UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS AGRÁRIAS E ENGENHARIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM PRODUÇÃO VEGETAL

RODRIGO MONTE LORENZONI

ANÁLISES FENOTÍPICAS E MOLECULARES NA IDENTIFICAÇÃO DE SNPs ASSOCIADOS AOS CONTEÚDOS DE PROTEÍNAS TOTAL E DE RESERVA DE SOJA [*Glycine max* (L.) Merril]

ALEGRE, ES

2020

RODRIGO MONTE LORENZONI

ANÁLISES FENOTÍPICAS E MOLECULARES NA IDENTIFICAÇÃO DE SNPs ASSOCIADOS AOS CONTEÚDOS DE PROTEÍNAS TOTAL E DE RESERVA DE SOJA [*Glycine max* (L.) Merril]

Tese apresentada a Universidade Federal do Espírito Santo, como parte das exigências do Programa de Pós-Graduação em Produção Vegetal, para obtenção do título de *Doctor Scientiae*.

Orientadora: Profª. Drª. Taís Cristina Bastos Soares.
Coorientador: Prof. Dr. Maximiller Dal-Bianco Lamas Costa

ALEGRE, ES

2020

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Lorenzoni, Rodrigo Monte, 1990-
L869a Análises fenotípicas e moleculares na identificação de SNPs associados aos conteúdos de proteínas total e de reserva de soja [Glycine max (L.) Merril] / Rodrigo Monte Lorenzoni. - 2020.
65 f. : il.

Orientadora: Taís Cristina Bastos Soares.
Coorientador: Maximiller Dal-Bianco Lamas Costa.
Tese (Doutorado em Produção Vegetal) - Universidade Federal do Espírito Santo, Centro de Ciências Agrárias e Engenharias.

1. Plantas - Melhoramento genético. 2. Biotecnologia vegetal. 3. Proteínas de soja. 4. Soja. I. Soares, Taís Cristina Bastos. II. Costa, Maximiller Dal-Bianco Lamas. III. Universidade Federal do Espírito Santo. Centro de Ciências Agrárias e Engenharias. IV. Título.

CDU: 63

RODRIGO MONTE LORENZONI

Análises fenotípicas e moleculares na identificação de SNPs associados aos conteúdos de proteínas total e de reserva de soja [*Glycine max* (L.) Merril]

Tese apresentada a Universidade Federal do Espírito Santo, como parte das exigências do Programa de Pós-Graduação em Produção Vegetal, para obtenção do título de Doutor em Produção Vegetal.

Aprovada em 28 de fevereiro de 2020.

COMISSÃO EXAMINADORA

Profª. Drª. Taís Cristina Bastos Soares
Universidade Federal do Espírito Santo
Orientadora

Prof. Dr. Maximiller Dal-Bianco Lamas Costa
Universidade Federal de Viçosa

Dr. Newton Deniz Piovesan
Universidade Federal de Viçosa

Prof. Dr. Moisés Zucoloto
Universidade Federal do Espírito Santo

Prof. Dr. Eduardo Frizzera Meira
Universidade Federal do Espírito Santo

Aos meus pais, Osmar e Marilene,

pelo exemplo, apoio e amor,

Dedico

“A resposta certa não importa nada: o essencial é que as perguntas estejam certas.”

(Mario Quintana)

AGRADECIMENTOS

A Deus pelo dom da vida e por me guiar e proteger em todos os momentos;

A Universidade Federal do Espírito Santo e ao Programa de Pós-Graduação em Produção Vegetal, pela oportunidade de realização do Curso;

A Universidade Federal de Viçosa (UFV) por permitir parcerias e realização dos experimentos;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa e apoio financeiro. E, a empresa Caramuru Alimentos S.A. pela parceria e viabilizar nossos estudos;

À professora Taís Cristina Bastos Soares, por tantos ensinamentos, confiança, profissionalismo, conselhos e amizade. E, acima de tudo por me acolher desde a iniciação científica da graduação até a conclusão do doutorado, muito obrigado por tudo!

Ao coorientador, professor Maximiller Dal-Bianco Lamas Costa, pela colaboração, suporte, auxílios e aconselhamento ao longo do desenvolvimento do trabalho;

A toda equipe do Laboratório de Bioquímica e Biologia molecular da UFES, em especial Adelson, Carla, Franciele, Liliana, Lucimara e Rodrigo Dadalto, pelas trocas de experiências, convivência, amizade, pelos momentos de descontração;

Às técnicas, Soninha, Jerusa, Sandra e Magda, pelo suporte e pelos divertidos momentos de brincadeiras e descontração;

Às secretárias da pós-graduação, Madalena e Alessandra, por serem sempre solícitas e atenciosas;

Aos Laboratórios de Bioquímica Genética de Plantas (BIOAGRO-UFV), Genética Molecular de Plantas (BIOAGRO-UFV) e Análises Bioquímicas (BIOAGRO-UFV) pelo suporte e por permitir a realização das análises;

A toda equipe do Laboratório de Bioquímica Genética de Plantas: Cal, Naldo, Rafael Bueno, Tiago, Yan, Rafael Aguiar, Lenon, Walter, Teresinha e Guilherme. Em especial, ao Bruno, Natália e Arthur pela parceria nas análises de laboratório e condução

dos experimentos de campo. E, sobretudo a Marina e Jéssica, pelo auxílio na rotina de laboratório, amizade, confiança e companheirismo em todos os momentos;

Ao Dr. Newton Deniz Piovesan pelo suporte e acompanhamento nos experimentos;

Ao Prof. Moises Zucoloto e Prof. Eduardo Frizzera Meira, meus agradecimentos pela contribuição e disponibilidade de avaliar o trabalho;

Aos amigos que fiz ao longo da vida e que carrego comigo mesmo distantes geograficamente: Ludymilla, Rafael, José Dias, Kati, Fran, Ane, Edilson, Paulo, Ana Carolina, Leonardo, Samuel, Tafarel e Andréia;

Aos novos amigos que fiz em Viçosa: Warley, Hallef, Jorge, Roberta e Darlan. Quantos momentos de descontração, ensinamentos e discussões tivemos, os quais levarei para a vida toda!

Aos companheiros de república Edilson, Jorge e Kaique, e Paulo por todo companheirismo, descontração e suporte ao longo dessa caminhada;

Ao meu melhor amigo e parceiro para a vida, Warley, pela compreensão, incentivo, respeito e por compartilhar tristezas e alegrias, sendo sempre apoio um do outro;

Àqueles que são a base dessa conquista e não hesitaram em medir esforços para que eu pudesse alcançar essa etapa, meus pais: Osmar e Marilene. Pelo amor incondicional, dedicação, apoio constante e por todo carinho;

À minha irmã Rafaela, pelo carinho e apoio ao longo desses anos;

E, a todos que de alguma forma, fizeram ou fazem parte da minha trajetória, me apoiaram e impulsionaram.

Muito Obrigado!

BIOGRAFIA

Rodrigo Monte Lorenzoni, filho de Osmar Lorenzoni e Marilene Monte Lorenzoni, nasceu em Venda Nova do Imigrantes-ES em 18 de dezembro de 1990.

Em 2009, entrou para a Universidade Federal do Espírito Santo (UFES), em Alegre-ES, onde obteve o título de Bacharel em Agronomia em Março de 2014. Durante o período de graduação foi bolsista de iniciação científica, onde desenvolveu atividades de pesquisa em biotecnologia, biologia molecular e melhoramento de plantas.

Entre março e julho de 2014 foi bolsista de apoio técnico no laboratório de Genética e Melhoramento Vegetal da UFES.

Em agosto de 2014, ingressou no mestrado no Programa de Pós-Graduação em Produção Vegetal, em que utilizou ferramentas moleculares e citogenéticas na caracterização genética-molecular de acessos de *Annona mucosa*, obtendo o título de mestre em fevereiro de 2016.

Em março de 2016, ingressou no doutorado no Programa de Pós-Graduação em Produção Vegetal - Biotecnologia e Ecofisiologia do Desenvolvimento de Plantas, em parceria com a Universidade Federal de Viçosa, submetendo-se à defesa da tese em fevereiro de 2020.

SUMÁRIO

INTRODUÇÃO GERAL

A soja [*Glycine max* (L.) Merr.] é uma das culturas agrícolas com maior expressão na economia brasileira e mundial. De acordo com as estimativas de safra 2019/20, o Brasil se tornará o maior produtor mundial com 120,86 milhões de toneladas (CONAB, 2020) enquanto, os Estados Unidos produzirão cerca de 96,84 milhões de toneladas. Nesse panorama, o Brasil terá um acréscimo de 5,13% na produção e os EUA uma redução de 19,64% (USDA, 2020).

A representatividade da cultura é devido, principalmente, aos subprodutos óleo e farelo. Este último constitui uma das principais fontes proteicas para alimentação animal. Além do uso na nutrição animal, o consumo de grãos inteiros ou alimentos derivados como leite, tofu e proteína hidrolisada, vem sendo cada vez mais adicionado como fonte proteica na alimentação humana (ZHANG et al., 2015). A maior demanda por grãos e produtos derivados de soja faz com que haja necessidade de promover melhorias na qualidade nutricional do produto final.

As proporções dos constituintes dos grãos de soja (proteínas, óleo, carboidratos e minerais) sofreram influência dos extensivos programas de melhoramento que visavam obter cultivares de soja mais produtivas. A correlação negativa entre produtividade e conteúdo proteico promoveu reduções significativas na composição proteica dos grãos (PATIL et al., 2017).

O conteúdo proteico é composto principalmente das proteínas de reserva (65 a 80%), as quais são classificadas de acordo com o coeficiente de sedimentação (2S, 7S, 11S e 15S). Dentre as proteínas de reserva, as frações 7S e 11S representam cerca de 70%. A fração 7S é composta principalmente de β-conglicinina (92%), além de β-amilase e lipoxigenase. Já a fração 11S é constituída pela globulina glicinina (QI et al., 2016).

A β-conglicinina é constituída de três subunidades, α', α, e β com massa molecular estimada em 150-170kDa. E, a glicinina, consiste em seis polipeptídios ácidos (A1a, A1b, A2, A3, A4 e A5) e cinco básicos (B1a, B1b, B2, B3 e B4) associados, de modo específico, por meio de ligações bissulfídricas com massa molecular total de 350kDa (CHEN et al., 2014). O que difere a glicinina e β - conglicinina é a composição dos aminoácidos de cada uma, principalmente, na quantidade de aminoácidos sulfurados como metionina, onde β – conglicinina possui

um nível reduzido (0,61%) enquanto, glicinina tem quantidades superiores de metionina (2,23%) e cisteína (1,83%) (TASKI-AJDUKOVIC et al., 2010).

Diante disso, a qualidade nutricional em relação a aminoácidos pode ser melhorada ao modular a proporção de glicinina e β – conglicinina por meio de cruzamentos convencionais ou programas de engenharia genética (CHEN et al., 2014; ZARKADAS et al., 2007).

Estratégias moleculares tornaram-se ao longo do tempo alternativas eficazes e promissoras nos programas de melhoramento, as quais juntamente com análises fenotípicas fornecem robustez aos programas de seleção, permitindo selecionar genótipos com características de interesse de forma mais eficiente e rápida.

O uso de marcadores moleculares constitui uma importante ferramenta nos programas de melhoramento, pois permitem a identificação de genes relacionados a diferentes características de importância agronômica. As novas tecnologias de genotipagem permitem acelerar os programas de melhoramento pela genotipagem de SNPs (*single nucleotide polymorphisms*) em larga escala (BOEHM et al., 2018; HWANG et al., 2014; LI et al., 2018). Atualmente, os SNPs constituem a classe de marcadores mais indicadas para estudos de mapeamento e seleção genômica.

A compreensão de características quantitativas foi revolucionada pela disponibilidade da grande quantidade de SNPs que cobrem o genoma. Dados de genotipagem de SNPs combinados com avaliações fenotípicas têm sido usados para três propósitos: estudar a arquitetura genética de características quantitativas, mapear regiões do genoma que causam variação nessas características (*quantitative trait loci* ou QTL) e predizer os ganhos genéticos (KEMPER et al., 2018; MOSER et al., 2015).

Em soja, nos últimos anos vêm sendo relatados diversos trabalhos que utilizam os SNPs como ferramenta para genotipagem de materiais. Em 2010 foi publicado um mapa genético de alta densidade utilizando SNPs e que serviu de base para os novos estudos que vêm sendo realizados (HYTEN et al., 2010). (SONG et al., 2013) identificaram 209.903 SNPs mapeando leituras curtas de cada um dos oito acessos de soja que incluíam seis genótipos cultivados (*Glycine max*) e dois genótipos de soja selvagem (*G. soja*) e selecionaram 52.041 SNPs para o projeto do SoySNP50K Illumina Infinium BeadChip. O BeadChip foi utilizado com sucesso para genotipar toda a coleção de germoplasma de soja do USDA (*United States Department of Agriculture)***,** o banco de dados está disponível no Soybase, (www.soybase.org) e vem sendo utilizado para análises de associação (BOEHM et al., 2018; SONG et al., 2015); análise de loco

de características quantitativas (QTL) (CHANG et al., 2018; LEE et al., 2017; WARRINGTON et al., 2015) e GWAS (BOEHM et al., 2018; LEE et al., 2019; LI et al., 2019, 2018; MAMIDI et al., 2014; PHANSAK et al., 2016; SONAH et al., 2015; WEN et al., 2014; YAN et al., 2017; ZHANG et al., 2016) .

Apesar das inovações tecnológicas que caracterizam os genomas de maneira rápida e dos métodos computacionais que continuam a melhorar a análise de grandes conjuntos de dados, a capacidade de determinar com rapidez e precisão caracteres fenotípicos continua sendo um fator limitante no melhoramento genético de plantas. A fenotipagem de alto rendimento é uma ferramenta emergente com potencial para acelerar a descoberta genética e identificar combinações genéticas que permitirão uma seleção mais rápida de variedades de alto rendimento (TATTARIS; REYNOLDS; CHAPMAN, 2016).

Dentro do contexto da fenotipagem de alto rendimento, a espectroscopia de infravermelho próximo (NIR) é uma tecnologia que pode predizer valores das principais moléculas de armazenamento de sementes simultaneamente (OSBORNE, 2006). Com vantagens de ser um método analítico rápido e de baixo custo, permitindo analisar elevado número de amostras em curto espaço de tempo (MONTES; MELCHINGER; REIF, 2007).

Os espectros de absorção do NIR são devido aos grupos funcionais C-H, N-H, O-H e S-H, que permitem a previsão de diversos compostos orgânicos. Materiais biológicos apresentam diferentes padrões de absorbância sobrepostos devido à mistura complexa de compostos orgânicos e, por isso abordagens estatísticas multivariadas são necessárias para interpretar espectros NIR de amostras biológicas (SPIELBAUER et al., 2009; XU et al., 2020). Nesse sentido, é necessário estabelecimento de métodos de calibração e validação que sejam eficientes para estimar o caracter a ser avaliado. Diferentes componentes de sementes de soja já foram estimados por NIR, como, proteína total, proteínas solúveis, óleo, ácidos graxos e viabilidade (KUSUMANINGRUM et al., 2018; PATIL et al., 2010; XU et al., 2020; ZHU et al., 2018).

Os principais constituintes das sementes de soja são de características quantitativas e influenciados pelo ambiente. A grande maioria dos estudos mencionados acima foi realizada em países com diferentes condições climáticas do Brasil e uso de diferentes cultivares, o que faz necessário ajustar curvas e validar modelos para que

possam estimar com acurácia os constituintes encontrados nas condições e cultivares brasileiras e**,** dessa forma introgredir análises por NIR nos programas de melhoramento.

Ao propor a calibração de um novo modelo para estimar o conteúdo de proteínas de reserva, como o realizado neste estudo, é possível reduzir o tempo das análises, por ser uma ferramenta que permite mensurar de forma bastante ágil elevado número de amostras ao comparado com o método tradicional, sem diminuir a acurácia. Além disso, como já se tem modelos para proteínas totais, óleo e umidade, torna-se possível estimar o conteúdo de proteínas de reserva simultaneamente aos demais caracteres a partir de uma mesma leitura espectral.

Com a competitividade de mercado, empresas de melhoramento buscam lançar cultivares a frente de suas concorrentes, para isso sucessivos ciclos de seleção vêm sendo realizados ao longo do ano, o que fornece elevado número de genótipos a serem avaliados e selecionados para os ciclos subsequentes. Por isso, a fenotipagem de alto rendimento, como a realizada por NIR, vem sendo cada ves mais requerida dentro dos programas. Concomitante, as ferramentas de genotipagem estão cada vez mais robustas e fornecem grande número de informações que devem ser cruzadas com a fenotipagem com intuito de buscar marcadores associados. Ao observar associação e verificar a estabilidade de um marcador, o mesmo pode vir a ser utilizado na seleção assistida, o que vem a acelerar os ciclos de seleção por garantir que determinada marca está associada ao aumento da característica de interesse e**,** dessa forma auxiliar na seleção de individuos superiores para seguir no programa.

Diante do exposto, foi hipotetizado que a associação dos caracteres fenotípicos com a genotipagem por SNPs, permite identificar e mapear QTLs para teores de proteína total e de cada subunidade proteica e assim é possível selecionar marcadores para utilização em seleção assistida diminuindo o tempo de seleção de materiais em programas de melhoramento. Sob esta perspectiva foram desenvolvidos os trabalhos relacionados aos dois capítulos subsequentes da tese.

REFERÊNCIAS

BOEHM, J. D. et al. Genetic mapping and validation of the loci controlling 7S α′ and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. **Theoretical and Applied Genetics**, v. 131, n. 3, p. 659–671, 1 mar. 2018.

CHANG, H. X. et al. Integration of sudden death syndrome resistance loci in the soybean genome. **Theoretical and Applied Genetics**, v. 131, n. 4, p. 757–773, 1 abr. 2018.

