

JÉSSICA NAYARA BASÍLIO SILVA

**UMA ABORDAGEM MÚLTIPLA PARA INVESTIGAR A HERANÇA DO
CONTEÚDO DE ÓLEO E PROTEÍNA DA SOJA BRASILEIRA**

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Doctor Scientiae*.

Orientador: Maximiller Dal-Bianco Lamas Costa

Coorientador: Cléberson Ribeiro

**VIÇOSA - MINAS GERAIS
2023**

**Ficha catalográfica elaborada pela Biblioteca Central da
Universidade Federal de Viçosa - Campus**

T

S586a
2023
Silva, Jéssica Nayara Basílio, 1993-
Uma abordagem múltipla para investigar a herança do conteúdo
de óleo e proteína na soja brasileira: / Jéssica Nayara Basílio Silva. -
Viçosa, MG, 2023.
1 tese eletrônica (73 f.): il. (algumas color.).
Orientador: Maximiller Dal-Bianco Lamas Costa
Tese (doutorado) - Universidade Federal de Viçosa, Departamento
de Bioquímica e Biologia Molecular, 2023.
Inclui bibliografia.
DOI: <https://doi.org/10.47328/ufvbbt.2023.572>
Modo de acesso: World Wide Web.

1. Soja - Melhoramento genético; 2. Marcadores genéticos; 3.
Cultivos agrícolas - Rendimento; I. Costa, Maximiller Dal-Bianco
Lamas II. Universidade Federal de Viçosa.. Departamento de
Bioquímica e Biologia Molecular. Programa de Pós-Graduação em
Bioquímica Aplicada III. Título

CDD 22. ed. 631.52

Bibliotecário(a) responsável: EUZEBIO LUIZ PINTO CRB-6/3317

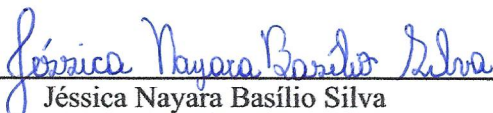
JÉSSICA NAYARA BASÍLIO SILVA

**UMA ABORDAGEM MÚLTIPLA PARA INVESTIGAR A HERANÇA DO
CONTEÚDO DE ÓLEO E PROTEÍNA DA SOJA BRASILEIRA**

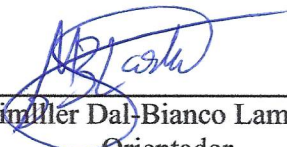
Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Doctor Scientiae*.

APROVADA: 31 de agosto de 2023.

Assentimento:



Jéssica Nayara Basílio Silva
Autora



Maximilner Dal-Bianco Lamas Costa
Orientador

**À minha Mãe Maria Aparecida,
Ao meu pai Antônio,
Aos meus demais familiares e amigos.**

AGRADECIMENTOS

Agradeço a Deus, por direcionar os meus passos.

Agradeço aos meus pais, Antônio e Maria Aparecida e ao meu irmão Jefferson, por todo o apoio e suporte que sempre me ofereceram.

Ao meu amigo e namorado Cleiton, por uma relação baseada no respeito e admiração.

À Universidade Federal de Viçosa (UFV) e ao Programa de Pós-Graduação em Bioquímica Aplicada, pela oportunidade.

Ao Laboratório de Bioquímica Genética de Plantas (LBGP – BIOAGRO-UFV) e ao Programa de Melhoramento de Qualidade de Soja (BIOAGRO/DBB/UFV), por todo o aprendizado.

Ao Laboratório de Genética Molecular de Plantas (BIOAGRO-UFV) e ao Prof. Cléberson Ribeiro, pelo suporte.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Fundação de Apoio à Pesquisa do Estado de Minas Gerais (FAPEMIG), pelo apoio financeiro.

Aos amigos, Roberta, Bruna e Thaís pelo apoio e amizade de uma vida inteira.

Aos funcionários José Carlos, Reginaldo e Luiz e aos estudantes Yan, Teresinha, Brizza, Uilton, Bruno, Bruna e Vitória do Laboratório de Bioquímica Genética de Plantas. Em especial, a Marina e Rodrigo, pela amizade, confiança e companheirismo em todos os momentos, dentro e fora do laboratório.

Ao pesquisador Dr. Newton Deniz Piovesan, pelo acompanhamento nos experimentos.

Ao Dr. Luíz Cláudio Costa Silva, pelo auxílio em análises e conselhos.

Ao coorientador Dr. Rafael Delmond Bueno, pelos conselhos e aprendizados.

Ao meu orientador Prof. Maximiller Dal-Bianco Lamas Costa, pelo respeito, pelos aprendizados, pela paciência infinita e pela confiança que sempre teve no meu trabalho.

Ao Dr. Thiago Ferreira, Prof. Luiz Cláudio Costa Silva, Prof. Dr. Cléberson Ribeiro, Prof. Dr. Pedro Reis e ao Prof. Dr. Humberto Ramos, meus agradecimentos pela contribuição e disponibilidade de avaliar meu trabalho.

RESUMO

SILVA, Jéssica Nayara Basilio, D.Sc., Universidade Federal de Viçosa, agosto de 2023. **Uma abordagem múltipla para investigar a herança do conteúdo de óleo e proteína da soja brasileira.** Orientador: Maximiller Dal-Bianco Lamas Costa. Coorientador: Cléberson Ribeiro.

A cultura da soja representa uma fonte valiosa de proteína e óleo na alimentação humana e animal. Os melhoristas de soja têm focado em produzir variedades com alto conteúdo de proteína e óleo, sem esquecer da melhoria do potencial de rendimento e produtividade. Essas características são muito desejadas e impulsionam o sucesso e o valor da soja. Neste sentido, a tese visa aprofundar os conhecimentos sobre marcadores moleculares para entender a herança e aumentar o conteúdo de óleo e proteína no grão da soja brasileira. No primeiro capítulo foram utilizados dados disponíveis na literatura para investigar marcadores SNP associados ao conteúdo de proteína e óleo na soja brasileira. Através desta pesquisa, vinte e dois novos QTLs (Quantitative Trait Loci) foram identificados. Destaca-se a influência substancial dos ancestrais Roanoke, Lee e Bragg na composição genética das variedades brasileiras de soja. No segundo capítulo foi desenvolvida uma população de 476 linhas endogâmicas recombinantes (RILs) a partir do cruzamento entre BARC-08 e SUPREMA e posteriormente avaliada em quatro diferentes ambientes. Dos marcadores moleculares avaliados, ss61, ss73, ss134 e ss185 associaram para o conteúdo de proteína em um dos ambientes. O marcador ss115 associou tanto para o conteúdo de óleo, quanto para proteína em todas as localidades, com provável efeito pleiotrópico. Além disso, foram apontados quatro genes candidatos (Glyma.15g034400, Glyma.15g050900, Glyma.15g063000 e Glyma.15g070700). No terceiro capítulo foi realizado o desenvolvimento de uma população de linhagens quase isogênicas (Near Isogenic Lines - NILs) com o objetivo de analisar um loci de características quantitativas (QTL) na linha de soja BR8014887 e posteriormente comprovar o possível gene causal. Os dados apresentados apresentaram importantes discussões sobre a região acerca do Glyma.20g084500, no cromossomo 20. O Glyma.20g084500 é o único gene presente na região alvo e pode ser o gene causal associado ao aumento do conteúdo da proteína.

Palavras-chave: Óleo. Proteína. SNP. Melhoramento genético.

ABSTRACT

SILVA, Jéssica Nayara Basilio, D.Sc., Universidade Federal de Viçosa, August 2023. **A multi-pronged approach to investigate the inheritance of oil and protein content of Brazilian soybeans.** Advisor: Maximiller Dal-Bianco Lamas Costa. Co-adviser: Cléberson Ribeiro.

Soybean cultivation represents a valuable source of protein and oil in human and animal nutrition. Soybean breeders have focused on producing varieties with high protein and oil content, without forgetting to improve yield and productivity potential. These characteristics are highly desired and drive the success and value of soybeans. In this sense, the thesis aims to deepen knowledge about molecular markers to understand the inheritance and increase the oil and protein content in Brazilian soybeans. In the first chapter, data available in the literature was used to investigate SNP markers associated with protein and oil content in Brazilian soybeans. Through this research, twenty-two new QTLs (Quantitative Trait Loci) were identified. The substantial influence of the Roanoke, Lee and Bragg ancestors on the genetic composition of Brazilian soybean varieties stands out. In the second chapter, a population of 476 recombinant inbred lines (RILs) was developed from the cross between BARC-08 and SUPREMA and subsequently evaluated in four different environments. Of the molecular markers evaluated, ss61, ss73, ss134 and ss185 were associated with protein content in one of the environments. The ss115 marker was associated with both oil and protein content in all locations, with a probable pleiotropic effect. Furthermore, four candidate genes were identified (Glyma.15g034400, Glyma.15g050900, Glyma.15g063000 and Glyma.15g070700). In the third chapter, the development of a population of almost isogenic lines (Near Isogenic Lines - NILs) was carried out with the aim of analyzing a quantitative trait loci (QTL) in the soybean line BR8014887 and subsequently proving the possible causal gene. The data presented important discussions about the region surrounding Glyma.20g084500, on chromosome 20. Glyma.20g084500 is the only gene present in the target region and may be the causal gene associated with the increase in protein content.

Keywords: Oil. Protein. SNP. Genetical enhancement.

LISTA DE ILUSTRAÇÕES

Capítulo 1

- Figure 1. Principal component analysis (PCA) showing the distribution of 620 accessions from the American continent available in the study developed by Bandillo et al. (2015)..... 28
- Figure 2. UPGMA phylogenetic tree of 177 genotypes using ancestors from the USA, Brazil, Japan and China.....29
- Figure 3. Manhattan plot showing significantly associated SNPs detected in genomic-wide association study (GWAS) for oil (A) and protein (B) content, using $p < 0.0001$ and $MAF > 0.05$. Each color represents a chromosome and each colored dot represents an SNP..... 30
- Figure 4. Allele frequency distribution for nine markers based on protein and oil content for genotypes from the Americas and modern Brazilian cultivars. Blue represents oil allele and orange represents protein allele.....31

Capítulo 2

- Figura 1. Distribuição do impacto da mutação para cada um dos genes estudados..... 55

Capítulo 3

- Figura 1. Gel de agarose mostrando testes com primers encontrados na literatura em três diferentes temperaturas de anelamento (62°C, 64°C e 66°C) para os genótipos de BR8014887 e NT12..... 65
- Figura 2. Gel de agarose mostrando primers encontrados na literatura e cultivares conhecidas pelo Programa PMQS (ANSC83022, UFVTN105AP). O indivíduo 16-18 representa um indivíduo controle entre o cruzamento de ANSC83022 e UFVTN105AP.....65

LISTA DE TABELAS

Capítulo 1

Table 1. Allelic frequency of SNP markers ss715621777 (SWEET30) and ss715637277 (POWR1) was evaluated in the 604 individuals used in this study and the 12 Brazilian ancestors.....	32
Table 2. SNPs markers identified through genome-wide association study (GWAS) associated with protein and oil content. The first columns represent the SNP markers and their respective chromosome and position. Then p-value, R ² and means for protein and oil content.....	33
Table 3. Polymorphic markers identified in this study and phenotyped mean data obtained by near-infrared (NIR) spectrometry on the 34 Brazilian genotypes. The first column represents the SNP markers. The second is averages for protein content and the third is averages for oil content.....	34

Capítulo 2

Tabela 1. Plantios e média dos dados fenotípicos dos 34 genótipos doadores do programa para conteúdo de e óleo (%) realizados por espectrometria de infravermelho próximo (NIR).....	56
Tabela 2. Genes diferencialmente expressos na região próxima ao marcador SNP ss115. Valores marcados em cinza representam FC>2.....	57

Capítulo 3

Tabela 1. Oligonucleotídeos iniciadores (primers) (R/F) utilizados neste estudo.....	61
Tabela 2. Marcadores utilizados no estudo para selecionar os nove indivíduos que possuíam combinações promissoras dentro da região alvo.....	67
Tabela 3. Marcadores SSR, SNP e da literatura utilizados neste estudo e indivíduos selecionados como promissores (163, 251, 499, 517).....	67
Tabela 4. Primers disponíveis na literatura corroboram os resultados que obtivemos do SNP 190.....	68
Tabela 5. Recombinações finais obtidas através do desenvolvimento das NILs até o momento pelo PMQS.....	69

SUMÁRIO

INTRODUÇÃO.....	11
REFERÊNCIAS.....	13
1. CAPÍTULO 1.....	15
1.1 INTRODUCTION.....	16
1.2 MATERIAL AND METHODS.....	17
1.2.1 Plant Materials Data.....	17
1.2.2 Genotyping and Phenotyping.....	17
1.2.3 Data Analysis.....	18
1.3 RESULTS AND DISCUSSION.....	18
1.3.1 Availability of SNP Marker Data for Oil and Protein Content in Soybeans.....	18
1.3.2 Representation of Brazilian Soybeans in GWAS Studies using SoySNP50K is Limited.....	19
1.3.3 Phylogenetic Analysis Reveals the Ancestry of Brazilian Soybeans.....	19
1.3.4 Low Frequency of Important QTLs for Protein and Oil Content in Brazilian Genotypes.....	20
1.3.5 GWAS Reveals New QTLs Associated with Oil and Protein Content.....	21
1.3.6 Polymorphism in Brazilian Cultivars.....	21
ACKNOWLEDGEMENT.....	23
REFERENCES.....	24
2. CAPÍTULO.....	35
2.1 INTRODUÇÃO.....	36
2.2 MATERIAL E MÉTODOS.....	37
2.2.1 Seleção de marcadores.....	37
2.2.2 Desenvolvimento da população.....	38
2.2.3 Ensaios: locais, épocas e tratos culturais.....	39
2.2.4 Delineamento experimental.....	39
2.2.5 Genotipagem da população.....	40
2.2.6 Fenotipagem da população.....	40
2.2.7 Análise de dados.....	40
2.2.8 Mapeamento dos reads e Análise de expressão diferencial.....	41
2.2.9 Identificação de genes candidatos.....	41
2.3 RESULTADOS E DISCUSSÃO.....	42
2.3.1 Associação de marcadores moleculares às características avaliadas.....	42
2.3.2 Genes diferencialmente expressos no cromossomo 15.....	44
AGRADECIMENTOS.....	48
REFERÊNCIAS.....	49
3. CAPÍTULO 3.....	58
3.1 INTRODUÇÃO.....	58
3.2 MATERIAL E MÉTODOS.....	60

3.2.1 Material Vegetal.....	60
3.2.2 Genotipagem.....	60
3.2.3 Desenho de primers.....	61
3.2.4 Desenvolvimento de população de linhagens quase isogênicas (Near Isogenic Lines - NILs).....	61
3.2.5 Análises Estatísticas.....	62
3.3 RESULTADOS E DISCUSSÃO.....	62
3.3.1 Desenvolvimento de população de linhagens quase isogênicas (Near Isogenic Lines - NILs).....	63
REFERÊNCIAS.....	70
CONSIDERAÇÕES FINAIS.....	73

INTRODUÇÃO

A soja (*Glycine max* (L.) Merr.) é uma das leguminosas mais importantes e cultivadas do mundo. Atualmente se destaca como a principal commodity agrícola brasileira, sendo o Brasil o maior produtor e exportador mundial. Representa uma fonte valiosa de proteína e óleo na nutrição humana, alimentação animal e matéria-prima industrial (Zhang et al. 2018). Essa cultura compreende cerca de 20% do conteúdo médio de óleo e 40% de proteína na composição do grão (Zhang et al. 2015).

Os melhoristas de soja têm o objetivo de produzir variedades com alto conteúdo de proteína e óleo (Li et al. 2018; Van et al. 2017; Tripathi e Khare, 2016). Aliado a isso, a melhoria do potencial de rendimento e produtividade é um dos objetivos mais importantes dos programas de melhoramento de soja (Liang et al. 2016). Essas características de interesse não são facilmente identificadas, além disso, são herdadas quantitativamente por múltiplos genes, que têm pequenos efeitos e são consideravelmente influenciadas pelo ambiente (Diers et al. 1992; Patil et al. 2018). Ambas as características são muito desejadas e impulsionam o sucesso e o valor da soja (Qi et al. 2016).

A partir de técnicas moleculares avançadas, muitos loci de características quantitativas (QTLs) de proteína e óleo de soja estão sendo estudados nos últimos anos. Identificar e estudar esses QTLs têm grande importância para o direcionamento do melhoramento molecular, identificando genes e funções que afetam o conteúdo dessas características (Zhang et al. 2019; Qi et al. 2016).

