MAIKON GUERITH BAPTISTELLA DA SILVA

ESTRATÉGIAS DE PREDIÇÃO DE CRUZAMENTOS DE SOJA COM BASE EM INFORMAÇÃO GENÔMICA

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

Orientador: Felipe Lopes da Silva

Coorientadores: Cosme Damião Cruz
Moysés Nascimento
Marcos Deon Vilela de Resende

VIÇOSA - MINAS GERAIS
2022

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

T
S586e
2022

Silva, Maikon Guerith Baptistella da, 1992-
Estratégias de predição de cruzamentos de soja com base em informação genômica / Maikon Guerith Baptistella da Silva. – Viçosa, MG, 2022.
1 tese eletrônica (75 f.): il. (algumas color.).

Orientador: Felipe Lopes da Silva.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Agronomia, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.531
Modo de acesso: World Wide Web.

1. Soja - Melhoramento genético. 2. Predição. 3. Genômica. I. Silva, Felipe Lopes da, 1981-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 633.342

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

MAIKON GUERITH BAPTISTELLA DA SILVA

ESTRATÉGIAS DE PREDIÇÃO DE CRUZAMENTOS DE SOJA COM BASE EM INFORMAÇÃO GENÔMICA

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

APROVADA: 25 de fevereiro de 2022.

Assentimento:

Maikon Guerith Baptistella da Silva
Autor

Felipe Lopes da Silva
Orientador

Aos meus avós in memorian, Matias, Nair e José que sempre torceram por mim.

Aos meus pais, Ademar e Marilene, e avó Maria pelo apoio em tudo, e serem meu exemplo de vida.

A todos os agricultores Brasileiros

DEDICO

AGRADECIMENTOS

Primeiramente, este trabalho não seria possível ser realizado sozinho, ele é a soma do esforço e dedicação de todos os envolvidos.

Agradeço imensamente a Deus, por sempre ter me abençoado e guiado minha vida.

À Universidade Federal de Viçosa (UFV) e ao Programa de Pós-Graduação em Genética e Melhoramento, aos secretários e funcionários do departamento de Genética, Marco Tulio e Odilon.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pela concessão da bolsa de estudos de Doutorado.

À GDM, por apoiar nossa pesquisa na UFV, e disponibilizar as informações necessárias para a realização deste estudo, em especial ao Nizio e Gaspar.

Estendendo a todos os professores da Universidade Federal de Viçosa onde tive a honra de adquirir conhecimento, gostaria de agradecer ao Professor Felipe Lopes da Silva, mais que orientador, e sim um amigo que levarei da UFV para a vida. Muito obrigado pela confiança e amizade ao longo destes anos. Aos coorientadores deste trabalho, Prof. Cosme Damião Cruz, Prof. Moysés Nascimento e Prof. Marcos Deon Vilela de Resende, obrigado por todas as sugestões que sem dúvida enriqueceram muito o trabalho.

A toda minha família, em especial aos meus irmãos Erick e Franciele, e sobrinhos Anthony, Luana e Vitor. À minha namorada, Julia de Araújo Rodrigues do Nascimento, obrigado por ser minha companheira em tudo.

Na UFV conheci pessoas excepcionais de diversas partes do Brasil, assim, agradeço imensamente a todos do Laboratório do Programa Soja, do Grupos de Estudos em Genética e Melhoramento de Plantas (GenMelhor) e aos amigos de todas as disciplinas cursadas. Grandes amigos que a UFV me presenteou e que carregarei ao longo de toda a vida.

Aos meus amigos de infância que sempre mandaram energia positiva.

A todos que me ajudaram direta e indiretamente, meus mais sinceros agradecimentos.

“A educação é a arma mais poderosa que você pode usar para mudar o mundo.”

(Nelson Mandela)

“Faça o seu melhor na condição que você tem, enquanto você não tem condições melhores, para fazer melhor ainda.”

(Mario Sério Cortella)

BIOGRAFIA DO AUTOR

Maikon Guerith Baptistella da Silva, nasceu em 11 de janeiro de 1992 no município de Ivaiporã – Paraná. Em dezembro de 2008 concluiu o ensino médio no Colégio Panamericano no mesmo município. Deu início a graduação em Agronomia pela Universidade Estadual de Maringá – Paraná no ano de 2009 com conclusão no final de 2013. Obteve o título de Mestre em Genética e Biologia Molecular no ano de 2016 pela Universidade Estadual de Londrina – Paraná sob orientação do Prof. Dr. Josué Maldonado Ferreira. Em março de 2016 iniciou o Doutorado em Genética e Melhoramento pela Universidade Federal de Viçosa – Minas Gerais, sob a orientação do Prof. Dr. Felipe Lopes Silva, sendo bolsista da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) por quase dois anos. Desde outubro de 2018 faz parte da equipe de Pesquisa da empresa GDM, atuando como Melhorista de Milho na região Subtropical – Sul do Brasil, com sede na cidade de Cambé-PR.