CHEN, Q. et al. The α′ subunit of β-conglycinin and the A1-5 subunits of glycinin are not essential for many hypolipidemic actions of dietary soy proteins in rats. **European Journal of Nutrition**, v. 53, n. 5, p. 1195–1207, 2014.

CONAB. **Acompanhamento safra brasileirade grãos**. 7. ed. Brasília: Companhia Nacional de Abastecimento, 2020.

HWANG, S. et al. Genetics and mapping of quantitative traits for nodule number, weight, and size in soybean (*Glycine max* L.[Merr.]). **Euphytica**, v. 195, n. 3, p. 419–434, 2014.

HYTEN, D. L. et al. High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. **BMC Genomics**, v. 11, n. 1, 15 jan. 2010.

KEMPER, K. E. et al. A multi-trait Bayesian method for mapping QTL and genomic prediction. **Genetics Selection Evolution**, v. 50, n. 1, 24 mar. 2018.

KUSUMANINGRUM, D. et al. Non-destructive technique for determining the viability of soybean (*Glycine max*) seeds using FT-NIR spectroscopy. **Journal of the Science of Food and Agriculture**, v. 98, n. 5, p. 1734–1742, 30 mar. 2018.

LEE, J. S. et al. Identification of SNPs tightly linked to the QTL for pod shattering in soybean. **Molecular Breeding**, v. 37, n. 4, 1 abr. 2017.

LEE, S. et al. Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. **Theoretical and Applied Genetics**, v. 132, n. 6, p. 1639–1659, 1 jun. 2019.

LI, D. et al. Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. **Genomics**, v. 111, n. 1, p. 90–95, 1 jan. 2019.

LI, Y. HUI et al. Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. **Plant Science**, v. 266, p. 95–101, 1 jan. 2018.

MAMIDI, S. et al. Genome-Wide Association Studies Identifies Seven Major Regions Responsible for Iron Deficiency Chlorosis in Soybean (*Glycine max*). **PLoS ONE**, v. 9, n. 9, p. e107469, 16 set. 2014.

MONTES, J. M.; MELCHINGER, A. E.; REIF, J. C. Novel throughput phenotyping platforms in plant genetic studies. **Trends in Plant Science**, v. 12, n. 10, p. 433–436, out. 2007.

MOSER, G. et al. Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. **PLoS Genetics**, v. 11, n. 4, 1 abr. 2015.

OSBORNE, B. G. Applications of near infrared spectroscopy in quality screening of early-generation material in cereal breeding programmes. **Journal of Near Infrared Spectroscopy**, v. 14, n. 1, p. 93–101, 2006.

PATIL, A. G. et al. Nondestructive estimation of fatty acid composition in soybean [*Glycine max* (L.) Merrill] seeds using Near-Infrared Transmittance Spectroscopy. **Food Chemistry**, v. 120, n. 4, p. 1210–1217, 15 jun. 2010.

PATIL, G. et al. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. **Theoretical and Applied Genetics**, v. 130, n. 10, p. 1975–1991, 2017.

PHANSAK, P. et al. Multi-population selective genotyping to identify soybean [*Glycine max* (L.) Merr.] seed protein and oil QTLs. **G3: Genes, Genomes, Genetics**, v. 6, n. 6, p. 1635–1648, 2016.

QI, Z. et al. Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. **Molecular Breeding**, v. 36, n. 5, 1 maio 2016.

SONAH, H. et al. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. **Plant Biotechnology Journal**, v. 13, n. 2, p. 211–221, 1 fev. 2015.

SONG, Q. et al. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLoS ONE**, v. 8, n. 1, 30 jan. 2013.

SONG, Q. et al. Fingerprinting soybean germplasm and its utility in genomic research. **G3: Genes, Genomes, Genetics**, v. 5, n. 10, p. 1999–2006, 2015.

SPIELBAUER, G. et al. High-throughput near-infrared reflectance spectroscopy for predicting quantitative and qualitative composition phenotypes of individual maize kernels. **Cereal Chemistry**, v. 86, n. 5, p. 556–564, set. 2009.

TATTARIS, M.; REYNOLDS, M. P.; CHAPMAN, S. C. A Direct Comparison of Remote Sensing Approaches for High-Throughput Phenotyping in Plant Breeding. **Frontiers in Plant Science**, v. 7, 3 ago. 2016.

USDA. **World agricultural production**. 1. ed. Washington: Foreign Agricultural Service/USDA, 2020.

WARRINGTON, C. V. et al. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. **Theoretical and Applied Genetics**, v. 128, n. 5, p. 839–850, 1 maio 2015.

WEN, Z. et al. Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. **BMC Genomics**, v. 15, n. 1, p. 809, 2014.

XU, R. et al. Use of near-infrared spectroscopy for the rapid evaluation of soybean [*Glycine max* (L.) Merri.] water soluble protein content. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 224, 5 jan. 2020.

YAN, L. et al. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. **BMC Genomics**, v. 18, n. 1, 12 jul. 2017.

ZARKADAS, C. G. et al. Protein quality and identification of the storage protein subunits of tofu and null soybean genotypes, using amino acid analysis, one- and two-dimensional gel electrophoresis, and tandem mass spectrometry. **Food Research International**, v. 40, n. 1, p. 111–128, jan. 2007.

ZHANG, J. et al. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). **Theoretical and Applied Genetics**, v. 129, n. 1, p. 117–130, 1 jan. 2016.

ZHANG, Y. H. et al. Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]. **Theoretical and Applied Genetics**, v. 128, n. 6, p. 1061–1072, 16 jun. 2015.

ZHU, Z. et al. Determination of soybean routine quality parameters using near-infrared spectroscopy. **Food Science and Nutrition**, v. 6, n. 4, p. 1109–1118, 1 jun. 2018.

Capítulo 1

ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO (NIR): UMA FERRAMENTA RÁPIDA E EFICIENTE DE ESTIMATIVA DA CONCENTRAÇÃO DE PROTEÍNAS DE RESERVA EM SEMENTES DE SOJA [*Glycine max* (L.) MERRIL]

RESUMO

LORENZONI, Rodrigo Monte. Universidade Federal do Espírito Santo, fevereiro de 2020. **Espectroscopia de infravermelho próximo (NIR): uma ferramenta rápida e eficiente de estimativa da concentração de proteínas de reserva em sementes de soja [*Glycine max* (L.) Merrill]**. Orientadora: Taís Cristina Bastos Soares. Coorientador: Maximiller Dal-Bianco Lamas Costa.

O conteúdo de proteínas de reserva em soja é uma característica que influencia diretamente na qualidade nutricional da proteína, uma vez que a proporção de β-conglicinina (7S) e glicinina (11S) determina o balanço de aminoácidos sulfurados. Programas de melhoramento genético voltados para qualidade da soja v**ê**m sendo desenvolvidos, e são necessárias avaliações de elevado número de genótipos nas etapas de seleção. O método tradicional de estimativa é laborioso e demorado, portanto, novas ferramentas devem ser utilizadas a fim de otimizar a caraterização de materiais. A espectroscopia de infravermelho próximo (NIR) é uma técnica que vem sendo utilizada na indústria de alimentos para determinação de diferentes constituintes e apresenta alta sensibilidade, elevado potencial preditivo, além de permitir analisar amostras em curto período. Devido à demanda para estimar conteúdo de proteínas de reserva objetivou-se com o presente trabalho desenvolver uma curva de calibração NIR a partir da correlação do método de referência e os espectros adquiridos por NIR e validá-la para estimar a concentração de proteínas de reserva em sementes de soja. Foi estabelecido um modelo de regressão de PLS com 105 amostras de calibração, 25 amostras de validação cruzada e 26 para validação externa, as quais foram previamente quantificadas por Kjeldahl e posteriormente obtidos os espectros de NIR. Os espectros foram corrigidos pela primeira derivada e então calibrado o modelo. Com base nas estimativas e correção dos espectros na calibração foi verificado coeficiente de determinação (R^2) de 0,965 e para a validação externa foi de 0,976. Os demais parâmetros de confiabilidade para validação foram suficientes para garantir a aplicabilidade do modelo ajustado, na estimativa do conteúdo proteico, onde foi obtido EPC de 0,293%, EPVC de 0,726% e EPP 0,562%. Foi ainda calculado o determinante preditivo da validação cruzada e da validação externa, com valores de 4,146 e 5,099, respectivamente. A capacidade preditiva do modelo desenvolvido neste trabalho foi considerada satisfatória e, portanto, foi validada a hipótese de que modelos preditivos pela regressão PLS é eficiente na estimativa do conteúdo de proteínas de reserva e, dessa forma análises de NIR podem ser utilizadas

em programas de melhoramento por propiciar redução de tempo e custos e, garantir acurácia nas estimativas.

ABSTRACT

LORENZONI, Rodrigo Monte. Universidade Federal do Espírito Santo, February 2020. **Near infrared spectroscopy (NIR): a fast and efficient tool for estimating the concentration of reserve proteins in soybean seeds [Glycine max (L.) Merrill].** Advisor: Taís Cristina Bastos Soares. Co-adivisor: Maximiller Dal-Bianco Lamas Costa.

Storage protein content in soybean seeds directly influences the nutritional quality of total protein, once the relation between glycinin (11S) e β-conglycinin (7S) fractions determine the rate of sulfur-containing essential amino acids. Soybean breeding programs that focus on soybean seed quality traits have been arising and thousands of genotypes must be evaluated every breeding cycle. The conventional method of estimating storage protein content is laborious and time-consuming, then, novel tools must be used in order to optimize the phenotyping process. Near-infrared spectroscopy (NIR) technique has been used in the food industry to determine different components. It has high sensitivity, predictive potential and allows rapid analysis of samples. To attend the high demand to estimate storage protein content, the present study aimed to validate a calibration curve through the correlation between the estimates from reference method and NIR spectra. Kjeldahl estimates and NIR spectra for storage protein content were obtained from 105 samples for calibration, 25 for cross-validation, and 26 for external validation, and then submitted to partial least squares analysis. The coefficient of determination between the two methods was 0.965 for calibration and 0.976 for external validation. The other parameters of reliability were estimated, then values of the standard error of calibration (0.293%), standard error of cross validation (0.726%) and standard error of prediction (0.562%) obtained were considered low. The relative predictive determinant was 4.146 and 5.099 for cross validation and external validation, respectively. The magnitudes of these parameters were adequate to ensure good predictive capacity and recommend the use of the adjusted model for estimating storage protein content through partial least squares regression. Therefore, low cost and time saving NIR analyses can be applied in soybean breeding programs for phenotyping seed component traits properly.

1. INTRODUÇÃO

A soja [*Glycine max* (L.) Merrill] é fonte primária de proteínas de origem vegetal na alimentação humana, e principal fonte proteica na nutrição animal. Seus grãos contêm mais proteína do que qualquer outra espécie vegetal utilizada comercialmente (KUSUMANINGRUM et al., 2018). Considerando a totalidade do conteúdo proteico, as proteínas de reserva compreendem a maior porção e são classificadas de acordo com o coeficiente de sedimentação, baseado na massa molecular que cada fração apresenta. As frações 11S (glicina) e 7S (β-conglicinina) são as principais unidades relacionadas à concentração de proteína na semente, por corresponderem de 70 a 85% da composição do grão (PANTALONE, 2012). O conteúdo, as proporções e a dinâmica da biossíntese das frações 11S e 7S variam com a cultivar e o ambiente, o que afeta diretamente a quantidade e qualidade de proteína nos grãos.

Na literatura é relatado correlação negativa entre produtividade e a concentração de proteína (PATIL et al., 2017). Entre as proteínas de reserva também é observado correlação negativa entre as frações 7S e 11S (KRISHNAN et al., 2007; WANG et al., 2014). Além da redução do teor proteico total, o acúmulo da fração 7S em detrimento da 11S resulta na diminuição da concentração de aminoácidos sulfurados (cisteína e metionina), portanto, perda de qualidade nutricional dos grãos de soja (BOEHM et al., 2018; KRISHNAN et al., 2007), uma vez que estes aminoácidos são essenciais para animais.

No mercado global, é crescente a comercialização de produtos classificados como de melhor qualidade nutricional. Com isso, caracteres relacionados à qualidade vêm se tornando um foco em programas de melhoramento. É importante ressaltar que se deve avaliar um elevado número de sementes para definir quais genótipos devem seguir no programa a partir da identificação e quantificação de um componente alvo. Consequentemente, as análises devem ser rápidas e acuradas (KUSUMANINGRUM et al., 2018; PATIL et al., 2017), para não comprometer o vigor das sementes e acelerar a obtenção de genótipos desejados. Convencionalmente, a estimativa de proteínas de reserva isolada é realizada pela quantificação de nitrogênio pelo método de Kjeldahl (LU et al., 2013). Entretanto, este método é demorado, laborioso, oneroso e poluente, fatores que restringem a avaliação de elevado número de amostras para determinação do conteúdo proteico dentro de um programa de melhoramento de soja. Com isso, é

imprescindível o desenvolvimento de métodos mais rápidos e econômicos para determinação da porcentagem de proteínas de reserva de soja.

A espectroscopia de infravermelho próximo (NIR) é uma técnica com elevado potencial analítico, por apresentar alta sensibilidade na determinação de multicaracteres em alimentos e permitir analisar grande número de amostras em curto período (CEN; HE, 2007; WANG; PALIWAL, 2007). A análise por NIR já vem sendo utilizada para mensurar a concentração de diferentes características e constituintes em sementes e grãos, dentre elas umidade, óleo e proteína total (ZHU et al., 2018), perfil de ácidos graxos (PATIL et al., 2010) e viabilidade de sementes (KUSUMANINGRUM et al., 2018). Contudo, o ponto crucial para utilização da técnica de NIR é a confiabilidade dos dados no processo de calibração da curva, a fim de garantir acurácia nas análises e fornecer resultados precisos e confiáveis.

Diante do exposto, o trabalho foi desenvolvido sob a hipótese de que é possível estimar a concentração de proteínas de reserva por meio da técnica NIR. Assim, objetivou-se desenvolver uma curva de calibração NIR a partir da correlação do método de referência e os espectros adquiridos por NIR e validá-la para estimar a concentração de proteínas de reserva em sementes de soja.

2. MATERIAL E MÉTODOS

2.1. Material vegetal

Para fins de calibração foram utilizadas 105 amostras de sementes de soja moídas, compreendendo 34 cultivares e 71 genótipos derivados de cruzamento em F5 entre duas cultivares de soja do programa de melhoramento para qualidade da soja do BIOAGRO-UFV (PMQS80 e PMQS12). Para validação cruzada foram utilizadas 25 amostras internas e para validação externa foram utilizadas um conjunto de 26 amostras provenientes de cultivares e linhagens do PMQS (BIOAGRO-UFV), as quais foram cultivadas em Capinópolis-MG/Brasil e Viçosa-MG/Brasil durante os anos de 2017 e 2018.

2.2. Análise de proteínas de reserva pelo método de referência

As análises foram realizadas no Laboratório de Bioquímica Genética de Plantas localizado no Instituto de Biotecnologia Aplicada à Agropecuária da Universidade Federal de Viçosa-MG.

Inicialmente, as proteínas de reserva glicinina (11S) e β-conglicinina (7S) foram extraídas com tampão fosfato salino (tampão fosfato de sódio 0,05 M, pH 7,6; NaCl 0,4 M; β-mercaptoetanol 0,28%) (SOARES et al., 2004). Para tal, foram utilizados aproximadamente 50 mg de soja moída e 1,5 mL do tampão de extração. Os tubos ficaram 45 minutos em banho de ultrassom, após foram centrifugados por 15 minutos a 14.000 rpm e retirado o sobrenadante.

Para análise de Kjeldahl foram utilizados 50 µL do sobrenadante extraído anteriormente contendo as frações 7S e 11S. Procedeu-se então a análise pelo método Kjeldahl modificado, segundo as normas analíticas do Instituto Adolfo Lutz (1985). Após a obtenção do material digerido com ácido sulfúrico, foi adicionado peróxido de hidrogênio 30%. Na fase de destilação, a amônia liberada foi recolhida em solução de ácido bórico 4%. Obtendo assim o teor de nitrogênio pela titulação da amônia com ácido clorídrico 0,05%. A partir do teor de nitrogênio, foi calculada a porcentagem de proteínas de reserva da amostra, empregando-se o fator de nitrogênio 6,25 no material analisado.

2.3. Aquisição dos espectros

Os espectros foram obtidos em espectrofotômetro modelo Antaris II FT-NIR Analyser™ (Thermo Fisher Scientific). O software usado para o NIR foi o TQ Analyst (Thermo Fisher Scientific), a região medida encontrava-se entre 9.995 e 4.025 cm^{-1} *wavenumber* e as médias de 20 varreduras sucessivas de cada uma das 105 amostras foram utilizadas para as análises.

Antes da calibração, os espectros de transmitâncias originais foram corrigidos pelo método matemático de primeira derivada.

2.4. Calibração e validação

A calibração foi realizada pelo método PLS (*Partial Least Square*). A técnica de calibração PLS é baseada no algoritmo de PLS1, que examina a região ou regiões especificadas dos espectros de calibração para determinar quais áreas variam estatisticamente em função da concentração do componente. A verificação da presença de outliers foi realizada verificando o valor real da concentração obtida pelo método de referência e o valor da predição pelo NIR. Extremos valores, acima e/ou abaixo, de duas vezes o desvio padrão foram removidos das análises (*outliers*) e a calibração realizada novamente. O erro padrão de calibração (EPC) e coeficiente de determinação (R^2) foram usados para selecionar a melhor equação de calibração.