Diante disso, esse estudo tem como objetivo aprofundar os conhecimentos sobre marcadores moleculares para entender a herança e aumentar o conteúdo de óleo e proteína no grão da soja brasileira. Para facilitar o entendimento, a tese foi dividida em três capítulos. O primeiro capítulo é o artigo submetido na revista *Bragantia*, abordando a diversidade dos genótipos brasileiros para encontrar novos marcadores e novos QTLs associados ao conteúdo

de proteína em soja. O segundo capítulo tem como objetivo identificar e validar marcadores SNP associados ao conteúdo de óleo em soja. O terceiro capítulo é a apresentação de análises e dados sobre o mapeamento de um QTL para identificação de um gene causal para o conteúdo de proteína em soja.

REFERÊNCIAS

- Diers, B. W., Keim, P., Fehr, W. R. and Shoemaker, R. C. (1992). RFLP analysis of soybean seed protein and oil content. *Theor Appl Genet*, 83, 608–612. <https://doi.org/10.1007/BF00226905>
- Liang, H., Xu, L., Yu, Y., Yang, H., Dong, W. and Zhang, H. (2016). Identification of QTLs with main, epistatic and QTL by environment interaction effects for seed shape and hundred-seed weight in soybean across multiple years. *Journal of genetics*, 95, 475-477. <https://doi.org/10.1007/s12041-016-0648-8>
- Li, Y. H., Reif, J. C., Hong, H. L., Li, H. H., Liu, Z. X., Ma, Y. S. and Qiu, L. J. (2018). Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant Science*, 266, 95-101. <https://doi.org/10.1016/j.plantsci.2017.04.013>
- Patil, G., Do, T., Vuong, T. D., Valliyodan, B., Lee, J. D., Chaudhary, J. and Nguyen, H. T. (2016). Genomic-assisted haplotype analysis and the development of high throughput SNP markers for salinity tolerance in soybean. *Scientific Reports*, 6, 19199. <https://doi.org/10.1038/srep19199>
- Qi, Z., Pan, J., Han, X., Qi, H., Xin, D., Li, W. and Chen, Q. (2016). Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. *Molecular Breeding*, 36, 55. <https://doi.org/10.1007/s11032-016-0475-x>
- Tripathi, N. and Khare, D. (2016). Molecular approaches for genetic improvement of seed quality and characterization of genetic diversity in soybean: a critical review. *Biotechnology letters*, 38, 645-1654, 2016. <https://doi.org/10.1007/s10529-016-2154-8>
- Van, K. and McHale, L. K. (2017). Meta-analyses of QTLs associated with protein and oil contents and compositions in soybean seed. *International journal of molecular sciences*, 18, 1180, 2017. <https://doi.org/10.3390/ijms18061180>
- Zhang, J., Wang, X., Lu, Y., Bhusal, S. J., Song, Q., Cregan, P. B. and Jiang, G. L. (2018). Genome-wide scan for seed composition provides insights into soybean quality improvement and the impacts of domestication and breeding. *Molecular plant*, 11, 460-472. <https://doi.org/10.1016/j.molp.2017.12.016>
- Zhang, D., Zhang, H., Hu, Z., Chu, S., Yu, K., Lv, L., Zhang, D., Zhang, H., Hu, Z., Chu, S., Yu, K., Lv, L., Yang, Y., Zhang, X., Chen, X., Kan, G., Tang, Y., An, Y. Q.C. and Yu, D. (2019). Artificial selection on GmOLEO1 contributes to the increase in seed oil during

soybean domestication. PLoS Genetics, 15, e1008267.
<https://doi.org/10.1371/journal.pgen.1008267>

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., Hou, W., Wu, C., Han, T. and 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, 5915.
<https://doi.org/10.3390/ijms20235915>

1. CAPÍTULO 1

Exploring SoySNP50K and USDA Germplasm collection data to find new QTLs associated with protein and oil content in Brazilian genotypes

ABSTRACT: Genetic diversity within a germplasm collection plays a vital role in the success of breeding programs. However, comprehending this diversity and identifying accessions with desirable traits pose significant challenges. This study utilized publicly available data to investigate SNP markers associated with protein and oil content in Brazilian soybeans. Through this research, twenty-two new QTLs (Quantitative Trait Loci) were identified, and we highlighted the substantial influence of Roanoke, Lee and Bragg ancestor on the genetic makeup of Brazilian soybean varieties. Our findings demonstrate that certain markers are being lost in modern cultivars, while others maintain or even increase their frequency. These observations indicate genomic regions that have undergone selection during soybean adaptation and could be valuable in breeding programs aimed at enhancing protein or oil content.

Key words:: *Glycine max*, protein, SNP, plant breeding

1.1 INTRODUCTION

Soybean (*Glycine max* (L.) Merrill) is a highly significant crop due to its economic value as a source of protein and oil (Clemente and Cahoon, 2009). It is utilized in animal feed, human food, and biodiesel production (Warrington et al. 2015). Originating from Southeast Asia, soybean was domesticated approximately 5000 years ago from its wild parent (Carter et al. 2004; Hymowitz et al. 1970). The domestication process led to the selection of certain alleles over others, resulting in the loss of beneficial alleles in cultivated species (Salgotra et al. 2022). According to Hyten et al. (2006), wild soybean species exhibit low levels of diversity, which explains the limited nucleotide diversity in elite cultivars. During domestication, approximately 50% of genetic diversity and 81% of rare alleles were lost, while around 60% of genes experienced significant changes in allele frequency (Hyten et al. 2006).

A reduction in diversity is an inevitable consequence of breeding. The development of new cultivars begins with a small set of parental lines selected from the entire population, aiming to generate higher-yielding lines but at the cost of narrowing the available genetic pool. Consequently, this process, combined with self-fertilization, leads to a decrease in variability (Viana et al. 2022).

One of the primary goals of modern soybean breeding programs is to identify genomic regions responsible for protein and oil accumulation in soybean seeds. Significant effects have been observed on chromosome 20 (Gm20) and chromosome 15 (Gm15), both displaying a negative correlation with these two traits. To date, only the genes SWEET39 and POWR1 have been characterized as associated with both traits, exhibiting pleiotropic effects (Bandillo et al. 2015; Bolon et al. 2010; Chung et al. 2003; Fliege et al. 2022; Goettel et al. 2021; Kim et al. 2023; Lee et al. 2019).

Despite the wealth of information in the literature regarding SNP markers associated with oil and protein content in ancestral genotypes, there is limited knowledge about the determinants of protein content in cultivars used in soybean-producing countries. In order to gain a better understanding of this topic, we conducted a genome-wide association study (GWAS) using public data. Our aim was to identify new SNPs present in lines from the American continent, which could serve as potential targets for use in marker-assisted selection programs to enhance protein or oil content in Brazilian soybean cultivars.

1.2 MATERIAL AND METHODS

1.2.1 Plant Materials Data

Data from Bandillo et al. (2015) were utilized for this study, specifically focusing on materials from the American continent that included genotyping and phenotyping data. Additionally, we included 34 genotypes obtained from the germplasm bank of Programa de Melhoramento da Qualidade da Soja (PMQS/UFV).

1.2.2 Genotyping and Phenotyping

At the V4 stage, young and healthy leaves were collected for DNA extraction, following the protocol described by Dellaporta et al. (1983). Each sample was quantified using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), and DNA quality was verified by 0.8% agarose gel electrophoresis. Subsequently, samples were diluted to a working concentration of 10 ng/ μ L for SNP genotyping. Genotyping was performed using the KASP genotyping methodology (LGC Genomics, 2013) on an Applied Biosciences 7500 equipment (ABI7500). Oil and protein phenotyping were conducted using

near-infrared (NIR) spectrometry with an Antaris II FT-NIR analyzer (Thermo Fisher Scientific Brasil Instrumentos de Processo Ltda., São Paulo, SP, Brazil).

1.2.3 Data Analysis

The TASSEL statistical software was employed for generating principal component analysis (PCA), phylogenetic clustering, and genome-wide association studies (GWAS) analyses. GWAS was performed by filtering SNPs with a minimum allele frequency (MAF) > 0.05 and utilizing the mixed linear model (MLM).

Genetic-statistical analyses were conducted using the software programs GENES (Cruz, 2013), GQMOL (Cruz and Schuster, 2004), and TASSEL 5.0 (Bradbury et al. 2007).

1.3 RESULTS AND DISCUSSION

1.3.1 Availability of SNP Marker Data for Oil and Protein Content in Soybeans

In recent years, there has been a growing interest in data related to SNP markers associated with oil and protein content in soybeans. However, the majority of the available data is based on older genotypes or ancestral lines. The primary focus has been on identifying QTLs that were lost during soybean domestication and mapping candidate genes, while there has been relatively little exploration of QTLs that have been indirectly incorporated into breeding programs over the years. Consequently, the identification of QTLs associated with protein and oil content that are more prevalent in elite cultivars can significantly expedite the development of new soybean varieties.

1.3.2 Representation of Brazilian Soybeans in GWAS Studies using SoySNP50K is Limited

Through an initial PCA, we investigated the distribution of genotypes based on their geographical origin (Figure 1). The results showed a substantial number of genotypes from the USA, while the representation from countries like Argentina and Brazil was limited. The majority of studies utilizing these public databases, which serve as the cornerstone of current knowledge, predominantly focus on USA lines, leaving limited information available for other countries. In our study, upon evaluating the genotypes used, we found that only 15 Brazilian genotypes were included, indicating a low representation of Brazil, despite its status as one of the world's major soybean-producing regions.

1.3.3 Phylogenetic Analysis Reveals the Ancestry of Brazilian Soybeans

To enhance our understanding of Brazilian soybeans ancestry, we constructed a phylogenetic tree using and included 177 genotypes from Brazil available in SoySNP50K and the ancestral genotypes from USA (Bandillo et al. 2015; Ude et al. 2013), Brazil (Wysmierski and Vello 2013), Japan (Zhou et al. 2000) and China (Cui et al. 2000; Li et al. 2001; Liu et al. 2017) (Figure 2). The analysis revealed the formation of four distinct clusters. Group IV consisted of Chinese and Brazilian ancestors with no Brazilian genotypes present. Group III included 2 Brazilian genotypes with American and Chinese ancestors, as well as several Brazilian ancestors. Group I included 13 Brazilian genotypes that clustered with Chinese, American, Japanese, and Brazilian ancestors. Subgroup II comprised the majority of Brazilian genotypes (161), which clustered with American ancestors and the common ancestor Roanoke, found in both the USA and Brazil ancestry. Interestingly, two test samples widely reported as ancestors of Brazilian soybeans, Lee and Bragg, were also grouped in this cluster. Previous studies have indicated that PI 548485 (Roanoke), PI 548445 (CNS), PI 548493

(Tokyo), and PI 548488 (S-100) are the most important Brazilian ancestors, accounting for 55.3% of the Brazilian soybean germplasm. However, our data suggest that the ancestry of Brazilian soybeans may be even more limited than previously reported. It appears that the genetic foundation of Brazilian soybeans was established from a small number of genotypes, with a notable influence from Roanoke, Lee, and Bragg.

1.3.4 Low Frequency of Important QTLs for Protein and Oil Content in Brazilian Genotypes

In recent years, new discoveries have shed light on important QTLs associated with protein and oil accumulation, such as the POWR1 gene and the SWEET39 gene. In our research group, we utilized SNP markers (ss715621777 for SWEET39 and ss715637277 for POWR1) to introgress these QTLs. Assessing the presence of these two SNPs in the 604 individuals used in our study, we observed that they are considered rare (Table 1). The allele frequency of ss715621777 was approximately 4.14%, while ss715637277 had an allele frequency of approximately 1%. Notably, none of the Brazilian ancestors carried the protein allele in SNP ss715621777, while only the Arksoy ancestor carried the protein allele in SNP ss715621777. These findings indicate that Brazilian ancestors contribute alleles favorable to oil accumulation, suggesting that increased protein content could be achieved through the introgression of protein alleles from other sources. It is important to emphasize that the protein content inheritance is complex and multigenic. In Brazil, it is possible to find cultivars containing more than 40% but missing these important alleles, and for this reason, it is interesting to find other allele sources to increase the protein content.

1.3.5 GWAS Reveals New QTLs Associated with Oil and Protein Content

A genome-wide association study (GWAS) for oil and protein content was conducted, resulting in the identification of 22 SNPs possibly associated with both traits (Table 2). The observed genetic contribution (R^2) of these SNPs ranged from 1.796% to 4.859%. Notably, the marker ss715637399 was identified in both protein and oil analyses. Several markers exhibited considerable additive effects on protein (ss715597897, ss715606597, and ss715637399) and oil (ss715589137, ss715598567, ss715602783, ss715599966, ss715612467, ss715617832, and ss715626186) content (Table 2).

1.3.6 Polymorphism in Brazilian Cultivars

Limited information is available in the literature regarding important QTLs associated with protein content in Brazilian soybeans. The identification of 22 new QTLs for American genotypes in our study could provide valuable sources for use in breeding programs, particularly if these QTLs are present in cultivars or lines maintained in germplasm banks. To test this hypothesis, we selected nine SNPs with favorable additive effects and genotyped 34 Brazilian genotypes, including high-protein donors from our breeding program and cultivars from different companies.

Among the nine evaluated SNPs, two were monomorphic (ss715589137 and ss715602783), while four were polymorphic (ss715606597, ss715599966, ss715626186, and ss715598567). The presence of monomorphic SNPs suggests that these markers may already be well established and associated with important traits related to soybean adaptability in Brazilian environments. However, for ss715637399, only the cultivar SUPREMA displayed a different genotype, while the genotype VX04 5692-62 was heterozygous. For ss715612467, only the individual TMG 2181 IPRO was heterozygous, and for ss715602783, five individuals from the company TMG were identified as heterozygotes (Data not shown).

When analyzing the polymorphic materials, we observed that the SNP marker ss715606597 led to an average increase in protein content from 39.694% to 42.326% and a corresponding decrease in oil content from 21.346% to 20.544% (Table 3). Similarly, the marker ss715598567 exhibited an increase in protein content from 41.445% to 42.845% and a decrease in oil content from 21.346% to 20.544%. When calculating the allele frequencies for the nine SNPs, it became evident that monomorphic or heterozygous markers are being excluded in modern cultivars according to our sampling, potentially due to selection for oil and/or productivity in Brazilian breeding programs. However, among the polymorphic markers, their frequencies are maintained or increasing, as observed in the case of the ss715598567 marker (Figure 4). It is worth noting that this protein marker was found in the cultivars M7739IPRO, M8372IPRO, TMG1176RR, and TMG2181IPRO, as well as in eight protein donors from our breeding program, making it an important marker worthy of further investigation.

In conclusion, our data clearly demonstrate that out of the nine markers identified, four are polymorphic and can be readily utilized in breeding programs aimed at increasing protein or oil content. However, we emphasize the importance of validating the effect of these markers in structured segregating populations before their application in breeding programs. Among these markers, the two most promising ones were ss715606597 and ss715598567. The marker ss715606597, located on chromosome 10 at position 38301886, exhibited a significant difference in protein content among genotypes (Table 3). This marker demonstrates stability and represents a viable alternative for use in breeding programs targeting increased protein content. On the other hand, the ss715598567 marker, located on chromosome 7 at position 7031910, did not show a significant difference in protein content among genotypes (Table 3). However, it exhibited a considerable increase and displayed an

elevated frequency in the 34 Brazilian cultivars, making it an important marker to be further investigated.

ACKNOWLEDGEMENT

The work was funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (312641/2021-4 to MD), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (APQ-01416-16, PPM-640-18 and RED-00124-22 to MD) and Caramuru Alimentos (Grant to MD).

