shook JM, Zhang J, Jones SE, Singh A, Diers BW, Singh AK. Meta-GWAS for quantitative trait loci identification in soybean. G3. 2021 Jul;11(7):jkab117.

Siczek A, Lipiec J. Soybean nodulation and nitrogen fixation in response to soil compaction and surface straw mulching. Soil and Tillage Research. 2011;114(1):50-56.

Sinclair TR, Rufty TW, Lewis RS. Increasing photosynthesis: unlikely solution for world food problem. Trends in Plant Science. 2019;24(11):1032-1039.

Sinclair TR. Improved carbon and nitrogen assimilation for increased yield. Soybeans: Improvement, production, and uses. 2004;16:537-568.

Smith LH. Seed development, metabolism, and composition. In: Tesar MB (Ed.). Physiological bases of crop growth and development. Madison: American Society of Agronomy: Crop Science Society of America: Soil Science Society of America. 1984. p.13-52.

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant biotechnology journal. 2015 Feb;13(2):211-21.

Tanya P, Srinives P, Toojinda T, Vanavichit A, Lee S. Identification of SSR Markers Associated with N2-Fixation Components in Soybean [Glycine max (L.) Merr.]. Korean Journal of Genetics. 2005;27(4):351.

Telles TS, Nogueira MA, Hungria M. Economic value of biological nitrogen fixation in soybean crops in Brazil. Environmental Technology & Innovation. 2023;103158.

Thorne JC, Fehr WR. Incorporation of high-protein, exotic germplasm into soybean populations by 2-and 3-way crosses 1. Crop Science. 1970;10(6):652-655.

Thu SW, Lu MZ, Carter AM, Collier R, Gandin A, Sitton CC, Tegeder M. Role of ureides in source-to-sink transport of photoassimilates in non-fixing soybean. Journal of Experimental Botany. 2020;71(15):4495-4511.

Torkamaneh D, Lemay MA, Belzile F. The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. Plant biotechnology journal. 2021 Sep;19(9):1852-62.

Torkamaneh D, Chalifour FP, Beauchamp CJ, Agrama H, Boahen S, Maaroufi H, et al. Genome-wide association analyses reveal the genetic basis of biomass accumulation under symbiotic nitrogen fixation in African soybean. Theoretical and Applied Genetics. 2020;133:665-676.

Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC bioinformatics. 2017;18(1):1-7.

Torres AR, Grunvald AK, Martins TB, Santos MAD, Lemos NG, Silva LAS, Hungria M. Genetic structure and diversity of soybean germplasm considering biological nitrogen fixation and protein content. Scientia Agricola. 2015;72:47-52..

Trněný O, Vlk D, Macková E, Matoušková M, Řepková J, Nedělník J, Hofbauer J, Vejražka K, Jakešová H, Jansa J, Piálek L. Allelic variants for candidate nitrogen fixation genes revealed by sequencing in red clover (Trifolium

pratense L.). International Journal of Molecular Sciences. 2019 Nov 2;20(21):5470.

United States Departament of Agriculture – USDA. Soybean Explorer - United States (usda.gov). Available from: <https://soybeanexplorer.usda.gov/>. Accessed on: 2023-04-25.

United States Soybean Export Council - USSEC. Available from: <https://ussec.org/>. Accessed on: 2021-08-20.

Van K, McHale LK. Meta-analyses of QTLs associated with protein and oil contents and compositions in soybean [Glycine max (L.) Merr.] seed. Int J Mol Sci. 2017;18(6):1180.

Vargas MAT, Hungria M. Fixação biológica do nitrogênio na cultura da soja. In: Biologia dos solos dos cerrados. 1997.

Vest G. Rj3 A gene conditioning ineffective nodulation in soybean 1. Crop Sci. 1970;10(1):34-35.

Vest G, Caldwell BE. Rj4—A gene conditioning ineffective nodulation in soybean 1. Crop Sci. 1972;12(5):692-693.

Vollmann J, Fritz CN, Wagentristl H, Ruckenbauer P. Environmental and genetic variation of soybean seed protein content under Central European growing conditions. J Sci Food Agric. 2000;80(9):1300-1306.

Vuong TD, Nickell CD, Harper JE. Genetic and allelism analyses of hypernodulation soybean mutants from two genetic backgrounds. Crop Sci. 1996;36(5):1153-1158.

Wang J, et al. Genetic mapping high protein content QTL from soybean 'Nanxiadou 25' and candidate gene analysis. BMC Plant Biol. 2021;21(1):1-13.

Wang L, et al. Use of CRISPR/Cas9 for symbiotic nitrogen fixation research in legumes. Prog Mol Biol Transl Sci. 2017;149:187-213.

Warembourg FR, Fernandez MP. Distribution and remobilization of symbiotically fixed nitrogen in soybean (Glycine max). Physiol Plant. 1985;65(3):281-286.

Wilcox JR. Soybean protein and oil quality. In: IV World Soybean Research Conference. 1989.

Wilcox JR, Cavins JF. Backcrossing high seed protein to a soybean cultivar. Crop Sci. 1995;35(4):1036-1041.

Williams LF, Lynch DL. Inheritance of a non-nodulating character in the soybean. Agron J. 1954;46(1):28-29.

Wysmierski PT, Vello NA. The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. Genet Mol Biol. 2013;36(4):547-555.

Yang Q, et al. Genetic analysis and mapping of QTLs for soybean biological nitrogen fixation traits under varied field conditions. Front Plant Sci. 2019;10:75.

Yang Y, et al. Characterization of genetic basis on synergistic interactions between root architecture and biological nitrogen fixation in soybean. Front Plant Sci. 2017;8:1466.

Yao Y, et al. Quantitative trait loci analysis of seed oil content and composition of wild and cultivated soybean. BMC Plant Biol. 2020;20(1):1-13.

Yu J, Buckler ES. Genetic association mapping and genome organization of maize. Curr Opin Biotechnol. 2006;17(2):155-160.

Zapata F, Danso F, Hardarson G, Fried M. Nitrogen Fixation and Translocation in Field-Grown Fababean1. Agron J. 1987;79(3):505-509.

Zhang S, et al. Linkage and association study discovered loci and candidate genes for glycinin and β-conglycinin in soybean (Glycine max L. Merr.). Theor Appl Genet. 2021;134:1201-1215.

Capítulo 2

Title

Unraveling Genomic Regions Controlling Biological Nitrogen Fixation and Protein Content in Soybean based on Genome-wide Association Study (GWAS)

Journal

Molecular Breeding (A2)

Authors

Paloma Helena da Silva Liborio[a], Mariangela Hungria[b], Antonio Eduardo Pípolo[b], Francismar Corrêa Marcelino Guimarães[b], Ivani de Oliveira Negrão Lopes[b], Flávia Raquel Bender[d], , Vincent-Thomas Boucher St-Amour[c], François Belzile[c], Marco Antonio-Nogueira[b]*

Affiliations

[a]Universidade Estadual de Londrina, PO Box 10.011, 86057-970, Londrina, PR, Brazil.
[b]Embrapa Soja, PO Box 231, 86001-970, Londrina, PR, Brazil.
[c]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec City, QC, Canada.
[d]Tropical Melhoramento & Genética (TMG), Celso Garcia Road, Cambe, PR, Brazil.

E-mail addresses

Liborio, P.H.S. paloma_liborio@hotmail.com
Hungria, M. mariangela.hungria@embrapa.br
Pípolo, A.E. antonio.pipolo@embrapa.br
Marcelino-Guimarães, F.C. francismar.marcelino@embrapa.br
Lopes, I.O.N. ivani.negrao@embrapa.br
Bender, F.R. flaviaabender@gmail.com
St-Amour, V.T.B. vtbos@ulaval.ca
Belzile, F. francois.belzile@fsaa.ulaval.ca
Nogueira, MA. marco.nogueira@embrapa.br

Correspondence

*Marco Antonio Nogueira
Embrapa Soja, PO Box 231, 86001-970, Londrina, PR, Brazil. Phone: +55 43 3371-6215; E-mail: marco.nogueira@embrapa.br

Unraveling Genomic Regions Controlling Biological Nitrogen Fixation and Protein Content in Soybean based on Genome-wide Association Study (GWAS)

Abstract

Increasing yield and at the same time increase or keep the protein content in grains is one of the major challenges for soybean (*Glycine max* L.) breeding programs. Nitrogen is highly demanded for development and formation of protein, and in soybean is mainly supplied by biological nitrogen fixation (BNF). Our objective was to identify regions in the genome involved in controlling BNF and protein content by performing a genome-wide association study. For that, 200 soybean accessions were initially selected, but during the association analyses, accessions that did not meet the requirements were removed. Thus, for the GWAS a modest but high-quality control panel containing 100 soybean accessions involving genetic lines, commercial cultivars, and plant introduction (PI) dated from 1948 to 2019 was used. Experiments were conducted under field conditions in the 2019/2020 and 2020/2021 crop seasons in Ponta Grossa-PR and Londrina-PR, respectively. The phenotypic traits evaluated were: shoot dry weight (g per plant); root dry weight (g per plant) and their ratio; number of nodules (number per plant); nodule dry weight (mg per plant); nodules specific weight (mg per nodule); N concentration (g kg^{-1}); mass of thousand grains (g); protein content in grains (%), oil content in grains (%), ureide content in petioles – UR (μmol g^{-1}) expected protein in the meal – EPM (%) and protein content (%). Due to absence of genotype-environment interaction, data from the two experiments were analyzed together. The 100 soybean accessions were genotyped by Sequencing Genotyping. The identification of candidate genes was based on significant SNPs for five traits (oil content, protein content in grains, nodule dry weight, nitrogen content and root dry weight). Candidate genes were associated with the five traits related to BNF and protein content in grains. Some genomic regions associated with the traits had already been reported by other studies, however, most of the found regions are new, being candidates to be confirmed in the control of these traits in future studies.

Key-words: *Glycine max*; *Bradyrhizobium*, *Azospirillum*, SNP, soybean high protein content

1. Introduction

Soybean (*Glycine max* L.) best illustrates the successful use of symbiosis with *Bradyrhizobium* for N supply in a rational and economical way. Biological nitrogen fixation (BNF) is largely responsible for soybean expansion in Brazil, the biggest world producer. In the 2022/23 crop season, the average Brazilian soybean yield was 3,527 kg ha^{-1} in an area of 43.5 Mha (CONAB, 2023).

The average protein and oil contents in the Brazilian soybean grains are 37.6% and 22.2%, respectively (Lorini, 2018). Despite that, obtaining grains with high protein contents has been challenging for soybean breeding programs. Zhang et al. (2018) demonstrated the impact of domestication and selection on the decrease of protein contents in soybean grains. One of the possible causes is the massive use of a few high-yielding genotypes, forming a narrow genetic basis. In addition, soybean grain protein content is a quality trait controlled by multiple genes (Wang et al. 2021).

Legumes are reported to fix about 22 million metric tons of N annually (Peoples et al. 2009; Ferguson et al. 2010). Some studies have focused on selecting elite germplasm with greater BNF ability in different legumes, and their results have shown that there are significant differences in BNF capacity among genotypes (Dwivedi et al. 2015). Considering the increasing demand for soybean protein, identifying molecular markers associated with quantitative trait loci (QTL) controlling grain yield, and protein and oil contents is required for breaking the negative correlations between these traits (Wang et al. 2021). The lack of a simple, rapid, and economical methodology to evaluate BNF traits has limited the genetic breeding progress. In addition, breeders generally select the best grain yield potential whereas minor attention has been given to protein or oil contents (Lee et al., 2019).

It has been hypothesized that domestication and breeding processes may have led to loss of genes related to the symbiosis with the N-fixing rhizobia, reducing the ability of soybean to achieve the maximum BNF potential. However, testing this hypothesis requires the characterization of BNF in many populations, and only few studies have been dedicated to this important trait (Liu et al., 2020). QTLs related to BNF coincide with the QTLs for protein content in grains (Torres et al., 2015). These complex relationships have not been completely elucidated yet, and more studies are requested to achieve markers that can be used in marker-assisted selection.

Genome-wide association studies (GWAS) have shown excellent results with single nucleotide polymorphisms (SNPs) associated with phenotypic traits of interest (Zhang et al., 2015; Torkamaneh et a., 2020; Wang et al., 2020; Chao et al., 2021). Elite bacterial germplasm has been selected, however, the genetic basis of BNF still needs to be further investigated. The progress has been hampered due to the complexity of evaluating BNF traits. Protein content and BNF are traits highly influenced by genotypes (G), environment (E), and their interaction (G × E). In these situations, the explanation of variations based only on phenotypic data is a difficult task and, in many cases, not conclusive. Thus, obtaining SNPs and bioinformatics tools for analyzing population structure and associating markers with phenotypes through GWAS studies are promising tools to support soybean breeders.

Torkamaneh et al. (2020) identified limitations in using only a single strain (*Bradyrizobium japonicum* 532C) in the phenotyping of BNF traits for GWAS. The authors emphasize the difficulty to cover all possible host genotypes due to plant × bacteria interaction. Burghardt et al. (2018) reported that different strains and mixed

inoculants exhibited different interactions with soybean genotypes, resulting in different phenotypes for BNF traits, what make hard to establish an association with QTL related to BNF. Co-inoculation of *Bradyrhizobium* and *Azospirillum* promotes an earlier and more abundant nodulation in soybean, resulting in more consistent traits on soybean BNF, including grain yield (Hungria et al., 2013) and protein contents in grains (Liborio et al., 2021)

We hypothesize that more promising symbioses may be associated with a higher protein content in grains. Additionally, if these regions are in linkage disequilibrium, traits that favor nodulation and protein content in soybeans can be inherited together. Therefore, we utilized phenotypic data on biological nitrogen fixation (BNF) and protein content traits from soybeans co-inoculated with Bradyrhizobium and Azospirillum under field conditions for GWAS.

2. Material and Methods

2.1 Soybean accessions

To carry out the study, 200 soybean accessions were selected based on protein and oil contents in grains, and BNF capacity, when this information was available. Accessions were selected from the Soybean Germplasm Collection, USDA-ARS, based on data from the Germplasm Resource Information Network - GRIN (Information Network on Germplasm Resources) (https://npgsweb.ars-grin.gov/gringlobal/search) and obtained from the Active Germplasm Bank (AGB) of Embrapa Soja, Londrina-PR.

The selected soybean accessions were initially tested for the ability to generate viable seeds to be sown in the field for phenotypic and genotypic analysis. Thus, accessions that did not meet the phenotypic analysis were excluded of the association analysis. Thus, for GWAS a lower but high-quality control panel including 100 soybean accessions including genetic lines, commercial cultivars, and plant introduction (PI) dated from 1948 to 2019 was used (Table 1).

Table 1. Soybean panel (*Glycine max*) containing the accession denomination, collection origin, breeding status, technology, maturity group, growth habit, and year of register/release.