2.5. Validação cruzada

O desempenho da equação obtida na calibração foi determinado a partir da validação cruzada utilizando 25 amostras. O erro padrão de calibração (EPC), o coeficiente de determinação (R^2) e o erro padrão de validações cruzadas (EPVC) foram utilizados para selecionar a melhor equação de calibração (WINDHAM; MERTENS; BARTON II, 1989). O determinante preditivo relativo para validação cruzada (DPRc = DP dos dados de referência / EPVC) foi calculado para determinar a capacidade preditiva de cada modelo.

2.6. Validação externa

Diferente da validação cruzada, a qual utiliza amostras de dentro do conjunto original, a validação externa deve ser realizada com amostras diferentes das originais. Portanto, um conjunto de testes independentes, que representavam uma faixa dentro da amplitude do conteúdo proteico predeterminada pela calibração foi utilizado para validação de cada modelo. Como parte da validação para determinar a precisão da estimativa foram calculados: coeficiente de determinação (R^2), erro padrão da predição (EPP) e determinante preditivo relativo (DPRv = DP dos dados do conjunto de testes / EPP) (WILLIAMS, 2001). A estatística de regressão de validação foi realizada conforme (GOMEZ; GOMEZ, 1984) .

3. RESULTADOS

3.1. Análise de espectroscopia

Com intuito de desenvolver uma curva de calibração para proteínas de reserva, inicialmente foram selecionadas 105 amostras para calibração e 26 para validação. Todas as amostras foram previamente quantificadas pelo método de Kjeldahl e os resultados evidenciaram que existe variação no conteúdo de proteínas de reserva. Entre as amostras de calibração foi observada uma amplitude de 26,41 a 43,63% com média de 35,62% e nas amostras de validação a amplitude foi de 26,86 a 42,81% e média de 34,93% (Tabela 1).

Posteriormente, os farelos de todas as amostras foram utilizados para as leituras no NIR e assim obter os espectros brutos (Figura 1a) e os espectros após a correção da linha de base pelo tratamento matemático da primeira derivada (Figura 1b) aplicada na faixa espectral de 9995,0 a 4025,0 $^{cm-1}$. Ao verificar a figura 1b é possível observar maiores diferenças espectrais dentro da faixa de 7000 $^{cm-1}$ a 5000 $^{cm-1}$.

Tabela 1. Estatística descritiva do conteúdo de proteínas de reserva em amostras de soja utilizadas na calibração e validação externa obtidas pelo método de referência (Kjeldahl).

	Nº. de amostras	**Nº. de outliers**	**Mín. (%)**	**Máx. (%)**	**Média (%)**	**Desvio padrão**
Cal.	105	13	26,41	43,63	35,62	3,01
Val.	26	-	26,86	42,81	34,93	2,87

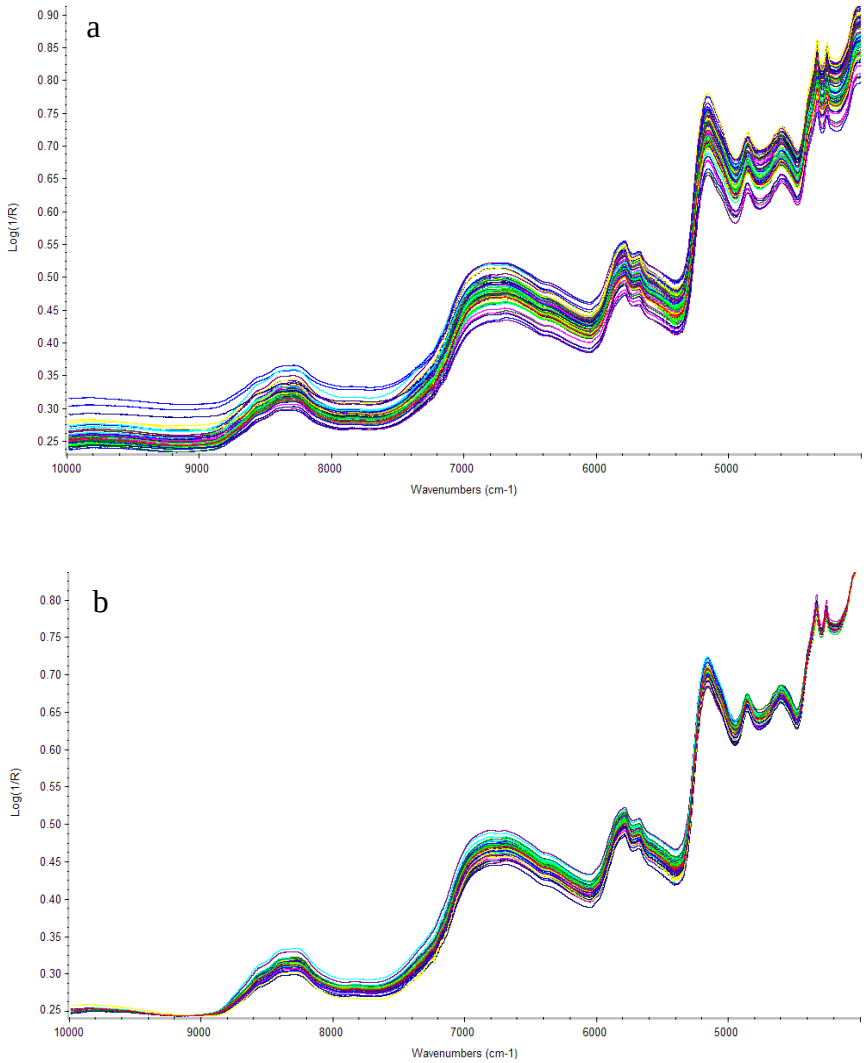

Figura 1. Espectros NIR originais do farelo de soja: (a) espectros brutos e (b) espectros corrigidos na linha de base obtidos a partir dos espectros brutos pelo tratamento matemático da primeira derivada.

3.2. Desenvolvimento e eficiência preditiva da equação de calibração

Após aquisição dos espectros procedeu-se análises de variáveis para predizer a eficiência do modelo desenvolvido pelo método de regressão PLS. Os coeficientes de determinação da calibração e da validação externa foram de 0,965 e 0,976, respectivamente (Tabela 2).

O procedimento de calibração do modelo apresentou erro padrão (EPC) de 0,293%, resultado inferior em comparação com as validações cruzada (0,726%) e externa, com erro padrão da predição (EPP) de 0,562%. Os determinantes preditivos, tanto da validação cruzada quanto externa, foram considerados elevados, sendo de 4,146 e 5,099 para as respectivas validações (Tabela 2).

A fim de observar o padrão de distribuição das amostras relacionando os valores obtidos por NIR e os valores de referência foram gerados gráficos de dispersão para as amostras de calibração (Figura 2a) e de validação externa do método (Figura 2b). Em ambos os gráficos é possível verificar uma distribuição contínua das concentrações de proteínas de reserva, com cobertura de pontos próximos o suficiente para a saturação dos valores dentro da amplitude do conteúdo de proteínas de reserva encontrada. Nas amostras de validação externa, a qual foi realizada para verificar a precisão do método, os valores preditos e os obtidos por NIR não apresentam valores discrepantes, uma vez que os pontos se encontram em cima ou muito próximos da linha de tendência, caracterizando nenhuma ou pequena variação entre os valores preditos e estimados.

Tabela 2. Estatísticas das equações de calibração, validação cruzada e validação externa para proteínas de reserva.

Calibração		**Validação cruzada**		**Validação externa**		
R^2	**EPC(%)**	**EPVC(%)**	**DPRc**	**R^2**	**EPP(%)**	**DPRv**
0,965	0,293	0,726	4,146	0,976	0,562	5,099

Análise de regressão PLS: R^2 - coeficiente de determinação; EPC – Erro padrão da calibração; EPVC – Erro padrão da validação cruzada; DPRc – Determinante preditivo da validação cruzada; EPP – Erro padrão da predição; DPRv – Determinante preditivo da validação externa.

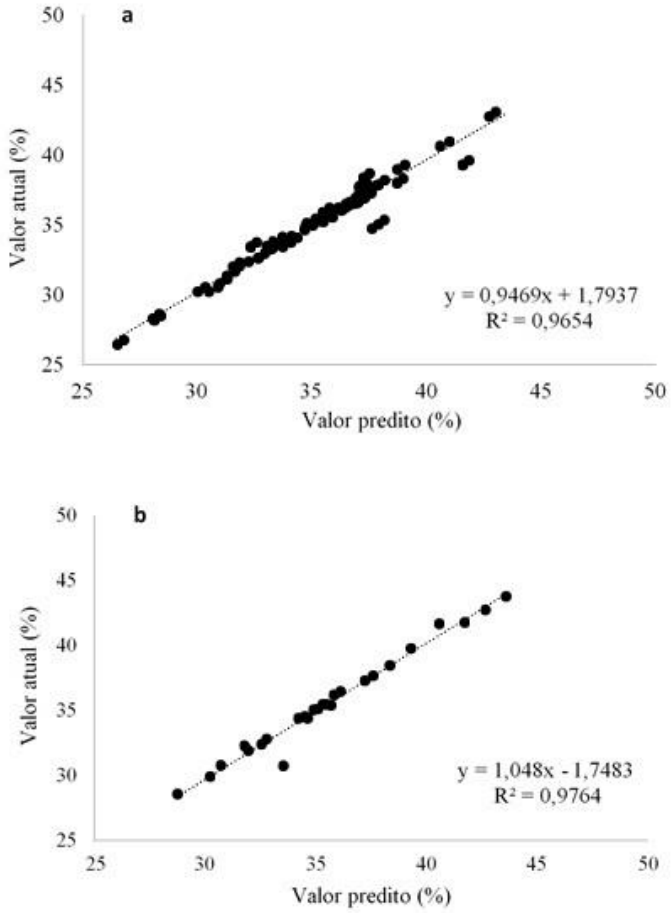

Figura 2. Gráficos de dispersão dos valores obtidos por NIR (valor atual) vs. valores de referência (preditos) de proteínas de reserva no conjunto de amostras de farelo de soja para calibração (a) e para validação externa do método (b).

4. DISCUSSÃO

4.1. Análise espectral

O conteúdo de proteínas de reserva das 105 amostras obtido pelo método de referência – Kjeldahl (Tabela 1) variou de 26,41 a 43,63%. Essa amplitude foi suficiente para inferir que é possível o desenvolvimento de modelos preditivos a partir da espectroscopia por NIR.

Os espectros brutos obtidos por NIR (Figura 1a) apresentam diferença na linha de base entre as amostras, essa característica pode levar a sobreposições e valores sub ou superestimados. Para minimizar os erros devem ser utilizados tratamentos matemáticos, como a primeira derivada, que permite separar sobreposições e eliminar backgrounds causados pela presença de outros constituintes na amostra (CEN; HE, 2007; VIDOTTI; ROLLEMBERG, 2006).

Na figura 1b, são observados os espectros corrigidos pela primeira derivada. Com a correção, as linhas de base foram suavizadas e acentuaram a resposta espectral à composição química, e**,** dessa forma permitiu inferir que a variação encontrada entre as amostras é devido à diferença do conteúdo de proteínas de reserva presentes em cada amostra. Quanto maior a absorção, maior valor de proteínas de reserva foi observado. As diferenças na absorção espectral entre as amostras, são devido à absorção pelos grupos funcionais (CH, CH_2 e NH) presentes na estrutura das proteínas (XU et al., 2020)**.**

4.2. Desenvolvimento, desempenho e acurácia da equação de calibração

O método de regressão PLS foi utilizado no desenvolvimento do modelo de calibração por ser uma técnica de decomposição espectral que utiliza informações de concentração durante o processo de decomposição (BAIANU; GUO, 2011). Como os conjuntos estão relacionados por uma regressão, é possível construir o modelo de calibração que posteriormente será validado para classificar amostras desconhecidas em uma aplicação real.

Os parâmetros de validação e precisão do método para os conjuntos de dados de calibração, validação cruzada e validação externa usando PLS estão resumidos na Tabela 2 e a correlação dos valores do método de referência e o estimado por NIR na figura 2**.**

A melhor equação prevista para quantificar conteúdo de proteína de reserva foi selecionada embasada no coeficiente de determinação (R^2) e na minimização do erro padrão de validação cruzada (EPCV) (WINDHAM; MERTENS; BARTON II, 1989). O valor R^2 da calibração foi de 0,9654 (tabela 2), indicando alto grau de ajuste dos valores preditos e estimados (Figura 2). Os resultados corroboram com os de diferentes autores que desenvolveram eficientes modelos de calibração para quantificação de outros constituintes presentes em sementes de soja por NIR. Como, por exemplo, para proteína total, foram descritos R^2 de 0,810 (FERREIRA; PALLONE; POPPI, 2013), 0,886 (INGLE et al., 2016) e 0,942 (ZHU et al., 2018).

Para avaliar o desempenho preditivo do modelo, deve ser realizado o método interno de validação cruzada e então verificar a acurácia do modelo pelos parâmetros de EPVC e DPRc. Quanto menor o EPVc, maior o DPRc e melhor o desempenho preditivo do método (ZHU et al., 2018). Os valores encontrados para EPVc e DPRc foram de 0,726 e 4,146 respectivamente (Tabela 2). A capacidade preditiva (DPR) do modelo

desenvolvido neste trabalho foi considerada satisfatória**,** uma vez que eficazes modelos de calibração são aqueles que apresentam valores de DPRc acima de 2,4, entre 1,5 e 2,4 são eficazes e valores abaixo de 1,5 não são adequados (BOTELHO; MENDES; SENA, 2013). Patil et al. 2010 ao desenvolver modelos para frações de ácidos graxos em soja encontraram EPVc com valores de 0,08; 0,21; 0,29; 1,05 e 1,10 e DPRc de 2,12; 2,34; 2,89; 3,14 e 3,20 e concluíram que todos os modelos foram eficientes para estimar o conteúdo de ácidos graxos.

4.3. Validação externa do modelo

O conjunto de amostras de validação foi representado por amostras com valores variando de 26,86 a 42,81% com desvio padrão de 2,87 (Tabela 1). A amplitude dos valores é fundamental para garantir estimativa eficiente do método. Modelos que utilizam amostras com baixa amplitude de resultados não satisfazem os parâmetros mínimos de sensibilidade, pois fornecem baixos valores de R^2 e DPR (PATIL et al., 2010).

Baseado nos parâmetros de R^2, EPP e DPRv (Tabela 2) e no gráfico de dispersão com as correlações entre valor de referência e estimados (Figura 2b)**,** o modelo previu com precisão o conteúdo de proteínas de reserva. O valor de R^2 foi de 0,9764, acima dos descritos por Zhu et al., (2018), Patil et al., (2010) e Xu et al., (2020) com valores de 0,958, 0,89 e 0,830, respectivamente. Modelos com R^2 próximo de 1 indicam alto grau de ajuste na regressão dos valores de referência e os estimados (ZHU et al., 2018). O EPP foi de 0,5628. Esse valor é considerado baixo, e por ter sido menor que o erro padrão da calibração é confirmada a robustez da estimativa pelo método criado. O DPRv encontrado foi de 5,099, valor bem acima do mínimo considerado adequado (BOTELHO; MENDES; SENA, 2013).

Após análise abrangente dos resultados da validação externa, o desempenho preditivo do modelo de calibração usando o conjunto de validação externa pôde ser validado com garantia da capacidade preditiva.

5. CONCLUSÕES

A calibração de modelos preditivos pelo método de regressão PLS é eficiente para propiciar estimativas de forma acurada. A partir da criação do modelo é possível utilizar a espectroscopia NIR para quantificar conteúdo de proteínas de reserva de soja

em um elevado número de amostras, reduzindo tempo e custos dentro de programas de melhoramento.

6. REFERÊNCIAS

BAIANU, I.; GUO, J. NIR Calibrations for Soybean Seeds and Soy Food Composition Analysis: Total Carbohydrates, Oil, Proteins and Water Contents. **Nature Precedings**, 16 nov. 2011.

BANDILLO, N. et al. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. **Plant Genome**, v. 8, n. 3, 1 nov. 2015.

BOEHM, J. D. et al. Genetic mapping and validation of the loci controlling 7S α′ and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. **Theoretical and Applied Genetics**, v. 131, n. 3, p. 659–671, 1 mar. 2018.

BOTELHO, B. G.; MENDES, B. A. P.; SENA, M. M. Implementação de um método robusto para o controle fiscal de umidade em queijo minas artesanal. Abordagem metrológica multivariada. **Quimica Nova**, v. 36, n. 9, p. 1416–1422, 2013.

CEN, H.; HE, Y. Theory and application of near infrared reflectance spectroscopy in determination of food quality. **Trends in Food Science and Technology**, v. 18, n. 2, p. 72–83, fev. 2007.

FERREIRA, D. S.; PALLONE, J. A. L.; POPPI, R. J. Fourier transform near-infrared spectroscopy (FT-NIRS) application to estimate Brazilian soybean [*Glycine max* (L.) Merril] composition. **Food Research International**, v. 51, n. 1, p. 53–58, abr. 2013.

GOMEZ, K. A.; GOMEZ, A. A. **Statistical procedures for agricultural research**. 2. ed. New York: John Wiley and Sons Ltd, 1984.

INGLE, P. D. et al. Determination of protein content by NIR spectroscopy in protein powder mix products. **Journal of AOAC International**, v. 99, n. 2, p. 360–363, 1 mar. 2016.

KRISHNAN, H. B. et al. Identification of glycinin and β-conglycinin subunits that contribute to the increased protein content of high-protein soybean lines. **Journal of Agricultural and Food Chemistry**, v. 55, n. 5, p. 1839–1845, 7 mar. 2007.

KUSUMANINGRUM, D. et al. Non-destructive technique for determining the viability of soybean (*Glycine max*) seeds using FT-NIR spectroscopy. **Journal of the Science of Food and Agriculture**, v. 98, n. 5, p. 1734–1742, 30 mar. 2018.