REFERENCES

- Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J. and Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*, 8, 1-13. <https://doi.org/10.3835/plantgenome2015.04.0024>
- Bolon, Y. T., Joseph, B., Cannon, S. B., Graham, M. A., Diers, B. W., Farmer, A. D. and Vance, C. P. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biology*, 10, 1-24. <https://doi.org/10.1186/1471-2229-10-41>
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatic*, 23, 2633-2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Carter Jr, T. E., Nelson, R. L., Sneller, C. H. and Cui, Z. (2004). Genetic diversity in soybean. *Soybeans: Improvement, production, and uses*, 16, 303-416. <https://doi.org/10.2134/agronmonogr16.3ed.c8>
- Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B., Specht, J. E. (2003). The seed protein, oil, and yield QTL on soybean linkage group I. *Crop science*, 43, 1053-1067. <https://doi.org/10.2135/cropsci2003.1053>
- Clemente, T. E. and Cahoon, E. B. (2009). Soybean Oil: Genetic Approaches for Modification of Functionality and Total Content. *Plant physiology*, 151, 1030-1040. <https://doi.org/10.1104/pp.109.146282>
- Cruz, C. D. and Carneiro, P. C. S. (2003). Modelos biométricos aplicados ao melhoramento de plantas. Viçosa: Editora da UFV.
- Cruz, C. D. (2013). Programa GENES: Estatística experimental e matrizes. Viçosa: Editora da UFV.
- Cruz, C. D. and Schuster, I. (2004). GQMOL: aplicativo computacional para análise de dados moleculares e de suas associações com caracteres quantitativos. Viçosa: Editora da UFV.
- Cui, Z., Carter, T. E. and Burton, J. W. (2000). Genetic base of 651 Chinese soybean cultivars released during 1923 to 1995. *Crop Science*, 40, 1470-1481. <https://doi.org/10.2135/cropsci2000.4051470x>
- Dellaporta, S. L., Wood, J. and Hicks, J. B. (1983). A plant DNA miniprep: version II. *Plant molecular biology reporter*, 1, 19-21.

Fliege, C. E., Ward, R. A., Vogel, P., Nguyen, H., Quach, T., Guo, M. and Diers, B. W. (2022). Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. *The Plant Journal*, 110, 114-128. <https://doi.org/10.1111/tpj.15658>

Genomics, L. G. C. (2013). KASP genotyping chemistry User guide and manual. LGC Limited, Teddington, UK.

Hiromoto, D. M. (1986). The genetic base of Brazilian soybean (*Glycine max* L. Merrill) cultivars. *Rev. Bras. Genet*, 9, 295-306.

Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D. and An, Y. Q. C. (2022). POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nature Communications*, 13, 3051. <https://doi.org/10.1038/s41467-022-30314-7>

Gwinner, R., Setotaw, T. A., Pasqual, M., Santos, J. B. D., Zuffo, A. M., Zambiazzi, E. V. and Bruzi, A. T. (2017). Genetic diversity in Brazilian soybean germplasm. *Crop Breeding and Applied Biotechnology*, 17, 373-381. <https://doi.org/10.1590/1984-70332017v17n4a56>

Hymowitz, T. (1970). On the domestication of the soybean. *Economic botany*, 24, 408-421. <https://doi.org/10.1007/BF02860745>

Hyten, D. L., Song, Q., Zhu, Y., Choi, I. Y., Nelson, R. L., Costa, J. M. and Cregan, P. B. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences*, 103, 16666-16671. <https://doi.org/10.1073/pnas.0604379103>

Kim, W. J., Kang, B. H., Moon, C. Y., Kang, S., Shin, S., Chowdhury, S. and Ha, B. K. (2023). Quantitative Trait Loci (QTL) Analysis of Seed Protein and Oil Content in Wild Soybean (*Glycine soja*). *International Journal of Molecular Sciences*, 24, 4077. <https://doi.org/10.3390/ijms24044077>

Lee, S., Van, K., Sung, M., Nelson, R., LaMantia, J., McHale, L. K. and Mian, M. R. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theoretical and Applied Genetics*, 132, 1639-1659. <https://doi.org/10.1007/s00122-019-03304-5>

Li, Z., Qiu, L., Thompson, J. A., Welsh, M. M. and Nelson, R. L. (2001). Molecular genetic analysis of US and Chinese soybean ancestral lines. *Crop Science*, 41, 1330-1336. <https://doi.org/10.2135/cropsci2001.4141330x>

Liu, Z., Li, H., Wen, Z., Fan, X., Li, Y., Guan, R., Guo, Y., Wang, S., Wang, D. and Qiu, L. (2017). Comparison of genetic diversity between Chinese and American soybean (*Glycine max* (L.)) accessions revealed by high-density SNPs. *Frontiers in plant science*, 8, 2014. <https://doi.org/10.3389/fpls.2017.02014>

- Maldonado dos Santos, J. V., Sant'Ana, G. C., Wysmierski, P. T., Todeschini, M. H., Garcia, A. and Meda, A. R. (2022). Genetic relationships and genome selection signatures between soybean cultivars from Brazil and the United States after decades of breeding. *Scientific Reports*, 12, 1-12. <https://doi.org/10.1038/s41598-022-15022-y>
- Mendonça, H. C., Pereira, L. F. P., Maldonado dos Santos, J. V., Meda, A. R. and Sant'Ana, G. C. (2022). Genetic Diversity and Selection Footprints in the Genome of Brazilian Soybean Cultivars. *Frontiers in Plant Science*, 13, 549. <https://doi.org/10.3389/fpls.2022.842571>
- Miao, L., Yang, S., Zhang, K., He, J., Wu, C., Ren, Y., Gai, J. and Li, Y. (2020). Natural variation and selection in GmSWEET39 affect soybean seed oil content. *New Phytologist*, 225, 1651-1666. <https://doi.org/10.1111/nph.16250>
- Salgotra, R. K., Thompson, M. and Chauhan, B. S. (2022). Unravelling the genetic potential of untapped crop wild genetic resources for crop improvement. *Conservation Genetics Resources*, 14, 109–124. <https://doi.org/10.1007/s12686-021-01242-3>
- Ude, G. N., Kenworthy, W. J., Costa, J. M., Cregan, P. B. and Alvernaz, J. (2003). Genetic diversity of soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. *Crop Science*, 43, 1858-1867. <https://doi.org/10.2135/cropsci2003.1858>
- Viana, J. P. G., Fang, Y., Avalos, A., Song, Q., Nelson, R. and Hudson, M. E. (2022). Impact of multiple selective breeding programs on genetic diversity in soybean germplasm. *Theoretical and Applied Genetics*, 135, 1591-1602. <https://doi.org/10.1007/s00122-022-04056-5>
- Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H. and Yu, D. (2016). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Scientific reports*, 6, 20728. <https://doi.org/10.1038/srep20728>
- Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S. and Boerma, H. R. (2015). QTL for seed protein and amino acids in the Benning× Danbaekkong soybean population. *Theoretical and applied genetics*, 128, 839-850. <https://doi.org/10.1007/s00122-015-2474-4>
- Wysmierski, P. T. and Vello, N. A. (2013). The genetic base of Brazilian soybean cultivars: evolution over time and breeding implications. *Genetics and molecular Biology*, 36, 547-555. <https://doi.org/10.1590/S1415-47572013005000041>
- Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M. L. and An, Y. Q. C. (2020). Selection of GmSWEET39 for oil and protein improvement in soybean. *PLoS Genetics*, 16, e1009114. <https://doi.org/10.1371/journal.pgen.1009114>

Zhou, X., Carter, T. E., Cui, Z., Miyazaki, S. and Burton, J. W. (2000). Genetic base of Japanese soybean cultivars released during 1950 to 1988. *Crop Science*, 40, 1794-1802.<https://doi.org/10.2135/cropsci2000.4061794x>

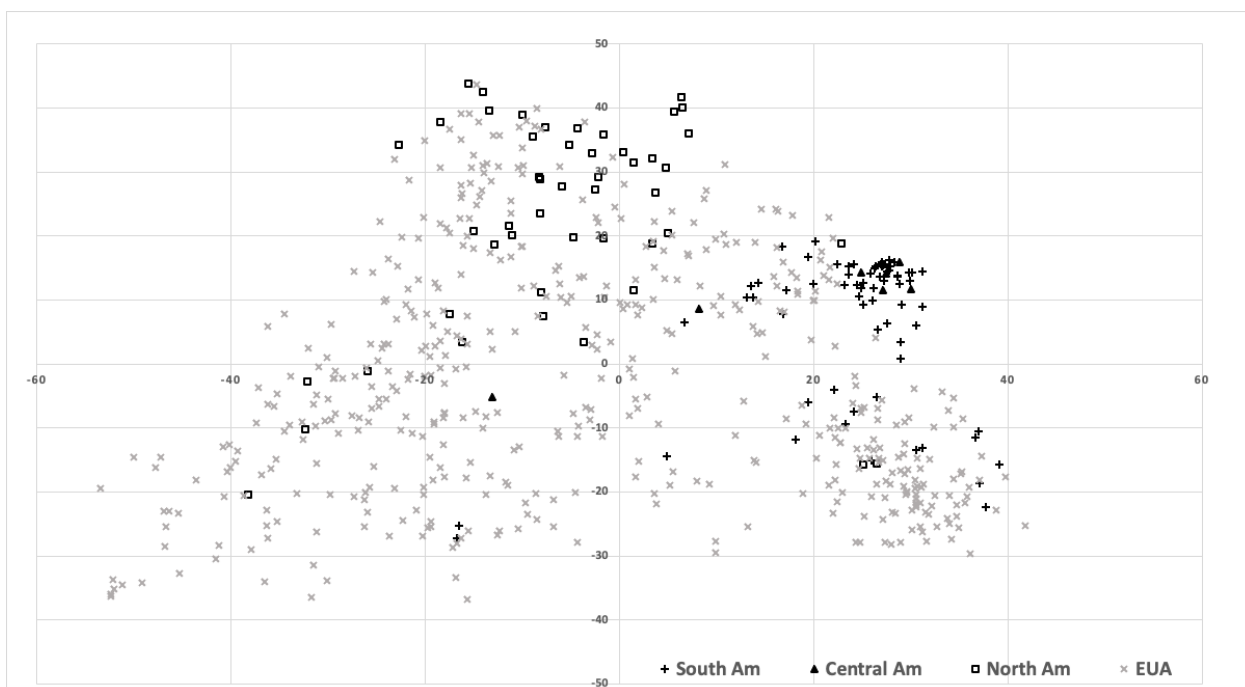


Figure 1. Principal component analysis (PCA) showing the distribution of 620 accessions from the American continent available in the study developed by Bandillo et al. (2015).

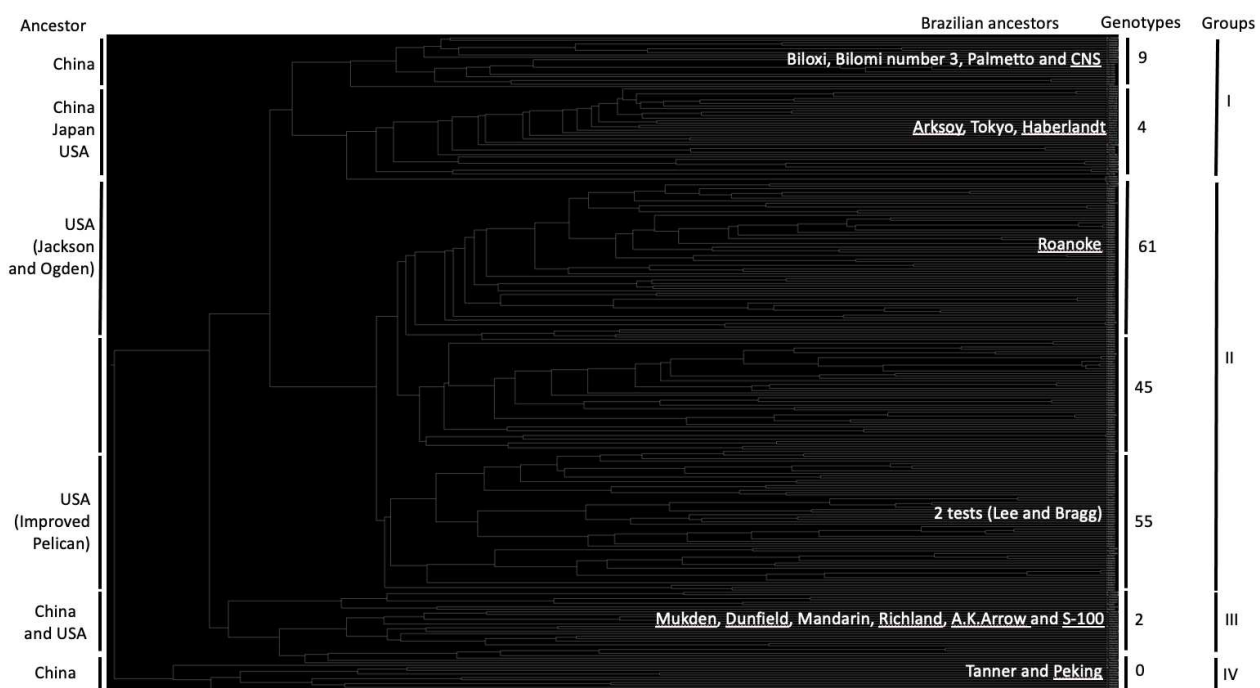


Figure 2. UPGMA phylogenetic tree of 177 genotypes using ancestors from the USA, Brazil, Japan and China.

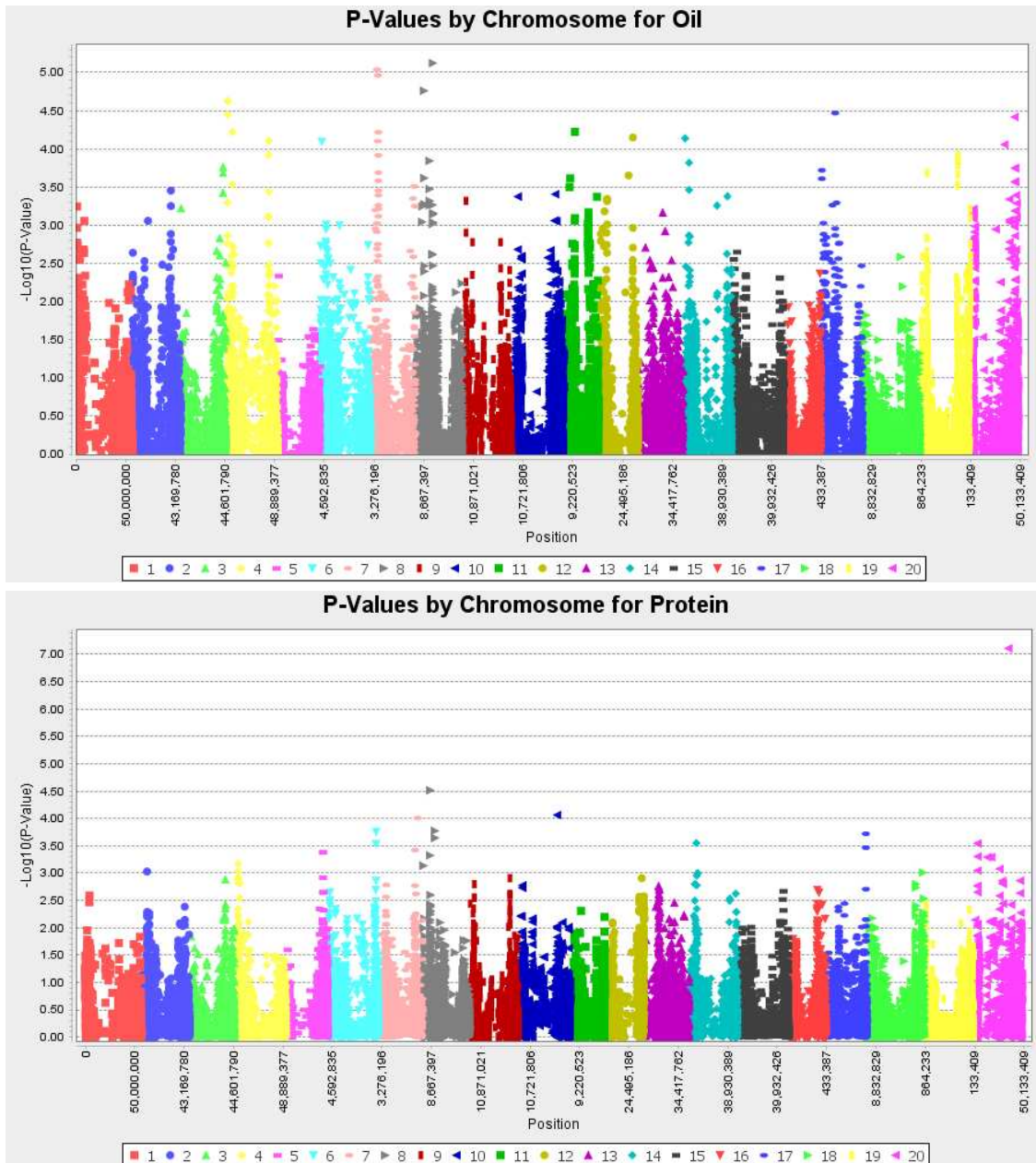


Figure 3. Manhattan plot showing significantly associated SNPs detected in genomic-wide association study (GWAS) for oil (A) and protein (B) content, using $p < 0.0001$ and $MAF > 0.05$. Each color represents a chromosome and each colored dot represents an SNP.