Accession	Collection	Year*
41S31	Australia	1952[3]
Haberlandt	Australia	1980[3]
Melrose	Australia	1999[3]
17X-1337-173-7 (195)	Brazil	-
17X-1337-25-3 (306)	Brazil	-
17X-1337-25-3 (199)	Brazil	-
17X-1345-151-11 (178)	Brazil	-
17X-1345-151-11 (298)	Brazil	-
17X-1345-234-6	Brazil	-
17X-1345-234-6 (70)	Brazil	-
17X-1354-109-8 (114)	Brazil	-
17X-1354-98-3 (109)	Brazil	-
17X-1354-98-3 (92)	Brazil	-
A014	Brazil	2015[1]
A015	Brazil	2013[1]
A016	Brazil	2015[1]
A017	Brazil	2012[2]
A018	Brazil	2011[1]
A019	Brazil	2016[1]
A020	Brazil	2008[1]
A021	Brazil	2013[1]

BR-29	Brazil	1988[2]
BR-3	Brazil	1977[2]
A024	Brazil	2014[1]
A025	Brazil	2015[1]
A026	Brazil	2015[1]
A027	Brazil	2015[1]
A028	Brazil	2015[1]
A029	Brazil	2007[1]
A030	Brazil	2019[1]
A031	Brazil	2015[1]
A032	Brazil	2015[1]
A033	Brazil	2014[1]
A034	Brazil	2014[1]
A035	Brazil	2014[1]
A036	Brazil	2014[1]
A037	Brazil	2015[1]
A038	Brazil	2015[1]
A039	Brazil	-
A040	Brazil	2014[1]
A041	Brazil	2013[2]
A042	Brazil	2013[2]
A043	Brazil	2013[2]
Embrapa 48	Brazil	1996[2]
A045	Brazil	2018[1]
A046	Brazil	2016[1]
A047	Brazil	2012[1]
A048	Brazil	2017[1]
A049	Brazil	2015[1]
A050	Brazil	2017[1]
A051	Brazil	2017[1]
A052	Brazil	2014[2]
A053	Brazil	2015[1]
A054	Brazil	2011[1]
A055	Brazil	2014[1]
A056	Brazil	2013[1]
A057	Brazil	2013[1]
A058	Brazil	2011[1]
A059	Brazil	2013[1]
A060	Brazil	2009[1]
A061	Brazil	2017[2]
A062	Brazil	2016[1]
A063	Brazil	2015[2]
A064	Brazil	2016[2]
A065	Brazil	-
A066	Brazil	-
A067	Brazil	2017[2]
A068	Brazil	-
A069	Brazil	2016[2]
A070	Brazil	2018[2]
A071	Brazil	2016[2]
A072	Brazil	2016[2]
A073	Brazil	2016[1]
A074	Brazil	-
A075	Brazil	2011[1]
A076	Brazil	2017[2]
(Long zhou dong feng dou)	China	1993[3]
Da bai shui dou No. 1	China	1996[3]
Ding an qing pi dou	China	-
Jing huang 18	China	1994[3]
Nanking 332	China	1948[3]

Pan-San	China	1948[3]
PI 594811	China	1996[3]
S-100	China	-
Shu yang hong mao qiu yi	China	2000[3]
Davis	EUA	1967[3]
Hampton	EUA	1978[3]
PI 281888	Indonesia	1962[3]
Okute	Japan	1952[3]
Shimotsu Ura	Japan	1952[3]
Yonekadake	Japan	1952[3]
E.G. 1	Philippines	1957[3]
KAERI 630-8	South Korea	1976[3]
KAS 390-4	South Korea	1975[3]
Ringgit strain 19 51	Suriname	1952[3]
Ringgit strain 23 51	Suriname	1952[3]
Hernon	Tanzania	1969[3]
Tung Tam	Thailand	1953[3]
USDA- ARD A	Thailand	1957[3]
PI 331795	Vietnam	1968[3]

* Year= [1]Register; [2]Release or [3]Received by NPGS-GRIN USDA.

2.2 Multiplication of accessions in the greenhouse and DNA extraction, library preparation, and genotyping-by-sequencing (GBS).

The accessions were multiplied in the greenhouse in 5 L pots containing soil taken from the 0-20 cm top layer of an agricultural soil. The experimental design was completely randomized, with three replications. Each pot was sown with 3 seeds, which were thinned out at 15 days to one plant per pot. Plants were then grown up to 30 days after germination, when the trefoil with just expanded leaflets of each accession with/without? petioles was collected and immediately frozen in liquid N2 for genomic DNA extraction Genomic DNA was extracted from 100 mg of trefoils of each accession using the DNeasy® Plant Mini Kit (Qiagen) according to the manufacturer's instructions and subsequently quantified in a Qubit® fluorometer (Thermo Fisher Scientific). DNA integrity was checked by electrophoresis in an agarose gel (1%), followed by quantification in a NanoDrop® ND-1000 spectrophotometer (Uniscience) and diluted with water to a concentration of 10 ng/μL.

The 96 well plates with the samples were sent to sequencing at the *Institut de Biologie Intégrative et des Systèmes* (IBIS), at the University of Laval, Quebec, QC, Canada. GBS libraries were produced using the protocol described by Elshire et al. (2011) and modified by Sonah et al. (2013). Briefly, DNA was digested with the ApeKI enzyme, followed by the ligation of adapters with barcodes in a pool of 96 samples per library, and sequenced in an Ion Torrent® sequencer (Thermo Fisher Scientific).

2.3 Field experiments

Two field experiments were carried out: in the 2019/20 crop season in Ponta Grossa-PR (25°09'31"S, 50°04'22"W, 886 m altitude) and repeated in the 2020/21 crop season in Londrina-PR (23°11'37" S, 51°11'03" W, 630 m altitude).

Sowings were carried out in a no-tillage system on wheat (*Triticum aestivum*) straw in Ponta Grossa and on corn (*Zea mays*) straw in Londrina. Plots in the experimental fields were previously randomized with the SAS software (SAS Inst. Inc., Cary, NC). Plots were arranged according to an incomplete balanced block design with a relative efficiency of 96.7%. The plots consisted of four lines of 3 m (Ponta Grossa) or 5 m (Londrina) in length,

spaced 0.5 m apart. Seeds were co-inoculated with liquid commercial inoculants containing *Bradyrhizobium* spp. and *Azospirillum brasilense* applied in the sowing furrow using a device coupled to a mechanized seeder-fertilizer, according to the manufacturer's recommendations.

The soil in Ponta Grossa is classified as Rhodic Hapludox with a clayey texture; in Londrina the soil is classified as Rhodic Eutrudox, with a very clayey texture according to Soil Taxonomy (USDA) (Soil Survei Saff, 1973). The results of soil chemical analyses are shown in Table 2. The climate in Ponta Grossa-PR is classified as Cfb (mesothermal and temperate), while in Londrina-PR is Cfa (mesothermal and subtropical), according to Köppen-Geiger (Caviglione et al., 2000).

Rainfall and temperature during the field trials were provided by the Agrometeorology Laboratory of Embrapa Soja, Londrina-PR. In the 2019/20 crop season (Ponta Grossa) the accumulated precipitation between December and May was 419.4 mm, while the average temperature was 20.4 °C. Potential evapotranspiration (PET) was higher than rainfall in the late December and early March, as well as occurrence of higher temperatures. In the 2020/21 crop season (Londrina) the accumulated precipitation between October and April was 663.3 mm, while the average temperature was 23.2°C. The PET was higher in early October and March when lower rainfall and higher temperatures were recorded (Figure 1).

Table 2. Soil chemical attributes at the 0-20 cm topsoil layer of the Ponta Grossa (2019/20) and Londrina (2020/2021) experimental fields before sowing.

Locality	P - resin	OM	pH	K	Ca	Mg	H+Al	Al	SB	S-SO_4^{2-}	CEC	V	Ca/CEC	Mg/CEC	m	B	Cu	Fe	Mn	Zn
	mg dm^{-3}	g dm^{-3}	---------- mmolc dm^{-3} ----------									%					---------- mg dm^{-3} ----------			
Ponta Grossa	35	17.0	5.0	1.9	13.0	7.0	22.0	1.0	21.9	12.0	43.9	50.0	30.0	16.0	4.0	0.3	0.8	50.0	35.2	0.9
Londrina	46	20.0	4.6	2.9	17.0	6.0	47.0	2.0	25.9	8.0	72.9	36.0	23.0	8.0	7.0	0.2	1.6	48.0	2.2	1.1

Extractors: Resin (P, K^+, B, Cu, Fe, Mn, Zn); 1M KCl (Ca^{+2}, Mg^{+2}, Al^{+3}); SMP (H + Al); CEC: Cation exchange capacity at pH 7.0; V: Base saturation.

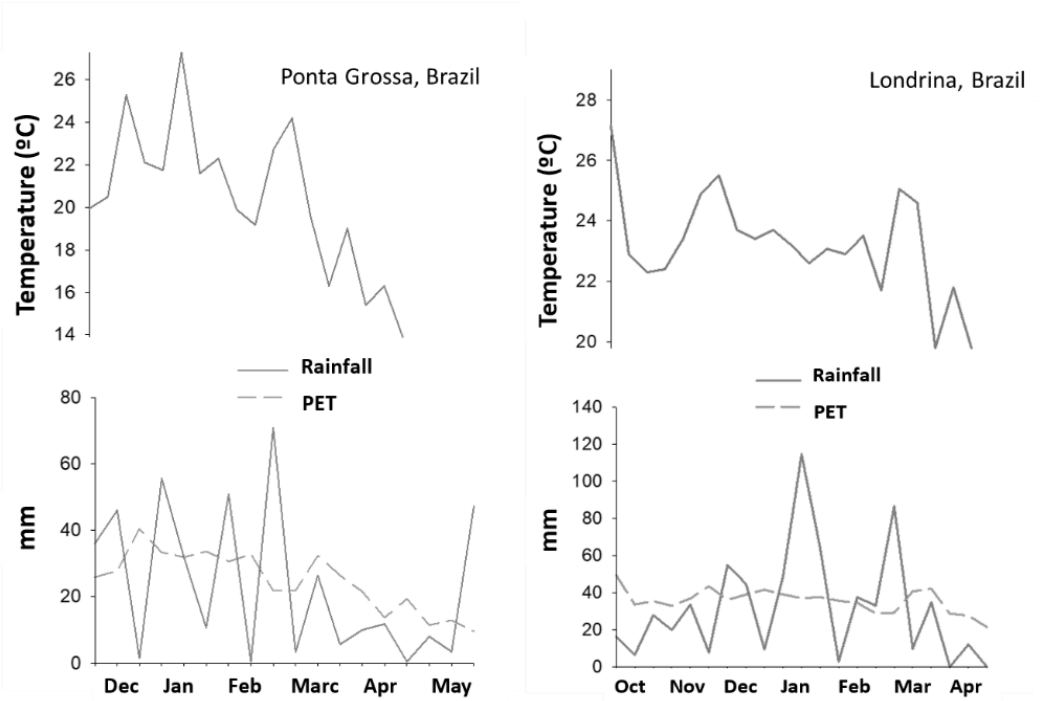

Figure 1. Rainfall (mm), potential evapotranspiration - PET (mm), and temperature (°C). Ponta Grossa-PR, Brazil, 2019/20 crop season (left) Londrina-PR, Brazil, 2020/21 crop season (right).

2.4 Phenotyping

For phenotyping of the accessions five plants were taken per plot, and assessed for the following traits: shoot dry weight – SDW (g per plant); root dry weight – RDW (g per plant); SDW/RDW ratio; number of nodules per plant - NN; nodule dry weight - NDW (mg per plant); nodules specific weight – NN/NDW (mg per nodule); N concentration in the shoot (g kg^{-1}); mass of thousand grains – MTG (g); protein – PROT (%) and oil contents in grains - OIL (%) (Nicolai et al., 2007), ureide content in the petioles – UR (µmol g^{-1}), expected protein in meal – EPM (%) (Pípolo et al., 2015), and protein content PROT (%).

2.5 Experimental design and statistical analyses of the phenotypical data

The phenotypes of the soybean cultivars were evaluated through an experiment conducted using a Balanced Incomplete Block Design (BIBD) across two distinct environments. The design was created using the optex procedure within the SAS/STAT software, Version 9.4®, copyrighted by SAS Institute Inc. (2016). With a total of 191 genotypes, the design selection was based on the Bayesian optimal design criteria (DuMouchel, 1994). This selection process involved exploring design variations in a grid with block counts ranging from 30 to 48 and plots per block varying between 8 and 12. The final configuration, comprising 44 blocks with 12 plots each, was chosen for its optimal balance, achieving the lowest total plot count while maximizing efficiency. As a result, the consolidated dataset encompassed 1,056 rows and 16 columns, representing three descriptive attributes (Environment, Genotype, and Block), alongside 13 response attributes (phenotypes).

The statistical analyses utilized generalized mixed models, a versatile collection of methodologies adept at capturing relationships between response attributes and a combination of fixed and random effects within descriptive attributes. These models were particularly suitable for this study, given the distinctive interactions observed between response attributes and descriptive attributes. Consequently, tailored model formulations were employed. A comprehensive overview of the adopted strategies is provided in Table 3 for reference. These analyses were performed using the SAS/STAT software, utilizing the glimmix procedure for model estimation and obtaining phenotypic estimates through the lsmeans statement, incorporating the blup and ilink options (Table 3).

Table 3. Summary of approaches applied in the modeling of phenotypic attributes. G=Genotype; E=Environment; B=Block.

Phenotype	Random effect*	Fixed effects	Statistical distribution
OIL - Oil content (%)	B(E)	G	Gaussian
PROT - Protein content (%)			
NDW - Nodule dry weight (mg pl^{-1})			
RDW - Root dry weight (g pl^{-1})	B(E)	G + E + GE	Gamma
SDW - Shoot dry weight (g pl^{-1})			
TSW - Thousand seed weight (g pl^{-1})			
N Content - Nitrogen content (g kg pl^{-1})	-		
MG - Maturation group (*No meaningful variability*)	-	-	-

*Random effect statement: *random intercept/subject=B(E) G s type=cs*.

2.6 Analysis of GBS sequences.

The SNPs were called using the Fast-GBS pipeline (Torkamaneh et al., 2017) and aligned against the soybean Williams 82 reference genome - Glyma.Wm82.a2 (Gmax_275_v2). Heterozygous genotypes were replaced by missing data, and any accession with >80% of missing data were removed from the dataset. The imputation of missing genotypic data was supported by the software BEAGLE v4.1 (Browning and Browning 2007).

2.7 GBS data quality control

Filters were applied on the GBS dataset obtained from the 100 accessions containing a total of 39,000 SNPs using VCFtools. SNPs without a known physical position on any of the 20 soybean chromosomes were excluded from the analyses. Data were filtered for removal of accessions with excess heterozygosity, removal of indels, minor allele frequency (MAF) ≥ 0.05, and those existing in minor states as only two alleles were segregating at each SNP locus. Heterozygous SNPs were also treated as absent since they were rare (< 2%). After the filtering steps, 36,200 high-quality SNPs were kept for further analysis.