LU, W. et al. Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. **Theoretical and Applied Genetics**, v. 126, n. 2, p. 425–433, 2013.

PANTALONE, V. R. Modern breeding approaches for enhancing soybean protein quality. In: WILSON, R. F. (Ed.). . **Designing Soybeans for 21st Century Markets**. 1. ed. Illinois: Boulder, Urbana, IL, 2012. p. 189–218.

PATIL, A. G. et al. Nondestructive estimation of fatty acid composition in soybean [*Glycine max* (L.) Merrill] seeds using Near-Infrared Transmittance Spectroscopy. **Food Chemistry**, v. 120, n. 4, p. 1210–1217, 15 jun. 2010.

PATIL, G. et al. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. **Theoretical and Applied Genetics**, v. 130, n. 10, p. 1975–1991, 2017.

SOARES, T. C. B. et al. Quantitative genetic analysis of storage proteins in soybean. **Cropp Breeding and Applied Biotechnology**, v. 4, n. 3, p. 317–324, 30 set. 2004.

VIDOTTI, E. C.; ROLLEMBERG, M. D. C. E. Espectrofotometria derivativa: Uma estratégia simples para a determinação simultânea de corantes em alimentos. **Quimica Nova**, v. 29, n. 2, p. 230–233, mar. 2006.

WANG, J. et al. A Dominant Locus, qBSC-1, Controls β Subunit Content of Seed Storage Protein in Soybean (*Glycine max* (L.) Merri.). **Journal of Integrative Agriculture**, v. 13, n. 9, p. 1854–1864, 1 set. 2014.

WANG, W.; PALIWAL, J. Near-infrared spectroscopy and imaging in food quality and safety. **Sensing and Instrumentation for Food Quality and Safety**, v. 1, n. 4, p. 193–207, dez. 2007.

WILLIAMS, P. **Near-infrared technology: in the agricultural and food industries**. 2. ed. Minessota: Amer Assn of Cereal Chemists, 2001.

WINDHAM, W. R.; MERTENS, D. R.; BARTON II, F. E. Protocol for NIRS calibration: Sample selection and equation development and validation. In: MARTEN,

G. C. (Ed.). . **Near infrared reflectance spectroscopy (NIRS): Analysis of forage quality**. 1. ed. Washington: USDA-ARS Agriculture Handbook, 1989. p. 96–103.

XU, R. et al. Use of near-infrared spectroscopy for the rapid evaluation of soybean [*Glycine max* (L.) Merri.] water soluble protein content. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 224, 5 jan. 2020.

ZHU, Z. et al. Determination of soybean routine quality parameters using near-infrared spectroscopy. **Food Science and Nutrition**, v. 6, n. 4, p. 1109–1118, 1 jun. 2018.

Capítulo 2

IDENTIFICAÇÃO E VALIDAÇÃO DE *LOCI* ASSOCIADOS AO CONTEÚDO DE PROTEÍNA TOTAL E DE RESERVA EM SOJA [*Glycine max* (L.) MERRIL]

RESUMO

LORENZONI, Rodrigo Monte. Universidade Federal do Espírito Santo. Fevereiro de 2020. **Identificação e validação de *loci* associados ao conteúdo de proteína total e de reserva em soja [*Glycine max* (L.) Merr.]**. Orientadora: Taís Cristina Bastos Soares. Coorientador: Maximiller Dal-Bianco Lamas Costa.

As proteínas de reserva correspondem a aproximadamente 70% do conteúdo total de proteínas nas sementes de soja e as proporções entre as suas principais frações (7S e 11S) influenciam a qualidade nutricional. Essa característica afeta a indústria de alimentos a base de soja voltada para consumo humano e também a utilização do farelo na ração animal. Em virtude dessa dinâmica entre as proteínas de reserva vêm sendo desenvolvidos programas de melhoramento com intuito de modular a qualidade nutricional da soja. A fim de contribuir com a identificação de marcadores úteis para seleção assistida, objetivou-se com este trabalho estimar o efeito de marcadores SNPs em uma população de RILs derivada do cruzamento entre dois genótipos elite pré-selecionados do 'Programa de Melhoramento para Qualidade da Soja' do BIOAGRO-UFV (PMQS-80 e PMQS-12) voltado para aumento do conteúdo de proteína e qualidade nutricional, e identificar genes candidatos presentes na região proximal ao marcador validado e que apresentem efeito sobre as características de interesse. Para isto, 271 RILs derivadas de F5 foram testadas em ensaios de campo em Capinópolis e Viçosa, nas safras 2017/18 e 2018/19. Esse material foi analisado para as características: conteúdo de proteína total em sementes de soja, conteúdo das proteínas de reserva 7S e 11S e de suas respectivas subunidades. Posteriormente os indivíduos foram genotipados e realizada análise associativa entre os caracteres morfológicos e moleculares. O conteúdo médio de proteína total nos ambientes de Capinópolis foram de 48,947 e 47,651%, sendo maiores que os observados em Viçosa**,** os quais foram 45,078 e 45,629%. O mesmo padrão foi observado para proteínas de reserva com 40,225 e 38,019% em Capinópolis e 33,091 e 31,010% em Viçosa. Percentualmente, a relação PTN_Res/PT foi maior em Capinópolis do que Viçosa. Foram observadas herdabilidades acima de 70% para PT, enquanto para PTN_Res os valores ficaram entre 50 e 60%. Foram verificadas as correlações entre PT e as proteínas de reserva e suas subunidades. PT apresentou correlação positiva com todas as variáveis, enquanto as subunidades da fração 7S apresentaram correlação positiva entre elas e negativa com as subunidades da fração 11S, o inverso também foi verificado. A correlação entre as proteínas 7S e 11S foi próxima de 0, mas ambas apresentaram correlações positivas com o somatório 7S+11S. Na análise associativa o SNP 190 foi

associado nos quatro ambientes para as variáveis PT, PTN_Res, subunidades ácidas e básicas, proteína 11S e 7S+11S e apresentou a explicação máxima de 31,042% da variação fenotípica no cultivo de CAP 2, assim como os maiores efeitos aditivos tanto para PT e PTN_Res. Por fim, para buscar DEGs relacionados ao acúmulo de proteínas nas sementes de soja foi realizado mapeamento das reads dos genitores PMQS80 e PMQS12 contra o genoma de referência e então selecionados quatro DEGs dentro da região próxima ao SNP190, entre eles o Glyma.20g084500 que está relacionado ao desenvolvimento do endosperma e, portanto pode estar associado ao acúmulo de proteínas em sementes de soja. Diante de todos os resultados observados, o SNP190 é um potencial marcador para ser utilizado na seleção assistida em programas de melhoramento de qualidade da soja, uma vez que ele apresenta estabilidade e, o conteúdo proteico e a composição subunitária das proteínas de reserva são fortemente influenciados por fatores ambientais e efeitos pleiotróficos.

ABSTRACT

LORENZONI, Rodrigo Monte. Universidade Federal do Espírito Santo, February 2020. **Identification and validation of loci associated with total and storage protein content in soybean [*Glycine max* (L.) Merr.].** Advisor: Taís Cristina Bastos Soares. Co-adivisor: Maximiller Dal-Bianco Lamas Costa

Storage proteins correspond to approximately 70% of the total protein content in soybean seeds and the proportions between its main fractions (7S and 11S) influences nutritional quality. This characteristic affects the soy-based food industry for human consumption and also the use of bran in animal feed. Due to this dynamic among the stotage proteins, breeding programs has been developed in order to modulate the nutritional quality of soybeans. In order to contribute to the identification of useful markers for assisted selection, the objective of this study was to estimate the effect of SNPs markers in a population of RILs derived from the crossing between two elite pre-selected genotypes of the ‘breeding program for soybean quality' from BIOAGRO-UFV (PMQS-80 and PMQS-12), aimed at increasing protein content and nutritional quality, and identifying candidate genes present in the region proximal to the validated marker and that have an effect on the characteristics of interest. For this, 271 F5-derived RILs were tested in field trials in Capinópolis and Viçosa-MG (2017/18 and 2018/19). This material was analyzed for characteristics: total protein content in soybean seeds, content

of storage proteins, 7S and 11S, and their respective subunits. Subsequently, the individuals were genotyped and an associative analysis was carried out between the morphological and molecular characters. The average total protein content in the Capinópolis environments was 48.947 and 47.651%, being higher than those observed in Viçosa, which were 45.078 and 45.629%. The same pattern was observed for reserve proteins with 40.225 and 38.019% in Capinópolis and 33.091 and 31.010% in Viçosa. As a percentage, the PTN_Res / PT ratio was higher in Capinópolis than Viçosa. Heritabilities above 70% were observed for PT, while for PTN_Res the values were between 50 and 60%. Correlations between PT and the storage proteins and their subunits were verified. PT showed a positive correlation with all variables, while the subunits of the 7S fraction showed a positive correlation between them and negative with the subunits of the 11S fraction, the reverse was also verified. The correlation between 7S and 11S proteins was close to 0, but both showed positive correlations with the sum of 7S + 11S. In the associative analysis, SNP 190 was associated in the four environments for the variables PT, PTN_Res, acidic and basic subunits, 11S and 7S + 11S protein and presented the maximum explanation of 31.042% of the phenotypic variation in CAP 2 cultivation, as well as the largest additive effects for both PT and PTN_Res. Finally, to search for DEGs related to protein accumulation in soybean seeds, the readings of the PMQS80 and PMQS12 readings were mapped against the reference genome and then four DEGs were selected within the region close to the SNP190, among them the Glyma.20g084500 which is related to the development of the endosperm and, therefore, may be associated with the accumulation of proteins in soybean seeds. In view of all the observed results, SNP190 is a potential marker to be used in assisted selection in soybean breeding programs, since it has stability and, the protein content and the subunit composition of the storage proteins are strongly influenced by environmental factors and pleiotrophic effects.

1. INTRODUÇÃO

A soja [*Glycine max* (L.) Merrill] é uma das culturas agrícolas de maior importância para a economia mundial. Ela destaca-se como uma das principais fontes proteicas presente no farelo utilizado na alimentação animal, assim como uma rica fonte de proteína para humanos por meio do consumo de grãos inteiros e alimentos derivados como tofu, proteína hidrolisada e leite (ZHANG et al., 2015). O consumo crescente de alimentos à base de soja nos últimos anos resultou em maior demanda por produtos

derivados e grãos inteiros de melhor qualidade (YANG et al., 2016). A produção nas diversas regiões produtoras do país aponta para uma safra 2019/2020 de 120,86 milhões de toneladas (CONAB, 2020).

O grão de soja é composto basicamente de proteínas (40%), carboidratos (30%), lipídios (20%) e minerais (10%). Grande parte das proteínas é descrita como proteínas de reserva, as quais foram classificadas de acordo com o coeficiente de sedimentação (S), sendo elas: 2S, 7S, 11S e 15S (BARAC et al., 2004). As frações 11S (glicina) e 7S (β-conglicinina) são as principais unidades que impactam na concentração de proteína na semente de soja e correspondem a aproximadamente 70% das proteínas de reserva totais (YANG et al., 2016). Ambas as subunidades têm um efeito direto na qualidade e quantidade proteica, de modo que o conteúdo, as proporções e a dinâmica da biossíntese das frações 11S e 7S podem variar com a cultivar e o ambiente (PANTALONE, 2012).

Na literatura é relatada a existência de uma correlação negativa entre produtividade e o conteúdo de proteína (PATIL et al., 2017). O melhoramento praticado nas últimas décadas, em que a produtividade foi favorecida, levou à redução acentuada do teor proteico no farelo do grão, criando a necessidade de substituição de cultivares, de modo que a tendência seja contida e, se possível, revertida.

Tem sido demonstrado por alguns trabalhos uma correlação negativa entre as frações 7S e 11S e que o maior acúmulo da subunidade β da β-conglicinina (7S) pode resultar em redução da fração 11S, a qual tem correlação positiva com conteúdo total (KRISHNAN, 2007; WANG et al., 2014). Além da redução do teor proteico total, o acúmulo da fração 7S em detrimento da 11S resulta em diminuição da concentração de aminoácidos sulfurados (cisteína e metionina), promovendo perda da qualidade nutricional dos grãos de soja (KRISHNAN, 2007; BOEHM et al., 2018), sendo, portanto, desejável uma maior relação 11S:7S.

Esforços visando aumentar o conteúdo de proteínas, assim como alterar as proteínas de armazenamento 7S e 11S na soja são impulsionados principalmente pelas indústrias de processamento de alimentos (BOEHM et al., 2018). Para isto, programas de melhoramento que buscam aumento do teor proteico e grãos com qualidade nutricional são fundamentais. Esse fato tem aplicação direta na indústria de derivados da soja, visto que a obtenção de um produto com alto conteúdo e qualidade proteica irá diminuir os custos da cadeia de processamento**,** favorecendo toda a indústria dependente da sojicultura brasileira.

O surgimento de técnicas de genotipagem e de sequenciamento mais acuradas levou a uma diminuição dos custos de genotipagem por *SNPs*, e isto permitiu que estudos para identificação de QTLs (*Quantitative Trait Loci*) e GWAS (*Genome-Wide Association Study*) tornassem mais acessíveis em programas de melhoramento. A utilização de marcadores *SSR* e *SNPs* para identificação de *QTLs* fica evidenciada em diferentes estudos como o de Panthee et al. (2004), que utilizaram uma população de RILs (*Recombinant Inbred Lines*) para mapear os *QTLs* que estavam associados às frações 7S e 11S das proteínas de armazenamento da soja. Teng et al. (2017) fizeram uso de 727 marcadores *SSR*, genotiparam 129 acessos e a análise associativa permitiu mapear oito *QTLs* em sete grupos de ligação diferentes para conteúdo de proteína total nos grãos. Ma et al. (2016) mapearam um total de 35 *QTLs* usando 184 RILs e 221 *SSRs* em quatro ambientes e foram identificadas cinco grandes regiões genômicas, localizadas nos cromossomos 1, 4, 6, 10 e 20, que se sobrepunham com múltiplos *QTLs* para diferentes características. Boehm et al. (2018) utilizaram os *SNPs* polimórficos entre os bulks caracterizados com o Illumina SoySNP50K iSelect BeadChips (SONG et al. 2013) e identificaram *QTLs* para subunidades de proteína de reserva no cromossomo 3 (11S A1), cromossomo 10 (7S α ′ e 11S A4) e cromossomo 13 (11S A3).

Atualmente, são descritos 275 *QTLs* relacionados a proteína total e subunidades de proteína de reserva em soja (SOYBASE, 2020). A maioria dos *QTLs* para conteúdo de proteína foi identificada usando populações F2 e RILs (YESUDAS et al., 2013; WANG et al., 2015; QI et al., 2016). A validação de marcadores associados e o mapeamento de *QTLs* para composição proteica podem fornecer informações sobre relações genéticas e identificar os genes envolvidos no controle do processo (BOEHM et al., 2018). E, dessa forma, elucidar a arquitetura genética intrínseca às complexas famílias de multigenes das subunidades glicina e β-conglicinina.

Diante do exposto, objetivou-se com este estudo estimar o efeito de marcadores *SNPs* em uma população de RILs derivada do cruzamento entre dois genótipos elite pré-selecionados do ‘Programa de Melhoramento para Qualidade da Soja’ do BIOAGRO-UFV (PQMS-80 e PQMS-12) voltado para aumento do conteúdo de proteína e qualidade nutricional, e identificar genes candidatos presentes na região proximal ao marcador validado e que apresentem efeito sobre as características de interesse.

2. MATERIAL E MÉTODOS

2.1. Material biológico

Foram selecionados dois acessos utilizados no Programa de Melhoramento da Qualidade da Soja (BIOAGRO/DBB/UFV) que apresentavam alto conteúdo proteico e então realizado um cruzamento dos genitores PMQS-80 e PMQS-12, 47,007 e 45,780% de conteúdo proteico, respectivamente. A partir dos indivíduos F1, obteve-se 271 sementes que foram conduzidas em casa de vegetação pelo método SSD, da geração F2 até F5. Posteriormente, cada planta foi colhida individualmente e gerou as 271 RILs que foram submetidas aos ensaios de campo.

2.2. 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 e dezembro/2018. Em Capinópolis, o plantio foi realizado na Central de Experimentação, Pesquisa e Extensão do Triângulo Mineiro (CEPET), em novembro/2017 e novembro/2018. Assim, os genótipos foram avaliados em quatro ambientes distintos (VIC 1, VIC 2, CAP 1 e CAP 3). 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. A irrigação via aspersão foi administrada de acordo com a demanda até o momento em que os grãos atingiram o ponto de maturação fisiológica.

2.3. Delineamento experimental

O delineamento utilizado foi o de blocos casualizados com duas repetições. 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 final de 12 plantas por metro. Após a maturidade fisiológica (R8), cada parcela foi colhida separadamente em *bulk* de sementes e amostradas para fenotipagem.

3. Fenotipagem

3.1. Estimativa do conteúdo de proteína total e proteínas de reserva

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 proteínas de reserva 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). Os valores obtidos foram convertidos para percentuais em base seca.

3.2. Análise das concentrações das proteínas 11S e 7S e de suas respectivas subunidades

Para as análises de concentração foram utilizadas 236 amostras (2 genitores + 234 RILs) de cada ambiente avaliado (CAP 1, CAP 2, VIC 1 e VIC 2) com duas repetições.

3.2.1. Extração das proteínas de reserva

As proteínas de reserva glicinina (11S) e β-conglicinina (7S) foram extraídas em duplicatas utilizando tampão fosfato salino (fosfato de sódio 0,05 mol.L^{-1}, pH 7,6; NaCl 0,4 mol.L^{-1}; β-mercaptoetanol 0,28%) (SOARES et al., 2004). Foram utilizados aproximadamente 50 mg de soja moída e 1,5 mL do tampão de extração. As amostras foram submetidas a banho de ultrassom por 45 minutos, após isso foram centrifugada**s** por 15 minutos a 14.000 rpm e o sobrenadante com o extrato proteico transferido para um novo tubo.