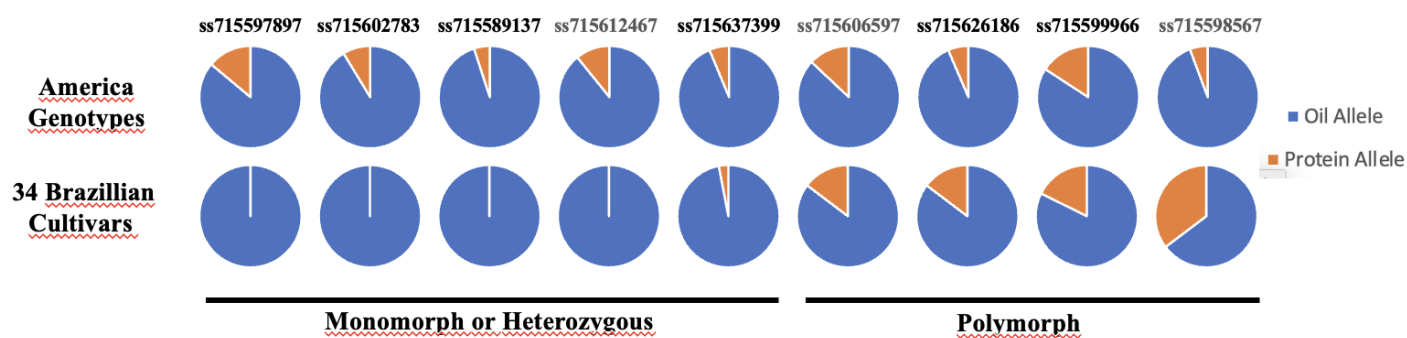


Figure 4. Allele frequency distribution for nine markers based on protein and oil content for genotypes from the Americas and modern Brazilian cultivars. Blue represents oil allele and orange represents protein allele.

		Allele Oil	Allele Protein	f (protein)
SWEET30 (ss715621777)	604 America Genotypes	577	25	4,139
	12 Brazillian Ancestors	11	1	0,166
POWR1 (ss715637277)	604 America Genotypes	598	6	0,993
	12 Brazillian Ancestors	12	0	0,000

Table 1. Allelic frequency of SNP markers ss715621777 (SWEET30) and ss715637277 (POWR1) was evaluated in the 604 individuals used in this study and the 12 Brazilian ancestors.

SNPcod	Chm	Position	Trait	p-value	R2	Protein Content			Oil Content		
						11	22	11-22	11	22	11-22
ss715587241	4	1775055	Oil	2,35E-05	0,02883	-	-	-	19,37	17,73	1,64
ss715587257	4	1826798	Oil	3,58E-05	0,02759	-	-	-	19,36	17,77	1,59
ss715587261	4	1834955	Oil	2,40E-05	0,02891	-	-	-	19,36	17,70	1,66
ss715589137	4	6314609	Oil	6,03E-05	0,02561	-	-	-	19,31	16,11	3,20
ss715587939	4	42790951	Oil	7,79E-05	0,02472	-	-	-	19,26	18,24	1,03
ss715593870	6	2104129	Oil	8,07E-05	0,02451	-	-	-	19,42	18,04	1,38
ss715598461	7	6273131	Oil	9,17E-06	0,03152	-	-	-	19,31	17,75	1,56
ss715598567	7	7031910	Oil	7,93E-05	0,02463	-	-	-	19,29	17,11	2,18
ss715598577	7	7181748	Oil	1,09E-05	0,03128	-	-	-	19,38	17,90	1,48
ss715598624	7	7536158	Oil	6,06E-05	0,02534	-	-	-	19,22	17,77	1,45
ss715597897	7	39631062	Protein	9,87E-05	0,02526	42,06	44,29	-2,23	-	-	-
ss715602692	8	7577565	Protein	3,08E-05	0,02922	42,47	41,05	1,42	-	-	-
ss715602783	8	8629168	Oil	1,74E-05	0,02936	-	-	-	19,45	16,26	3,19
ss715599966	8	17653606	Oil	7,57E-06	0,03245	-	-	-	19,56	17,12	2,44
ss715606597	10	38301886	Protein	8,70E-05	0,02577	42,07	44,37	-2,30	-	-	-
ss715609447	11	11214137	Oil	6,00E-05	0,02585	-	-	-	19,59	18,45	1,13
ss715612467	12	34710107	Oil	7,13E-05	0,02516	-	-	-	19,37	17,36	2,01
ss715617832	14	1629337	Oil	7,30E-05	0,0253	-	-	-	18,97	19,78	-0,82
ss715626186	17	14065524	Oil	3,41E-05	0,02708	-	-	-	19,36	16,27	3,09
ss715637399	20	34111616	Protein	7,78E-08	0,04859	42,13	46,02	-3,89	-	-	-
ss715637399	20	34111616	Oil	8,79E-05	0,02428	-	-	-	19,30	16,87	2,43
ss715638506	20	44272285	Oil	3,84E-05	0,02708	-	-	-	19,35	18,49	0,86

Table 2. SNPs markers identified through genome-wide association study (GWAS) associated with protein and oil content. The first columns represent the SNP markers and their respective chromosome and position. Then p-value, R2 and means for protein and oil content.

	Protein		Oil	
	Mean 11	Mean 22	Mean 11	Mean 22
ss715606597	42,326 b	39,6947 a	20,544 a	21,3463 a
ss715599966	42,082 a	41,2698 a	20,4972 a	21,4341 a
ss715626186	41,113	40,931	20,625	20,876
ss715598567	41,445	42,845	20,544	21,346

Table 3. Polymorphic markers identified in this study and phenotyped mean data obtained by near-infrared (NIR) spectrometry on the 34 Brazilian genotypes. The first column represents the SNP markers. The second is averages for protein content and the third is averages for oil content.

2. CAPÍTULO

Identificação e validação de marcadores SNP associados ao conteúdo de óleo em soja

[*Glycine max* (L.) Merr.]

RESUMO: O óleo de soja é uma fonte valiosa para a agroindústria, e aumentar essa característica tem sido um alvo importante no melhoramento da soja. No presente estudo desenvolvemos uma população de 476 linhas endogâmicas recombinantes (RILs) a partir do cruzamento entre BARC-08 e SUPREMA, e avaliamos os indivíduos em quatro diferentes ambientes. As RILs foram genotipadas pela metodologia KASP e fenotipadas para conteúdo de proteína e óleo por espectroscopia de infravermelho próximo (NIR). Dos marcadores moleculares avaliados, ss61, ss73, ss134 e ss185 associaram para o conteúdo de proteína em um dos ambientes. O marcador ss115 associou tanto para o conteúdo de óleo, quanto para proteína em todas as localidades, com provável efeito pleiotrópico. Foi realizada uma análise de RNASeq com os genitores para gerar uma lista de genes diferencialmente expressos (DEGs) na região alvo do cromossomo 15, e apontamos quatro genes candidatos (Glyma.15g034400, Glyma.15g050900, Glyma.15g063000 e Glyma.15g070700). Nossos dados demonstram que o marcador ss115 pode ser utilizado em programas de melhoramento visando aumentar o conteúdo de óleo e aponta novos genes que podem ser importantes para a característica.

Palavras-chave: *Glycine max*, óleo, SNP, gene candidatos.

2.1 INTRODUÇÃO

A soja (*Glycine max* (L.) Merr.) é uma das culturas oleaginosas mais importantes do mundo, sendo que o óleo de soja representa cerca de 18 a 21% do peso seco da semente de soja e é usado para consumo humano, bem como matéria-prima renovável para vários produtos industriais (Kumar et al. 2021).

Do ponto de vista econômico, o óleo de soja ocupa a segunda colocação em produção e oferta entre os óleos vegetais no mercado mundial com participação de 27% do total (USDA, 2023). O Brasil é o terceiro maior produtor de óleo de soja e segundo maior exportador (USDA 2023). Previsões demonstram que haverá demanda elevada no consumo de óleo de soja, cuja produção está estimada em torno de 10,7 milhões de toneladas, apresentando um aumento de 7,6% em comparação ao período passado (Abiove, 2023).

O conteúdo de óleo tem chamado muita atenção nos programas de melhoramento devido ao aumento da demanda por óleos vegetais. Com o processo de domesticação da soja houve seleção de sementes maiores e maior conteúdo de óleo, que resultaram em menor conteúdo de proteína (Hwang et al. 2014). O conteúdo de óleo correlaciona-se negativamente com o conteúdo de proteínas e positivamente com rendimento e tamanho de sementes (Bandillo et al. 2015; Chung et al. 2003; Hyten et al. 2004), o que dificulta a obtenção de linhagens com alto conteúdo de óleo e proteínas simultaneamente. Como a associação entre essas características é difícil, os programas de melhoramento estão buscando formas de aumentar uma característica, sem que ocorra diminuição em outra, e ainda levando em consideração o aumento de produtividade (Akond et al. 2013).

Uma estratégia promissora para acelerar o desenvolvimento de cultivares é a identificação de marcadores moleculares, locos de características quantitativas (QTLs) e genes funcionais associados a características de óleo (Eskandari et al. 2013; Phansak et al. 2016; Prioli et al. 2019). Até o momento, cerca de 322 locos de características quantitativas

(QTL) relacionados ao conteúdo óleo foram identificados usando populações biparentais (Kim et al. 2023). No total, 16 QTLs foram confirmados para conteúdo de proteína e óleo e estes estão distribuídos em 11 cromossomos, incluindo os cromossomos 15 e 20 (Kim et al. 2023).

Estudos recentes também descrevem genes funcionais importantes para controlar o conteúdo de óleo de semente, como FAD2-1A e FAD2-1B (Haun et al. 2014), GmDREBL (Zhang et al. 2016), GmZF351 (Li et al. 2017), B1 (Zhang et al. 2018), GmOLEO1 (Zhang et al. 2019), GmPDAT (Liu et al. 2020), GmSWEET39 (Miao et al. 2020; Wang et al. 2020; Zhang et al. 2020). Esses estudos auxiliam na compreensão do mecanismo do metabolismo lipídico da soja e fornecem informações e ferramentas úteis para programas de melhoramento de soja. Entretanto, pouco ainda se sabe sobre o acúmulo de óleo no grão de soja e em se tratando de uma característica multigênica, muita informação ainda é desconhecida.

No presente estudo, um total de 476 RILs foram desenvolvidas e avaliadas em quatro diferentes ambientes. Os dados fenotípicos foram associados com a genotipagem de seis marcadores SNP com o objetivo de estimar os efeitos desses marcadores para aplicação em programa de melhoramento voltado para o conteúdo de óleo em soja. Além disso, tentamos identificar genes candidatos por RNASeq dentro de loci de associação significativos que estavam potencialmente envolvidos na regulação da características.

2.2 MATERIAL E MÉTODOS

2.2.1 Seleção de marcadores

Trabalhos anteriores do Programa de Melhoramento da Qualidade da Soja (PMQS/BIOAGRO/DBB/UFV) selecionaram 269 SNPs na literatura e na plataforma online Soybase (<https://www.soybase.org/>). A seleção foi realizada com base em SNPs publicados

em estudos de associação genômica ampla (GWAS) e SNPs próximos a microssatélites com associação ao conteúdo de proteína, óleo e resistência à nematoides, principais linhas de pesquisa desenvolvidas pelo grupo de pesquisa. Posteriormente, foi realizada a genotipagem dos marcadores SNPs selecionados em 34 genótipos utilizados no programa de melhoramento do BIOAGRO/UFV. A empresa LGC Genomics realizou todo o procedimento de genotipagem com base nos 269 SNPs selecionados. Dos 269 SNPs selecionados, 230 foram utilizados e 39 SNPs foram inconclusivos (Bueno et al. 2022).

A fenotipagem dos 34 genótipos para óleo foi realizada por espectrometria de infravermelho próximo (NIR) (Tabela 1). Esses dados de fenotipagem e genotipagem foram utilizados para a associação nos programas QMOL (como marca simples) e TASSEL (como estudo de seleção genômica ampla-GWAS). A partir disso foram selecionados marcadores SNPs para serem utilizadas como possíveis marcadores para o óleo. Dentre estes marcadores para óleo, foram selecionados e analisados aqueles polimórficos entre SUPREMA e BARC-08 (Tabela 2).

2.2.2 Desenvolvimento da população

Foi realizado o cruzamento dos genitores SUPREMA (36,32% de proteína e 24,54% de óleo) e BARC-08 (45,08% de proteína e 21,21% de óleo). As plantas F1 foram autofecundadas gerando 476 sementes F2, que foram conduzidas pelo método SSD (Single Seed Descendent) até a F4 em casa de vegetação. Os indivíduos F4 tiveram irrigação, fertilização e iluminação de forma com que cada planta proporcionasse quantidade elevada de sementes para serem posteriormente submetidas a ensaios de campo em anos consecutivos.

2.2.3 Ensaios: locais, épocas e tratos culturais

Os ensaios foram conduzidos em duas localidades, Viçosa-MG (20° 45' 14" S 42° 52' 55" O) e Capinópolis-MG (18° 40' 55" S 49° 34' 12" O). Em Viçosa, o plantio foi realizado na Unidade de Ensino, Pesquisa e Extensão (UEPE) do Departamento de Fitotecnia, em duas diferentes épocas de semeadura, sendo elas novembro/2017 (geração F4) e abril/2018 (geração F4:5), correspondentes às safras de verão e inverno, respectivamente. Em janeiro/2019 foi conduzido em casa de vegetação (geração F4:6). Em Capinópolis, o plantio foi realizado na Central de Experimentação, Pesquisa e Extensão do Triângulo Mineiro (CEPET), em novembro/2018 (geração F4:6). Dessa forma, os genótipos foram avaliados em quatro ambientes distintos.

Os tratos culturais seguiram as recomendações para a cultura da soja, incluindo fertilização do solo no momento do plantio, controle químico de plantas daninhas, insetos-praga e fitopatógenos. Irrigação via aspersão foi administrada de acordo com a demanda até o momento em que os grãos atingiram o ponto de maturação fisiológica. A colheita das plantas foi realizada manualmente e o material armazenado em casa de vegetação por uma semana. Posteriormente, foi devidamente trilhado e organizado para análises de fenotipagem.

2.2.4 Delineamento experimental

O delineamento adotado foi o de blocos casualizados a exemplo de Zhang et al. (2015). A dimensão das parcelas foi de 1 m de comprimento, 0,5 m de largura entre parcelas e 1 m de intervalo entre blocos subsequentes. Foram semeadas 20 sementes por metro, com posterior desbaste pós-emergência para uma população de 12 plantas por metro. Após a maturidade fisiológica (R8), cada parcela foi colhida separadamente em bulk de sementes e amostradas para fenotipagem.

2.2.5 Genotipagem da população

Folhas jovens e sadias foram coletadas em estadio V4, em janeiro de 2018 das plantas F4. A extração de DNA foi feita de acordo com o protocolo de Dellaporta et al. (1983). Cada amostra foi quantificada por meio do espectrofotômetro NanoDrop (NanoDrop Technologies, Wilmington, DE, EUA). A verificação da qualidade do DNA foi realizada através de eletroforese em gel de agarose 0,8%. Posteriormente foi realizada diluição para a concentração de trabalho de 15 ng.µL⁻¹ para a genotipagem.

Para a validação dos marcadores, foi utilizada a população do cruzamento entre os dois genitores selecionados (SUPREMA e BARC-08). A genotipagem dos marcadores SNP selecionados seguiu a metodologia de genotipagem KASP (LGC Genomic, 2003), executada no equipamento Real Time PCR 7500 Systems (Applied Biosciences). A discriminação alélica foi executada no programa computacional AB 7500 Software v.2.3A.

2.2.6 Fenotipagem da população

Após colheita manual das plantas nos ensaios de campo, uma quantidade mínima de 30 grãos por parcela foi moída em um moinho industrial modelo MA020 (Marconi Equipamentos para Laboratório, Piracicaba, SP, Brasil), de forma a gerar granulometria adequada. O farelo obtido foi analisado para o conteúdo de proteína e óleo por espectrometria de infravermelho próximo (NIR), utilizando-se o equipamento Antaris II FT-NIR analyzer (Thermo Fisher Scientific Brasil Instrumentos de Processo Ltda., São Paulo, SP, Brasil).

2.2.7 Análise de dados

As análises estatísticas foram realizadas com auxílio dos aplicativos computacionais GENES (Cruz, 2013), GQMOL (Cruz e Schuster, 2004) e TASSEL (Bradbury et al. 2007).