RESUMO

SILVA, Maikon Guerith Baptistella, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. **Estratégias de predição de cruzamentos de soja com base em informação genômica.** Orientador: Felipe Lopes da Silva. Coorientadores: Cosme Damião Cruz, Moysés Nascimento e Marcos Deon Vilela de Resende.

Utilizar das melhores estratégias para a seleção dos melhores cruzamentos é fundamental para o sucesso no desenvolvimento de cultivares de soja. Atualmente, estratégias como o uso da média dos genitores e a distância genética vem sendo empregadas como ferramenta nas escolhas das combinações. Com o avanço da computação e da genômica, outras metodologias vêm sendo desenvolvidas para auxiliar o melhorista. Porém, de acordo com alguns estudos, a predição da variância genética é mais difícil e de baixas correlações com a distância genética, o que causa um viés na seleção dos genitores a englobar os blocos de cruzamentos. Assim, para esta presente tese, foram desenvolvidos dois capítulos. O primeiro utilizando de simulação computacional, teve como objetivo, validar as relações entre as estimativas de média e variância e distância genética, resultantes da metodologia de seleção de cruzamentos biparentais por meio da predição genômica via progênies simuladas e a metodologia de particionamento das distancias com base nos efeitos dos marcadores genéticos. Foram simulados 300 genitores e 886 marcadores codominantes. A seleção genômica foi aplicada tendo como população de treinamento as informações fenotípicas e genotípica dos parentais, e os efeitos dos marcadores foram divididos em quatro grupos de acordo com as suas magnitudes. Os resultados demonstraram que o uso da distância total não foi totalmente informativo para encontrar populações com alta variabilidade genética. A captação da divergência nas regiões de maiores efeitos, mediante o particionamento dos marcadores, proporcionou um melhor entendimento da variância predita. Para o segundo capítulo, foram utilizados dados reais, e proposto uma nova metodologia de seleção de cruzamentos com base na predição genômica e índice de seleção de postos aplicados na soma dos efeitos dos marcadores particionados. Para isto, foram utilizadas 102 linhagens da macrorregião 3 do Brasil para constituir a população de treinamento. Do total de 5151 combinações possíveis, 10 foram realizadas para a validação do modelo proposto. Os resultados indicaram uma melhor acurácia preditiva da seleção de cruzamentos com base no

índice de seleção, pois caracterizou melhor a complementariedade dos genes entre os parentais.

Palavras-chave: Predição de cruzamentos. Predição de progênies. Simulação. Seleção genômica. Valor genético genômico. *Glycine max.*

ABSTRACT

SILVA, Maikon Guerith Baptistella, M.Sc., Universidade Federal de Viçosa, February, 2022. **Strategies for prediction crosses of soybean based on genomic information.** Adviser: Felipe Lopes da Silva. Co-advisers: Cosme Damião Cruz, Moysés Nascimento and Marcos Deon Vilela de Resende.

Using the best strategies for the selection of the best crosses is essential for the success in the development of soybean cultivars. Currently, strategies such as the use of the average of the parents and the genetic distance have been used as a tool in the choice of combinations. With the advancement of computing and genomics, other methodologies have been developed to help the breeder. However, according to some studies, the prediction of genetic variance is more difficult and of low correlations with genetic distance, which causes a bias in the selection of parents to encompass the blocks of crosses. Thus, for this present thesis, two chapters were developed. The first one, using computer simulation, aimed to validate the relationships between the estimates of mean and variance and genetic distance, resulting from the methodology of selection of biparental crosses through genomic prediction via simulated progenies and the methodology of partitioning the distances based on the effects of genetic markers. 300 parents and 886 codominant markers were simulated. The genomic selection was applied having the phenotypic and genotypic information of the parents as the training population, and the effects of the markers were divided into four groups according to their magnitude. The results showed that the use of total distance was not fully informative to find populations with high genetic variability. The capture of the divergence in the regions of greatest effects, through the partitioning of the markers, provided a better understanding of the predicted variance. For the second chapter, real data were used, and a new methodology for cross selection based on genomic prediction and rank selection index applied to the sum of the effects of partitioned markers was proposed. For this, 102 strains from macroregion 3 of Brazil were used to constitute the training population. Of the total of 5151 possible combinations, 10 were performed to validate the proposed model. The results indicated a better predictive accuracy of the selection of crosses based on the selection index, as it better characterized the complementarity of genes between the parents.