3.2.2. Eletroforese em gel SDS-PAGE 1D

A uma alíquota de 5 µL do extrato proteico foram adicionados 15 µL do tampão de amostra 3X (Tris 0,1875 mol.L^{-1}; SDS 6,9%; glicerol 30%; pH 6,8), 10 µL do corante azul de bromofenol 0,05% e 30 µL de β-mercaptoetanol. Antes da aplicação no gel os tubos eram colocados em banho-maria a 100ºC por 3 minutos e 10 µL da solução foram aplicados em gel SDS-PAGE (5% - 12,5%). A corrida eletroforética foi conduzida a 80 Volts até o corante atingir o gel de separação e em seguida a voltagem era elevada para 120 Volts por 3h30min. Terminada a corrida, os géis eram colocados

em solução corante (1,5 g de Coomassie Brilliant Blue G, 90 mL de ácido acético, 450 mL de metanol e 460 mL de água) por 12 horas e depois em solução descorante (75 mL de ácido acético, 250 mL de metanol e 675 mL de água) por 12 horas. Os géis, depois de descorados, foram armazenados em solução de glicerol 10% até o momento da análise por densitometria (SOARES et al., 2004).

Em todos os géis foi utilizada a mesma amostra padrão (M7739 IPRO), aplicada no início e no fim gel, as quais foram quantificadas junto com as demais amostras para corrigir possíveis diferenças entre as extrações e eletroforeses. Para possibilitar a quantificação via densitometria foram utilizadas 4 concentrações de BSA (0,1; 0,2; 0,4 e 0,6 µg). E, como forma de garantir que as bandas analisadas correspondiam às proteínas 7S e 11S e suas respectivas subunidades foi utilizado marcador LMW-SDS Marker Kit (GE Healthcare Life Sciences).

3.2.3. Densitometria e estimativa da concentração

Para proceder a análise densitométrica os géis foram escaneados e as imagens obtidas pelo software ImageMaster (GE Healthcare Life Sciences). A quantificação foi realizada pelo software ImageJ pela diferença entre a densidade ótica observada em cada banda e a densidade ótica do background.

A porcentagem final das subunidades α', α e β da β-conglicinina, e das subunidades ácidas e básicas da glicinina foi obtida segundo Soares et al. (2004).

1. As porcentagens das subunidades foram obtidas de todas as amostras analisadas, inclusive dos padrões de extração:

$$\%\ subunidade = \frac{\%S\ x\ PE}{100}$$

Em que:
% Subunidade = porcentagem da subunidade em questão em 100 mg de proteína;
%S = porcentagem da subunidade presente no gel;
PE = porcentagem de proteína do extrato obtida por espectroscopia NIR.

2. Cada subunidade dos padrões de extração foi corrigida a partir da média dos padrões:

$$E = \varepsilon p - \varepsilon m$$

Em que:

E = Desvio de cada subunidade para cada gel;

εp = % de cada subunidade no padrão de extração para cada gel;

εm = % média de cada subunidade dos padrões de extração.

3. Cada subunidade das proteínas foi obtida a partir de:

$$\%final = \%Subunidade - E$$

Os valores de porcentagem final de cada subunidade foram corrigidos para 70%, o que corresponde ao valor médio esperado da soma das frações 7S e 11S para proteínas de reserva em soja (WANG et al., 2014).

4. Genotipagem

4.1. Extração de DNA

Foram coletadas folhas jovens e completamente expandidas das plantas em estádio V4 cultivadas em Viçosa em janeiro de 2018. O DNA foi extraído de acordo com o protocolo proposto por KING et al. (2014). Após a extração foi verificado a quantidade e qualidade em espectrofotômetro NanoDrop (NanoDrop Technologies, Wilmington, DE, EUA) e em gel de agarose 0,8%, respectivamente.

4.2. Seleção de marcadores

Em projeto realizado anteriormente pelo Programa de Melhoramento da Qualidade da Soja (PMQS/BIOAGRO/DBB/UFV) foram selecionados 269 *SNPs* no banco de dados soybase (*https://www.soybase.org*), presentes em regiões associadas a *QTLs* que estão associados ao aumento do conteúdo de proteína, óleo e resistência a nematoides. Estes marcadores *SNPs* foram selecionados com base em: 1) *SNPs* publicados em estudos de associação genômica ampla (*GWAS*); 2) *SNPs* próximos a regiões de características de interesse e; 3) *SNPs* próximos a microssatélites sabidamente associados com as características mencionadas e já utilizados no programa. Posteriormente, foi realizada genotipagem dos *SNPs* selecionados em 34 genótipos utilizados no PMQS pela empresa LGC Genomics.

Para seleção dos marcadores mais promissores foi realizada análise associativa entre os marcadores e os dados fenotípicos para conteúdo de proteína. Entre os marcadores associados foram selecionados aqueles polimórficos entre os genitores PMQS80 e PMQS12. Os marcadores selecionados são descritos na Tabela 1.

Tabela 1. Marcadores SNPs selecionados e usados na validação.

SNP	Cromossomo	GL	Referência
46	1	D1a	SONG et al., 2013
56	18	G	SONG et al., 2013
62	20	I	SONG et al., 2013
94	8	A2	ZHANG et al., 2015
115	15	E	BANDILLO et al., 2015
190	20	I	BANDILLO et al., 2015

*4.3.*Amplificação

A genotipagem dos 6 *SNPs* seguiu a metodologia KASP (LGC Genomics) e foi realizada no BIOAGRO-UFV em equipamento Applied Biosciences 7500 (AB7500). A reação de amplificação compreendeu 1 ciclo de 94 ºC por 15 minutos; 10 ciclos de 94 ºC por 20 segundos, com gradiente de 61-55 ºC, decaindo 0,6 ºC a cada 60 segundos; 30 ciclos de 94 ºC por 20 segundos e 55 ºC por 60 segundos; e um ciclo de 37 ºC por 60 segundos. Cada reação consistiu em 2,5 µL de DNA a 10 ng. µL-1, 2,5 µL de 2x Master Mix e 0,14 µL de Primer Mix. A discriminação alélica foi executada no programa computacional AB 7500 Software v.2.3.

4.4. Análises genético-estatísticas

4.4.1. Associação marcador x características

O estudo de associação marcador-características foi realizado pela análise de marca simples (SCHUSTER; CRUZ, 2008), por meio de regressão linear.

O valor do coeficiente de determinação da regressão (R^2) é a proporção da variação dos valores fenotípicos explicados pelo marcador. Assim, a contribuição de cada loco é estimada pelo valor de R^2 da análise de regressão.

4.4.2. Estimativa da herdabilidade

A herdabilidade para proteína total (PT); proteínas de reserva (PTN_Res); β-conglicinina (7S) e suas subunidades α', α e β; Glicinina (11S) e suas subunidades Ácidas e Básicas; e o somatório das frações 7S e 11S foi estimada como proposto por Zhang et al. (2015).

em que:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma^2}{r}}$$

σ_g^2 é o componente de variância do efeito aleatório dos genótipos;

σ^2 é o componente de variância associado ao erro;
r é o número de repetições.

4.4.3. Correlações entre variáveis

Foi realizado um estudo de correlação genotípica entre todas as variáveis avaliadas, por meio da estimação do coeficiente de correlação (r) (CRUZ et al., 2012).

4.4.4. Análises Estatísticas

As análises estatísticas foram realizadas com auxílio dos softwares GENES (CRUZ, 2013), GQMOL (CRUZ; SCHUSTER, 2004) e TASSEL (BRADBURY et al., 2007).

4.5. Mapeamento das reads

Trabalhos anteriores desenvolvidos pelo grupo de pesquisa PMQS/UFV realizaram uma comparação do perfil transcricional de três variedades, contendo alto conteúdo de proteína e baixo teor de óleo, com três variedades contendo baixo conteúdo de proteína e alto teor de óleo no estádio de desenvolvimento R5 (SILVA, 2018). Com o intuito de identificar genes diferencialmente expressos dentro de uma possível região que contém um importante QTL relacionado ao aumento do conteúdo de proteínas em sementes de soja, foram analisados os dados de RNASeq para as duas variedades utilizadas naquele estudo (*Sequence Read Archive* (SRA) número SRP155599).

4.6. Análise de expressão diferencial

As *reads* dos genitores PMQS-80 e PMQS-12 foram mapeadas com base no genoma de referências da soja Williams 82 na região do cromossomo 20 utilizando o *software STAR* (DOBIN et al., 2013). A contagem das *reads* mapead**a**s e a análise de expressão diferencial foram realizada**s** com o *software* R a partir das contagens de alinhamento com o genoma de referência. Foi utilizada uma comparação contrastando a linhagem PMQS-80 e PMQS-12. Os genes diferencialmente expressos (DEGs) encontrados na região selecionada foram anotados e investigados nas plataformas *Soybase*, *Phytozome* e PUBMED para a identificação de prováveis genes candidatos a propiciar aumento do conteúdo de proteínas.

5. RESULTADOS

5.1.Características gerais da população

Com intuito de selecionar materiais para o desenvolvimento de estudos associativos entre marcadores moleculares e caracteres fenotípicos, inicialmente foram realizados procedimentos de fenotipagem para avaliação do conteúdo proteico total das sementes de soja e da composição das proteínas de reserva β-conglicinina e glicinina. Na figura 1 é apresentado o padrão eletroforético das proteínas de reserva e na tabela 2 são demonstrados os desempenhos dos genitores e das RILs para conteúdo de PT, PTN_Res, frações 7S e 11S, e suas respectivas subunidades nos quatro ambientes avaliados. Os conteúdos médios de proteína total (PT) e de reserva (PTN_Res) foram maiores nas *RILs* cultivadas em Capinópolis, com média dos dois anos consecutivos de 48,30% para PT e 39,12% para PTN_Res. Nas RILs cultivadas em Viçosa foram observados valores de 45,35% e 32,05%. Para todas as demais variáveis foi observado o mesmo padrão, em que os cultivos em Capinópolis tiveram maiores médias do que os realizados em Viçosa.

Para verificar a composição do conteúdo proteico encontrados nas sementes foi observada a razão entre PTN_Res e PT e (Tabela 2) e foi verificado que tanto os genitores quanto as RILs cultivadas em Capinópolis apresentaram maiores valores de proteínas de reserva em relação aos ambientes de Viçosa. Contudo, ao avaliar a dinâmica entre as principais proteínas de reserva pela relação da fração 11S/7S não foi encontrada diferença estatística entre os ambientes.

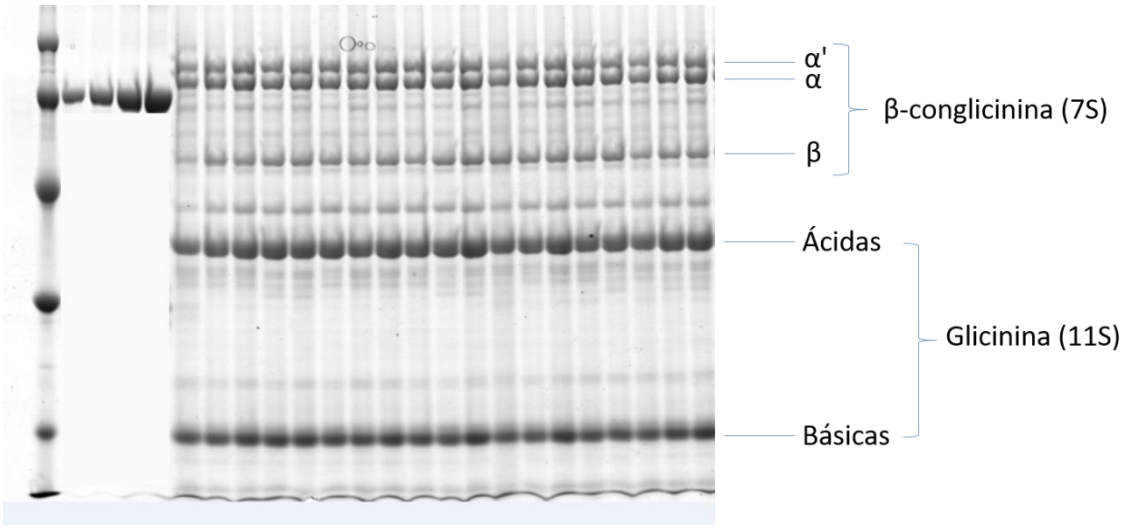

Figura 1. Gel de poliacrilamida desnaturante 12% (SDS-PAGE) de algumas RILs utilizadas neste trabalho. No primeiro poço é observado o padrão LMW-SDS Marker - GE Healthcare Life Sciences, nos quatro poços subsequentes são padrões de BSA (0,1; 0,2; 0,4 e 0,6 µg, respectivamente) e em seguida é verificado o padrão de separação das proteínas de reserva β-conglicinina (7S) e glicinina (11S) em alguns dos genótipos avaliados.

Tabela 2. Médias fenotípicas dos conteúdos de proteína total (PT), proteínas de reserva (PTN_Res), β-conglicinina (7S), glicinina (11S) e suas respectivas subunidades da população de RILs e genitores nos quatro ambientes avaliados.

Amb,	Genótipos	PT (%)	PTN_Res (%)	α' (%)	α (%)	β (%)	Ácidas (%)	Básicas (%)	7S (%)	11S (%)	11S/7S	PTN_Res/PT
CAP 1	BR80	48,768 ±1,276	39,827±1,987	3,580±2,021	2,335±0,783	2,576±0,871	9,097±1,103	10,290±1,976	8,491±2,014	19,388±1,993	2,283	81,666
	NT12	46,676±2,027	37,309±1,874	2,820±2,879	2,280±0,679	2,447±0,582	8,383±1,004	10,186±2,307	7,547±1,967	18,569±2,303	2,460	79,932
	RILs	48,947±2,130	40,225±2,545	3,089±1,069	2,819±0,911	2,786±1,098	8,771±1,421	9,991±2,079	8,693±1,814	18,762±2,480	2,158	82,180
CAP 2	BR80	47,906±2,109	37,376±3,762	3,359±0,998	2,517±1,006	2,107±0,981	7,827±1,361	9,226±1,231	7,983±2,093	17,053±2,322	2,136	78,020
	NT12	47,011±2,653	37,618±1,682	4,178±0,876	2,636±1,005	1,966±1,026	7,662±1,203	9,891±1,532	8,780±1,025	17,552±2,211	1,999	80,020
	RILs	47,651±2,409	38,019±3,217	2,989±1,063	2,719±0,922	2,692±1,128	8,464±1,499	9,629±2,137	8,400±1,954	18,093±2,661	2,154	79,578
VIC 1	BR80	46,470±2,132	33,998±1,788	2,649±0,724	2,049±0,891	2,016±1,013	8,316±1,008	8,769±0,998	6,714±1,112	17,085±2,087	2,545	73,161
	NT12	45,17±2,415	33,232±1,814	3,260±0,658	2,883±0,989	1,363±0,997	7,751±1,105	8,006±2,187	7,505±0,999	15,757±2,166	2,099	73,568
	RILs	45,078±2,793	33,091±3,184	2,703±0,988	2,170±0,967	2,236±1,102	7,479±1,893	8,557±2,380	7,109±2,084	16,036±3,313	2,256	73,408
VIC 2	BR80	46,761±2,224	23,932±2,128	1,715±0,968	1,164±0,937	1,375±1,215	4,911±1,923	7,197±1,782	4,254±2,081	12,108±3,151	2,846	51,179
	NT12	45,947±3,213	34,231±1,814	2,827±0,898	3,081±0,769	2,001±0,892	7,575±1,874	8,479±2,541	7,908±1,786	16,053±3,546	2,030	74,501
	RILs	45,629±3,806	31,010±2,364	2,514±1,110	2,032±1,027	2,075±1,089	6,902±2,160	7,964±2,734	6,621±2,138	14,866±4,012	2,245	67,961

5.2.Parâmetros genéticos e fenotípicos

Pelo teste de normalidade todas as variáveis puderam ser analisadas por meio da distribuição normal, uma vez que todos os caracteres apresentaram distribuição contínua dos dados pelo teste de Lilliefors (p= 0,01).

Ao avaliar todas as variáveis foram observadas diferenças significativas (p<0,01) entre genótipos e na interação genótipo x ambiente (G x A), o que demonstra comportamento diferenciado nos ambientes avaliados. Na seleção é fundamental conhecer se a variação observada é devido a causas genéticas ou ambientais, por isso foram realizadas estimativas das herdabilidades (h^2) e observada a razão CVg/CVe (Tabela 3). Onde foram verificados menores valores de h^2 para todas as variáveis nos cultivos de Viçosa em relação a Capinópolis, sendo que o conteúdo de PT foi a característica que apresentou maior herdabilidade (73 a 84%) e apresentou razão CVg/CVe acima de 1, enquanto todas as demais variáveis apresentaram valores inferiores a 1. Um terceiro parâmetro foi avaliado a fim de garantir a robustez dos resultados e as relações >QMR/<QMR apresentaram valores entre 1,548 (PT) e 2,99 (PTN_Res). A confiabilidade dos resultados pode ser observada pelos coeficientes de variação, os quais ficaram abaixo de 10% para todas as características. O resumo das análises é apresentado na tabela 3.

5.3.Correlação entre as variáveis

No quadro 1 estão demonstradas as correlações entre os caracteres. Proteína total apresentou correlação positiva com todas as variáveis, enquanto as subunidades α', α e β demonstraram correlação positiva com a fração 7S (0,665; 0,728 e 0,564, respectivamente) e negativa com a 11S (-0,020; -0,025 e -0,052, respectivamente). De forma semelhante, as subunidades ácidas e básicas obtiveram correlações positivas com a fração 11S (0,740 e 0,850) e negativas com a 7S (-0.257 e -0.188). As proteínas 7S e 11S apresentaram correlações positivas com o somatório 7S+11S, entretanto foi verificado grande diferença entre os valores estimados de cada uma, onde a correlação foi mais alta para a fração 11S (0,732) do que a 7S (0,397). A correlação entre as proteínas 7S e 11S foi próxima de 0.