2.8 GmHapMap imputation and filtering

Global HapMap (GmHapMap) provides a unique worldwide resource for soybean genomics and breeding. This is a worldwide haplotype map for soybean constructed using Whole Genome Sequencing (WGS) sequences from 1007 soybean (*G. max*) individuals (Torkamaneh et al., 2021). Untyped variants from GmHapMap were imputed in the association panel. The imputation was made separately chromosome by chromosome. After imputation with GmHapMap, the panel had 2,078,660 SNPs. The software BEAGLE v4.1 (Browning, 2016) was used for GMHapMap imputation.

After this procedure, the filter (MAF) ≥ 0.05 was applied, resulting in 1,924,227 SNPs to be submitted to GWAS. Thus, variants with an allele frequency of less than 5% in the study population will be excluded from the analysis. This helps to focus on the most common and relevant variants, which are most likely to be associated with traits.

2.9 Pruning for Structure analysis and GWAS

Pruning is employed for systematic removal of SNPs with high linkage disequilibrium (LD) each other before the genetic analysis, aiming to select a subset of SNPs that capture a representative portion of the genetic diversity. Pruning assists to mitigate issues such as multicollinearity and reduces the computational complexity in downstream analyses, ensuring a more efficient and meaningful exploration of genetic associations. For that, PLINK 1.9 software (Chang etal.,2015) was used, based on the following parameters: window size in SNPs = 200, number of SNPs to shift the window at each step = 10, and threshold (r^2) = pairwise SNP-SNP (genotype correlation) = 0.3.

After this, step MAF 5% was applied again, and kept 16,187 SNPs for population structure and GWAS analyses.

2.10 Population structure analysis

Population structure was estimated using a variational Bayesian inference implemented using the fastSTRUCTURE software (Raj et al. 2014). The most likely K value was determined by the natural log probability of the data (LnP(D)) and delta K, based on the rate of change in LnP(D) between successive K values. Admixture proportion plot of 100 soybean accessions was plotted using Distruct v2.3.

2.11 GWAS analysis

The association analysis between phenotypes and SNPs markers for the 13 traits related to BNF and protein content in grains was carried out using the R software, using the GAPIT – Genome Association and Prediction Integrated Tool package (Lipka et al., 2012) to verify the significant associations between the loci and the traits of interest, the significance levels between the markers and the variables. The analyses of the blocks in linkage disequilibrium were estimated using the (r2) by the software TASSEL 5.2 (Glaubitz et al., 2014).

For the GWAS, the models CLM, GLM, MLM, MLMM, CMLM, BLINK and FarmCPU were used. CLM (Composite Likelihood Method), analyzes interactions between genetic loci using composite likelihood. GLM (Generalized Linear Model), models relationships between variables, accommodating various distributions. MLM

(Mixed Linear Model), considers population structure and relatedness in genomic studies. MLMM (Multi-Locus Mixed Model), simultaneously models multiple loci interactions in GWAS. CMLM (Compressed Multi-Locus Mixed Model), captures interactions between loci with reduced complexity. BLINK (Linkage-disequilibrium Iteratively Nested Keyway), identifies genetic variants and interactions in GWAS using Bayesian approach. FarmCPU (Fixed and random model Circulating Probability Unification), addresses population structure and relatedness with fixed and random effects in GWAS using Bayesian approach (Lipka et al., 2012).

2.12 Identification of Candidate Genes

The search for candidate genes was based on significant SNPs from sequences available in the SoyBase (http://www.soybase.org) and Phytozome (http://phytozome.jgi.doe.gov) databases. Only candidate genes closest to the significant SNP locus were considered.

3. Results

3.1 Population structure

A panel of 100 soybeans co-inoculated with *Bradyrhizobium* spp. and *Azospirillum brasilense* evaluated under field conditions was used for GWAS. The panel was 73% composed of accessions from Brazil and 27% from other countries, encompassing maturity groups from 5 to 8 (see Table 1). The structure analysis of the accessions based on fastSTRUCTURE revealed five populations (K=5), where everyone (from 1 to 100) is represented by a single vertical line and each color represents one cluster (Figure 2).

Populations 1, 2, 4, and 5 predominantly consist of accessions from Brazil, while population 3 contains accessions from different countries. Population 1 mostly consists of Brazilian commercial cultivars containing IPRO technology, but also includes some conventional non-GM cultivars. Population 2 includes accessions comprising lines and commercial cultivars (conventional, RR and IPRO) and one cultivar from EUA. Population 4 is composed of older cultivars from Brazil and one from the United States. Population 5 consists only of RR and IPRO GM-cultivars. Population 3 consists of a small group, but a diverse set of accessions from Oceania (Australia), Southeast Asia (Thailand, Vietnam, and Indonesia), East Asia (China, Japan, and South Korea), South America (Suriname), and East Africa (Tanzania). Populations 2 and 5 exhibit admixtures, likely attributed to the limited number of parentals used in the genetic crosses in Brazilian commercial cultivars (Figure 2).

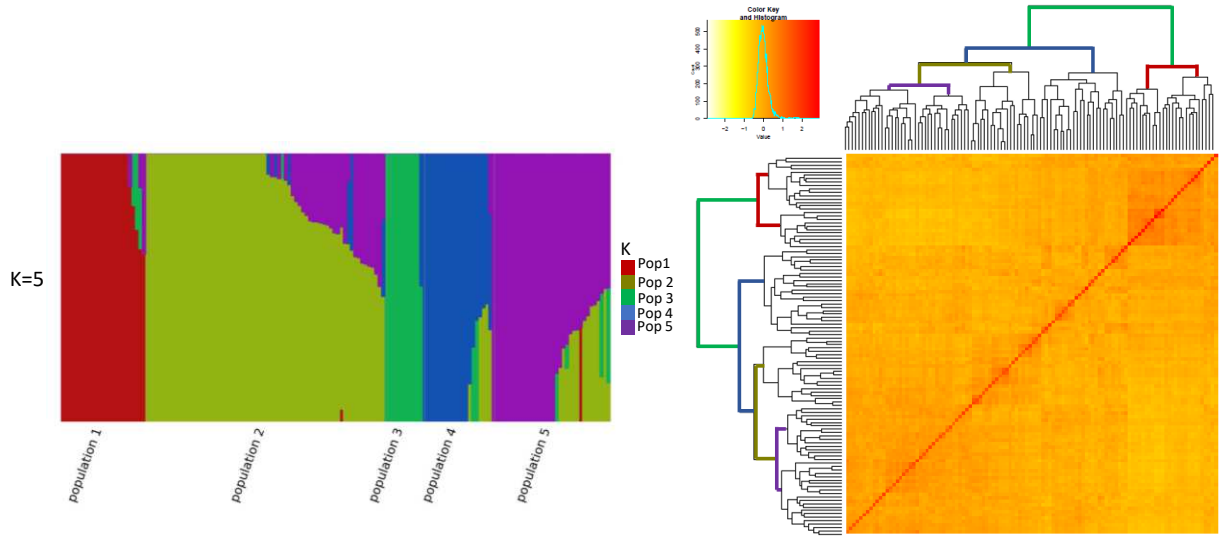

Figure 2. Population structure in a panel consisting of 100 selected soybean accessions (*Glycine max*). Results are based on fastSTRUCTURE and Distruct v2.3.

The traits protein and oil contents in grains, nodule dry weight, N content in shoots, and root dry weight showed significant associations in GWAS, and their phenotypic distributions are shown in Figure 3. There was high phenotypic variability among accessions. Furthermore, significant correlations were observed between oil (OIL%) and protein (PROT%) contents in grains (-0.90**), Nodule dry weight (NDW) and OIL% (-0.40*), NDW and PROT% (0.35**), root dry weight (RDW) and OIL% (-0.30**), RDW and NDW (0.52**).

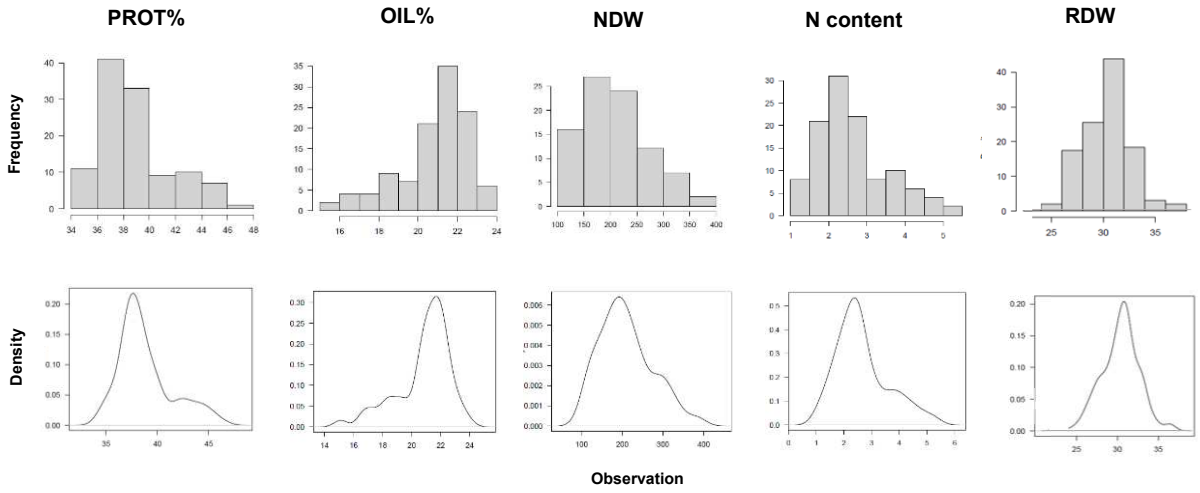

Figure 3. Phenotypic distribution of five traits evaluated for Genome wide-association studies (GWAS) in a panel containing 100 soybean accessions (*Glycine max*) derived from the analysis of two field trials (Ponta Grosa 2020/21 and Londrina 2021/22). Left to right: protein content in grains (PROT%); Oil content in grains (OIL%); Nodule dry weight (NDW); Nitrogen content in the plant shoots (N content) and Root dry weight (RDW).

Accession registration or year of release (when available) was arranged in regression analyses for protein and oil contents in grains. A stagnant trend in protein contents can be inferred within the panel used for GWAS containing accessions released from 1948 to 2019 (Figure 4a). Some accessions from population 3 (originated from different countries) showed protein content ≥ 45%, and were registered or released from 1952 to 1994. In the context of protein content ≤ 36%, some accessions from population 1 had the lowest protein content in grains and were registered or released from 2014 to 2018 (Figure 4a). The accession Jing huang 18 showed the highest protein content (PROT= 45.46%; OIL%=17.43%; year = 1994) (Figure 4a).

For oil content, there is a trend to increase over the years within the panel containing accessions released from 1948 to 2019 (Figure 4b). The cultivars from populations 1 and 5 identified in the structure analysis containing RR and IPRO technologies showed higher oil contents ≥ 23%, and were registered or released from 2012 to 2018. The accession A045 had the highest oil content (OIL =23.96% and PROT= 34.82%, year = 2018) (Figure 4b).

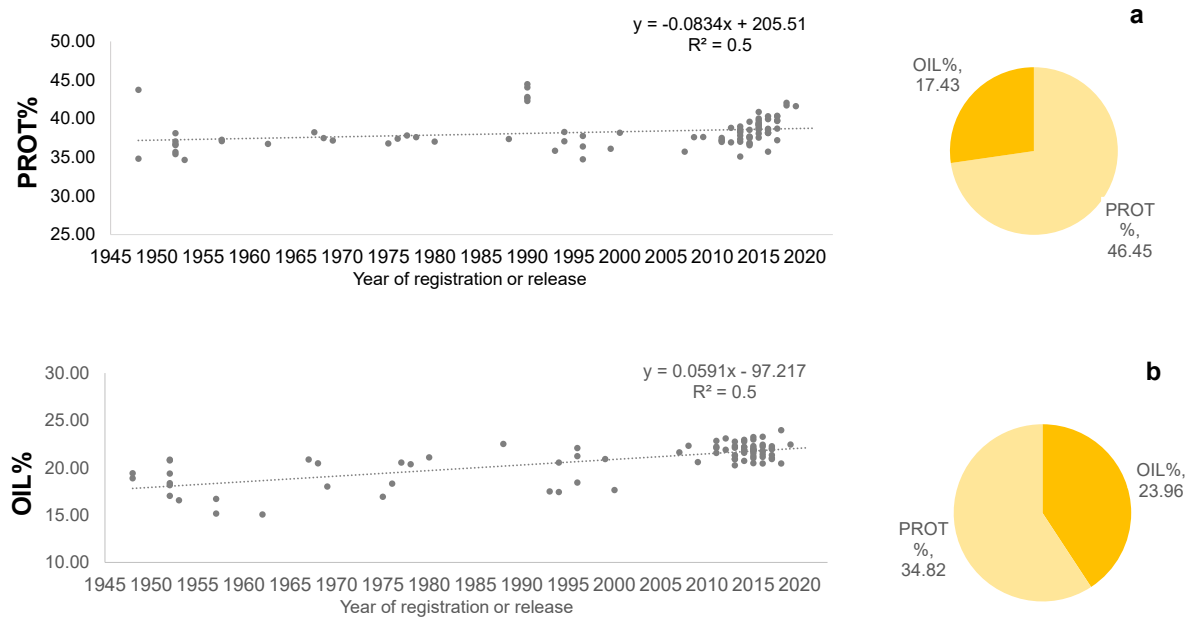

Figure 4. a. (Left) Protein content (PROT%) in soybean grains × Year of registration or release from 1948 to 2019 for 100 accessions of soybean (*Glycine max*) and (Right) plot of the accession with the highest protein content in grains (Jing huang 18); b. (Left) Oil content (OIL%) in soybean grains × Year of registration or release from 1948 to 2019 for 100 accessions of soybean (*Glycine max*) and (Right) plot of the accession with the highest oil content in grains (A045).

The pool of 100 accessions composing the panel for the GWAS exhibited a clear downward trend ($r^2 = 0.8$) on NDW over the 70 years of release years from 1948 to 2019 (Figure 5). In agreement with the protein content in grains, accessions belonging to population 3 generally presented higher NDW (760.5 to 292.2 mg per plant). On the other hand, accessions belonging to group 5 showed in general lower NDW (287.0 to 93.5 mg per plant).