Keywords: Cross prediction. Progeny prediction. Simulation. Genomic selection. Genomic genetic value. *Glycine max*.

LISTA DE ILUSTRAÇÕES

CAPÍTULO 1

CAPÍTULO 2

LISTA DE TABELAS

CAPÍTULO 1

CAPÍTULO 2

SUMÁRIO

INTRODUÇÃO GERAL

Dentre as culturas mais produzidas no mundo, a soja [Glycine Max (L.) Merr.] vem se destacando por ser uma excelente fonte de proteína e óleo, possuindo usos tanto para alimentação humana como animal. No Brasil, desde a expansão da soja da região Sul para o Cerrado, a área cultivada com esta leguminosa vem crescendo a cada ano, alcançando na última safra de verão 2021/2022 um total de aproximadamente 40,3 milhões de hectares, tornando-a, a cultura de maior importância econômica para o país, e proporcionando ao Brasil o maior produtor e exportador mundial (CONAB, 2022; Silva, 2017).

Assim como o avanço de área cultivada, a produtividade nos últimos anos vem aumentando, em que, segundo os primeiros dados da série histórica das safras da Companhia Nacional de Abastecimento (CONAB) da safra 1976/1977 até a safra de 1987/1988 a produtividade de grãos de soja não superou a barreira dos 2000 kg ha^{-1}, sendo alcançado somente na safra de 1991/1992. E quase 20 safras depois, foi superado os 3000 kg ha^{-1}. Para a safra atual a previsão é de aproximadamente 3500 kg ha^{-1} (CONAB, 2021).

Juntamente com as novas tecnologias de manejo agrícola, como por exemplo a implementação do sistema de plantio direto, o melhoramento genético da soja teve e ainda tem um papel fundamental para esse incremento de produtividade, desenvolvendo anos após anos, cultivares de alto potencial produtivo e adaptadas ao ambiente de produção para os produtores de todas as regiões do Brasil.

A produtividade de plantas é um caráter quantitativo, e, portanto, controlado por vários genes e muito influenciado pelo ambiente. Logo, a incorporação e o acúmulo de alelos favoráveis são imprescindíveis para a obtenção de genótipos superiores (Silva, 2017). E para que isso ocorra, a definição dos parentais e posterior formação de populações segregantes não é simples, porém, de extrema importância para o sucesso de um programa de melhoramento genético de soja (Fehr, 1987). A escolha assertiva dos genitores representa, além de econômica de tempo e recursos genéticos, um grande progresso de ganho genético (Silva, 2017: Nass, 2001).

Uma população segregante (ou também definida como população base), de destaque, é aquela que apresenta uma alta média e variância genética para aquela característica de interesse ao melhorista, ou seja, a população possui uma alta frequência de alelos favoráveis e diversidade genética entre os parentais. Para isso,

há algumas metodologias que auxiliam o melhorista a escolher quais os melhores parentais.

Visando principalmente um caráter quantitativo, Baenzinger e Peterson (1991) classificou os métodos de seleção de parentais em duas categorias: a primeira com base nas informações dos pais, e a segunda utilizando o desempenho das progênies geradas. Na primeira categoria, a metodologia mais usual é a utilização da média dos parentais, porém, neste caso, não estamos levando em consideração a complementariedade genética, pois, uma combinação com alta média pode não apresentar uma variabilidade suficiente para a ocorrência de segregação transgressiva, e, portanto, não gerar um genótipo com produtividade superior à dos pais. Outra ferramenta proposta para auxiliar o melhorista é a utilização da diversidade genética entre os parentais por meio de marcadores molecular (Nass, 2001).

Na segunda categoria, utilizando o desempenho das progênies, os cruzamentos dialélicos são os mais usuais, onde são obtidas as informações das capacidades gerais (CGC) e especificas de combinação (CEC). Entretanto, esta metodologia apresenta uma limitação quando o número de genitores é elevado, pois, considerando cruzamentos biparentais com n genitores em um dialelo completo sem recíprocos, o número de populações a serem testadas será de n(n-1)/2, tornando a avaliação das progênies onerosa para o programa. Outras metodologias como Jinks e Pooni (1976) e de Vencovsky (1987) também podem ser empregadas, porém, também demandam de avaliações em populações segregantes, tornando a necessidade de se trabalhar com número limitados de combinações.