Para os estudos de associação genômica ampla utilizando o programa Tassel, foram utilizados os modelos mistos (MLM) e parâmetros “default”.

2.2.8 Mapeamento dos reads e Análise de expressão diferencial

Trabalhos anteriores desenvolvidos pelo grupo de pesquisa PMQS/UFV realizou uma comparação do perfil transcricional de três variedades contendo alta proteína e baixo óleo com três variedades contendo baixa proteína e baixo óleo no estágio de desenvolvimento R5 (Silva, 2018). Com o intuito de identificar genes diferencialmente expressos dentro da possível região ancorando um importante QTL no Gm15 para o aumento do conteúdo de óleo em sementes de soja, foram analisados os dados de RNASeq para as duas variedades utilizadas neste estudo (Sequence Read Archive (SRA) número SRP155599).

Os reads dos genitores SUPREMA e BARC-08 foram mapeados com base no genoma de referências da soja Williams 82 na região do cromossomo 15 utilizando o software STAR (Dobin et al. 2013). A contagem dos reads mapeados e análise de expressão diferencial foi realizada com o software R a partir das contagens de alinhamento com o genoma de referência. Foi utilizada uma comparação contrastando a linhagens de alto conteúdo de óleo (SUPREMA) e baixo conteúdo de óleo (BARC-08). Os genes diferencialmente expressos (DEGs) encontrados na região selecionada foram anotados e investigados nas plataformas Soybase, Phytozoma e PUBMED para a identificação dos prováveis candidatos.

2.2.9 Identificação de genes candidatos

As reads de SUPREMA e BARC-08 foram baixadas do Sequence Read Archive (SRA) número SRP155599, foram convertidas em arquivos BAM e alinhadas pelo pacote de softwares BWA (Li e Durbin, 2009). Para remover duplicatas em cada arquivo com extensão .bam usamos o MarkDuplicate do Picard. Para isso, foram utilizados arquivo de alinhamento

(formato BAM) e o genoma de referência (no formato FASTA). Foram selecionados apenas os loci sequenciados com cobertura acima de 20 ($DP > 20$). Após a execução em todos os dados de alinhamento foi gerado um VCF. Foram recuperadas as informações detalhadas dos loci e eliminados os genótipos indesejados. Os resultados foram integrados em um único conjunto de chamadas e o efeito dos *SNPs* foi determinado com a ferramenta Ensembl Variant Effect Predictor (VEP) (Ensembl.org). (McLaren et al. 2016). As mutações conhecidas forneceram um nível de confiabilidade em relação às variantes.

2.3 RESULTADOS E DISCUSSÃO

2.3.1 Associação de marcadores moleculares às características avaliadas

Entre os objetivos deste estudo está a identificação de novos marcadores associados com o conteúdo de óleo em soja. Com o intuito de atender a este objetivo, 34 acessos previamente genotipados foram fenotipados para o conteúdo de óleo (Tabela 1) e esses dados utilizados para associação. Foram identificados um total de nove marcadores SNPs associados por marca simples com um p-valor menor do que 5% (134, 129, 153, 73, 115, 61, 164, 185 e 190) e três SNPs identificados por estudo de associação genômica ampla com um p-valor menor do que 5% (134, 115, 190).

Dentre os marcadores SNPs associados para conteúdo de óleo, selecionamos para estudo apenas sete que se apresentaram polimórficos para a população desenvolvida (Tabela 2). Devido a aproximação dos marcadores ss61 e ss164, ambos localizados no cromossomo 20 da soja, o ss61 foi escolhido por ter melhores dados de associação. Sendo assim, seis marcadores (ss61, ss73, ss115, ss129, ss134 e ss185) foram utilizados para genotipar a população F4 (Viçosa/2017) e esses dados utilizados com os dados de fenotipagem dos ambientes para a realização das análises de associação.

Apenas um marcador apresentou associação significativa para o conteúdo de óleo nos quatro ambientes avaliados, sendo este o ss115, localizado na posição 3,828,587 pb do cromossomo 15 (Gm15). O efeito aditivo desse QTL para a característica de óleo variou de 0,4081 a 1,285 nos quatro ambientes. Para os ambientes Viçosa/2017, Capinópolis/2018 e Viçosa/2019 a variação foi de 0,9958 a 1,285. O alelo favorável do SNP em questão foi doado pelo genitor SUPREMA. A análise de dados para o conteúdo de proteína, também revela associação do marcador ss115 para os quatro ambientes avaliados. No entanto, o alelo favorável é doado pelo genitor BARC-08 e tem efeito aditivo variando de 1,0761 a 1,5276. Isso evidencia a correlação genética negativa entre a herança de proteína e óleo conforme já relatado na literatura (Bandillo et al. 2015; Lee et al. 2019).

Em nosso trabalho ressaltamos a importância do SNP ss115, localizado no cromossomo 15 (Gm15). O marcador em questão se mostrou comum tanto para conteúdo de óleo quanto para conteúdo de proteína na semente e apresenta valores significativos. Essa forte associação indica um possível efeito pleiotrópico desta marca ou QTLs fortemente ligados. O efeito pleiotrópico é mais provável, pois essas características quase sempre co-segregam e a tentativa de separá-las ainda não obteve êxito (Wang et al. 2014). Outro ponto relevante é a forte estabilidade encontrada ao longo dos quatro ambientes testados, apresentando um valor menor de associação no ambiente de Viçosa/2018, que revelou menor efeito aditivo e contribuição genética (R^2) para conteúdo de óleo, por isso a interação tende a ser pequena e consequentemente, a associação também. Além disso, estudos anteriores mostram que o efeito dos loci no cromossomo 15 é ambientalmente dependente (Pathan et al. 2013; 21 Tajuddin et al. 2003), o que evidencia que apesar da sua estabilidade, o efeito ambiental determinará o quão influente este marcador será.

Kim et al. (2016) avaliou uma população de linhagens puras para proteína e óleo e através de mapeamento fino do cromossomo 15 encontrou uma região de 535kb no intervalo

entre BARCSOYSSR_15_0161 e BARCSOYSSR_15_0194, que se sobrepõe parcialmente à faixa determinada por ss715622170 e ss115. Zhang et al. (2018) encontrou um QTL de grande efeito identificado em 3,8-4,8 Mb no cromossomo 15, sendo liderado por ss715622170 e ss115. Esse locus foi associado simultaneamente com o conteúdo de proteína, óleo, ácido linolênico e diversos aminoácidos, explicando 22% e 14% da variação genética para teor de proteína e óleo, respectivamente. Lee et al. (2019) aponta para quatro marcadores no cromossomo 15, dentre eles ss115, posicionados dentro de um intervalo de 139 kb. Essas informações revelam o efeito e estabilidade do loco estudado e são extremamente úteis na seleção assistida por marcadores. Todos esses dados juntamente com os obtidos neste projeto corroboram para a aplicabilidade do marcador ss115 para programas de seleção assistida.

2.3.2 Genes diferencialmente expressos no cromossomo 15

O marcador SNP ss115 (Gm15) se mostrou significativo e estável no aumento do conteúdo de óleo nos grãos de soja. Com o objetivo de investigar os genes expressos próximos a essa região de interesse, foi conduzido análise de RNASeq para os genitores utilizados neste estudo e foram identificados 78 genes diferencialmente expressos (DEGs-Differential Expressed Genes) no cromossomo 15 ao utilizar um p-valor ajustado (padj) menor do que 0,05. Filtramos estes genes para selecionar apenas aqueles localizados em um intervalo de mais ou menos 2Mb do SNP ss115 (3,828,587: intervalo de 1,828,587 a 5,828,587) e foi gerado uma lista de 15 DEGs no cromossomo 15 da soja (Tabela 4). A partir de uma segunda análise mais rigorosa (FDR $q < 0,05$ e $FC > 2$) apenas quatro genes passariam a ser considerados (Glyma.15g034400, Glyma.15g050900, Glyma.15g063000 e Glyma.15g070700) (Tabela 4). Genes com maior variação de expressão gênica poderiam estar associados a variação das características e por isto utilizamos a filtragem mais rigorosa para identificar tais genes.

O gene Glyma.15g034400 é uma proteína da família das aldeído desidrogenases que foi indicada como possivelmente importante para o acúmulo de proteína e óleo nos estudos de Zhang et al. (2019). Neste trabalho, foi mapeado a região alvo do QTL *qPro15-1/qOil15-1* no cromossomo 15 no intervalo de 2,60 até 3,50 Mb e selecionadas proteínas que poderiam estar envolvidas no metabolismo de óleo e proteínas. Apesar de mencionar o gene Glyma.15g034400, ele não foi diferencialmente expresso em seus trabalhos, diferente do que nós observamos em nosso estudo. A diferença pode ter sido devido a diferença de parentais utilizados, mas novos estudos precisam ser conduzidos para provar a importância desse gene para a característica.

O gene Glyma.15g050900 corresponde à família de proteínas de montagem de clatrina. Chu et al. (2020) identifica esse gene dentre os possivelmente associados a regiões diferencialmente metiladas (DMR) em resposta ao estresse de baixo fósforo. Não temos como fazer uma associação direta desse gene com o conteúdo de proteínas.

Glyma.15g063000 corresponde a proteínas basic/helix-loop-helix (bHLH), uma superfamília de fatores de transcrição. Wang et al. (2017) identifica esse gene pelo seu papel potencial na resposta ao estresse hídrico. Foi identificado também no estudo de Atencio et al. (2021) como um homólogo de AtILR3 (bHLH105), que interage com AtPYE e AtBTS para regular a expressão de genes envolvidos na aquisição e homeostase do ferro. Yao et al. (2020) identifica esse gene em uma lista de possíveis genes envolvidos no metabolismo lipídico. Além disso, é encontrado nos estudos de Turquetti-Moraes et al. (2022) na região de varredura seletiva de genes candidatos envolvidos na biossíntese de óleo em soja. Uma melhor caracterização do envolvimento deste gene nas heranças de proteína e óleo deve ser posteriormente investigada, mas nos chamou a atenção o fato da grande diferença de expressão desse gene em relação aos parentais SUPREMA e BARC-08 e o fato de que o

efeito de um fator de transcrição poderia facilmente explicar o efeito pleiotrópico observado para a característica. Portanto, este é um gene que seria interessante de ser investigado.

O gene Glyma.15g070700 contém pouquíssima informação na literatura, sendo anotada no Soybase como possivelmente uma Disease Resistance Protein (classe TIR). Entretanto, análises de sequência de cultivares utilizados na região de Huang-Huai-Hai (Liu et al. 2022) foi identificado que o gene possivelmente sofreu variações acarretando o aparecimento de InDels do tipo “conservative inframe insertion” que podem gerar a adição de no mínimo 1 códon na sequência da proteína. Apesar deste Indel poder indicar uma explicação para a diferença de fenotipagem, primeiramente seria importante identificar se ela existe nos parentais, assim como posteriormente validar o efeito no conteúdo de óleo e proteína.

Além das informações descritas anteriormente, é importante mencionar que entre os genes não selecionados, mas com $FC > 1,5$, encontramos 2 genes interessantes de ter a sua função investigada se são os genes causais associados com o aumento de óleo do doador SUPREMA: Glyma.15g068700 e o Glyma15g049200.

O gene Glyma.15g068700 é responsável pelo transporte de aminoácidos e pertence à família Lysine histidine transporter-like 8. Descrito como proteínas diferencialmente fosforiladas (DPPs) envolvidas com transporte no estudo de Han et al. (2020). Zhang et al. (2019) aponta esse gene como genes alvo regulados pelo alelo Dt2 e associados à manutenção do meristema, tempo de floração, densidade estomática e respostas ao estresse.

O segundo gene é o Glyma15g049200, um transportador de açúcar específico da capa da semente com uma correlação positiva com o conteúdo de óleo em sementes e pertencente a família de transportadores SWEET (Sugars Will Eventually be Exported Transporters). Ele é encontrado na literatura com os nomes SWEET10a ou SWEET39, sendo que diferentes trabalhos apontam este gene como a causa do efeito nesta região do cromossomo 15 e

associados com o conteúdo de óleo, conteúdo de proteínas, peso de 100 sementes e tamanho da semente (Jin et al. 2023; Kumar et al. 2022; Miao et al. 2019; Yang et al. 2029; Wang et al. 2020; Zhang et al. 2020; Zhang et al. 2021). É importante mencionar que é necessário validar a diferença entre os acessos SUPREMA e BARC-08 antes de concluir que o efeito associado a esta região alvo está diretamente relacionada com este gene. Entretanto, existe a possibilidade de o gene alvo ser ou o mesmo ou uma variante de alelos descritos nos trabalhos da literatura.

Foi observada a presença de SNPs causais no cromossomo 15 nos 15 genes em destaque (Tabela 2). As mutações podem ser classificadas quanto ao impacto em quatro níveis: alto, moderado, baixo e modificador. A figura 1 mostra a distribuição do impacto em cada um dos genes. Embora tenha sido encontrado um número elevado de mutações, avaliamos o impacto de cada uma em cada gene e selecionamos as mais relevantes para nosso estudo, ou seja, alto e moderado impacto.

Quanto a mutações de impacto moderado, os genes que possuem maior quantidade de mutações são: 70700 (109), 69500 (96), 55700 (31) e 53700 (26). Todas com uma alteração na posição de aminoácido e que afete a função da proteína (mutação missense).

Quanto a mutações de alto impacto, representam 36 no total, porém a grande maioria está em região de íntron. Dois genes se mostram interessantes na região codificadora. O gene 53700 possui 3 mutações de alto impacto, no entanto, duas representam uma variação de sequência em que pelo menos uma base de um códon é alterada, resultando em um códon de parada prematuro, podendo levar a um polipeptídeo encurtado. O gene 70700 também possui 3 mutações de alto impacto e uma delas causa uma variação da sequência de codificação e um possível erro no processo de splicing. Essas mudanças podem afetar a estrutura final da proteína e, consequentemente, sua função

Finalizando, o presente trabalho validou o efeito do marcador SNP ss115 sobre o aumento do conteúdo de óleo e proteína em grãos de soja, apresentando uma alternativa viável para o uso deste marcador em programas de seleção assistida no Brasil e utilizando a cultivar SUPREMA como doador. As informações com base em quatro ambientes distintos revelam o efeito e a estabilidade do marcador, muito útil na seleção assistida por marcadores. Genes diferencialmente expressos nesta região foram apresentados para maior compreensão dos efeitos observados e revelando alguns possíveis candidatos. Os genes precisam de validação funcional adicional, entretanto o conhecimento sobre esses materiais e genes será útil para futuras pesquisas no programa de melhoramento da soja.

AGRADECIMENTOS

O trabalho foi financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico [312641/2021-4 para MD), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (APQ-01416-16, PPM-640-18 e RED-00124- 22 a MD) e Caramuru Alimentos (Outorga a MD).

REFERÊNCIAS

- ABIOVE – ASSOCIAÇÃO BRASILEIRA DAS INDÚSTRIAS DE ÓLEOS VEGETAIS. Estatística Mensal do Complexo Soja (2023) Disponível em: <https://abiove.org.br/estatisticas/>. Acessado em 23/03/2023.
- Akond, A. M., Ragin, B., Bazzelle, R., Kantartzi, S. K., Meksem, K. and Kassem, M. A. (2012). Quantitative trait loci associated with moisture, protein, and oil content in soybean [*Glycine max* (L.) Merr.]. *Journal of Agricultural Science*, 4, 16-25. <http://dx.doi.org/doi:10.5539/jas.v4n11p16>
- Atencio, L., Salazar, J., Lauter, A. N. M., Gonzales, M. D., O'Rourke, J. A. and Graham, M. A. (2021). Characterizing short and long term iron stress responses in iron deficiency tolerant and susceptible soybean (*Glycine max* L. Merr.). *Plant Stress*, 2, 100012. <https://doi.org/10.1016/j.stress.2021.100012>
- Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J. and Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*, 8, 1-13. <https://doi.org/10.3835/plantgenome2015.04.0024>
- Bueno, R. D., Lorenzoni, R. M., Bernardeli, A., Silva, L. D. C., Silva, L. C. C., Silva, J. N. B., Vidigal, P. M. P., Piovesan, N. D., Ribeiro, C and Dal-Bianco, M. (2022). Fine-mapping of a major QTL for protein content in BR8014887 soybean line. <https://doi.org/10.21203/rs.3.rs-1598069/v1>
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatic*, 23, 2633-2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Chu, S., Zhang, X., Yu, K., Lv, L., Sun, C., Liu, X., Zhang, J., Jiao, Y. and Zhang, D. (2020). Genome-wide analysis reveals dynamic epigenomic differences in soybean response to low-phosphorus stress. *International Journal of Molecular Sciences*, 21, 6817. <https://doi.org/10.3390/ijms21186817>
- Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B. and Specht, J. E. (2003). The seed protein, oil, and yield QTL on soybean linkage group I. *Crop science*, 43, 1053-1067. <https://doi.org/10.2135/cropsci2003.1053>
- Cruz, C. D. and Schuster, I. (2004). GQMOL: aplicativo computacional para análise de dados moleculares e de suas associações com caracteres quantitativos. Versão 2.1. Viçosa: UFV. 2004.