Tabela 3. Estimativas de parâmetros genéticos para conteúdo de proteína total, proteína de reserva, β-conglicinina (7S) e glicinina (11S) com suas respectivas subunidades nos quatro ambientes avaliados.

Variável	Ambiente	h^2	>QMR/ <QMR	CVg/ Cve	CV%	F_G	$F_{G \times A}$
PT	CAP 1	0,846	1,548	1,220	4,112	13,367**	9,862**
	CAP 2	0,825					
	VIC 1	0,737					
	VIC 2	0,776					
PTN_Res	CAP 1	0,620	2,990	0,488	5,044	11,022**	6,217**
	CAP 2	0,599					
	VIC 1	0,510					
	VIC 2	0,571					
α'	CAP 1	0,708	1,672	0,482	7,442	2,648**	1,690**
	CAP 2	0,725					
	VIC 1	0,354					
	VIC 2	0,401					
α	CAP 1	0,783	1,836	0,475	7,634	2,468**	1,845**
	CAP 2	0,730					
	VIC 1	0,411					
	VIC 2	0,484					
β	CAP 1	0,790	1,957	0,564	8,118	3,082**	1,834**
	CAP 2	0,805					
	VIC 1	0,327					
	VIC 2	0,436					
Ácidas	CAP 1	0,757	2,576	0,610	6,889	2,709**	2,616**
	CAP 2	0,793					
	VIC 1	0,704					
	VIC 2	0,708					
Básicas	CAP 1	0,604	1,720	0,520	7,249	2,443**	2,252**
	CAP 2	0,701					
	VIC 1	0,714					
	VIC 2	0,660					
7S	CAP 1	0,711	1,900	0,493	7,090	12,451**	5,139**
	CAP 2	0,753					
	VIC 1	0,440					
	VIC 2	0,630					
11S	CAP 1	0,664	2,253	0,613	5,558	2,462**	3,090**
	CAP 2	0,774					
	VIC 1	0,597					
	VIC 2	0,567					

** significativo a p<0,01**.**

h^2 = herdabilidade no sentido amplo;
QMR = quadrado médio do resíduo;
CVg/CVe = relação entre coeficiente de variação genotípica e ambiental;
CV = Coeficiente de variação;
Fg e Fg x a = valores de F para genótipo e interação genótipo x ambiente, respectivamente.

Quadro 1. Estimativas das correlações genotípicas entre proteína total (PT) e as proteínas 7S e 11S com suas respectivas subunidades.

	α'	α	β	Ácidas	Básicas	7S	11S	7S+11S
PT	0,318	0,126	0,071	0,627	0,461	0,263	0,666	0,795
α'		0,450	-0,097	0,416	-0,035	0,665	-0,020	0,385
α			0,077	0,331	-0,294	0,728	-0,025	0,248
β				-0,013	-0,065	0,564	-0,052	0,148
Ácidas					0,273	-0,257	0,740	0,645
Básicas						-0,188	0,850	0,541
7S							0,064	0,397
11S								0,732

5.4. Associação de marcadores moleculares às características avaliadas

5.4.1. SNPs associados

Dentre os seis marcadores associados ao conteúdo de proteína no estudo de *GWAS* (Dados não publicados) e polimórficos entre os genitores selecionados para a genotipagem das RILs (Tabela 1), quatro apresentaram associação com PT em pelo menos um dos ambientes avaliados (Tabela 4).

Dois marcadores merecem destaque, SNPs 62 e 190, ambos encontrados no cromossomo 20. O SNP 190 foi associado nos quatro ambientes para as variáveis PT, PTN_Res, subunidades ácidas e básicas, proteína 11S e 7S+11S e apresentou as maiores variações fenotípicas explicadas pelo marcador, chegando a 31,042% no cultivo de CAP 2, assim como os maiores efeitos aditivos tanto para PT e PTN_Res. Outro marcador de destaque foi o SNP 62 o qual, associou em três ambientes (CAP 1, CAP 3 e VIC 1) para PT e PTN_Res. Ao avaliar as subunidades da β-conglicinina foi verificado associação do SNP 46 para a subunidade α nos ambientes VIC 1 e VIC 2 e O SNP 56 foi associado com a subunidade α' nos ambientes CAP 1, CAP 2 e VIC 1, com variação de 2,416 a 3,212 %, contudo o efeito aditivo foi negativo.

A variação fenotípica explicada pelos marcadores foi maior nos cultivos realizados em Capinópolis. CAP 1 apresentou 34,36% para PT e 13,79% para PTN_Res. No ambiente CAP 2 foram observados valores de 39,56% (PT), 15,11% (PTN_Res), 9,78% (11S) e 13,99% (7S+11S). Enquanto menores valores foram estimados nos cultivos de Viçosa. Ao analisar os dados de VIC 1 foram verificadas variações de 30,60% (PT), 14,65% (PTN_Res), 14,13 (11S) e 15,56 (7S+11S). Apesar de menores em comparação com os cultivos de Capinópolis, o cultivo realizado no ano

de 2017/18 (VIC 1) apresentou maiores valores em relação ao cultivo de 2018/19 em Viçosa (VIC 2), onde foram observados 8,45% (PT), 2,92% (PTN_Res), 6,23% (11S) e 2,92% (7+11S) (Tabela 4).

As subunidades da proteína glicinina apresentaram maiores associações com os marcadores em comparação com as subunidades de β-conglicinina. As subunidades ácidas foram associadas em pelo menos um ambiente com os marcadores 46 e 62, enquanto, com o SNP 190 foram associadas nos quatro ambientes. As subunidades básicas foram associadas em um ambiente com o SNP 62 e em todos com o SNP 190. Ao avaliar as subunidades da β-conglicinina é possível observar que a subunidade α' associou em CAP 1 e VIC 1 com o SNP 56 que apresenta efeito aditivo negativo, a α foi associada em VIC 1 e VIC 2 com o SNP 46 e a β não apresentou nenhuma associação.

Tabela 4. SNPs associados aos conteúdos de proteína total, proteína de reserva, β-conglicinina (7S) e glicinina (11S) com suas respectivas subunidades nos quatro ambientes avaliados, avaliados por meio do mapeamento por marca simples em RILs.

SNP	Cromossomo/ GL	Ambiente	Característica	R^2* (%)	p-Valor	Efeito aditivo	Fonte do alelo favorável
46	1 / D1a	VIC 1	α	3,617	0,0039	0,262	PMQS-12
			Ácidas	1,733	0,0471	0,387	
		VIC 2	α	1,747	0,0462	0,1841	
56	18 / G	CAP 1	α'	2,577	0,0146	-0,253	PMQS-80
		CAP 2	α'	2,416	0,0181	-0,042	
		VIC 1	PT	2,595	0,0142	-0,908	
			α'	3,212	0,0063	-0,242	
			7S	1,744	0,0449	-0,418	
			7S + 11S	2,367	0,0193	-1,289	
62	20 / I	CAP 1	PT	4,741	0,0011	1,068	PMQS-80
			PTN_Res	3,354	0,0021	0,993	
		CAP 2	PT	8,523	0,0011	1,468	
			PTN_Res	3,454	0,0021	0,993	
			Básicas	2,555	0,0174	0,535	
			11S	3,406	0,0059	0,8	
			7S + 11S	2,339	0,0223	0,993	
		VIC 1	PT	3,873	0,0039	1,103	
			PTN_Res	3,385	0,0033	0,972	
			Ácidas	4,431	0,0016	0,638	
			7S	2,099	0,0313	0,465	
			11S	3,313	0,0067	1,024	
			7S + 11S	1,928	0,0391	1,521	
115	15 / E	CAP 1	PT	4,403	0,0014	0,893	PMQS-80
190	20 / I	CAP 1	PT	25,224	0	3,935	PMQS-80
			PTN_Res	10,445	0	2,217	
			Ácidas	1,913	0,0372	0,296	
			Básicas	2,116	0,0281	0,229	
			11S	9,425	0,0002	1,212	
			7S + 11S	10,446	0,0001	1,556	
		CAP 2	PT	31,042	0	2,347	
			PTN_Res	11,656	0,0002	1,561	
			Ácidas	5,202	0,0005	0,547	
			Básicas	2,614	0	0,534	
			11S	6,375	0,0001	1,082	
			7S + 11S	11,656	0,0001	1,572	
		VIC 1	PT	24,136	0	4,26	
			PTN_Res	11,27	0,0002	1,517	
			Ácidas	9,834	0	0,944	
			Básicas	5,796	0,0003	0,927	
			11S	10,824	0	1,872	
			7S + 11S	11,27	0,0004	1,95	
		VIC 2	PT	8,45	0,0013	1,746	
			PTN_Res	2,925	0,0214	0,659	
			Ácidas	4,797	0,0009	0,645	
			Básicas	4,427	0,0014	0,844	
			11S	6,237	0,0001	1,492	
			7S + 11S	2,925	0,0163	1,735	

* Proporção da variância fenotípica explicada pelo marcador.

5.4.2. Combinação de loci *favoráveis*

A fim de verificar a contribuição dos marcadores e seus efeitos nos conteúdos de PT, PTN_Res e das frações 7S e 11S, os SNPs associados foram agrupados com todas as possíveis combinações alélicas encontradas nas RILs (Tabela 5).

Ao analisar os genótipos que continham apenas um marcador, o SNP 190 foi quem mais contribuiu com PT, tendo efeito aditivo de 2,20. Para PTN_Res o mesmo marcador apresentou o segundo melhor ganho (1,21). Quem mais contribuiu para PTN_Res foi o SNP 94 (1,52%), no entanto ele promoveu redução de PT. A proteína 7S teve ganho de 0,712% com o SNP 46, e esse marcador foi quem mais contribuiu negativamente com o conteúdo de PT (-2,4%). A presença do SNP 62 favoreceu PT, PTN_Res e 11S e reduziu os conteúdos de 7S e 7S+11S.

Quando analisados os genótipos que apresentam dois marcadores, a combinação de SNP46 e 190 apresentou maiores ganhos, com efeitos aditivos de 3,04% (PT), 3,95% (PTN_Res), 1,96% (11S) e 1,64% (7S+11S). Nas linhagens com três marcadores, a combinação mais favorável foi a dos SNP 46, 56 e 190, a qual propiciou aumento em todas as características avaliadas, sendo 3,31% (PT), 1,97% (PTN_Res), 0,11% (7S), 1,79% (11S) e 1,90% (7S+11S). Acima de três marcadores, as contribuições apresentam valores próximos ou menores que as demais combinações alélicas.

Dentre todos os marcadores, o SNP 190, além de ter apresentado maior contribuição para aumento do conteúdo proteico total, esteve presente em todas as combinações que favoreciam o aumento das características avaliadas, com exceção da β-conglicinina.

Tabela 5. Comparação dos conteúdos de proteína total, proteína de reserva, β-conglicinina (7S) e glicinina (11S) em grãos de soja em função do número e combinação de loci favoráveis para a característica e seus respectivos valores de efeito aditivo (EA).

SNP	% da Pop.	PT	EA	PTN_Res	EA	7S	EA	11S	EA	7S+11S	EA
46	0.427	42.757	-2.400	32.521	-2.473	8.830	0.712	14.562	-1.823	23.392	-1.111
56	0.427	43.854	-1.303	34.498	-0.497	7.207	-0.911	17.302	0.917	24.509	0.006
62	0.855	45.755	0.599	35.320	0.325	7.409	-0.709	16.769	0.384	24.178	-0.324
94	0.427	44.245	-0.911	36.520	1.526	8.581	0.463	16.115	-0.270	24.695	0.193
115	1.282	45.364	0.208	35.321	0.327	7.220	-0.898	18.231	1.846	25.451	0.948
190	4.701	47.360	2.204	36.207	1.213	8.081	-0.037	17.212	0.827	25.293	0.790
46 e 56	1.282	44.172	-0.984	34.139	-0.855	7.814	-0.304	16.269	-0.116	24.082	-0.421
46 e 62	0.427	44.638	-0.518	34.302	-0.692	7.326	-0.791	15.883	-0.502	23.210	-1.293
46 e 115	2.137	44.592	-0.565	34.628	-0.366	7.593	-0.525	16.561	0.176	24.153	-0.349
46 e 190	0.855	48.196	3.040	38.953	3.959	7.797	-0.321	18.351	1.966	26.148	1.645
56 e 115	0.855	43.453	-1.703	32.788	-2.206	7.734	-0.384	15.745	-0.640	23.479	-1.023
56 e 190	2.991	46.758	1.602	35.810	0.816	7.878	-0.239	17.333	0.948	25.211	0.708
62 e 56	0.855	44.137	-1.019	32.609	-2.385	7.723	-0.395	15.760	-0.625	23.483	-1.020
62 e 94	0.427	42.518	-2.638	31.856	-3.138	6.835	-1.283	15.368	-1.017	22.203	-2.300
62 e 115	2.137	45.095	-0.061	34.155	-0.839	7.397	-0.720	16.329	-0.056	23.726	-0.777
62 e 190	1.282	46.628	1.472	35.915	0.921	7.651	-0.466	16.500	0.115	24.151	-0.352
94 e 56	0.427	45.516	0.360	34.870	-0.125	6.983	-1.134	17.205	0.820	24.188	-0.314
94 e 115	0.427	45.555	0.398	35.685	0.690	7.804	-0.313	16.508	0.123	24.312	-0.191
94 e 190	1.709	46.529	1.372	36.919	1.924	7.579	-0.538	18.024	1.639	25.603	1.101
115 e 190	2.137	46.309	1.153	35.588	0.593	7.871	-0.246	17.193	0.808	25.065	0.562
46, 56 e 62	0.855	44.780	-0.376	34.354	-0.640	7.853	-0.265	16.320	-0.065	24.173	-0.330
46, 56 e 94	0.855	44.015	-1.141	33.755	-1.239	7.882	-0.236	15.784	-0.601	23.665	-0.837
46, 56 e 190	1.709	45.696	0.540	35.263	0.269	7.567	-0.550	15.578	-0.807	23.146	-1.357
46, 115 e 190	1.709	48.469	3.313	36.971	1.977	8.233	0.115	18.178	1.793	26.410	1.908
46, 62 e 94	2.137	44.426	-0.730	33.915	-1.080	8.063	-0.055	16.722	0.337	24.784	0.282
46, 62 e 115	0.855	44.703	-0.454	34.270	-0.724	7.461	-0.657	16.517	0.132	23.978	-0.525
46, 62 e 190	0.855	48.467	3.311	24.891	-10.104	8.092	-0.026	17.424	1.039	25.516	1.013
46, 94 e 115	0.427	42.989	-2.167	30.748	-4.246	7.827	-0.291	14.878	-1.507	22.704	-1.798
46, 94 e 190	2.991	46.396	1.240	36.124	1.130	7.689	-0.429	17.287	0.902	24.976	0.473
56, 62 e 94	2.991	45.022	-0.134	34.820	-0.174	7.488	-0.629	16.433	0.048	23.922	-0.581
56, 62 e 115	0.855	44.113	-1.044	33.526	-1.468	6.944	-1.174	17.045	0.660	23.989	-0.514
56, 62 e 190	0.427	46.133	0.976	33.464	-1.531	6.851	-1.267	16.382	-0.003	23.233	-1.270
56, 94 e 115	0.855	45.045	-0.111	34.951	-0.044	7.646	-0.472	17.008	0.623	24.654	0.151
56, 94 e 190	2.991	45.403	0.247	34.583	-0.411	7.480	-0.637	16.658	0.273	24.139	-0.364
56, 115 e 190	1.282	47.370	2.214	36.966	1.971	7.860	-0.257	17.734	1.349	25.594	1.092
62,94 e 115	1.709	44.170	-0.987	34.036	-0.959	7.584	-0.533	16.737	0.352	24.322	-0.181
94, 115 e 190	2.564	48.005	2.849	36.909	1.915	8.110	-0.007	17.722	1.337	25.833	1.330
46, 56, 62 e 94	0.427	42.721	-2.435	34.915	-0.080	7.440	-0.678	16.383	-0.002	23.823	-0.680
46, 56, 62 e 115	0.855	45.720	0.564	35.221	0.227	7.142	-0.976	17.287	0.902	24.429	-0.074
46, 56, 62 e 190	0.855	46.291	1.135	35.503	0.509	6.963	-1.155	17.011	0.626	23.974	-0.529
46, 56, 94 e 115	0.427	44.543	-0.613	37.386	2.391	8.577	0.459	17.314	0.929	25.891	1.388
46, 56, 94 e 190	2.564	46.252	1.096	35.849	0.854	8.121	0.004	17.369	0.984	25.491	0.988
46, 62, 94 e 115	2.564	45.505	0.349	35.773	0.779	7.819	-0.299	16.655	0.270	24.474	-0.029
46, 62, 94 e 190	0.427	44.774	-0.382	34.585	-0.409	8.320	0.202	16.128	-0.257	24.448	-0.055
46, 94, 115 e 190	2.137	46.643	1.487	35.054	0.060	7.618	-0.500	17.536	1.151	25.154	0.651
56, 62, 94 e 115	2.564	45.283	0.126	33.853	-1.141	7.725	-0.393	16.752	0.367	24.477	-0.025
56, 94, 115 e 190	1.282	47.811	1.654	36.309	1.315	7.363	-0.755	17.496	1.111	24.860	0.357
46, 56, 62, 94 e 115	2.137	45.349	0.193	35.803	0.809	8.213	0.095	16.154	-0.231	24.367	-0.135
46, 56, 62, 94 e 190	0.427	45.492	0.336	34.915	-0.080	7.894	-0.224	16.708	0.323	24.602	0.099
46, 56, 94, 115 e 190	2.137	46.594	1.438	37.683	2.689	7.639	-0.479	18.002	1.617	25.641	1.139
56, 62, 94, 115 e 190	1.709	47.042	1.886	36.630	1.636	7.513	-0.604	17.668	1.283	25.181	0.678
46, 56, 62, 94, 115 e 190	0.855	45.025	-0.132	34.997	0.002	6.956	-1.161	17.599	1.214	24.555	0.052
Heterozigotos	20.513	45.774	0.618	35.104	0.110	7.702	-0.416	16.817	0.432	24.518	0.016
Nenhum	0.427	45.156	-	34.994	-	8.118	-	16.385	-	24.503	-

5.5. Genes diferencialmente expressos no cromossomo 20

Diante da influência do SNP190 no conteúdo de proteínas foi realizado o mapeamento das *reads* dos genitores PMQS80 e PMQS12. Após o mapeamento as *reads* foram confrontadas com o cromossomo 20 do genoma de referência da soja (Williams 82 Assembly 2), fornecendo uma lista de 1862 genes diferencialmente expressos (DEGs- *Differencial Expressed Genes*).