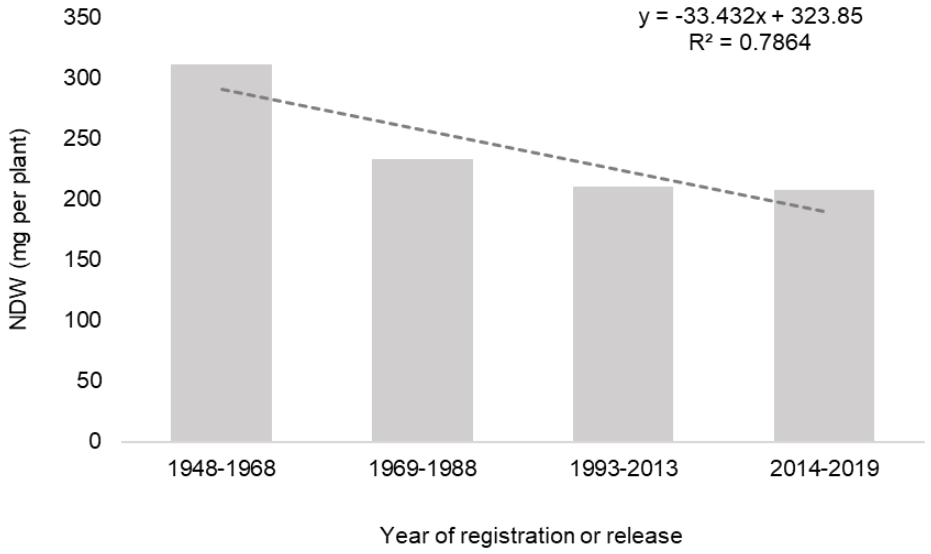

Figure 5. Nodule dry weight (NDW) × year of registration or release from 1948 to 2019 of 100 selected soybean accessions to compose a panel for Genome wide-association studies (GWAS) for protein content in grains and biological nitrogen fixation traits.

3.2 GWAS

To identify genomic regions that are involved in the control of protein content in grains and BNF in soybean, a GWAS was performed based on eight models in our panel of 100 accessions using the selected 16,187 SNPs markers and the data on the five phenotypic traits obtained from the pooled analysis of the two field experiments. A total of 22 QTL regions (− log10 P ≥ 5.5 and q value ≤ 0.05) were significantly associated with the five traits: oil (OIL%) and protein (PROT%) contents in grains, nodule dry weight (NDW), nitrogen content in the shoots (N content), and root dry weight (RDW). The remaining traits we could not found significant associations. The found regions explained 1.84 to 58.96% of the phenotypic variation observed among the selected genotypes of the panel.

For OIL%, we identified significant peak SNPs using BLINK and CMLM models. Among them, all markers showed consistent associations across the two statistical models. These markers were located on Chr04:6032360, Chr06:19398261, Chr06:29876176, Chr10:6522129 and Chr18:1231616 (Figure 6). The higher MAF values for OIL% were obtained for the SNPs found on chromosomes 6, 10, and 18, with values of 0.49, 0.40 and 0.48, respectively. Chr06:19398261, Chr06:29876176, and Chr18:51048046 presented positive effect on OIL%. These three SNPs together explained 16.78% of this phenotypic variation. The remaining three SNPs showed a low negative effect on OIL% of -0.39 and -0.37, respectively and explained 24.74% of the phenotypic variation of the panel (Table 3).

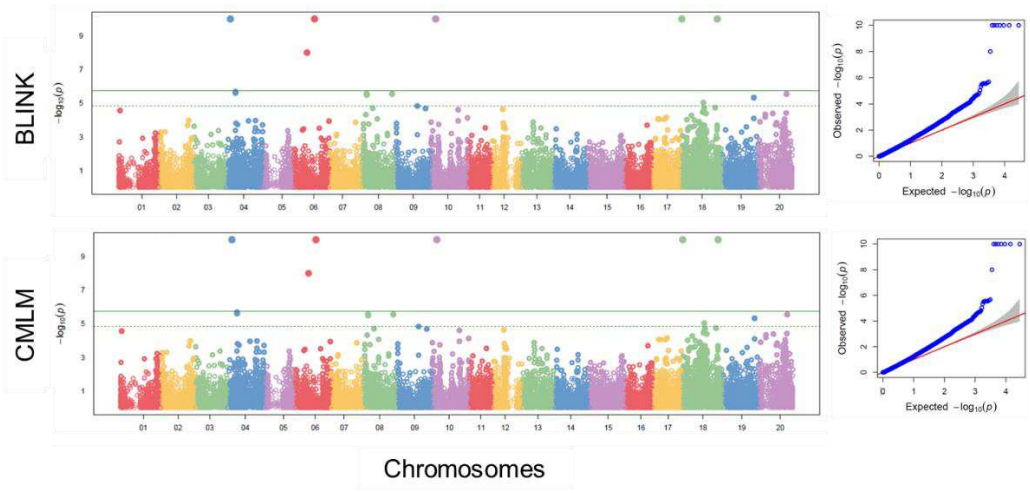

Figure 6. Genome-wide association analysis (GWAS) in a soybean population containing 100 accessions. Manhattan plot results from different GWAS models to evaluate the relationship with oil content in grains (OIL%). The dotted line indicates the significance threshold (false discovery rate ≤ 0.05). Left side: significantly associated markers. Right side: comparison of quantile-quantile (QQ) plots of different models for GWAS. The y-axis is the observed negative base-10 logarithm of the p values, and the x-axis is the expected observed negative base-10 logarithm of the p-value.

Focusing on the SNPs with positive effect on oil content, SNP markers at Chr06:19398261, Chr06:29876176, and Chr18:51048046 performed this function and had positive effect on OIL%. Among these loci, the lowest p-values were obtained by Chr06:29876176 and Chr18:51048046 (p = 1.01E-10) (Table 4), showing positive relationship with OIL%.

Accessions that showed "TT" in the position 19398261 of chromosome 6 which explained 8.36% of the phenotypic variation of the considered traits among the genotypes in the panel. The oil content among these accessions was generally >20%. However, accessions that had <20% oil content were associated with "GG" at this locus (Okute, Pan-San, 17X-1354-98-3 (109), Ringgit strain 19 51, S-100, 17X-1345-234-6 70, Nanking 332, 17X-1337-25-3 (306), 17X-1345-234-6, Da bai shui dou No. 1, KAERI 630-8, Shimotsu Ura, Hernon, Shu yang hong mao qiu yi, Long zhou dong feng dou, Jing huang 18, Yonekadake, KAS 390-4, E.G.1, Ding an qing pi dou, Tung Tam, USDA- ARD A, and AA025). It can be inferred that the most exotic, older, and high-protein accessions presented this haplotype, i.e., lower OIL%, most of them belonging to the group 3 of the population structure analysis.

Accessions that showed "CC" at position 29876176 of chromosome 6 also presented oil contents in grains >20%. The remaining accessions that had <20% oil content were associated with "AA" at this locus (Ding an qing pi dou, Tung Tam, USDA- ARD A, and AA025).

Most of the accessions showed "GG" at position 51048046 of chromosome 18 and also had oil contents in grains >20%. Few accessions were associated with "AA" in this locus (A072, A056, Tung Tam, USDA- ARD A, PI 281888, and A073), and their oil contents in grains ranged from 15 to 20%.

Considering these loci together, there were three of the four possible haplotypes, in which "T+C+G" type was observed in the accessions with higher OIL%, whereas only two of the four possible haplotypes ("G+A") were found in accessions with low OIL% (Figure 7).

Table 4. Significant SNP associations with oil content in grains (OIL%) in a soybean population containing 100 accessions using BLINK and CLMLM models.

	Peak SNP position	*p*-value	MAF	Effect	PVE (%)
BLINK	Chr04:6032360	1.01E-10	0.09	-1.23	15.21
	Chr06:19398261	1.01E-08	0.49	0.66	8.36
	Chr06:29876176	1.01E-10	0.10	0.31	6.57
	Chr10:6522129	1.01E-10	0.40	-0.39	2.68
	Chr18:1231616	1.01E-10	0.48	-0.74	6.85
	Chr18:51048046	1.01E-10	0.07	0.27	1.84
CLMLM	Chr04:6032360	1.01E-10	0.09	-1.23	15.21
	Chr06:19398261	1.01E-08	0.49	0.66	8.36
	Chr06:29876176	1.01E-10	0.10	0.31	6.57
	Chr18:1231616	1.01E-10	0.48	-0.74	6.85
	Chr10:6522129	1.01E-10	0.40	-0.39	2.68
	Chr18:51048046	1.01E-10	0.07	0.27	1.84

Peak SNP position: Genomic position of the peak Single Nucleotide Polymorphism (SNP) associated with oil content in soybean grains; p-value: indicates the statistical significance of the association between the SNP and the phenotypic trait (OIL%); MAF: minor allele frequency, represents the frequency of the less common allele at the SNP locus, indicating its prevalence in the population; PVE (%): proportion of phenotypic variation explained by the SNP, indicating the relative contribution of the respective SNP to the observed variation on the phenotypic trait.

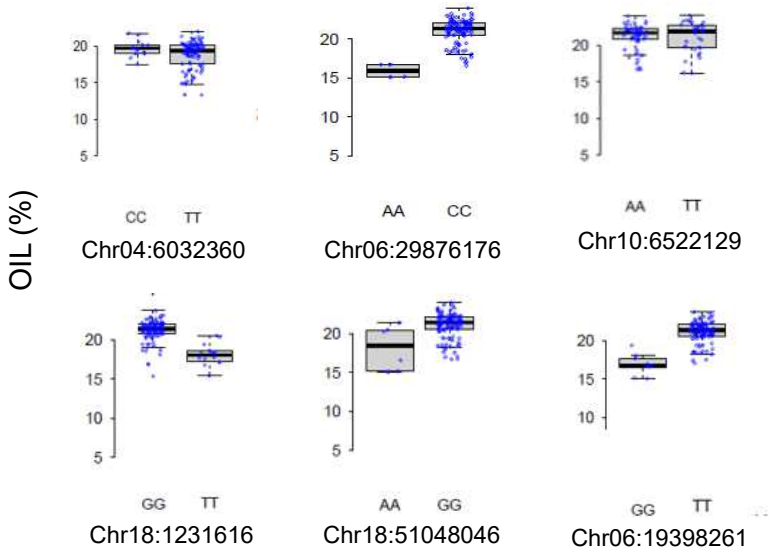

Figure 7. Oil content (OIL%) in grains of a soybean population containing 100 accessions with contrasting alleles in six significant loci.

For PROT%, significant SNP peaks were identified using BLINK, CMLM and FarmCPU models. Among them, the SNP markers at Chr04:6032360 and Chr18:1231616 showed consistent associations across the statistical models BLINK and FarmCPU (Figure 8), assigning the reliability to the findings.

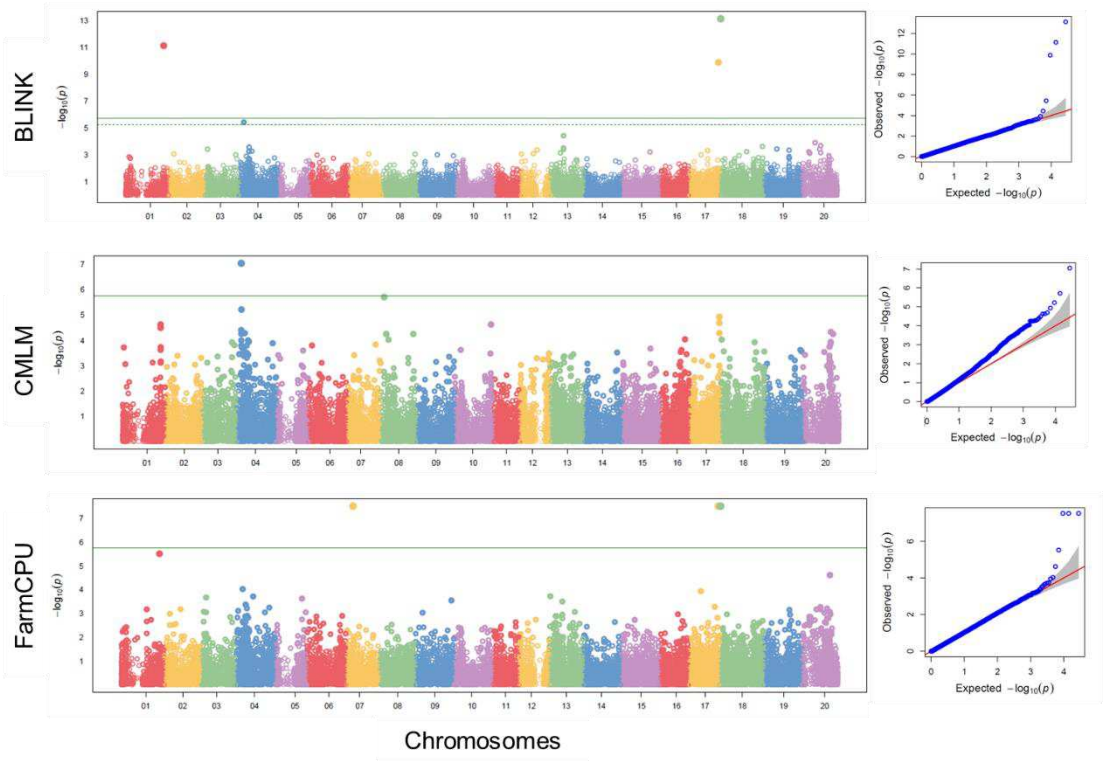

Figure 8. Genome-wide association analysis (GWAS) in a soybean population containing 100 accessions. Manhattan plot results from different GWAS models to evaluate the relationship with protein content in grains (PROT%). The dotted line indicates the significance threshold (false discovery rate ≤ 0.05). Left side: significantly associated markers, identified by association models. Right side: comparison of quantile–quantile (QQ) plots of different models for GWAS. The y-axis is the observed negative base-10 logarithm of the p values, and the x-axis is the expected observed negative base-10 logarithm of the p value.

In addition to Chr04:6032360 and Chr18:1231616, Chr17:39602638 was also addressed for discussion for although having a negative effect on the phenotype, also had a significative contribution on the phenotype variation of 30.27% in the average of BLINK and FarmCPU models. Among these loci, Chr18:1231616 showed the lowest p-values (7.40E-14) based on BLINK model. This SNP was also identified by the FarmCPU model, reinforcing the consistency of the marker (Table 5).

Regarding the SNP Chr04:6032360, 29 accessions that presented "TT" showed protein content ≥40% [Okute, Jing huang 18, KAS 390-4, Yonekadake, PI 281888, USDA- ARD A, KAERI 630-8, Ding an qing pi dou, Da bai shui dou No. 1, Shu yang hong mao qiu yi.sort, E.G.1, Tung Tam, 17X-1337-25-3 (306), Shimotsu Ura, Long zhou dong feng dou, 17X-1345-234-6, S-100, 17X-1345-234-6 70, Nanking 332, Hernon, 41 S 31, 17X-1337-25-3 199, BR-3, Ringgit strain 19 51, A057, 17X-1354-98-3 92, Pan-San, 17X-1337-173-7 (195), and 17X-1354-98-3 (109)]. The other accessions had "TT" and presented a lower range of variation in protein content in grains from 39.6 to 34.67%. Still for Chr04:6032360 the "CC" haplotype showed the highest oil content.