Cruz, C. D. (2013) GENES – a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum*, p.271-276.

Dellaporta, S. L., Wood, J., Hicks, J. B. (1983). A plant DNA miniprep: Version II. *Plant Molecular Biology Reporter*, 1, 19-21.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21. <https://doi.org/10.1093/bioinformatics/bts635>

Eskandari, M., Cober, E. R. and Rajcan, I. (2013). Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. *Theoretical and Applied Genetics*, 126, 1677-1687. <https://doi.org/10.1007/s00122-013-2083-z>

Genomics LGC. (2013). KASP genotyping chemistry User guide and manual. LGC Limited, Teddington, UK.

Han, R., Wei, Y., Xie, Y., Liu, L., Jiang, C. and Yu, Y. (2020). Quantitative phosphoproteomic analysis provides insights into the aluminum-responsiveness of Tamba black soybean. *Plos one*, 15, e0237845. <https://doi.org/10.1371/journal.pone.0237845>

Haun, W., Coffman, A., Clasen, B. M., Demorest, Z. L., Lowy, A., Ray, E., Retterath, A., Stoddard, T., Juillerat, A., Cedrone, F., Mathis, L., Voytas, D. F. and Zhang, F. (2014) Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant biotechnology journal*, 12, 934-940. <https://doi.org/10.1111/pbi.12201>

Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J. and Cregan, P. B. (2014) A genome-wide association study of seed protein and oil content in soybean. *BMC Genom*, 15, 1. <https://doi.org/10.1186/1471-2164-15-1>

Hyten, D. L., Pantalone, V. R., Sams, C. E., Saxton, A. M., Landau-Ellis, D., Stefaniak, T. R. and Schmidt, M. E. (2004) Seed quality QTL in a prominent soybean population. *Theoretical and Applied Genetics*, 109, 552-561. <https://doi.org/10.1007/s00122-004-1661-5>

Jin, H., Yang, X., Zhao, H., Song, X., Tsvetkov, Y. D., Wu, Y., Gao, Q., Zhang, R. and Zhang, J. (2023). Genetic analysis of protein content and oil content in soybean by genome-wide association study. *Frontiers in Plant Science*, 14, 1182771. <https://doi.org/10.3389/fpls.2023.1182771>

Kim, M., Schultz, S., Nelson, R. L. and Diers, B. W. (2016) Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. *Crop Science*, 56, 219-225. <https://doi.org/10.2135/cropsci2015.06.0340>

Kim, W. J., Kang, B. H., Moon, C. Y., Kang, S., Shin, S., Chowdhury, S., Choi, M. S., Park, S. W., Moon, J. K. and Ha, B. K. (2023) Quantitative Trait Loci (QTL) Analysis of Seed Protein and Oil Content in Wild Soybean (*Glycine soja*). International Journal of Molecular Sciences, 24, 4077. <https://doi.org/10.3390/ijms24044077>

Kumar, V., Goyal, V., Mandlik, R., Kumawat, S., Sudhakaran, S., Padalkar, G., Rana, N., Deshmukh, R., Roy, J., Sharma, T. R. and Sonah, H. (2022). Pinpointing genomic regions and candidate genes associated with seed oil and protein content in soybean through an integrative transcriptomic and QTL meta-analysis. Cells, 12, 970. <https://doi.org/10.3390/cells12010097>

Kumar, V., Vats, S., Kumawat, S., Bisht, A., Bhatt, V., Shivaraj, S. M., Padalkar, G., Goyal, V., Zargar, S., Gupta, S., Kumawat, G., Chandra, G., Chalam, V.C., Ratnaparkhe, M.B., Gill, B.S., Jean, M., Patil, G. B., Vuong, T., Rajcan, I., Deshmukh, R., Belzile, F., Sharma, T.R., Nguyen, H.T. and Sonah, H. (2021). Omics advances and integrative approaches for the simultaneous improvement of seed oil and protein content in soybean (*Glycine max* L.). Crit. Rev. Plant Sci, 40, 398–421. <https://doi.org/10.1080/07352689.2021.1954778>

Lee, S., Van, K., Sung, M., Nelson, R., LaMantia, J., McHale, L. K. and Mian, M. R. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. Theoretical and Applied Genetics, 132, 1639-1659. <https://doi.org/10.1007/s00122-019-03304-5>

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>

Li, Q. T., Lu, X., Song, Q. X., Chen, H. W., Wei, W., Tao, J. J., Bian, X. H., Shen, M., Ma, B., Zhang, W. K., Bi, Y.D., Li, W., Lai, Y. C., Lam, S. M., Shui, G. H., Chen, S. Y. and Zhang, J. S. (2017). Selection for a zinc-finger protein contributes to seed oil increase during soybean domestication. Plant Physiology, 173, 2208-2224. <https://doi.org/10.1104/pp.16.01610>

Liu, J. Y., Zhang, Y. W., Han, X., Zuo, J. F., Zhang, Z., Shang, H., Song, Q. and Zhang, Y. M. (2020). An evolutionary population structure model reveals pleiotropic effects of GmPDAT for traits related to seed size and oil content in soybean. Journal of experimental botany, 71, 6988-7002. <https://doi.org/10.1093/jxb/eraa426>

Liu, J., Xie, H., Lin, T. (2022). Putative variants, genetic diversity and population structure among Soybean cultivars bred at different ages in Huang-Huai-Hai region. Sci Rep, 12, 2372. <https://doi.org/10.1038/s41598-022-06447-6>

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. and Cunningham, F. (2016). The ensembl variant effect predictor. Genome biology, 17, 1-14. <https://doi.org/10.1186/s13059-016-0974-4>

- Miao, L., Yang, S., Zhang, K., He, J., Wu, C., Ren, Y., Gai, J. and Li, Y. (2020). Natural variation and selection in GmSWEET39 affect soybean seed oil content. *New Phytologist*, 225, 1651-1666. <https://doi.org/10.1111/nph.16250>
- Pathan, S. M., Vuong, T., Clark, K., Lee, J. D., Shannon, J. G., Roberts, C. A., Ellersieck, M. R., Burton, J. W., Cregan, P. B., Hyten, D. L., Nguyen, H. T. and Sleper, D. A. (2013). Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. *Crop Science*, 53, 765-774. <https://doi.org/10.2135/cropsci2012.03.0153>
- Phansak, P., Soonsuwon, W., Hyten, D. L., Song, Q., Cregan, P. B., Graef, G. L. and Specht, J. E. (2016). Multi-population selective genotyping to identify soybean [*Glycine max* (L.) Merr.] seed protein and oil QTLs. *G3: Genes, Genomes, Genetics*, 6, 1635-1648. <https://doi.org/10.1534/g3.116.027656>
- Priolli, R. H. G., Carvalho, C. R. L., Bajay, M. M., Pinheiro, J. B. and Vello, N. A. (2019). Genome analysis to identify SNPs associated with oil content and fatty acid components in soybean. *Euphytica*, 215, 1-14. <https://doi.org/10.1007/s10681-019-2378-5>
- Silva, L. A. C. (2018). Caracterização do perfil de expressão em variedades de soja contrastantes para o teor de proteína. Tese de Doutorado. Universidade Federal de Viçosa.
- Tajuddin, T., Watanabe, S., Yamanaka, N. and Harada, K. (2003) Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. *Breeding Science*, 53, 133-140.
- Turquetti-Moraes, D. K., Moharana, K. C., Almeida-Silva, F., Pedrosa-Silva, F. and Venancio, T. M. (2022). Integrating omics approaches to discover and prioritize candidate genes involved in oil biosynthesis in soybean. *Gene*, 808, 145976. <https://doi.org/10.1016/j.gene.2021.145976>
- USDA - United States Department of Agriculture (2023) Oilseeds: world markets and trade. Disponível em: <https://www.fas.usda.gov/data/oilseeds-world-markets-and-trade>. Acessado em: 25/03/2023.
- Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y. C. Liu, Z., Frommer, W. B., Ma, J. F., Chen, L. Q., Guan, Y., Shou, H. and Tian, Z (2020) Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. *National Science Review* 7: 1776-1786. <https://doi.org/10.1093/nsr/nwaa110>
- Wang, X., Jiang, G. L., Green, M., Scott, R. A., Song, Q., Hyten, D. L. and Cregan, P. B. (2014). Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. *Mol Genet Genomics*, 289, 935-949, 2014. <https://doi.org/10.1007/s00438-014-0865-x>

Wang, W., Jiang, W., Liu, J., Li, Y., Gai, J. and Li, Y. (2017). Genome-wide characterization of the aldehyde dehydrogenase gene superfamily in soybean and its potential role in drought stress response. *BMC genomics*, 18, 1-17. <https://doi.org/10.1186/s12864-017-3908-y>

Yang, H., Wang, W., He, Q., Xiang, S., Tian, D., Zhao, T. and Gai, J. (2019). Identifying a wild allele conferring small seed size, high protein content and low oil content using chromosome segment substitution lines in soybean. *Theor Appl Genet*, 132, 2793–2807. <https://doi.org/10.1007/s00122-019-03388-z>

Yao, Y., You, Q., Duan, G., Ren, J., Chu, S., Zhao, J., Li, X., Zhou, X. and Jiao, Y. (2020) Quantitative trait loci analysis of seed oil content and composition of wild and cultivated soybean. *BMC plant biology*, 20, 1-13. <https://doi.org/10.1186/s12870-019-2199-7>

Zhang, D., Sun, L., Li, S., Wang, W., Ding, Y., Swarm, S. A., Li, L., Wang, X., Tang, X., Zhang, Z., Tian, Z., Brown, P. J., Cai, C., Nelson, R. L. and Ma, J. (2018). Elevation of soybean seed oil content through selection for seed coat shininess. *Nature plants*, 4, 30–35. <https://doi.org/10.1038/s41477-017-0084-7>

Zhang, D., Wang, X., Li, S., Wang, C., Gosney, M. J., Mickelbart, M. V., & Ma, J. (2019) A post-domestication mutation, Dt2, triggers systemic modification of divergent and convergent pathways modulating multiple agronomic traits in soybean. *Molecular plant*, 12, 1366-1382. <https://doi.org/10.1016/j.molp.2019.05.010>

Zhang, D., Zhang, H., Hu, Z., Chu, S., Yu, K., Lv, L., Zhang, D., Zhang, H., Hu, Z., Chu, S., Yu, K., Lv, L., Yang, Y., Zhang, X., Chen, X., Kan, G., Tang, Y., An, Y. Q.C. and Yu, D. (2019). Artificial selection on GmOLEO1 contributes to the increase in seed oil during soybean domestication. *PLoS Genetics*, 15, e1008267. <https://doi.org/10.1371/journal.pgen.1008267>

Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M. L. and An, Y. Q. C. (2020) Selection of *GmSWEET39* for oil and protein improvement in soybean. *PLoS Genet*, 16, e1009114. <https://doi.org/10.1371/journal.pgen.1009114>

Zhang, S., Hao, D., Zhang, S., Zhang, D., Wang, H., Du, H., Kan, G. and Yu, D. (2021). Genome-wide association mapping for protein, oil and water-soluble protein contents in soybean. *Molecular Genetics and Genomics*, 296, 91-102. <https://doi.org/10.1007/s00438-020-01704-7>

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., Hou, W., Wu, C., Han, T. and 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, 5915. <https://doi.org/10.3390/ijms20235915>

Zhang, Y. H., Liu, M. F., He, J. B., Wang, Y. F., Xing, G. N., Li, Y., Yang, S. P., Zhao, T. J. and Gai, J. Y. (2015). Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]. *Theor Appl Genet*, 128, 1061-1072. <https://doi.org/10.1007/s00122-015-2490-4>

Zhang, Y. Q., Lu, X., Zhao, F. Y., Li, Q. T., Niu, S. L., Wei, W., Zhang, W. K., Ma, B., Chen, S. Y. and Zhang, J. S. (2016) Soybean GmDREBL increases lipid content in seeds of transgenic Arabidopsis. *Scientific Reports*, 6, 34307. <https://doi.org/10.1038/srep34307>

Distribuição do impacto de mutação para cada gene

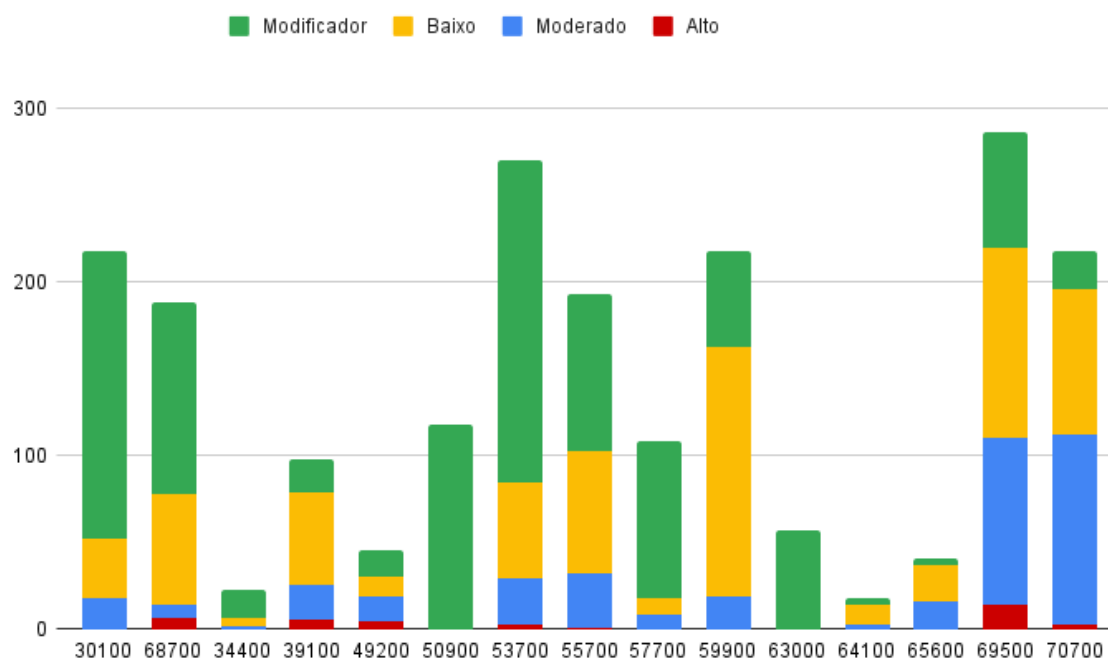


Figura 1. Distribuição do impacto da mutação para cada um dos genes estudados.

Genótipo	Plantio 1	Plantio 2	Plantio 3	Média
ANS 83022	22,41692901	24,046825	22,67251535	23,04542312
ANS 89109	20,71370301	20,82633301	21,65119718	21,0637444
B3 PTA 21619	19,71483962	17,36783955	18,39865119	18,49377679
BARC - 8	21,37659635	20,7537763	21,52642901	21,21893389
BR80 14887	18,41823968	18,4340884	17,17726329	18,00986379
BRS 284	21,95874315	19,88545295	23,67663143	21,84027584
BRS 7980	20,35418432	19,17548666	19,17859887	19,56942328
CD 201	20,80776198	18,62480381	20,08465293	19,8390729
CD 213 PTA	20,63793052	20,80548301	20,27139139	20,57160164
CD 237 RR	23,07566706	24,75143702	22,38735652	23,4048202
CD 2737 RR	22,81786585	21,86068149	21,61642134	22,0983229
CD TN 4	18,59559396	19,29265576	19,49513479	19,12779484
CONQUISTA	21,32977422	20,19128794	20,10824938	20,54310384
CS 303 PTA 137 - 4-10	18,33796363	17,82153697	16,12044838	17,42664966
CS 303 PTA 182	21,15007743	19,80438578	19,80566182	20,25337501
CS 303 PTA 276 1-2	19,64028537	21,1162003	21,36805851	20,70818139
CS 303 PTA 276 3-4	16,38113134	18,81381491	16,84969059	17,34821228
DESAFIO	21,22232101	21,63013491	21,54504953	21,46583515
M 7739 IPRO	22,8574044	21,69159618	21,23884265	21,92928108
M 8372 IPRO	20,6528805	20,39358033	19,80392548	20,2834621
N 7901	21,82191857	21,3487589	20,75965828	21,31011192
NS 7227	20,63018488	21,86345384	22,36911065	21,62091646
NT - 12	18,89076705	19,68923901	19,15466654	19,24489086
PINTADO	21,34834055	21,23307229	19,40203259	20,66114848
SUPREMA	24,46527024	24,52046084	24,642886	24,54287236
SYN 1060	18,92436694	17,3760805	17,50949826	17,93664856
TMG 1176	19,50380799	19,49734911	19,2798449	19,42700067
TMG 1180	21,03806048	20,18116085	20,47773387	20,56565173
TMG 1288	21,7106523	22,54909905	20,06340693	21,44105276
TMG 2181 IPRO	23,25264145	24,18020466	23,80503525	23,74596045
TMG 4182	20,92920074	23,25851816	22,79622174	22,32798021
TMG 4185	22,21275511	21,97393979	20,86866176	21,68511888
TMG 7262 RR	21,57249077	22,12986751	20,70468764	21,46901531
UFV TN 105 AP	19,4114134	19,37425896	18,41558456	19,06708564
VX - 04 5692 - 62	18,86481292	19,4447916	18,92849406	19,0793662

Tabela 1. Plantios e média dos dados fenotípicos dos 34 genótipos doadores do programa para conteúdo de e óleo (%) realizados por espectrometria de infravermelho próximo (NIR).