A fim de validar o perfil de expressão e garantir confiabilidade dos mapeamentos foi realizada análise de componentes principais (PCA – *Principal Componente Analysis*) a partir da contagem das *reads* mapeadas, onde as repetições dos genótipos avaliados foram separadas pelo PCA 1 em dois grupos com explicação de 87% da variação (Figura 2). De forma semelhante no gráfico de heatmap (Figura 3), é verificado um perfil de expressão diferencial entre PMQS-80 e PMQS-12 e as repetições biológicas das variedades foram agrupadas da forma esperada para que as análises fossem acuradas e validadas.

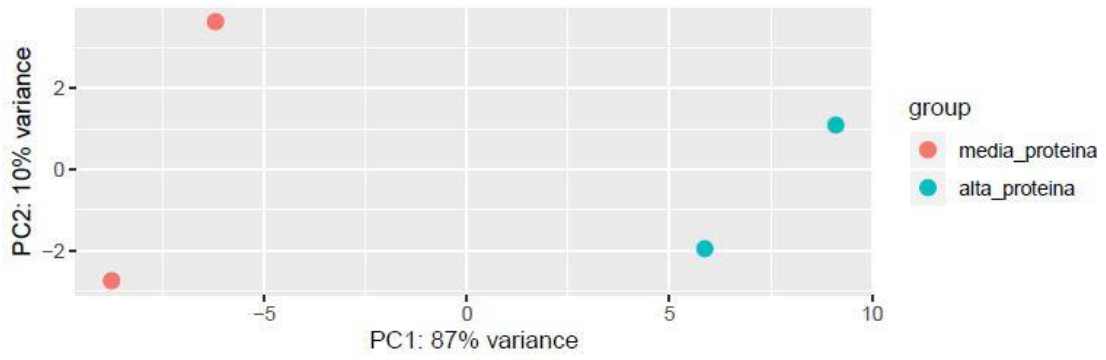

Figura 2. Análise de componentes principais (PCA) dos perfis de expressão das 2 réplicas das variedades PMQS-80 e PMQS-12.

Com a finalidade de propor de forma mais precisa a contribuição da região cromossômica do SNP190 e suas proximidades foram selecionados aqueles localizados em um intervalo de 31.502.249 a 32.282.623 pb no cromossomo 20, que engloba o SNP190. Essa região apresenta 15 DEGs, entretanto, ao levar em consideração o p-valor e selecionar os genes estatisticamente significativos em 5%, foram selecionados quatro genes possivelmente associados com a característica (Tabela 6). Destes, um está up-regulado (Glyma.20g084500)**,** enquanto os demais são genes down-regulados (Tabela 6).

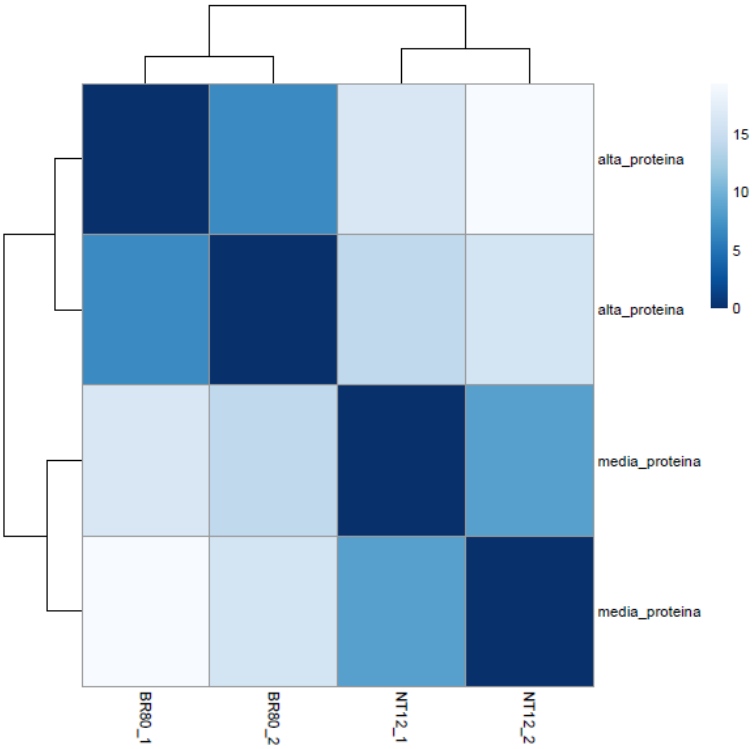

Figura 3. Heatmap com as diferenças no perfil das amostras nas condições de alto (PMQS-80) e médio (PMQS-12) conteúdo proteico.

Tabela 6. Genes diferencialmente expressos na região proximal ao SNP 190.

Genes	log2 FC	FC	p-valor	q-valor	Regulação	Função
Glyma.20g084500	0.21	1.15	0.05	0.19	Up	Splicing mRNA / Complexo de ribonucleoproteínas / Proteínas DWD
Glyma.20g084600	-0.6	0.66	0.05	0.19	Down	Não anotado
Glyma.20g085700	-2.63	0.16	0	0	Down	Atividade catalítica / aldolase tipo TIM barrel
Glyma.20g086200	-0.85	0.56	0.01	0.04	Down	Não anotado

6. DISCUSSÃO

6.1.Parâmetros fenotípicos e genotípicos

Em programas de melhoramento genético é fundamental o conhecimento da estrutura genética das populações e assim determinar o peso de cada caracter avaliado e as metodologias mais apropriadas para seleção de genótipos de interesse (LI et al., 2018). Para que seja possível acessar a estrutura genética, devem ser obtidas estimativas de parâmetros fenotípicos e genéticos das populações a serem avaliadas.

Um dos critérios para utilização de determinada população em estimativa de QTLs é que ela apresente distribuição normal para as variáveis avaliadas. Essa imposição indica a existência de variação contínua e, geralmente, é o padrão esperado para uma herança quantitativa. Em todas as variáveis analisadas foi observad**a** normalidade ($p<0,01$), fato que permitiu proceder com as demais análises de parâmetros fenotípicos e genéticos. Outra variável a ser considerada é o coeficiente de variação (CV), utilizado para indicar a precisão do experimento, para características quantitativas são aceitos valores até 30% (NOGUEIRA et al., 2012). Com os valores de CV obtidos (abaixo de 10%) para as variáveis avaliadas (Tabela 3) é possível inferir que os dados fenotípicos apresentam precisão e homogeneidade. Trabalhos que apresentam maiores CVs que os verificados neste trabalho para PT, PTN_Res, frações 7S, 11S e suas subunidades foram eficientes para identificação de QTLs (MA et al., 2016; SOARES et al., 2004). Da mesma forma, os valores das relações >QMR/<QMR foram considerados baixos, inferiores aos encontrados por (RODRIGUES et al., 2014) e reafirmam a eficiência das RILs utilizadas para estudo de avaliação do conteúdo proteico e a composição das proteínas de reserva.

Os resultados do presente trabalho demonstram que PT é em sua maioridade controlada por fatores genéticos, entretanto PTN_Res e suas subunidades são mais influenciadas por fatores ambientais. Isso pode ser decorrente das complexas famílias multigenes que controlam a dinâmica dessas proteínas com menor efeito e**,** dessa forma o fenótipo é mais influenciado pelo ambiente.

As herdabilidades foram consideradas altas em todas as variáveis, exceto para PTN_Res, ao avaliar os cultivos de Capinópolis (CAP 1 e 2). Enquanto os cultivos em Viçosa (VIC 1 e 2) apresentaram herdabilidade mediana para todas as variáveis, exceto as subunidades ácidas e básicas (Tabela 3). Ao considerar as variáveis, levando em conta a h^2 e a relação entre CVg:CVa, é possível inferir que PT é menos influenciada por fatores ambientais do que o conteúdo de PTN_Res e suas subunidades. Na literatura é descrito a elevada h^2 para PT (LEE et al., 2019; WANG et al., 2014; ZHANG et al., 2015) e valores de baixo a medianos são observados para as frações 7S e 11S e suas respectivas subunidades (MA et al., 2016; PANTHEE et al., 2004). Foram encontrados resultados semelhantes para todas as variáveis ao considerar o ambiente de Viçosa, porém, quando é analisado o ambiente de Capinópolis é verificado que as h^2 são maiores do que as descritas por (MA et al., 2016; PANTHEE et al., 2004). Estes resultados evidenciam a influência ambiental na composição proteica das sementes de

soja. Considerando os ambientes testados, uma das hipóteses é que as diferentes respostas podem ser provenientes da diferença de temperatura entre os mesmos. Como existe correlação positiva entre conteúdo proteico e temperatura (PATIL et al., 2017), o fato de Capinópolis apresentar temperatura média (25,38ºC) maior do que Viçosa (20,6ºC), pode ter interferido no conteúdo total de proteínas e sua composição de proteínas de reserva (Tabela 2).

O conteúdo total de proteínas em sementes de soja é determinado principalmente pelo acúmulo das proteínas de reserva. As estimativas encontradas para essas proteínas (67,96 a 82,18%) estão em consonância com os valores descritos na literatura, onde são observadas médias entre 65 a 80% (KRISHNAN et al., 2007; YANG et al., 2016). Diante dos resultados encontrados, é verificada a influência do ambiente na composição final das proteínas nas sementes. Uma vez que, os cultivos de Capinópolis apresentaram maiores concentrações de PTN_Res, β-conglicinina e glicinina em comparação a Viçosa.

Dentre as proteínas de reserva a β-conglicinina e glicinina correspondem a aproximadamente 70% do total. O conteúdo e as proporções das subunidades que constituem essas proteínas estão intimamente ligados a características das sementes que são determinantes no processamento de alimentos, como as propriedades de formação e emulsificação de gel, que podem vir a interferir na qualidade e rendimento dos alimentos derivados de soja, como, por exemplo o tofu (YANG et al., 2016). Nesse âmbito, as relações, proporções e composição das proteínas de reserva de soja vêm demonstrando sua importância e sendo cada vez mais estudadas (IPPOUSHI et al., 2018; JAMES e YANG, 2016; PANTHEE et al., 2004; YANG et al., 2016; ZHANG et al., 2017; BOEHM et al., 2018).

Um exemplo da importância de se conhecer a composição das proteínas de reserva é a concentração de aminoácidos sulfurados, a fração 11S apresenta melhor qualidade nutricional em detrimento a 7S por possuir maiores níveis dos aminoácidos cisteína e metionina (KRISHNAN et al., 2007; BOEHM et al., 2018). Além da qualidade nutricional, o acúmulo da fração 11S está diretamente ligada ao aumento do conteúdo total de proteínas (KRISHNAN e NESON, 2011). Esses fatores demonstram a importância de se ter maiores valores de relações 11S:7S. Ao observar a relação 11S:7S na tabela 1 percebe-se que em todos os ambientes a relação foi semelhante, ou seja, mesmo nos ambientes de Capinópolis que apresentaram maiores conteúdos de proteínas de reserva em relação a Viçosa, o balanço final das frações permaneceu praticamente

inalterado, com concentração de glicinina mais de duas vezes maior que a de β-conglicinina. Outro problema descrito relacionado a glicinina é que a perda de subunidades da fração 11S, como a subunidade ácida 11SA4, promove aumento de proteínas induzidas por condições de estresse, como superóxido dismutase, SAM e proteínas de choque térmico classe I, o que pode levar a planta a apresentar sintomas de estresse mesmo não estando sob essa condição (YANG et al., 2016).

6.2. Correlações

As correlações genotípicas observadas no quadro 1, corroboram com o esperado na composição subunitária das proteínas de reserva. As subunidades α', α e β apresentaram correlação positiva com a fração 7S e as ácidas e básicas com a fração 11S. A baixa correlação entre a fração 7S com PT e a alta correlação da 11S com PT demonstram que apesar das duas proteínas contribuírem com o maior percentual do conteúdo total de proteínas, a forma como cada subunidade é controlada geneticamente é diferente. LU et al. (2013) ao encontrarem baixas correlações entre proteínas solúveis e proteína total relataram a diferença nos controles genéticos dessas proteínas.

A fração 11S apresenta maiores correlações com o conteúdo de PT do que a fração 7S. Esse resultado confirma a informação de KRISHNAN e NESON (2011)**,** em que quando se tem maiores porcentagens de 11S em relação a 7S é verificado aumento do conteúdo de proteínas totais. A alta correlação do conjunto das proteínas 7S+11S com PT sugere a existência de genes com efeitos pleiotrópicos, em que um gene pode ter efeito simultaneamente em duas ou mais características ou efeitos de ligação gênica. Quando a correlação é positiva, o desenvolvimento de cultivares que combinem o aumento dessas características simultaneamente é facilitado. O contrário é descrito quando se tem correlações negativas, como óleo e proteína, o que dificulta a combinação de altos conteúdos de óleo e proteínas simultaneamente (LI et al., 2018).

6.3. Associação de marcadores moleculares às características avaliadas

Os SNPs 46, 56 e 115 associaram em apenas um ou dois ambientes, portanto, não apresentaram estabilidade e o efeito aditivo desses marcadores foi pequeno. O uso dessas marcas na seleção assistida por marcadores seria útil apenas no desenvolvimento de variedade para ambientes específicos. A interação genótipo x ambiente pode ser uma das explicações para a baixa consistência dos loci identificados. Assim como o fato do conteúdo proteico e sua composição serem de herança quantitativa, onde múltiplos

genes estão envolvidos e cada um apresenta pequeno efeito (CONTRERAS-SOTO et al., 2017).

A estabilidade dos valores de R^2 e efeito aditivo (Tabela 4) ao longo dos ambientes indicam a relevância do uso dos SNPs 62 e 190, sendo ambos encontrados no cromossomo 20. Neste mesmo cromossomo foram descritos outros QTLs que apresentam alto R^2 e valor aditivo, como no estudo de WARRINGTON et al. (2015) em que o QTL qProt_Gm20 contribui cerca de cinco vezes mais que QTLs encontrados em outros cromossomos. É verificado principalmente a associação desses marcadores com PT, PTN_Res, 11S e 7S+11S, indicando o efeito pleiotrópico dos genes próximos a estas marcas, ou que seus *QTLs* estão fortemente ligados (WANG et al., 2014). A variação fenotípica explicada pelos marcadores (Tabela 4) para PT, PTN_Res e 7S+11S sugere que o conteúdo de PT é uma característica complexa, controlada por vários genes, mas os efeitos observados pelos marcadores são maiores em PT do que os observados para o conteúdo de proteínas de reserva e suas subunidades. Isto demonstra que a estrutura genética das proteínas de reserva é mais complexa e seu conteúdo é controlado por vários QTLs de efeito menor. Resultados semelhantes foram encontrados por ZHANG et al. (2017) ao estudarem a arquitetura genética das proteínas solúveis encontradas na semente de soja. Os autores identificaram 18 QTLs associados ao conteúdo de proteína com explicação da variação fenotípica entre 17,4% e 29,2%, enquanto para as proteínas solúveis foram encontrados 10 QTLs com variação entre 7 e 19,3%.

Apesar do elevado número de QTLs localizados no cromossomo 20 relacionados ao conteúdo de proteínas totais descritos na literatura (PANTHEE et al., 2004, LU et al., 2012; PANDURANGAN et al., 2012; MAO et al., 2013; SONG et al., 2013; BANDILLO et al., 2015; WARRINGTON et al., 2015; HACISALIHOGLU et al., 2018), para proteínas de reserva e suas subunidades apenas dois são descritos (PANTHEE et al., 2004; MA et al., 2016). Essa informação demonstra a importância da validação do SNP 190 neste trabalho, pois além de explicar até 31% (CAP 1) da variação encontrada e com efeito aditivo de até 4,26% (VIC 1) para PT, ele está associado com as frações ácidas, básicas da proteína 11S, com o somatório de 7S e 11S e conteúdo total de proteínas de reserva. Portanto, os resultados consistentes de alto efeito aditivo e associação aos caracteres avaliados, e a estabilidade da marca entre ambientes justificam a aplicabilidade desse marcador em programas de melhoramento

que busquem promover aumento de proteínas totais e melhora na qualidade nutricional por meio das proteínas de reserva, em especial a fração 11S.