For Chr18:1231616 the same 29 accessions above mentioned presented "TT" in this position and showed ≥40% protein content in grains, whereas the remaining accessions with <40% protein had "GG" in that position. This SNP was also verified as having effect on oil content, where Chr18:1231616 containing the haplotype "GG" was associated with higher oil content, whereas the haplotype containing "TT" in this position was associated with higher

protein content in grains.

Accessions that presented “CC” in the Chr17:39602638 position, had protein contents ≥40%, whereas when “TT” was instead (accessions: Shu yang hong mao qiu yi, Ringgit strain 23 51, and PI 331795), the protein content in grains ranged in a lower range from 35.09 to 38.18%.

Considering these loci together, two of the four possible haplotypes were found, where “T+C” were observed in the accessions with high PROT%, and three of the four possible haplotypes “G+C+T” were found in the accessions with lower PROT% (Figure 9).

Table 5. Significant SNP associations with protein content in grains (PROT%) in a soybean population containing 100 accessions using BLINK, CMLM, and FarmCPU models.

	Peak SNP position	***p*-value**	**MAF**	**Effect**	**PVE (%)**
BLINK	Chr01:50419030	7.40E-12	0.22	-0.81	6.62
	Chr17:39602638	1.30E-10	0.06	-0.97	30.69
	Chr18:1231616	7.40E-14	0.48	0.71	7.20
CLMLM	Chr04:6032360	9.07E-08	0.09	2.30	6.99
FarmCPU	Chr07:9510621	3.06E-08	0.16	-0.66	9.33
	Chr17:39602638	3.06E-08	0.06	-0.98	29.85
	Chr18:1231616	3.06E-08	0.48	0.59	7.74

Peak SNP position: Genomic position of the peak Single Nucleotide Polymorphism (SNP) associated with protein content in soybean grains; p-value: indicates the statistical significance of the association between the SNP and the phenotypic trait (PROT%); MAF: minor allele frequency, represents the frequency of the less common allele at the SNP locus, indicating its prevalence in the population; PVE (%): proportion of phenotypic variation explained by the SNP, indicating the relative contribution of the respective SNP to the observed variation on the phenotypic trait.

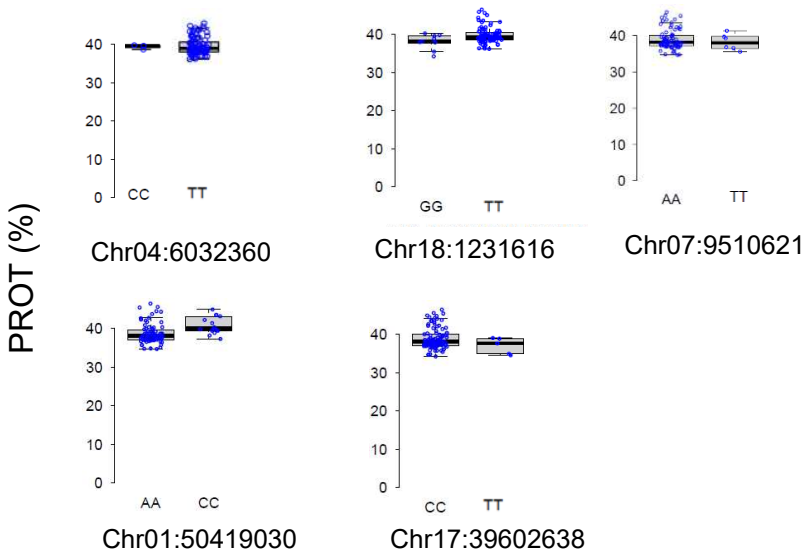

Figure 9. Protein content (PROT%) in grains of a soybean population containing 100 accessions with contrasting alleles in five significant loci.

Our GWAS for this panels based on 100 accessions also revealed significant SNPs associated with nodule dry weight (NDW). Among them, Chr18:9704961 and Chr18:30985378 exhibited a consistent association between CLM e MLMM models, providing robustness to the results (Figure 10).

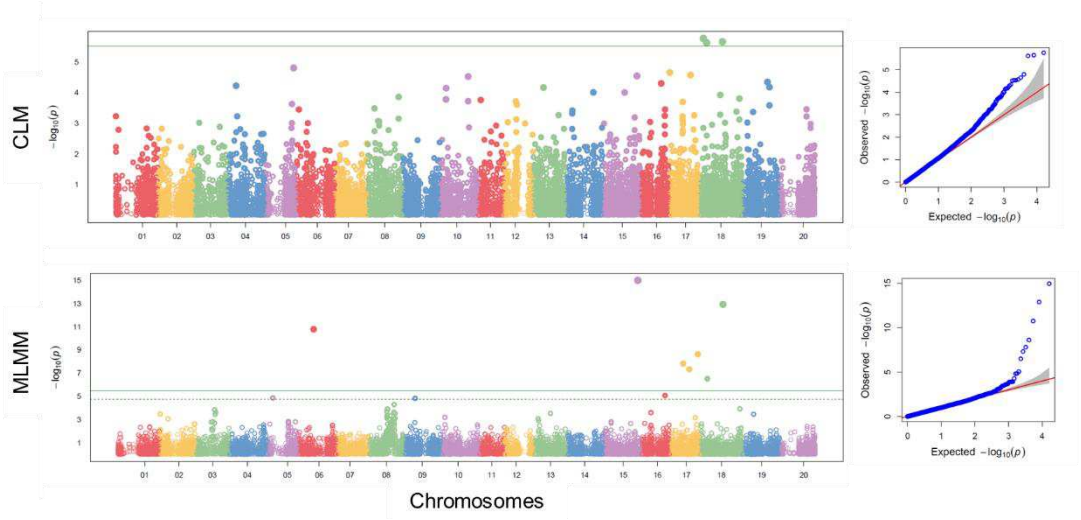

Figure 10. Genome-wide association analysis (GWAS) in a soybean population containing 100 accessions. Manhattan plot results from different GWAS models to evaluate the relationship with nodule dry weight (NDW). The dotted line indicates the significance threshold (false discovery rate ≤ 0.05). Left side: significantly associated markers, identified by association models. Right side: comparison of quantile–quantile (QQ) plots of different models for GWAS. The y-axis is the observed negative base-10 logarithm of the p values, and the x-axis is the expected observed negative base-10 logarithm of the p value.

The low p-values (ranging from 1.74E-06 to 1.05E-15) indicate strong evidence of these associations (Table 5). MAF values for these SNPs ranged from 0.05 to 0.37, suggesting that some alleles associated with NDW may be relatively rare in this population. The effect shows negative values, indicating that certain alleles are linked to decrease in NDW, suggesting the presence of multiple genetic variants influencing this trait. The PVE values, ranging from 2.34% to 17.04%, suggest that these associated SNPs collectively explain a moderate to relatively high proportion of the phenotypic variation (Table 6). We will focus on the SNPs Chr18:9704961 and Chr18:30985378, which were common between CLM and MLMM models.

Regarding these SNPs, the presence of the haplotype “AA” was associated with NDW ranging from 250 to 760 mg per plant among the following accessions: Da bai shui dou No. 1, A024, Jing huang 18, Pan-San, A038, A062, A067, A048, A066, Tung Tam, and Okute. Accessions that stood out for NDW ranking were less prone to negative effects. The remaining accessions with the “GG” haplotype reached a maximum NDW of 500 mg per plant and the minimum 70.5 mg per plant. Thus, in the presence of AA, the NDW was more favored than in the presence of GG (Figure 11).

The presence of SNPs with negative effect is also useful for genetic breeding, because by eliminating accessions containing SNPs with a negative effect, breeding programs can direct crosses and selections to favor plants with beneficial genetic variations driving to higher NDW.

Table 6. Significant SNP associations with nodule dry weight (NDW) in a soybean population containing 100 accessions using CLM and MLMM models.

	Peak SNP position	*p*-value	MAF	Effect	PVE (%)
CLM	Chr18:5180517	1.74E-06	0.06	-171.41	15.24
	Chr18:9704961	2.41E-06	0.08	-118.89	13.45
	Chr18:30985378	2.23E-06	0.11	-90.62	13.66
MLMM	Chr06:20321946	1.71E-11	0.13	-62.37	9.81
	Chr15:46330139	1.05E-15	0.07	-85.72	14.76
	Chr17:18482247	1.53E-08	0.32	-36.74	2.34
	Chr17:26743711	4.71E-08	0.05	-83.58	17.04
	Chr17:38383587	2.36E-09	0.37	-50.44	7.60
	Chr18:9704961	3.08E-07	0.08	-60.73	14.09
	Chr18:30985378	1.21E-13	0.11	-70.57	8.68

Peak SNP position: Genomic position of the peak Single Nucleotide Polymorphism (SNP) associated with nodule dry weight in soybean roots; p-value: indicates the statistical significance of the association between the SNP and the phenotypic trait (NDW); MAF: minor allele frequency represents the frequency of the less common allele at the SNP locus, indicating its prevalence in the population; PVE (%): proportion of phenotypic variation explained by the SNP, indicating the relative contribution of the respective SNP to the observed variation on the phenotypic trait.

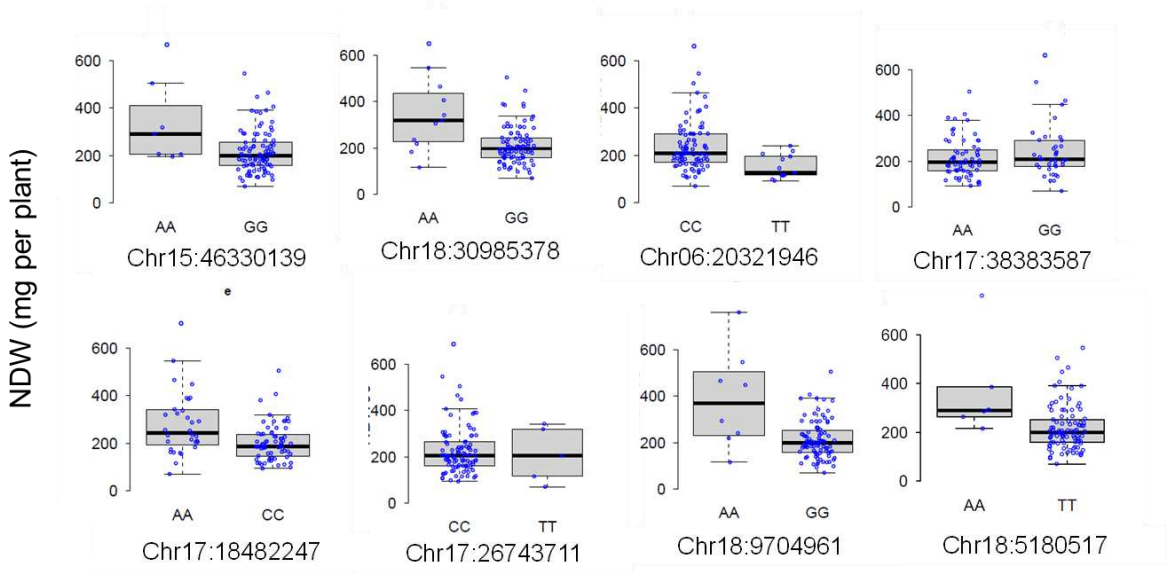

Figure 11. Nodule dry weight (NDW) in roots of a soybean population containing 100 accessions with contrasting alleles in eight significant loci.

The GWAS to explore the association between genetic markers and N content was significant using the MLM and MLMM models (Figure 12). The genetic marker located on chromosome 09 at position 19995675 (Chr09:19995675) emerged as statistically significant candidate associated with N content in soybean. Both MLM and MLMM models showed the same result, indicating a consistent and robust association. The low p-value of 2.77E-06 supports strong evidence on this marker controlling the associated trait.

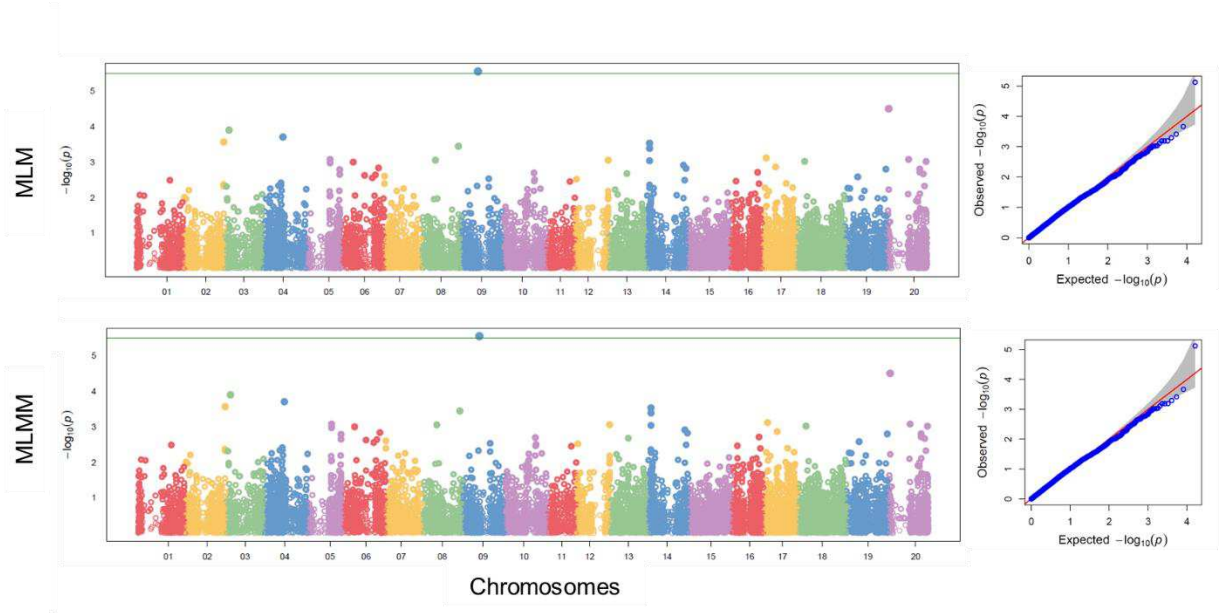

Figure 12. Genome-wide association analysis (GWAS) in a soybean population containing 100 accessions. Manhattan plot results of different GWAS models to evaluate the relationship with nitrogen content in shoots (N content). The dotted line indicates the significance threshold (false discovery rate ≤ 0.05). Left side: significantly associated markers, identified by association models. Right side: comparison of quantile–quantile (QQ) plots of different models for GWAS. The y-axis is the observed negative base-10 logarithm of the p values, and the x-axis is the expected observed negative base-10 logarithm of the p value.