	Start	End	padj	Fold Change
Glyma.15g030100	2420032	2426219	0,02074471	1,758799725
Glyma.15g034400	2740960	2746344	0,036901101	0,460267492
Glyma.15g039100	3078440	3081975	0,028260063	0,704824414
Glyma.15g049200	3874867	3876757	0,002493267	0,690947937
Glyma.15g050900	4003079	4006306	0,012796265	3,212511238
Glyma.15g053700	4224743	4229800	0,012718635	1,816634883
Glyma.15g055700	4356583	4362253	0,011683281	1,755970215
Glyma.15g057700	4440424	4444439	0,007319481	1,846523237
Glyma.15g059900	4634570	4638101	0,045851666	1,288458677
Glyma.15g063000	4825805	4829074	0,009011917	83,31856888
Glyma.15g064100	4877866	4879175	0,009982097	1,724352783
Glyma.15g065600	4981427	4985949	0,028716586	1,849899504
Glyma.15g068700	5226328	5232403	0,0000842	1,633236863
Glyma.15g069500	5304688	5312952	0,041375697	1,256053141
Glyma.15g070700	5424597	5430524	0,0000317	2,308538009

Tabela 2. Genes diferencialmente expressos na região próxima ao marcador SNP ss115. Valores marcados em cinza representam FC>2.

3. CAPÍTULO 3

Mapeamento fino de um QTL na linha de soja BR8014887 para identificação de gene causal para o conteúdo de proteína no grão de soja

3.1 INTRODUÇÃO

O conteúdo de proteína corresponde a cerca de 40% do peso seco da semente de soja (Zhang et al. 2019) e é uma característica importante que determina a qualidade, aumenta a competitividade e valor da cultura (Kim et al. 2016; Wang et al. 2014).

Nos últimos anos, o interesse centrou-se na proteína da soja por várias razões. Primeiro, pela mudança nos hábitos alimentares visando qualidade nutricional e segundo, pelos produtos derivados dessa importante cultura. Nesse caso, o interesse do cliente está na proteína oferecida pelo grão ao incrementá-la na alimentação tanto humana, quanto animal. Assim, é esperado que, em algum momento, a soja seja valorizada por sua qualidade e poderá ser comercializada não apenas pelo peso total, mas por critérios que valorizem os grãos de maneira diferente, com base no conteúdo e qualidade de proteína (Updew et al. 1976).

O melhoramento da soja no Brasil tem sido um processo contínuo e dinâmico. O Brasil é o maior produtor e exportador mundial de soja. Contudo, nos últimos anos têm sido frequentes os questionamentos sobre a redução do conteúdo de proteína nas novas cultivares e a qualidade do grão. Isso é reflexo do passado, em que nossos ancestrais selecionaram fenótipos e indivíduos que cultivaram e começaram a controlar essa reprodução (Orel et al. 1981). Nessa seleção, concentraram esforços em alto conteúdo de óleo e produtividade, mas ao custo de diminuir o conteúdo de proteína (Fonseca et al. 2020; Karikari et al. 2019; Li et al. 2019).

Patil et al. (2017) aponta que o objetivo de desenvolver cultivares de soja mais produtivas também promoveu essa redução significativa, devido em grande parte a correlação negativa entre produtividade e conteúdo proteico. O foco no rendimento levou a uma soja contendo em média 20 a 25% menos proteína no Brasil (Cotegna, 2021). Embora unir todas essas características seja o objetivo, o melhoramento simultâneo na mesma cultivar é uma tarefa desafiadora, uma vez que o conteúdo de proteína está geralmente correlacionado negativamente com o conteúdo de óleo, rendimento e produtividade (Karikari et al. 2019; Li et al. 2019; Patil et al. 2017).

A partir de técnicas moleculares avançadas, muitos loci de características quantitativas (QTLs) de proteína de soja estão sendo estudados nos últimos anos. Até o momento foram identificadas 255 e 322 QTLs relacionados ao conteúdo de proteína e óleo (Kim et al. 2023). Poucos foram confirmados e efetivamente usados no melhoramento, portanto, é desejável identificar QTLs com efeitos consistentes (He et al. 2021; Phansak et al. 2016; Zhang et al. 2019).

A compreensão dos mecanismos genéticos e moleculares que controlam a variação no controle de proteínas de soja ainda não são totalmente esclarecidos. Por isso, a identificação de locus genéticos que governem essa característica facilitam estudos nesta área, assim como no desenvolvimento de linhagens (Chung et al. 2003; Eskandari et al. 2013; Phansak et al. 2016). O desenvolvimento de cultivares com maior conteúdo de proteína aumentaria ainda mais o valor econômico da cultura e ajudaria a enriquecer toda a cadeia de valor, desde os agricultores até os processadores e usuários finais (Patil et al. 2017). Estudos usando vários ciclos de seleção assistida em linha quase isogênica (NILs) apontam que isso pode ser uma realidade no futuro (Prenger et al. 2019).

No presente estudo foi realizado o desenvolvimento de uma população de linhagens quase isogênicas (Near Isogenic Lines - NILs) com o objetivo de analisar um loci de

características quantitativas (QTL) na linha de soja BR8014887 e posteriormente comprovar o possível gene causal.

3.2 MATERIAL E MÉTODOS

3.2.1 Material Vegetal

Para os experimentos deste estudo foi utilizada a população de RILs desenvolvida pelo cruzamento de acessos de alto conteúdo de proteína (BR8014887 e NT12). Estes acessos pertencem ao banco de germoplasma do Programa de Melhoramento da Qualidade da Soja (PMQS/BIOAGRO/DBB/UFV).

3.2.2 Genotipagem

Materiais foliares foram coletados da população RILs em ensaios desenvolvidos em casa de vegetação em dezembro/2021. Foi extraído o DNA de cada progênie de acordo com Dellaporta et al. (1983). Após a verificação da qualidade do DNA através de eletroforese em gel de agarose 0,8 %, cada amostra foi quantificada por meio do espectrofotômetro NanoDrop (NanoDrop Technologies, Wilmington, DE, EUA) e diluída para a concentração de trabalho 15 ng.µL⁻¹ para a genotipagem. A genotipagem dos marcadores SNPs seguiu a metodologia de genotipagem KASP (LGC Genomics, 2013), executada no equipamento Applied Biosciences 7500 (AB7500). A discriminação alélica foi executada no programa computacional AB 7500 Software v.2.3

3.2.3 Desenho de primers

Foi realizada uma busca na literatura e os seguintes primers foram selecionados: POWR1-PCR (Goettel et al. 2022), NILS-PCR (Goettel et al. 2022) e 20g.85100 (Fliege et al. 2022) (Tabela 1).

Código do Primer	Nome da sequência	Sequência 5' - 3'	Referência
POWR1-PCR	POWR1-PCR-Fw	TGGCACCAGAAATCTCACACT	Goettel et al. 2022
POWR1-PCR	Rv-POWR1-PCR	TCCTTGTTTATGGCTCTCTCCA	Goettel et al. 2022
NIL-PCR	NIL-PCR-Fw	CACTTCAAGGGTGGCAGTGTT	Goettel et al. 2022
NIL-PCR	Rv-NIL-PCR	CGGGATGGGAAAAGTGTCTTA	Goettel et al. 2022
20g.85100	Glyma.20G85100-Fw	ACTGCATCAACCAAGCCTTATGC	Fliege et al. 2022)
20g.85100	Rv-Glyma.20G85100	TGTACGTTTCTAACTCACTTAACCTATTGG	Fliege et al. 2022)

Tabela 1. Oligonucleotídeos iniciadores (primers) (R/F) utilizados neste estudo.

3.2.4 Desenvolvimento de população de linhagens quase isogênicas (Near Isogenic Lines - NILs)

No ano de 2014 foi executado um cruzamento simples envolvendo dois acessos de alto conteúdo de proteína do PMQS (BR8014887 e NT12). Os indivíduos foram avançados ao longo dos anos pelo método de descendente de semente única (Single Seed Descendent - SSD) até a geração F5, onde as sementes de cada planta foram agrupadas gerando as RILs. Estes indivíduos foram utilizados para os 4 experimentos de campo conduzidos pelo nosso grupo de pesquisa em Capinópolis e em Viçosa nos anos de 2017 e 2018.

Em 2021, selecionamos um indivíduo de cada RILs e individualizamos a população. Essa população foi avançando em casa de vegetação com a finalidade de obter uma população de HIFs (Heterogeneous Imbred Families), que é um tipo de NILs (Near-Isogenic Lines). Materiais foliares foram coletados, extração de DNA realizada e a população foi devidamente genotipada por marcadores microsatélites (SSR) e SNPs selecionados. Posteriormente,

indivíduos promissores e marcadores foram selecionados para investigar a região de interesse deste trabalho.

3.2.5 Análises Estatísticas

As análises estatísticas desenvolvidas neste trabalho foram realizadas com auxílio dos aplicativos computacionais GENES (Cruz, 2013), GQMOL (Cruz, C. D; Schuster, 2004) e TASSEL (Bradbury et al. 2007). Figuras e tabelas foram criadas no Microsoft Office Excel e Microsoft Office PowerPoint.

3.3 RESULTADOS E DISCUSSÃO

O Programa de Melhoramento da Qualidade da Soja (PMQS/BIOAGRO/UFV) iniciou um projeto de seleção assistida de marcadores SNP para obtenção de cultivares com alto conteúdo de proteína e óleo. Nessa linha de pesquisa, trabalhos anteriores (Bueno et al. 2022) utilizaram uma população de linhagens endogâmicas recombinantes (RILs) desenvolvida entre BR8014887 e NT12 e analisaram uma região com marcadores SNP (Single Nucleotide Polymorphism) e SSR (microssatélites). Um QTL no cromossomo 20 da soja foi mapeado e associado com o conteúdo de óleo e proteínas em uma região de 764.992 pb. Os dados de mapeamento permitiram estreitar a região alvo para 116.736 pb entre os marcadores Satt667 (31.627.414) e Satt670 (31.744.150), e identificando o gene Glyma.20g084500 como o possível gene causal da característica.

Entre os grupos que trabalham nessa linha de pesquisa, o cromossomo 20 é considerado um *hotspot* para conteúdo de proteína e encontrar genes associados a essa característica é um desafio. Essa região foi caracterizada por outros trabalhos (Bandillo et al. 2015; Fliege et al. 2022; Hwang et al. 2014; Marsh et al. 2022; Patil et al. 2018; Zhang et al.

2019). Apesar disso, apenas o Glyma.20g085100 foi caracterizado como um gene causal no trabalho de Fliege et al. (2022).

Trabalhos conduzidos por Marsh et al. (2022) apresentam uma região de 173Kb e essa região compreende o gene Glyma.20g084500, Glyma.20g085100 e Glyma.20G085250, mas também acreditam na relevância de Glyma.20G085100. Esses trabalhos não encontraram uma mutação causal na região codificadora de Glyma.20g084500, por isso esse gene foi eliminado das análises.

Diante do que está disponível na literatura até o momento não se pode confirmar ou descartar a influência de Glyma.20g084500 no conteúdo de proteína. Esse gene não foi caracterizado e praticamente não há informações sobre este locus. No entanto, nosso grupo de pesquisa possui dados que apontam para a importância desse gene e decidi investigar-lo um pouco mais para compreender a caracterização funcional. Nesse sentido, seguimos o desenvolvimento de uma população de NILs entre BR8014887 e NT12 para obter blocos de recombinações na região estudada.

3.3.1 Desenvolvimento de população de linhagens quase isogênicas (Near Isogenic Lines - NILs)

Uma análise mais detalhada da região alvo identificada precisava ser conduzida. Além disso, seria importante verificar se o alelo presente na linhagem BR8014887 contém similaridades com a descrição dos estudos de Goettel (2022) e Fliege (2022), e para isto, os primers NILS-PCR (Goettel et al. 2022) e Glyma 20g.85100 (Fliege et al. 2022) foram selecionados.

No estudo de Goettel et al. (2022) é demonstrado que existe uma inserção TE (elemento transponível, 321 pb) que afeta o domínio CCT e isso provoca aumento do conteúdo de óleo, peso e rendimento e diminuição do conteúdo de proteína. Esse gene é

chamado de POWR1 e provavelmente controla pleiotropicamente essas características. Eles ainda apontam que o alelo de inserção TE é fixado na soja cultivada por causa de seleção durante o processo de domesticação.

Fliege et al. (2022) identifica variantes e caracteriza o gene Glyma.20G85100 como causal nesse QTL. Em seu estudo, ele trabalha com o primer Glyma.85100. Utilizamos estes 3 primers e os avaliamos em BR8014887 e NT12.

Foram realizados testes para encontrar temperaturas de anelamento ideais para cada primer (Figura 1). A melhor temperatura testada foi a de 66°C, eliminando bandas inespecíficas. Posteriormente, foram testados na temperatura de 66°C para variedades utilizadas em nosso programa (ANSC83022 e UFVTN105AP) para comprovar sua eficiência (Figura 2).

Neste estudo, observamos que BR8014887 demonstra o padrão POWR1-TE apontado por Goettel et al. (2022), ou seja, sem inserção de elemento transponível, o que explica o aumento no conteúdo de proteínas. Enquanto NT12 demonstra o padrão heterozigoto para esse marcador (Figura 2). Para os primers NILS-PCR e Glyma.85100, BR8014887 apresenta padrão sem inserção (-TE) e NT12 com inserção (+TE). Por isso, para avaliar a população nos próximos passos, optamos por utilizar os primers NILS-PCR e Glyma.85100.

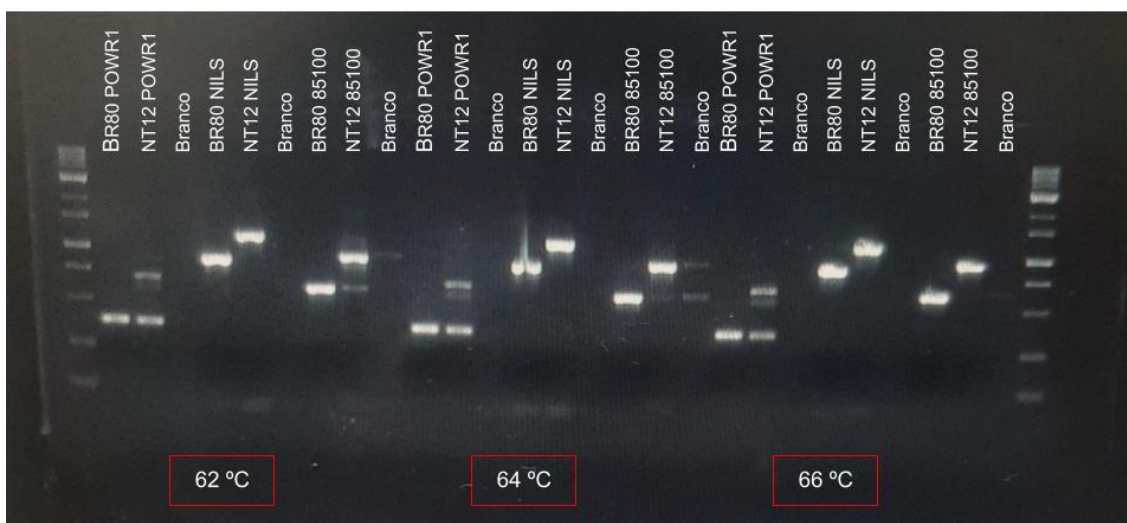


Figura 1. Gel de agarose mostrando testes com primers encontrados na literatura em três diferentes temperaturas de anelamento (62°C, 64°C e 66°C) para os genótipos de BR8014887 e NT12.