A fim de verificar os efeitos individuais e combinados dos marcadores, as possíveis combinações alélicas foram determinadas (Tabela 5). As maiores médias fenotípicas foram observadas na combinação entre SNP46, 115 e 190 em que cada marcador é encontrado em um cromossomo diferente e sugere a ocorrência de efeito pleiotrópico. No entanto ao avaliar os SNPs individualmente, o SNP46 apresenta elevado efeito aditivo negativo, sendo observado efeito positivo apenas para a fração 7S. Essa característica do marcador corrobora com o descrito por WANG et al. (2014), em que demonstram que o acúmulo de β-conglicina (7S) reduz o acúmulo da glicinina e tem efeito direto na redução do conteúdo total de proteínas. O SNP115 não apresentou efeito aditivo significativo para PT e PTN_Res. Mas, o SNP 190 foi quem apresentou maiores contribuições de PT e PTN_Res e ainda propiciou efeito aditivo negativo da fração 7S, e aumento da fração 11S. Como é desejável maiores concentrações de 11S, devido **à** sua qualidade nutricional, a inclusão do alelo favorável do progenitor PMQS80 ao SNP 190, por si só, favorece ganhos em quantidade e qualidade da proteína final. A combinação dos marcadores 46 e 190 também apresentou elevados acréscimos da composição de PT e principalmente de proteínas de reserva, nesse caso acredita-se que possa haver interação epistática entre os loci**,** ocasionando modificações de vias metabólicas relacionadas ao armazenamento de proteínas nas sementes.

Apesar da contribuição conjunta dos SNPs 46 e 115 com o 190, a utilização desses marcadores pode não ser eficiente nos programas de seleção assistida, uma vez que eles não apresentam estabilidade entre ambientes e a introgressão de maior número de alelos em uma cultivar é dificultada. Portanto, para fins de SAM, visando o aumento do conteúdo e concomitante aumento de qualidade proteica a utilização de apenas o SNP 190 é a alternativa mais vantajosa.

6.4. Expressão diferencial de genes no cromossomo 20

A identificação de regiões cromossômicas em que haja marcadores associados a características de interesse, é fundamental para predizer locais em que possa existir genes candidatos (ZHANG et al., 2014). A forte associação do SNP190, sugere que em regiões próximas a esse marcador podem existir genes relacionados ao conteúdo proteico e das proteínas de reserva 7S e 11S. Por isso, a análise de expressão diferencial foi realizada no cromossomo 20 e verificados os genes dentro do intervalo de

31.502.249 a 32.282.623 pb, região que compreende o SNP190. Na literatura são descritos quatro genes relacionados a proteínas de reserva no cromossomo 20, sendo dois para a subunidade α (Glyma.20g148400 e Glyam20g28660) e dois para subunidade β (Glyma.20g146200 e Glyma.20g148200) (YAMADA et al., 2014).

A lista de DEGs identificados no contraste entre PMQS80 e PMQS12 foi utilizada para buscar os genes que apresentavam anotação em bancos de dados e verificar a possibilidade de algum dos DEGs encontrados estarem relacionados com conteúdo proteico. Dos quatro genes identificados dentro da região (Tabela 6), Glyma.20g084600 e Glyma.20g086200, não apresentam anotação e, portanto, para verificar se estão relacionados a conteúdo proteico são necessários novos estudos. Uma das abordagens que podem ser utilizadas é o silenciamento gênico com posterior observação do efeito no fenótipo.

Glyma.20g084500 é um gene relacionado com o fator 19 do processamento de pré-mRNA (PRPF19) e faz parte da classe dos genes GmDWD (BIAN et al., 2017). O PRPF19 é um fator essencial no splicing mRNA e tem função de ativação e estabilização estrutural do spliceossomo (MAKAROVA et al., 2004). PRPF19 é um membro da família U-box da ubiquitina-ligase E3 (LEE et al., 2011; MARINO et al., 2013; YEE et al., 2009). Os diversos motivos e domínios das proteínas DWD sugerem diversidade funcional, e é consistente com indícios de que o complexo de ubiquitina E3 regula diversos processos, como processamento de RNA, montagem e degradação de proteínas, transdução de sinal, regulação epigenética e progressão do ciclo celular (BIAN et al., 2017). É descrito que as proteínas DWD desempenham funções em diversos processos, que incluem a regulação da fotomorfogênese e do tempo de floração (CHEN et al., 2010; CHEN et al., 2006), transdução de sinal (LEE et al., 2011; LEE et al., 2010), modificação de cromatina (PAZHOUHANDEH et al., 2011), resposta ao estresse (KIM et al., 2014; ZHANG et al., 2008), bem como desenvolvimento de gametófitos (DUMBLIAUSKAS et al., 2011), embriões e endospermas (BJERKAN et al., 2012). A influência sobre o desenvolvimento do endosperma pode estar relacionada com a síntese de proteínas de reserva. Em concordância com os dados fenotípicos, em que PMQS-80 apresenta maior conteúdo proteico em relação a PMQS-12, o fato desse gene estar up-regulado, e portanto sua expressão ser maior no genitor PMQS-80, permite inferir que quando ocorre maior expressão desse gene é propiciado um aumento do conteúdo de proteínas em sementes de soja.

O gene Glyma.20g085700 possui atividade catalítica e está relacionado a uma aldolase tipo TIM barrel. As proteínas da família TIM barrel do tipo aldolase são reguladas por estresses ambientais e em soja sob estresse hídrico foi verificado um aumento de sua atividade (KOMATSU et al., 2010). Uma das proteínas dessa classe é a anidrase carbônica, enzima que catalisa a hidratação reversível do CO_2, importante componente da maioria dos tecidos vegetais superiores (DIMARIO et al., 2017).

Diante da complexidade dos mecanismos genéticos e moleculares que regulam características quantitativas, a falta de informação relacionada a dois dos genes encontrados e as anotações já realizadas para os genes Glyma.20g084500 e Glyma.20g085700, são necessários novos trabalhos com intuito de verificar e comprovar a associação dos genes aqui descritos com o conteúdo proteico.

7. CONCLUSÕES

O conteúdo proteico e a composição subunitária das proteínas de reserva são fortemente influenciad**o**s por fatores ambientais e efeitos pleiotróficos.

A associação do SNP 190 com as diferentes características nos 4 ambientes, fornece informações imprescindíveis para que possam ser desenvolvidas novas linhagens comerciais com maiores teores proteicos e que apresentem melhoria da qualidade nutricional em um menor tempo, utilizando este marcador na SAM.

Genes diferencialmente expressos na região do SNP190 foram apresentados para maior compreensão dos efeitos observados, porém são necessários novos estudos para confirmar a influência desses genes no conteúdo proteico.

8. REFERÊNCIAS

BANDILLO, N. et al. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. **Plant Genome**, v. 8, n. 3, 1 nov. 2015.

BARAĆ, M. B. et al. Soy protein modification: A review. **Acta periodica technologica**, v. 35, p. 3–16, 2004.

BIAN, S. et al. Genome-wide analysis of DWD proteins in soybean (*Glycine max*): Significance of Gm08DWD and GmMYB176 interaction in isoflavonoid biosynthesis. **PLoS ONE**, v. 12, n. 6, 1 jun. 2017.

BJERKAN, K. N. et al. Arabidopsis WD REPEAT DOMAIN55 interacts with DNA DAMAGED BINDING PROTEIN1 and is required for apical patterning in the Embryo. **Plant Cell**, v. 24, n. 3, p. 1013–1033, 1 mar. 2012.

BOEHM, J. D. et al. Genetic mapping and validation of the loci controlling 7S α′ and 11S A-type storage protein subunits in soybean [*Glycine max* (L.) Merr.]. **Theoretical and Applied Genetics**, v. 131, n. 3, p. 659–671, 1 mar. 2018.

BORÉM, A.; MIRANDA, G. V; FRITSCHE-NETO, R. **Melhoramento de plantas**. 6. ed. Viçosa: Editora UFV, 2013.

BRADBURY, P. J. et al. TASSEL: Software for association mapping of complex traits in diverse samples. **Bioinformatics**, v. 23, n. 19, p. 2633–2635, 1 out. 2007.

CHEN, H. et al. Arabidopsis CULLIN4 forms an E3 ubiquitin ligase with RBX1 and the CDD complex in mediating light control of development. **Plant Cell**, v. 18, n. 8, p. 1991–2004, ago. 2006.

CONAB. **Acompanhamento safra brasileirade grãos**. 7. ed. Brasília: Companhia Nacional de Abastecimento, 2020.

CONTRERAS-SOTO, R. I. et al. A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. **PLOS ONE**, v. 12, n. 2, p. e0171105, 2 fev. 2017.

CRUZ, C. D. GENES - Software para análise de dados em estatística experimental e em genética quantitativa. **Acta Scientiarum - Agronomy**, v. 35, n. 3, p. 271–276, 2013.

CRUZ, C. D.; CARNEIRO, P. C. S.; REGAZZI, A. J. **Modelos Biométricos Aplicados ao Melhoramento Genético**. 1. ed. Viçosa: Editora UFV, 2012.

CRUZ, C. D.; SCHUSTER, I. **GQMOL**. Viçosa:UFV, 2004.

DIMARIO, R. J. et al. Plant carbonic anhydrases: structures, locations, evolution, and physiological roles. **Molecular Plant**, v. 10, n. 1, p. 30–46, 2017.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15–21, jan. 2013.

DUMBLIAUSKAS, E. et al. The Arabidopsis CUL4-DDB1 complex interacts with MSI1 and is required to maintain MEDEA parental imprinting. **EMBO Journal**, v. 30, n. 4, p. 731–743, 16 fev. 2011.

HACISALIHOGLU, G. et al. Quantitative trait loci associated with soybean seed weight and composition under different phosphorus levels. **Journal of Integrative Plant Biology**, v. 60, n. 3, p. 232–241, mar. 2018.

IPPOUSHI, K. et al. Absolute quantification of the α α′ and β subunits of β-conglycinin from soybeans by liquid chromatography/tandem mass spectrometry using stable isotope-labelled peptides. **Food Research International**, v. 116, p. 1223–1228, 1 fev. 2019.

JAMES, A. T.; YANG, A. Interactions of protein content and globulin subunit composition of soybean proteins in relation to tofu gel properties. **Food Chemistry**, v. 194, p. 284–289, 17 ago. 2016.

KIM, S. H. et al. Characterization of a novel DWD protein that participates in heat stress response in Arabidopsis. **Molecules and Cells**, v. 37, n. 11, p. 833–840, 1 nov. 2014.

KING, Z. et al. Non-toxic and efficient DNA extractions for soybean leaf and seed chips for high-throughput and large-scale genotyping. **Biotechnology Letters**, v. 36, n. 9, p. 1875–1879, 2014.

KOMATSU, S. et al. Identification of flooding stress responsible cascades in root and hypocotyl of soybean using proteome analysis. **Amino Acids**, v. 38, n. 3, p. 729–738, mar. 2010.

KRISHNAN, H. B. et al. Identification of Glycinin and β-Conglycinin Subunits that Contribute to the Increased Protein Content of High-Protein Soybean Lines. **Journal of Agricultural and Food Chemistry**, v. 55, n. 5, p. 1839–1845, mar. 2007a.

KRISHNAN, H. B.; NELSON, R. L. Proteomic Analysis of High Protein Soybean (*Glycine max*) Accessions Demonstrates the Contribution of Novel Glycinin Subunits. **Journal of Agricultural and Food Chemistry**, v. 59, n. 6, p. 2432–2439, 23 mar. 2011.

LEE, J. H. et al. DWA1 and DWA2, two Arabidopsis DWD protein components of CUL4-based E3 ligases, act together as negative regulators in ABA signal transduction. **Plant Cell**, v. 22, n. 6, p. 1716–1732, 2010.

LEE, J. H.; TERZAGHI, W.; DENG, X. W. DWA3, an Arabidopsis DWD protein, acts as a negative regulator in ABA signal transduction. **Plant Science**, v. 180, n. 2, p. 352–357, fev. 2011.

LEE, S. et al. Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. **Theoretical and Applied Genetics**, v. 132, n. 6, p. 1639–1659, 1 jun. 2019.

LI, Y. HUI et al. Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. **Plant Science**, v. 266, p. 95–101, 1 jan. 2018.

LU, W. et al. Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. **Theoretical and Applied Genetics**, v. 126, n. 2, p. 425–433, 2013.

MA, Y. et al. Quantitative trait loci (QTL) mapping for glycinin and β-conglycinin contents in soybean (*Glycine max* L. Merr.). **Journal of Agricultural and Food Chemistry**, v. 64, n. 17, p. 3473–3483, 4 maio 2016.

MAKAROVA, O. V. et al. A subset of human 35S U5 proteins, including Prp19, function prior to catalytic step 1 of splicing. **EMBO Journal**, v. 23, n. 12, p. 2381–2391, 16 jun. 2004.

MAO, T. et al. Identification of quantitative trait loci underlying seed protein and oil contents of soybean across multi-genetic backgrounds and environments. **Plant Breeding**, v. 132, n. 6, p. 630–641, 2013.

MARINO, D. et al. Arabidopsis ubiquitin ligase MIEL1 mediates degradation of the transcription factor MYB30 weakening plant defence. **Nature Communications**, v. 4, 2013.

NOGUEIRA, A. P. O. et al. Análise de trilha e correlações entre caracteres em soja cultivada em duas épocas de semeadura. **Bioscience Journal**, v. 28, n. 6, p. 877–888, 2012.

PANDURANGAN, S. et al. Relationship between asparagine metabolism and protein concentration in soybean seed. **Journal of Experimental Botany**, v. 63, n. 8, p. 3173–3184, maio 2012.

PANTALONE, V. R. Modern breeding approaches for enhancing soybean protein quality. In: WILSON, R. F. (Ed.). . **Designing Soybeans for 21st Century Markets**. 1. ed. Illinois: Boulder, Urbana, IL, 2012. p. 189–218.

PANTHEE, D. R. et al. Quantitative trait loci for β-conglycinin (7S) and glycinin (11S) fractions of soybean storage protein. **Journal of the American Oil Chemists' Society**, v. 81, n. 11, p. 1005–1012, nov. 2004.

PATIL, G. et al. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. **Theoretical and Applied Genetics**, v. 130, n. 10, p. 1975–1991, 2017.

PAZHOUHANDEH, M. et al. MSI4/FVE interacts with CUL4-DDB1 and a PRC2-like complex to control epigenetic regulation of flowering time in Arabidopsis. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 8, p. 3430–3435, 22 fev. 2011.

QI, Z. et al. Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. **Molecular Breeding**, v. 36, n. 5, 1 maio 2016.

RODRIGUES, J. I. DA S. et al. Biometric analysis of protein and oil contents of soybean genotypes in different environments. **Pesquisa Agropecuaria Brasileira**, v. 49, n. 6, p. 475–482, 2014.

SCHUSTER, I.; CRUZ, C. D. **Estatística Genômica**. 2. ed. Viçosa: Editora UFV, 2008.

SILVA, L. A. C. **Caracterização do perfil de expressão em variedades de soja contrastantes para o teor de proteína**. Tese. Bioquímica aplicada, Universidade Federal de Viçosa, 2019.

SOARES, T. C. B. et al. Quantitative genetic analysis of storage proteins in soybean. **Cropp Breeding and Applied Biotechnology**, v. 4, n. 3, p. 317–324, 30 set. 2004.

SONG, Q. et al. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. **PLoS ONE**, v. 8, n. 1, 30 jan. 2013.

TENG, W. et al. Identification of quantitative trait loci underlying seed protein content of soybean including main, epistatic, and QTL × environment effects in different regions of Northeast China. **Genome**, v. 60, n. 8, p. 649–655, 2017.

WANG, J. et al. A Dominant Locus, qBSC-1, Controls β Subunit Content of Seed Storage Protein in Soybean (*Glycine max* (L.) Merri.). **Journal of Integrative Agriculture**, v. 13, n. 9, p. 1854–1864, 1 set. 2014.

WANG, J. et al. Identification and mapping of stable QTL for protein content in soybean seeds. **Molecular Breeding**, v. 35, n. 3, p. 92, 28 mar. 2015.

WANG, X. et al. Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. **Molecular genetics and genomics**, v. 289, n. 5, p. 935–949, 2014.

WARRINGTON, C. V. et al. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. **Theoretical and Applied Genetics**, v. 128, n. 5, p. 839–850, 1 maio 2015.

YAMADA, T. et al. Knockdown of the 7S globulin subunits shifts distribution of nitrogen sources to the residual protein fraction in transgenic soybean seeds. **Plant Cell Reports**, v. 33, n. 12, p. 1963–1976, 13 nov. 2014.

YANG, A. et al. Rebalance between 7S and 11S globulins in soybean seeds of differing protein content and 11SA4. **Food Chemistry**, v. 210, p. 148–155, 1 nov. 2016.

YEE, D.; GORING, D. R. The diversity of plant U-box E3 ubiquitin ligases: From upstream activators to downstream target substrates. **Journal of Experimental Botany**, v. 60, n. 4, p. 1109–1121, mar. 2009.

YESUDAS, C. R. et al. Identification of germplasm with stacked QTL underlying seed traits in an inbred soybean population from cultivars Essex and Forrest. **Molecular Breeding**, v. 31, n. 3, p. 693–703, mar. 2013.

ZHANG, D. et al. Use of single nucleotide polymorphisms and haplotypes to identify genomic regions associated with protein content and water-soluble protein content in soybean. **Theoretical and Applied Genetics**, v. 127, n. 9, p. 1905–1915, 1 set. 2014.

ZHANG, D. et al. The genetic architecture of water-soluble protein content and its genetic relationship to total protein content in soybean. **Scientific Reports**, v. 7, n. 1, 1 dez. 2017.

ZHANG, Y. et al. Arabidopsis DDB1-CUL4 associated factor1 forms a nuclear E3 ubiquitin ligase with DDB1 and CUL4 that is involved in multiple plant developmental processes. **Plant Cell**, v. 20, n. 6, p. 1437–1455, jun. 2008.

ZHANG, Y. H. et al. Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr.]. **Theoretical and Applied Genetics**, v. 128, n. 6, p. 1061–1072, 16 jun. 2015.