The negative effect obtained from both models was small, which implies that multiple genes or factors may contribute collectively to the overall variation in this trait. Furthermore, the genetic marker under investigation demonstrated a considerable explanatory power, explaining approximately 58.96% of the phenotypic variance for N content (Table 7).

The accessions presenting the haplotype "AA" had N content that reached more than 30 g kg^{-1} [Accessions: A058, A043, A061, A018, 17X-1345-151-11 (178), A071, A035, A051, A030, Yonekadake, A021, KAS 390-4, 17X-1345-151-11 (298), A075, A034, BR-3, A027, A049, A068, Ding an qing pi dou, A076, A053, 17X-1354-98-3 92, A048, A046, A050, Jing huang 18, KAERI 630-8, A025, A026, A067, 17X-1337-173-7 (195), 17X-1337-25-3 (306), Embrapa 48, A041, A042, A059, Da bai shui dou No. 1, A045, A056, A069, A062, A016, A054, 17X-1345-234-6, BRS 713 IPRO, 17X-1337-25-3 199, Davis, Nanking 332, A036, Shu yang hong mao qiu yi, A035, A040, A039, A038, Tung Tam, A031, A070, A014, A052, and A055]. Part of these accessions coincides with the ones that showed higher levels of PROT% in grains.

The other part of the accessions presented the haplotype "GG," which reached a maximum of 25 g kg^{-1} of N in the shoots.

Table 6. Significant SNP associations with nitrogen content in shoots (N) in a soybean population containing 100 accessions using MLM and MLMM models.

	Peak SNP position	*p*-value	MAF	Effect	PVE (%)
MLM	Chr09:19995675	2.77E-06	0.1	-1.63	58.96
MLMM	Chr09:19995675	2.77E-06	0.1	-1.63	58.96

Peak SNP position: Genomic position of the peak Single Nucleotide Polymorphism (SNP) associated with N contents in soybean shoots; p-value: indicates the statistical significance of the association between the SNP and the phenotypic trait (N); MAF: minor allele frequency represents the frequency of the less common allele at the SNP locus, indicating its prevalence in the population; PVE (%): proportion of phenotypic variation explained by the SNP, indicating the relative contribution of the respective SNP to the observed variation on the phenotypic trait.

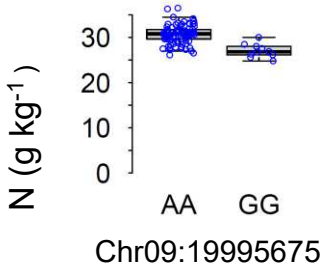

Figure 13. Nitrogen content (N) in shoots a of soybean population containing 100 accessions with contrasting alleles in one significant locus.

The GWAS for root dry weight (RDW) highlighted significative markers using GLM, MLMM and FarmCPU models. Moreover, all regions identified showed consistency among the models (Figure 14).

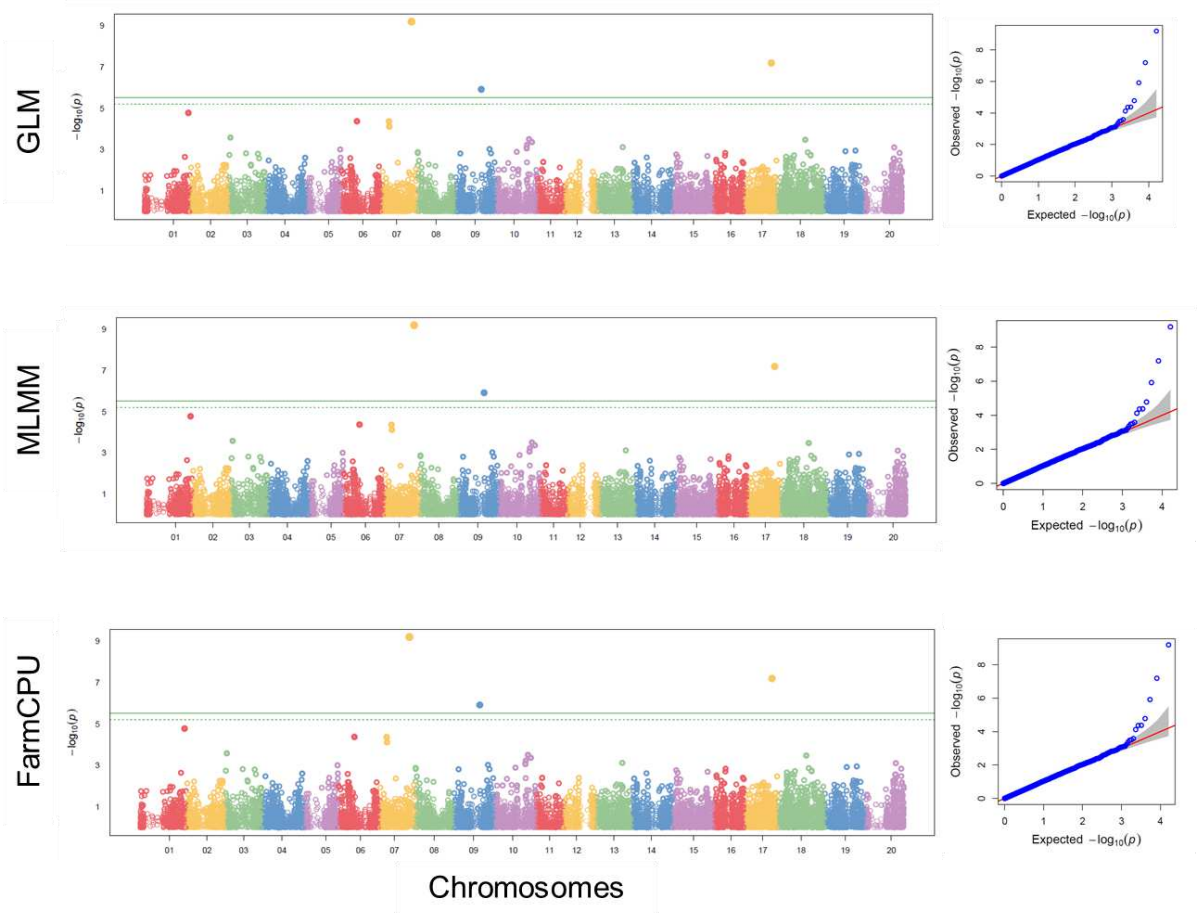

Figure 14. Genome-wide association analysis (GWAS) in a soybean population containing 100 accessions. Manhattan plot results of different GWAS models to evaluate the relationship with root dry weight (RDW). The dotted line indicates the significance threshold (false discovery rate ≤ 0.05). Left side: significantly associated markers, identified by association models. Right side: comparison of quantile–quantile (QQ) plots of different models for GWAS. The y-axis is the observed negative base-10 logarithm of the p values, and the x-axis is the expected observed negative base-10 logarithm of the p value.

The genetic marker Chr07:37193798 suggests a strong genetic effect on RDW as evidenced by its identical p-value (6.44E-10) in both GLM and FarmCPU models (Table 7). Similarly, Chr09:32113409 also exhibited consistent associations across the three methods, with a relatively high effect of 0.57 and a moderate MAF of 0.06. The PVE percentage of 21.27 indicates that this SNP explains a substantial proportion of the variation in soybean RDW. In general, the moderate MAF (0.08) supports the identified SNs as influencing RDW (Table 7).

Table 7. Significant SNP associations with root dry weight (RDW) in a soybean population containing 100 accessions using GLM, MLMM and FarmCPU models.

	Peak SNP position	*p*-value	MAF	Effect	PVE (%)
GLM	Chr07:37193798	6.44E-10	0.20	0.20	4.73
	Chr09:32113409	1.21E-06	0.06	0.57	21.27
	Chr17:34227184	6.44E-08	0.08	-0.70	42.58
MLMM	Chr07:37193798	1.62E-06	0.13	-0.48	10.89
	Chr09:32113409	2.78E-06	0.06	0.69	21.58
	Chr17:34227184	2.98E-09	0.08	-0.92	39.61
FarmCPU	Chr07:37193798	6.44E-10	0.20	0.20	4.73
	Chr09:32113409	1.21E-06	0.06	0.57	21.27
	Chr17:34227184	6.44E-08	0.08	-0.70	42.58

Peak SNP position: Genomic position of the peak Single Nucleotide Polymorphism (SNP) associated with root dry weight; p-value: indicates the statistical significance of the association between the SNP and the phenotypic trait root dry weight (RDW); MAF: minor allele frequency represents the frequency of the less common allele at the SNP locus, indicating its prevalence in the population; PVE (%): proportion of phenotypic variation explained by the SNP, indicating the relative contribution of the respective SNP to the observed variation on the phenotypic trait.

The markers Chr07:37193798 and Chr09:32113409 were associated with root dry weight (RDW) in soybean. For the SNP Chr07:37193798, the presence of the haplotype "TT" was associated the maximum root dry weight of 5.2 g per plant (Okute). On the other hand, the presence of "CC" haplotype was associated with RDW lower than 3.0 g per plant in three accessions.

Considering the SNP Chr09:32113409, the presence of the haplotype "CC" the accessions ranged from 1.08 g per plant (PI 594811) up to 5.2 g per plant (Okute). In the presence of the haplotype "TT" the accessions presented large variation for this trait, but ranging from 2.67 g per plant (A043) to 5.0 g per plant (A069).

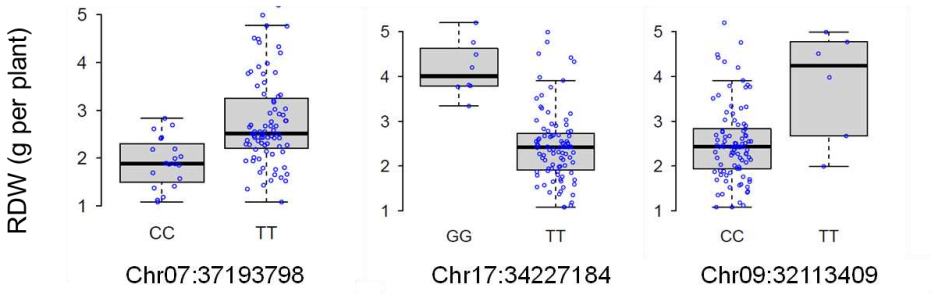

Figure 15. Root dry weight (RDW) of a soybean population containing 100 accessions with contrasting alleles in three significant loci.

Regarding the five traits (OIL%, PROT%, NDW, N and RDW) evaluated in the GWAS, the accession Jing huang 18 had the highest protein content in grains (46.45%) while the average for the 100 accessions was 38.93%. The lowest content was observed in A034 (34.67%). For oil, A045 showed the highest content (23.96%), while the average among the accessions was 20.77%. The lowest oil content was observed in A025 (15.07%).

For RDW and NDW, the accession Okute (RDW=5.20 g per plant and NDW=760.5 mg per plant) highlighted. The average for RDW among the accessions was 2.59 g per plant, whereas PI 594811 was the accession with the lowest root dry mass (1.08g per plant). For NDW the average among the soybean accessions was 227.28 mg

per plant, whereas PI 281888 was the accession with the lowest NDW (70.50 mg per plant).

For N content, the highlight was A058, with 36.50 g kg^{-1}, while the average among the accessions was 24.82 g kg^{-1}. The accession PI 281888 showed the lowest N content in shoots in the GWAS, with 24.82 g kg^{-1}, which also coincided with the lowest NDW.

Traits presenting peak SNPs on the same chromosome, as well as their position within the respective chromosome are presented in Table 8. The OIL% and PROT% traits shared 2 QTLs, one on chromosome 4 (6032360) and other on chromosome 18 (1231616), which represents a LD given by r2 equal to 1. Due to the high inverse correction between these traits (-0.9) as expected, the effects of these SNPs were opposite. These two SNPs found in the same positions by different models provided evidences that they are associated with the increase in protein and decrease in oil contents in the 100 soybean accessions constituting the panel. This demonstrates the consistency of these data and models since the inverse correlation between oil and protein has been widely evidenced.

Although some traits had peak SNPs on the same chromosome in the GWAS, some are physically very far from each other, despite these SNPs were also tested to check the LD. The best candidates for being in LD were located on chromosome 17 (PROT% and NDW), as they are physically closer. PROT% and NDW (Chr 17:39602638 and Chr 17:38383587) were separated by 297.9 kb, however they had a LD of 0.042, suggesting that there is a very low association between the alleles in this locus. The absence of significant association between protein content in grains and BNF traits in your study underscores the intricate nature of soybean genetics and the multifaceted regulatory networks that drive these traits. BNF is a complex process that involves intricate symbiotic relationships between the plant and the nitrogen-fixing bacteria, in addition to the sensitivity of both to environmental conditions and soil nutrient availability. Variability in these environmental factors across different accessions and experimental conditions can obscure the direct genetic associations with protein content in grains.

For OIL% and NDW in chromosome 6 (positions 19398261 and 20321946) are separated by 232.7 kb, despite an even smaller LD of 0.007 was found (Table 8). For the other less promising positions, the LD was even closer to 0 (not shown). Figure 16b, illustrates how LD decay for the 16,187 SNPs inside the 20 soybean chromosomes. The red color is associated with higher r^2 or stronger associations (close to 1), whereas blue represents weaker associations, therefore, r^2 is smaller (closer to 0)

As a result of the GWAS based on a soybean panel with 100 accessions aiming to identify causal genes or alleles responsible for the genetic portion of phenotypic variation related to BNF traits and grain contents of oil and protein, we implemented a systematic approach that successfully identified candidate genes (Table 8).

Table 8. Genetic loci and candidate genes associated with BNF and protein content traits located on chromosomes 1, 4, 6, 7, 17 and 18 based on GWAS.