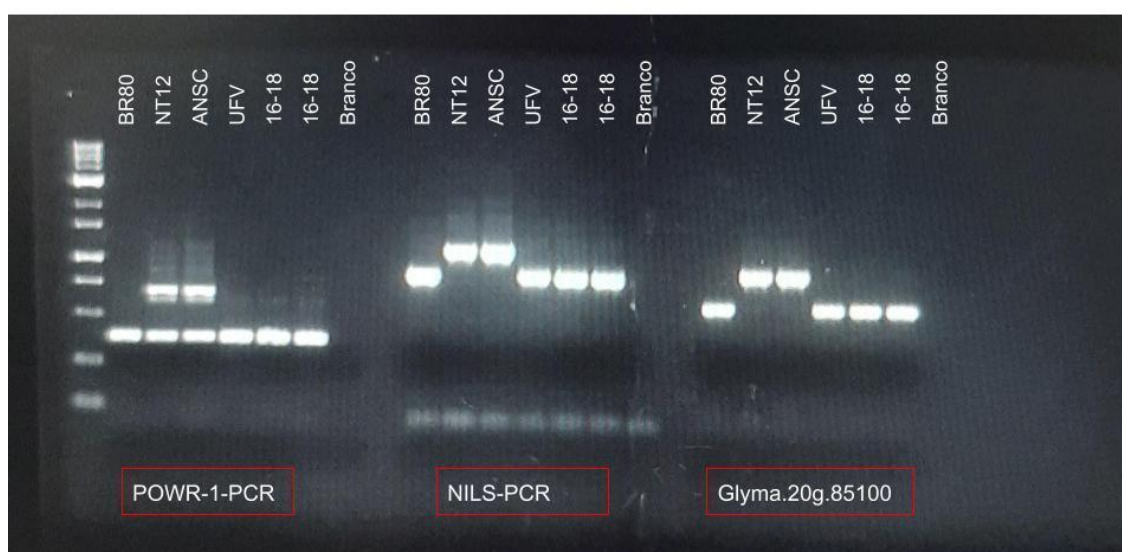


Figura 2. Gel de agarose mostrando primers encontrados na literatura e cultivares conhecidas pelo Programa PMQS (ANSC83022, UFVTN105AP). O indivíduo 16-18 representa um indivíduo controle entre o cruzamento de ANSC83022 e UFVTN105AP.

Em 2021, a população de RILs entre BR8014887 e NT12 possuía 549 genótipos. Cada genótipo foi individualizado por uma única semente e essas avançadas em casa de vegetação com o objetivo de ser individualizada para futuros estudos. Foram coletadas folhas jovens e realizada uma nova extração de DNA. Os marcadores SNP#115, SNP#190, Satt667, Satt668 e

Satt670 foram devidamente avaliados em toda a população. O objetivo era encontrar combinações entre marcadores SNP#115 (Chm15) e 190 (Chm20) e recombinações entre Satt667, Satt668 e Satt670 dentro da região de estudo. Não conseguimos as recombinações duplas esperadas no marcador Satt668 proximal ao gene Glyma.20g084500, por isso selecionamos 9 indivíduos que possuíam alguma combinação dentro da região alvo (Tabela 2).

A partir dos indivíduos dessa primeira seleção, avaliamos novamente os marcadores Satt667, Satt668 e Satt670 para confirmação de recombinação cromossômica. Além disso, analisamos os marcadores Satt663, Satt666, Satt667, Satt668, Satt670, Satt672 e Satt674 e SNP#115 e SNP#190. Uma nova seleção foi realizada e 4 novos indivíduos promissores foram selecionados (163, 251, 499 e 517) (Tabela 3). O indivíduo 163 foi avançado pela provável recombinação para SNP 115. O 251, 499 e 517 foi avançado pois era heterozigoto para SNP 190 e para SSR 667, 668 e 670. Ou seja, eram promissores para futuras recombinações, o que nos forneceria o bloco de recombinações dentro dessa região. Esses indivíduos foram novamente plantados em casa de vegetação, matéria foliar foi coletado para extração de DNA e uma análise de marcadores mais completa foi realizada utilizando marcadores SSR 663, 666, 667, 668, 670, 674, 678, SNP 115 e 190. Além disso, os primers encontrados na literatura (POWR-1-PCR, NILS-PCR e Glyma 20g.85100) foram analisados.

	SNP 115	663	666	667	668	670	672	674	SNP 190
BR80	0	0	0	2	2	0	0	2	0
NT12	2	2	2	0	0	2	2	0	2
57	0	0	0	2	2	0	0	2	0
163	1	0	0	2	0	2	2	0	2
211	2	0	0	2	2	0	0	2	0
251	2	2	1	1	1	1	2	1	1
292	0	0	0	2	2	0	0	2	0
408	0	0	0	2	2	0	0	2	0
483	0	0	0	2	2	0	0	2	0
499	2	2	1	1	1	1	2	1	1
417	2	2	1	1	1	1	2	1	1

Tabela 2. Marcadores utilizados no estudo para selecionar os nove indivíduos que possuíam combinações promissoras dentro da região alvo.

Marcador	SNP 115	663	666	667	668	670	672	674	SNP 190
BR80	0	0	0	2	2	0	0	2	0
NT12	2	2	2	0	0	2	2	0	2
163	1	0	0	2	0	2	2	0	2
251	2	2	1	1	1	1	2	1	1
499	2	2		1	1	1	2	1	1
417	2	2	1	1	1	1	2	1	1

Tabela 3. Marcadores SSR, SNP e da literatura utilizados neste estudo e indivíduos selecionados como promissores (163, 251, 499, 517).

Marcador	POWR1-PCR	NILS-PCR	20g.85100	SNP 190
BR80	0	0	0	0
NT12	2	2	2	2
163	2	2	2	2
251	1	1	1	1
499	1	1	1	1
417	1	1	1	1

Tabela 4. Primers disponíveis na literatura corroboram os resultados que obtivemos do SNP 190.

Observamos que para estudar a região do SNP 115 os indivíduos do genótipo 163 são muito interessantes, pois apresentam combinações em 115 e são fixados em 190 (NT-BR, BR-BR e Heterozigotos-BR. Além disso, encontramos as recombinações desejadas em 115 e 190: **NT-BR** (251, 499, 517) e **BR-NT** (163) (Tabela 5) e isto irá permitir ver o efeito deste gene na característica.

Quanto aos SSR 667, 668 e 670, apenas a recombinação **BR-NT-NT** foi encontrada fixada nos indivíduos do genótipo 163. Entretanto, os indivíduos 251, 499 e 517 possuem heterozigotos que podem ser novamente avançados e proporcionar as recombinações desejadas. Esse é o próximo passo realizado para desenvolvimento das NILs.

Concluindo, os dados apresentados neste estudo apresentam importantes discussões sobre a região acerca do Glyma.20g084500, no cromossomo 20. O Glyma.20g084500 é o único gene presente na região alvo e pode ser o gene causal associado ao aumento do conteúdo da proteína. Porém, experimentos futuros ainda serão necessários para finalizar a clonagem de Glyma.20g084500 e para o desenvolvimento de combinações na população NILS.

Marcador	115	663	666	667	668	670	674	678	190
BR80	0	0	0	2	2	0	2	0	0
NT12	2	2	2	0	0	2	0	2	2
163	1	0	0	2	0	2	0	2	2
10 indivíduos	1	0	0	2	0	2	0	2	2
8 indivíduos	2	0	0	2	0	2	0	2	2
1 indivíduo	0	0	0	2	0	2	0	2	2
251	2	1	1	1	1	1	1	1	1
6 indivíduos	2	1	1	1	1	1	1	1	1
3 indivíduos	2	0	0	2	2	0	2	0	0
1 indivíduo	2	2	2	0	0	2	0	2	2
499	2	1	1	1	1	1	1	1	1
8 indivíduos	2	1	1	1	1	1	1	1	1
2 indivíduos	2	2	2	0	0	2	0	2	2
1 indivíduo	2	0	0	2	2	0	2	0	0
517	2	1	1	1	1	1	1	1	1
2 indivíduos	2	2	2	0	0	2	0	2	2
6 indivíduos	2	0	0	2	2	0	2	0	0
13 indivíduos	2	1	1	1	1	1	1	1	1

Tabela 5. Recombinações finais obtidas através do desenvolvimento das NILs até o momento pelo PMQS.

REFERÊNCIAS

- Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J. and Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*, 8, 1-13. <https://doi.org/10.3835/plantgenome2015.04.0024>
- Bueno, R. D., Lorenzoni, R. M., Bernardeli, A., Silva, L. D. C., Silva, L. C. C., Silva, J. N. B., Vidigal, P. M. P., Piovesan, N. D., Ribeiro, C and Dal-Bianco, M. (2022). Fine-mapping of a major QTL for protein content in BR8014887 soybean line. <https://doi.org/10.21203/rs.3.rs-1598069/v1>
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatic*, 23, 2633-2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B. and Specht, J. E. (2003). The seed protein, oil, and yield QTL on soybean linkage group I. *Crop science*, 43, 1053-1067. <https://doi.org/10.2135/cropsci2003.1053>
- Cotegna. (2021). Brazilian Soybeans Exports – Market and Quality Insights. <https://www.cotegna.com/en/media/news/brazilian-soybeans-exports-market-and-quality-insights>. Accessed 21 Feb 2022.
- Cruz, C. D. and Schuster, I. (2004). GQMOL: aplicativo computacional para análise de dados moleculares e de suas associações com caracteres quantitativos. Versão 2.1. Viçosa: UFV. 2004.
- Cruz, C. D. (2013) GENES – a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum*, p.271-276.
- Dellaporta, S. L., Wood, J. and Hicks, J. B. (1983). A plant DNA miniprep: version II. *Plant molecular biology reporter*, 1, 19-21.
- Eskandari, M., Cober, E. R. and Rajcan, I. (2013). Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. *Theoretical and Applied Genetics*, 126, 1677-1687. <https://doi.org/10.1007/s00122-013-2083-z>
- Fonseca, A. F., Neves, E., Dipple, F. L. and Lima, T. C. (2020). Correlation between productivity and centesimal composition of soybean cultivars. *Scientific Electronic Archives*, 13(12), 54-58. <http://dx.doi.org/10.36560/131220201242>

Fliege, C. E., Ward, R. A., Vogel, P., Nguyen, H., Quach, T., Guo, M. and Diers, B. W. (2022). Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. *The Plant Journal*, 110, 114-128. <https://doi.org/10.1111/tbj.15658>

Genomics, L. G. C. (2013). KASP genotyping chemistry User guide and manual. LGC Limited, Teddington, UK.

Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D. and An, Y. Q. C. (2022). POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nature Communications*, 13, 3051. <https://doi.org/10.1038/s41467-022-30314-7>

He, Q., Xiang, S., Yang, H., Wang, W., Shu, Y., Li, Z. and Wang, S. (2021). A genome-wide association study of seed size, protein content, and oil content using a natural population of Sichuan and Chongqing soybean. *Euphytica*, 217, 1-14. <https://doi.org/10.1007/s10681-021-02931-8>

Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J. and Cregan, P. B. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC genomics*, 15, 1-12. <https://doi.org/10.1186/1471-2164-15-1>

Karikari, B., Chen, S., Xiao, Y., Chang, F., Zhou, Y., Kong, J. and Zhao, T. (2019). Utilization of interspecific high-density genetic map of RIL population for the QTL detection and candidate gene mining for 100-seed weight in soybean. *Frontiers in Plant Science*, 10, 1001. <https://doi.org/10.3389/fpls.2019.01001>

Kim, M., Schultz, S., Nelson, R. L. and Diers, B. W. (2016) Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. *Crop Science*, 56, 219-225. <https://doi.org/10.2135/cropsci2015.06.0340>

Kim, W. J., Kang, B. H., Moon, C. Y., Kang, S., Shin, S., Chowdhury, S., Choi, M. S., Park, S. W., Moon, J. K. and Ha, B. K. (2023) Quantitative Trait Loci (QTL) Analysis of Seed Protein and Oil Content in Wild Soybean (*Glycine soja*). *International Journal of Molecular Sciences*, 24, 4077. <https://doi.org/10.3390/ijms24044077>

Li, D., Zhao, X., Han, Y., Li, W. and Xie, F. (2019). Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. *Genomics*, 111, 90-95. <https://doi.org/10.1016/j.ygeno.2018.01.004>

Marsh, J. I., Hu, H., Petereit, J., Bayer, P. E., Valliyodan, B., Batley, J. and Edwards, D. (2022). Haplotype mapping uncovers unexplored variation in wild and domesticated soybean at the major protein locus *cqProt-003*. *Theoretical and Applied Genetics*, 135, 1443-1455. <https://doi.org/10.1007/s00122-022-04045-8>

Orel, V. and Wood, R. (1981). Early developments in artificial selection as a background to Mendel's research. *History and Philosophy of the Life Sciences*, 3, 145-170. <http://www.jstor.org/stable/23328309>.

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

Patil, G., Chaudhary, J., Vuong, T. D., Jenkins, B., Qiu, D., Kadam, S. and Nguyen, H. T. (2017). Development of SNP genotyping assays for seed composition traits in soybean. *International Journal of Plant Genomics*, 2017. <https://doi.org/10.1155/2017/6572969>

Patil, G., Vuong, T. D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C. and Nguyen, H. T. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. *Plant Biotechnology Journal*, 16, 1939-1953. <https://doi.org/10.1111/pbi.12929>

Prenger, E. M., Yates, J., Mian, M.A.R. et al (2019) Introgression of a High Protein Allele into an Elite Soybean Cultivar Results in a High-Protein Near-Isogenic Line with Yield Parity. *Crop Sci* 59:2498– 2508. <https://doi.org/10.2135/CROPSCI2018.12.0767>

Updaw, N. J., Bullock, J. B. and Nichols, T. E. (1976). Pricing soybeans on the basis of oil and protein content. *Journal of Agricultural and Applied Economics*, 8, 129-132. doi:10.1017/S0081305200013340

Wang, X., Jiang, G. L., Green, M., Scott, R. A., Song, Q., Hyten, D. L. and Cregan, P. B. (2014). Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. *Mol Genet Genomics*, 289, 935-949, 2014. <https://doi.org/10.1007/s00438-014-0865-x>

Zhang, T., Wu, T., Wang, L., Jiang, B., Zhen, C., Yuan, S., Hou, W., Wu, C., Han, T. and 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, 5915. <https://doi.org/10.3390/ijms20235915>

CONSIDERAÇÕES FINAIS

Este estudo identificou vinte e dois novos QTLs (Quantitative Trait Loci) que podem ser utilizados em programas de melhoramento que visam aumentar o conteúdo de proteína ou óleo em grãos de soja. Dentre esses marcadores, os dois mais promissores foram o ss715606597 e o ss715598567. Ressaltamos a importância dos ancestrais Roanoke, Lee e Bragg na composição genética das variedades brasileiras de soja. As descobertas deste trabalho demonstram que certos marcadores estão sendo perdidos nas cultivares modernas, enquanto outros mantêm ou até aumentam sua frequência. Estas observações indicam regiões genômicas que passaram por seleção e podem ser valiosas em programas de melhoramento de soja.

Marcadores moleculares SNPs foram avaliados em uma população de RILS e destes o SNP ss61, ss73, ss134 e ss185 associaram-se para o conteúdo de proteína em um dos ambientes. O marcador ss115 está associado tanto ao conteúdo de óleo quanto à proteína em todas as localidades e representa uma alternativa viável para o uso deste marcador em seleção assistida no Brasil e utilizando a cultivar SUPREMA como doador. Além disso, foram apontados quatro genes candidatos para aumento no conteúdo de óleo (Glyma.15g034400, Glyma.15g050900, Glyma.15g063000 e Glyma.15g070700).

A partir de estudos anteriores do Programa de Melhoramento de Soja (PMQS/UFV) foi apontado o gene Glyma.20g084500 como o único gene presente na região alvo e pode ser o gene causal associado ao aumento do conteúdo da proteína em soja. Experimentos futuros ainda serão necessários para finalizar a clonagem de Glyma.20g084500 e para o desenvolvimento de transferências na população NILS.