Trait	SNP	Associated trait	Candidate gene[a]	Functional annotation	Physical position[b]
PROT%	Chr01:50419030	-	*Glyma.01g166400*	PTHR11139:SF1 - TRANSFORMATION/TRANSCRIPTION DOMAIN-ASSOCIATED PROTEIN	Chr01:50404780..50436297 (-)[c]
PROT%	Chr04:6032360	OIL%	*Glyma.04g072400*	PTHR34457:SF1 - EMBRYO DEFECTIVE 2410 PROTEIN	Chr04:6026655..6056647 (-)
NDW	Chr06:20321946	-	*Glyma.06g208300*	PF00335 - Tetraspanin family (Tetraspannin)	Chr06:20321665..20323023 (+)
RDW	Chr07:37193798	-	*Glyma.07g202700*	PTHR33873:SF1 - TRANSCRIPTION FACTOR VOZ1	Chr07:37189616..37194333 (+)
PROT%	Chr07:9510621	-	*Glyma.07g100200*	PTHR24056:SF0 - CYCLIN-DEPENDENT KINASE 7	Chr07:9509360..9514936 (-)
NDW	Chr17:18482247	-	*Glyma.17g175900*	PTHR12864//PTHR12864:SF13 - RAN BINDING PROTEIN 9-RELATED	Chr17:18476563..18490094 (+)
RDW	Chr17:34227184	-	*Glyma.17g208500*	PTHR12446//PTHR12446:SF27 - TESMIN/TSO1-RELATED	Chr17:34224092..34230310 (-)
NDW	Chr17:38383587	-	*Glyma 17g20850.17g228800*	Glycine dehydrogenase (aminomethyl-transferring) / Glycine-cleavage complex P-protein	Chr17:38383183..38389776 (+)
PROT%	Chr17:39602638	-	*Glyma.17g240400*	Poly-adenylate binding protein, unique domain (PABP)	Chr17:39597802..39603019 (-)
PROT%	Chr18:1231616	OIL%	*Glyma.18g017300*	PTHR23354//PTHR23354:SF76 - NUCLEOLAR PROTEIN	Chr18: 1,230,462-1,234,426 (-)
OIL%	Chr18:51048046	-	*Glyma.18g223300*	PTHR13743:SF16 - PROTEIN T01H10.8	Chr18:51002334..51066341 (-)
NDW	Chr18:9704961	-	*Glyma.18g095100*	PTHR22835//PTHR22835:SF116 - ZINC FINGER FYVE DOMAIN CONTAINING PROTEIN	Chr18:9704194..9707617 (-)

[a]Genes ID from Williams 82 genome - Wm82.a2 (https://phyto zome.jgi.doe.gov/)
[b]Physical position in chromosomes 1, 4, 6, 7, 17 and 18 Glyma.Wm82.a2 (Gmax_275_v2)
[c] (-) Represents the antisense strand and (+) represents the sense strand of DNA.

4. Discussion

4.1 Population Structure analysis

Genetic diversity is essential for soybean breeding to improve several traits, including increase in protein content in grains and efficiency in the BNF process. The identification of genes of interest will allow for its incorporation during the development of new varieties breed for a specific or several traits. The diversity index showed that genetic diversity of soybean varieties cultivated in Brazil is narrow (Wysmierski and Vello 2013), due to use of few parentals in the genetic breeding programs. The use of a small set of genotypes can be considered key factor for the loss of genetic diversity in the current soybean cultivars. Most of the Brazilian soybean germplasm is derived only from four genotypes (CNS, S-100, Roanoke, and Tokyo), which contribute around 55% of the genetic basis for cultivars released in Brazil. Gwinner et al. (2017) highlighted the relevance of increasing the genetic variability to ensure genetic gains in soybean breeding, and for this, identification of genes related to traits of interest can support such programs.

A population structure analysis on the set of soybean cultivars showed that Brazil has a narrower genetic basis compared with the US cultivars (Maldonado dos Santos et al. 2022). In a study considering the population structure and diversity of soybean varieties in relation to BNF capacity and protein content in grains, two groups were found: one with approximately 50% of adhesions, encompassing varieties from Brazilian public and private companies; the other grouping 45% of the accessions, including Brazilian, exotic, and private germplasms. Some

accessions (5%) were not grouped in any cluster (Torres et al., 2015).

The analysis of the population structure of 343 soybean lines from Brazil, North America, and Asia, indicated the existence of three subpopulations originating from different geographic regions (Mendonça et al., 2022). The Asian genotypes were the most distinct group of the panel, with higher π values, as well as smaller linking blocks, indicating greater genetic diversity than the Brazilian genotypes.

An analysis on the population structure of 14,000 soybean accessions [*Glycine max* (L.) Merr. and *G. soja* Siebold & Zucc.] indicated that the ancestry of American accessions derived from two Chinese subpopulations, which reflects the genetic composition of American accessions. A GWAS represented by 12,000 soybean accessions conducted on protein and seed oil identified SNPs with strong signals on chromosomes 20 and 15 (Bandillo et al., 2015)

Our findings agree with the results presented in the above studies, where the more exotic cultivars, mainly from Asia, were grouped together in group 3. In addition, our panel indicated that older accessions showed NDW and lower OIL%. On the other hand, groups containing more recent accessions with RR and IPRO technologies exhibited higher OIL% and lower PROT%, whereas NDW varied among accessions.

4.2 GWAS

Around 200 and 300 QTLs for protein and oil contents, respectively, have been deposited in SoyBase (http://www.soybase.org) (Brown et al., 2021). GWAS and QTL analyses on diverse soybean populations have suggested that regions of chromosomes 10, 15, 18, and 20 have been often associated with these traits (Diers et al., 1992; Seboldt et al., 2000; Zhang et al., 2019; Wang et al. 2020). Our findings mapped SNPs for PROT% in chromosomes 1, 4, 7, 17, and 18 and for OIL% in chromosomes 4, 6, 10 and 18.

A GWAS for oil and protein content in grains based on a panel composed of 298 accessions identified 40 SNPs located in 17 different genomic regions in 12 out of the 20 soybean chromosomes (Hwang et al., 2014a), including a SNP in the position Chr07:9512225, close to Chr07:9510621 we found for this trait. Our results demonstrated 22 SNPs located in 9 out of the 20 chromosomes. The oil and protein contents in grains show a wide genetic variation among soybean accessions, in addition to the strong negative correlation each other, as already has been widely reported (Diers et al., 1992; Sediyama et al.,1993; Chung et al., 2003; Patil et al., 2017; Kumar et al., 2021). GWAS can find SNPs that have inverse effect on both traits, i.e., when present, these SNPs lead to decreased in that trait. Our study identified the SNPs Chr04:6032360 and Chr18:1231616 which had a positive effect on PROT% and a negative effect on OIL%.

Protein and oil contents in seeds are quantitative traits determined by the interaction among many genes with small to moderate genetic effects and their interactions with the environment (Hwang et al., 2014; Akond et al., 2014). Two main QTLs for protein and oil contents in grains have been mapped in many soybean populations, and they are located on chromosomes 15 and 20 (Wang et al. 2020). In our GWAS we did not identify SNPs in these chromosomes.

Protein content in grains is a typical quantitative trait controlled by multiple minor effects genes (Patil et al., 2017). However, these effects depend on the expression levels of the related genes. For example, the low protein parent Rongxiandongdou had more upregulated genes and fewer downregulated genes than the high protein genotype

Nanxiadou 25, showing that the accumulation of storage protein was mainly based on downregulated genes in the high protein genotype, while the low protein genotype had genes for accumulation of storage protein upregulated (Wang et al., 2021).

Zhang et al. (2019) identified the SNPs Chr6:5713084 ~ 5992538, Chr9:40301013, Chr20:34801441 ~ 35512580 for OIL%. For PROT% the SNPs were Chr6:5836780 ~ 5931027, Chr9:38117239 ~ 41020511, and Chr20:34990941 ~ 25578946. We found a SNP also on chromosome 06 for OIL%, but in a different position (Chr06:19398261). Clone sequencing showed different SNPs and indels between high and low protein genotypes in *Glyma. 20g088000* and *Glyma.16g066600* may be the cause of changes in this trait (Wang et al. 2021).

Soybean is a rich source of protein, so it is imperative to investigate genes responsible for protein accumulation in grains. β-conglycinin and glycinin are the prevailing proteins in soybean (Thanh and Shibasaki, 1976), and we have found a gene related to glycinin *Glyma.17g228800* in this study. Valliyodan et al. (2016) identified variations in candidate genes (*HSP, Glyma20g19680; ammonium transporter, Glyma20g21030; ethylene receptor, Glyma20g21780*) associated with protein content in Gm20. Lestari et al. (2013) reported that parts of Gm20 and Gm10 showed synteny, however, QTLs for grain protein content were detected only in Gm20.

The amount of nitrogen fixed by legume–rhizobia symbioses may be increased by as much as 300% by plant breeding and crop management (Vance et al., 1998). In red clover (*Trifolium pratense*) the analysis of candidate genes revealed the molybdate transporter 1 gene strongly associated with BNF (Vega et al., 2015). The number and mass of nodules can be predictors highly related to BNF (Pitumpe Arachchige et al., 2020). N_2-fixation activity seems to be more related to NDW than number of nodules, also implying that only successful nodule formation does not ensure high N_2-fixation activity (Martins et al., 2022).

GWAS and QTL analyzes indicated that regions of chromosomes 6, 17, 18, 19, and 20 were frequently associated with nodule dry weight in SoyBase (http://www.soybase.org). Our findings mapped SNPs associated with NDW in chromosomes 6, 15, 17, and 18. Hwang et al. (2014b) identified eight QTLs for number of nodules, six QTLs for specific nodule weight, seven QTLs for nodule volume, and five QTLs for total nodule weight per plant in a field experiment by using the RIL derived from 'KS4895' and 'Jackson' with 664 markers. Research revealed that biomass-related traits, including NDW, underwent changes during soybean evolution (Wang et al., 2023).

Ureide content is the BNF-related trait with the highest number of QTLs available in Soybase (http://www.soybase.org). However, in our study, NDW was the only BNF-related trait showing a significant association, but not ureides. Variation in NDW could be more sensitive to genetic differences and this complexity creates a richer genetic environment for identifying significant associations in GWAS. Torkamaneh et al. (2020) recorded similar results in a GWAS using 297 African soybean genotypes to evaluate BNF-related traits. However, the SNP the authors detected on Chr07:2419122 was not detected in our panel.

Huo et al. (2017) found four loci on Gm17, which included two SNP markers (ss715626633 and ss715626686), associated with soybean nodule fresh weight, nodule dry weight, and large nodule numbers under field conditions. These results indicated that a QTL on Gm17 was linked with number and mass of nodules. These findings agree with our results, where NDW SNPs were found on chromosome 17.

Santos et al. (2006) evaluated previously described QTL and identified new QTL, by developing a mapping

population of 157 F2:7 soybean lines varying in BNF capacity, Bossier (high) × Embrapa 20 (average). This study represented an initial step in the identification and confirmation of QTL associated with the N-fixing symbiosis in Brazilian soybeans. Seven markers previously identified in the F2:3 population resulting from BRS 133 × Embrapa 20 cross were confirmed in the F2:7 population resulting from Bossier × Embrapa 20. These SSR markers were the first indication for the use of molecular markers to increase the contribution of BNF in soybean breeding programs. In a study for mapping QTL related to BNF traits, a composite interval mapping, mapped two QTLs related to SDW (LGs E and L), three to NN (LGs B1, E, and I), and one for specific mass of nodules; all QTLs were of small effect (R values ranging from 1.7% to 10.0%) and explained 15.4%, 13.8%, and 6.5% of the total variation for these three traits, respectively (Santos et al., 2013).

The genetics of the host plant and the rhizobia strongly regulate the nodulation process, collectively determining the fixed N output (Wang et al., 2018; Ferguson et al., 2019). The high heritability of nodulation traits in different environments is an indicator of traits controlled by genetic loci (Yang et al., 2019). A significant correlation is known between the nodulation traits and fixed N in a symbiotic process (Pereira et al., 1993; Pazdernik et al., 1996). A positive correlation was also found between nodule and plant dry weight with grain yield (Burias and Planchon, 1990).

GWAS and QTL analyzes indicated that regions of chromosomes 9, 15, 16, 19, and 20 were frequently associated directly or indirectly with nitrogen content in soybean plants (http://www.soybase.org). We found SNPs related to this trait on chromosome 9. Kaler et al. (2020) found in a GWAS for the trait dark green color index (DGCI) three SNPs on chromosome 9 at positions 4612586, 12240541, and 46800908. Dhanapal et al. (2015) identified regions associated with $\delta^{13}C$, one of which is located at Chr09:2069867.

GWAS and QTL analyzes have indicated that regions of chromosomes 7, 15, 18, 19, and 20 were often associated with soybean root dry weight in SoyBase (http://www.soybase.org). We found SNPs to related to this trait on chromosomes 7, 9, and 17.

Root size and architecture are important for determining yield performance, particularly under conditions of water restriction (Price et al., 2002). Well-developed roots may allow the plant to respond better to the stress, maintaining the efficiency of BNF for longer. Thus, soybean cultivars of the future must have robust root systems that help to withstand climate changes. The development of a healthy and more efficient root system is also crucial for BNF. Nodule growth and efficiency are correlated with root capacity, which in turn can impact the root dry weight.

Price et al. (2015), evaluating root thickness in *G. soja*, found significant SNPs on Chromosome 7 (59884.1.S1_8 - 8398.1.S1_11) and (8398.1.S1_11 - 1900.1.S1_3). Kim et al. (2023), analyzing root average length in soybean, found SNPs on chromosome 7 at positions 7177786 and 7419551. We also found SNPs on chromosome 7 for this trait but in a different position (Chr07:37193798). Based on comparisons of significant SNPs detected in the present study and those previously reported for root traits, we identified significant SNPs on chromosomes 9 and 17 that differ from the SNPs previously reported for root traits. These results agree with the idea that multiple genes and pathways are involved in the root growth and development in response to genetic and environmental factors.

It is important to point out that the results may vary depending on the target plant population, the analytic techniques, growth conditions, and specific genetic traits of the soybean genetic cluster in question. A comparison of

our results with the reported in literature highlights the importance of these chromosomal regions and suggests potential targets for further studies. For most of the traits evaluated, our results are different from those previously described, which may mark the identification of new regions for these traits of great importance for breeding more focused on grain quality and N-fixation efficiency.

4.3 Candidate Genes

We successfully identified candidate genes associated with OIL%, PROT%, NDW, and RDW. According to the candidate genes found in Phytozome (https://phytozome-next.jgi.doe.gov/) and according to the predicted function of these genes in plants using SoyBase (https://www.soybase.org/), we inferred some scenarios of performance of these genes in the 100 soybean accessions evaluated in the GWAS for traits related to BNF and protein content in grains.

Candidate genes related to PROT% have different functions: *Glyma.01g166400* is associated with proteins that regulate gene expression in plants. Genes that contain this domain play roles in physiological processes such as organ development and growth regulation. *Glyma.04g07240* was the candidate gene simultaneously associated with PROT% and OIL%. This gene may be involved in essential processes for cell differentiation and organization during embryogenesis. *Glyma.07g100200* has a function linked to the cell cycle regulation, development, differentiation, and plant response to stress. *Glyma.17g240400* influences the synthesis of protein and mRNA translation. *Glyma.18g017300* was also simultaneously related to PROT% and OIL% in our study. This gene is associated with synthesis of ribosome and production of protein. Higher efficiency in the production of ribosomes can increase the synthesis of protein.

For oil content in grains, the candidate gene *Glyma.18g223300* can influence the distribution of proteins, lipids, and other cell materials in different cell compartments, affecting the functioning of organelles, especially related to storage, in addition to metabolic processes, protein and oil contents. For nodulation, the candidate gene *Glyma.06g208300* is related to cell differentiation, cell wall organization, hormone response, intercellular communication, and photosynthesis. *Glyma.17g175900* influences vital functions for development, differentiation, and response to stress in plants. *Glyma.17g228800* is involved in the conversion and metabolism of the amino acid glycine in plants. Glycine plays essential roles in several metabolic pathways, including synthesis of protein and metabolism of N-compounds. We can infer that if this gene is involved in regulating the metabolism of glycine and is necessary for proper nodule formation. A mutation or deletion in this gene could lead to negative effects on nodulation. *Glyma.18g095100* is involved in cell organization, regulation of transport, signaling, adaptation to stress, and plant development. Their interactions with cell membranes have direct impact on vital functions of plant cells.

These genes have potential to affect cell differentiation, intercellular communication, stress adaptation and amino acid metabolism, including those related to the nodular processes. In this way, we can infer that a gene associated with glycine metabolism may play a crucial role in the formation of nodules since glycine plays essential roles in metabolic pathways crucial for the symbiotic interaction. However, as nodulation is a complex process that involves many genes with small effects, we must be cautious in making inferences. In addition, environmental variations greatly affect nodulation. The negative effect of these genes may be related to deficiencies in nodule formation, cellular disorganization, dysregulated stress responses, less efficient communication between the plant and N-fixing symbiont,

dysregulation of amino acid metabolism and inefficiency in the signaling required for FBN. These adverse effects can be influenced by complex genetic, environmental, and symbiotic interactions.

For root dry weight, the candidate gene *Glyma.07g202700* is more intensely expressed in vascular tissues, such as roots and young stems, and is associated with the development of the xylem, which is responsible for transporting water and nutrients from roots to shoots. As this gene was found to be associated with RDW, we can infer its influence on root architecture, regulating its growth, formation, and branching. The soybean accessions in the panel had great variation for RDW (1.08 to 5.20 g per plant). The genes that had positive effect on RDW may be involved in cell differentiation in roots, leading to formation of specific cell types involved in nutrient uptake, defense, and interactions with *Bradyrhizobium* and *Azospirillum* provided via co-inoculation. It is also plausible that these genes are linked to the root response to environmental stresses, modulating adaptations to adverse conditions, and possibly contributing to the symbiotic interactions of roots with microorganisms, improving nutrient uptake and plant performance. The candidate gene *Glyma.17g208500* belongs to a family of genes conserved in several plant species and is associated with different processes in plant development.

Most of inferred functions of the candidate genes do not have a clear negative or positive effect on the respective trait. If there is an imbalance in the regulation or expression of these genes, they may impact plant physiology, resulting in changes in the traits they affect. Furthermore, the traits addressed in our study are governed by many genes and are highly influenced by the environment.

We chose to employ co-inoculation and conducted the association analysis in this context, using plants that received two strains of bacteria. We found that NDW, which is easier to measure than the number of nodules, was significant in the GWAS. So, this measure can be adopted as an efficient way to evaluate BNF in soybean plants. However, BNF traits, especially in the field need high-throughput phenotyping techniques to enable the evaluation of these traits in a wide range of genotypes.

Advancing the genetic understanding of increased protein content in grains and the efficiency of BNF is important to improve the nutritional quality and the symbiosis for BNF. The adoption of different statistical models and the consistency of associations in several models reinforce the robustness of the identified markers. These findings should encourage the consideration of these markers for marker-assisted selection, contributing to the development of soybean cultivars with increased protein in grains and more efficient BNF. However, these candidate genes must be validated to ensure the accuracy and reliability of the identified associations. Confirmation can be accomplished with approaches such as gene expression assays, targeted mutations, or the use of gene editing techniques such as CRISPR-Cas9 to assess the effects of these genes on a soybean trait of interest. Additionally, performing functional assays can provide a more complete understanding on how these genes affect these traits. The validation not only reinforces the credibility of the gene associations but also serves as a solid guide for further genetic breeding efforts in the search for more productive and nutritionally richer soybean cultivars.

5. Conclusions

The GWAS panel covering soybean releases from 1948 to 2019 showed a stagnant trend in protein content, an increase in oil content and a decrease in nodule dry weight.

Despite the lack of direct correlation between traits related to protein content and BNF, we were able to identified relevant genomic regions and candidate genes for both traits individually. Our investigation found SNPs linked to biological nitrogen fixation traits and grain protein content traits, along with the identification of candidate genes. Some SNPs were found in proximity to already documented genomic regions in soybean. However, a significant proportion of the located identified loci could indicate unexplored regions. The SNPs and candidate genes identified in our research hold the potential to contribute to further endeavors in marker-assisted selection for soybean genetic breeding, emphasizing grain quality and symbiotic efficiency within a framework of sustainability and food safety priorities.

References

Akond, M., Liu, S., Boney, M., Kantartzi, S. K., Meksem, K., Bellaloui, N., ... & Kassem, M. A. (2014). Identification of quantitative trait loci (QTL) underlying protein, oil, and five major fatty acids' contents in soybean. American Journal of Plant Sciences, 2014.

Avery, B. W. (1973). Soil classification in the Soil Survey of England and Wales. Journal of Soil Science, 24(3), 324-338.

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., & Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. The Plant Genome, 8(3), plantgenome2015-04.

Brown, A. V., Conners, S. I., Huang, W., Wilkey, A. P., Grant, D., Weeks, N. T., et al. (2021). A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 49, 1496–D1501. doi: 10.1093/nar/gkaa1107.

Burghardt, L. T., Epstein, B., Guhlin, J., Nelson, M. S., Taylor, M. R., Young, N. D., ... & Tiffin, P. (2018). Select and resequence reveals relative fitness of bacteria in symbiotic and free-living environments. Proceedings of the National Academy of Sciences, 115(10), 2425-2430.

Chao, S., Du, W., Lu, T., Yang, Y., Wang, K., Du, H., ... & Yu, D. (2021). Genome-wide association study of soybean (Glycine Max) phosphorus deficiency tolerance during the seedling stage. Plant Breeding, 140(2), 267-284.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4(1), s13742-015.

Companhia Nacional de Abastecimento - CONAB. (2023). Décimo Levantamento. Safra 2022/23. Recuperado de <https://www.conab.gov.br/info-agro/safras/graos/levantamento-de-graos>. Acesso em: 25 de abril de 2023.

Cregan PB et al. An integrated genetic linkage map of the soybean genome. Agronomy-Faculty Publications. 1999;20.

Dhanapal, A. P., Ray, J. D., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., ... & Fritschi, F. B. (2015). Genome-wide association analysis of diverse soybean genotypes reveals novel markers for nitrogen traits. The Plant Genome, 8(3), plantgenome2014-11.

DuMouchel, W. and Jones, B. (1994), "A Simple Bayesian Modification of D-Optimal Designs to Reduce Dependence on an Assumed Model," Technometrics, 36, 37–47.

Elshire RJ et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one. 2011;6(5):e19379.

Ferguson, B. J., Indrasumunar, A., Hayashi, S., Lin, M. H., Lin, Y. H., Reid, D. E., & Gresshoff, P. M. (2010). Molecular analysis of legume nodule development and autoregulation. Journal of integrative plant biology, 52(1), 61-76.

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high-capacity genotyping by sequencing analysis pipeline. PloS one, 9(2), e90346.

Hwang S et al. Genetics and mapping of quantitative traits for nodule number, weight, and size in soybean (Glycine max L.[Merr.]). Euphytica. 2014;195(3):419-434.

Kim, S. H., Tayade, R., Kang, B. H., Hahn, B. S., Ha, B. K., & Kim, Y. H. (2023). Genome-Wide Association Study for Biomass Accumulation Traits in Soybean. Molecular Breeding, 43(5), 33.

Lee S, Van K, Sung M, Nelson R, LaMantia J, McHale LK, Mian MR. Genome-wide association study of seed protein,

oil and amino acid contents in soybean from maturity groups I to IV. Theoretical and Applied Genetics. 2019;132:1639-59.

Lestari, P., Van, K., Lee, J., Kang, Y. J., & Lee, S. H. (2013). Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. Frontiers in plant science, 4, 176.

Liborio PH, et al. Co-inoculation of Bradyrhizobium japonicum and Azospirillum brasilense on the physiological quality of soybean seeds. Semina: Ciências Agrárias. 2020;41(6):2937-2950.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., ... & Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics, 28(18), 2397-2399.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X. Pan-genome of wild and cultivated soybeans. Cell. 2020;182.

Lorini I. Qualidade de sementes e grãos comerciais de soja no Brasil-safra 2016/17. Embrapa Soja-Documentos (INFOTECA-E). 2018.

Maldonado dos Santos, J. V., Sant'Ana, G. C., Wysmierski, P. T., Todeschini, M. H., Garcia, A., & Meda, A. R. (2022). Genetic relationships and genome selection signatures between soybean cultivars from Brazil and United States after decades of breeding. Scientific Reports, 12(1), 10663.

Martins, J. T., Rasmussen, J., Eriksen, J., Arf, O., De Notaris, C., & Moretti, L. G. (2022). Biological N fixation activity in soybean can be estimated based on nodule dry weight and is increased by additional inoculation. Rhizosphere, 24, 100589.

Mendonça, H. C., Pereira, L. F. P., Maldonado dos Santos, J. V., Meda, A. R., & Sant'Ana, G. C. (2022). Genetic Diversity and Selection Footprints in the Genome of Brazilian Soybean Cultivars. Frontiers in Plant Science, 13, 842571.

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. Postharvest biology and technology, 46(2), 99-118.

Patil G, et al. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theoretical and Applied Genetics. 2017;130(10):1975-1991.

Peoples ang L, et al. Use of CRISPR/Cas9 for symbiotic nitrogen fixation research in legumes. Prog Mol Biol Transl Sci. 2017;149:187-213.

Pitumpe Arachchige, P. S., Rosso, L. H. M., Hansel, F. D., Ramundo, B., Torres, A. R., Asebedo, R., ... & Jagadish, S. K. (2020). Temporal biological nitrogen fixation pattern in soybean inoculated with Bradyrhizobium. Agrosystems, Geosciences & Environment, 3(1), e20079.

Price AH, Townend J, Jones MP, Audebert A, Courtois B. Mapping QTLs associated with drought avoidance in upland rice grown in the Philippines and West Africa. Plant Mol Biol. 2002; 48: 683–695. pmid:11999843.

Prince, S. J., Song, L., Qiu, D., Maldonado dos Santos, J. V., Chai, C., Joshi, T., ... & Nguyen, H. T. (2015). Genetic variants in root architecture-related genes in a Glycine soja accession, a potential resource to improve cultivated soybean. BMC genomics, 16(1), 1-20.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics, 197(2), 573-589.

Reinprecht Y, et al. Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (Glycine max (L.) Merrill) germplasm. Genome. 2006;49(12):1510-1527.

Rodrigues JIS, et al. Mapping QTL for protein and oil content in soybean. Pesquisa Agropecuária Brasileira. 2010;45(5):472-480.

Santos MA, et al. Mapping of qtls associated with biological nitrogen fixation traits in soybean. Hereditas. 2013;150(2-3):

Santos MA, Nicolás MF, Hungria M. Identificação de QTL associados à simbiose entre Bradyrhizobium japonicum, B. Elkanii e soja. Pesquisa Agropecuária Brasileira. 2006;41:67-75.

Santos MS, Nogueira MA, Hungria M. Microbial inoculants: reviewing the past, discussing the present and previewing an outstanding future for the use of beneficial bacteria in agriculture. AMB Express. 2019;9(1):1-22.

Santos, M. A. D., Nicolás, M. F., & Hungria, M. (2006). Identificação de QTL associados à simbiose entre Bradyrhizobium japonicum, B. elkanii e soja. Pesquisa Agropecuária Brasileira, 41, 67-75.

Santos, M. A., Geraldi, I. O., Garcia, A. A. F., Bortolatto, N., Schiavon, A., & Hungria, M. (2013). Mapping of QTLs associated with biological nitrogen fixation traits in soybean. Hereditas, 150(2-3), 17-25.

Sebolt AM, Shoemaker RC, Diers BW. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Science. 2000;40(5):1438-1444.

Sediyama T, Pereira MG, Sediyama CS, Gomes JLL. Cultura da Soja, Parte I. UFV, Minas Gerais, 97p. 1993.

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., ... & Belzile, F. (2013). An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PloS one, 8(1), e54603.

Thanh, V. H., & Shibasaki, K. (1976). Major proteins of soybean seeds. A straightforward fractionation and their characterization. Journal of Agricultural and Food Chemistry, 24(6), 1117-1121.

Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC bioinformatics. 2017;18(1):1-7.

Torkamaneh D, Lemay MA, Belzile F. The pan-genome of the cultivated soybean (PanSoy) reveals an extraordinarily conserved gene content. Plant biotechnology journal. 2021 Sep;19(9):1852-62.

Torkamaneh, D., Chalifour, F. P., Beauchamp, C. J., Agrama, H., Boahen, S., Maaroufi, H., ... & Belzile, F. (2020). Genome-wide association analyses reveal the genetic basis of biomass accumulation under symbiotic nitrogen fixation in African soybean. Theoretical and Applied Genetics, 133, 665-676.

Torres, A. R., Grunvald, A. K., Martins, T. B., Santos, M. A. D., Lemos, N. G., Silva, L. A. S., & Hungria, M. (2015). Genetic structure and diversity of a soybean germplasm considering biological nitrogen fixation and protein content. Scientia Agricola, 72, 47-52.

Wang J, et al. Genetic mapping high protein content QTL from soybean 'Nanxiadou 25' and candidate gene analysis. BMC Plant Biol. 2021;21(1):1-13.

Wang, L., Yang, Y., Zhang, S., Che, Z., Yuan, W., & Yu, D. (2020). GWAS reveals two novel loci for photosynthesis-related traits in soybean. Molecular Genetics and Genomics, 295, 705-716.

Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y. C., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. Nat. Sci. Rev. 7, 1776–1786. doi: 10.1093/nsr/nwaa110

Wang, X., Zhou, S., Wang, J., Lin, W., Yao, X., Su, J., ... & Guan, Y. (2023). Genome-wide association study for

biomass accumulation traits in soybean. Molecular Breeding, 43(5), 33.

Wysmierski, P. T., & Vello, N. A. (2013). The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. Genetics and molecular Biology, 36, 547-555.

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., & Jiang, G. L. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. BMC genomics, 16, 1-11.

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., ... & Sun, S. (2019). A combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. International journal of molecular sciences, 20(23), 5915.

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., ... & Sun, S. (2019). A combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. International journal of molecular sciences, 20(23), 5915.

CONCLUSÃO GERAL

Nossos resultados revelaram a presença de loci genéticos de importância significativa para a fixação biológica de nitrogênio e o teor de proteína em soja em 100 acessos de soja datados de 1948 a 2019. As descobertas deste estudo têm implicações práticas para a melhoria da qualidade dos grãos e potencialização de desempenho simbiótico da soja. Esperamos que essas descobertas contribuam para o desenvolvimento de cultivares de soja mais produtivas, sustentáveis e nutritivas, beneficiando tanto a indústria quanto a segurança alimentar global.