Estudos de predição de médias de progênies derivadas de diversas combinações biparentais se iniciaram com os trabalhos de Schnell e Utz (1975) e posteriormente de Zhong e Jannink (2007), mediante a utilização de dados fenotípicos. Porém, a variância genética aditiva entre os genitores era fundamental para acurácia da metodologia proposta. Com o avanço da informática e principalmente em marcadores moleculares, a metodologia da predição genômica ampla (GWS) proposta por Meuwissen (2001) proporcionou a predição de genótipos não fenotipados, tendo como base apenas a sua informação genotípica. E, após o trabalho de Bernardo (2014), a ferramenta de GWS também foi inserida nos estudos de predição de cruzamentos.

A simulação computacional de progênies endogâmicas juntamente com a GWS é uma das metodologias que vem sendo empregada para a definição de

genitores e blocos de cruzamentos, tanto para dados reais como também em dados simulados para estudos genéticos. Com esta ferramenta é possível a predição de milhares de cruzamentos e a obtenção das estimativas das médias e variâncias de cada combinação, conforme trabalhos de Jean et al., (2021), Sant'Anna et al., 2019, Adeyemo e Bernardo (2019), Beckett et al., (2019), Neyhart e Smith, (2019), Osthushenrich et al., (2018), Yao et al., (2018), Tiede et al., (2015) e Mohammadi et al., (2015). Entretanto, nos estudos de Adeyemo e Bernardo (2019), Beckett et al., (2019) e Neyhart e Smith, (2019), foi possível observar uma baixa correlação na validação da predição da variância genética das populações e, baixa correlação entre a distância genética entre os parentais e a variância genética predita. Porém, é importante ressaltar que a distância genética utilizada para a comparação com a variância genética, foi com base em todas as marcas.

Diante do exposto, a presente tese teve como objetivo realizar a predição genômica de cruzamentos biparentais de soja mediante dados fenotípicos e genotípicos simulados e de dados reais, validando as relações entre as estimativas de média e variância resultantes da metodologia de predição genômica via progênies simuladas, e da distância genética entre os genitores. Logo, utilizou-se de uma nova abordagem tendo como base o particionamento da distância genética por meio das informações do efeito de cada marcador e agrupadas conforme a sua magnitude, além também, da seleção de cruzamentos com base em um índice de seleção aplicado após a execução do particionamento dos efeitos dos marcadores e as respectivas somas dentro de cada grupo pré-definido.

REFERÊNCIAS

ADEYEMO, E.; BERNARDO, R. Predicting genetic variance from genomewide marker effects estimated from a diverse panel of maize inbreds. **Crop Science**. 59:583–590. 2019.

BAENZIGER, P. S.; PETERSON, C. Genetic variation: Its origin and use for breeding self-pollinated species. In Plant Breeding in the 1990's, edited by H. Stalker and J. Murphy, pp. 69–100, **North Carolina State University**, North Carolina. 1991.

BECKETT, T. J.; ROCHEFORD, T. R.; MOHAMMADI, M. Re-imagining maize inbred potential: identifying breeding crosses using genetic variance of simulated progeny. Crop Science, vol. 59: 1-12. 2019.

BERNARDO, R. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. **Crop Science**. 54:2586–2595. 2014.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO – **Grãos por Produto**. Disponível em: < https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras>. Acesso em: 19 dez. 2021.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. Acompanhamento da Safra Brasileira de Grãos, Brasília, DF, v. 9, safra 2021/22, n. 4 quarto levantamento, jan. 2022.

FEHR, W. Principles of cultivar development. **Theory and technique**. Macmillan, New York. 1987.

JEAN, M.; COBER, E.; O'DONOUGHUE, L.; RAJCAN, I.; BELZILE, F. Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. **Crop Science**. 61. 2021.

JINKS, J. L.; POONI, H. S. Predicting the properties of recombinant inbreed lines derived by single seed descent. Heredity, Oxford, v. 36, n. 2, p. 243-266. 1976.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157: 1819–1829. 2001.

MOHAMMADI, M.; TIEDE, T.; SMITH, K. P. A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. **Crop Science**. 55: 2068–2077. 2015.

NASS, L. L.; VALOIS, A. C. C.; MELO, I. S.; VALADARES-INGLIS, M. C. Recursos genéticos e melhoramento – plantas. **Fundação MT**. 2001.

NEYHART, J. L.; SMITH, K. P. Validating genomewide predictions of genetic variance in a contemporary breeding program. **Crop Science**, vol. 59:1062-1072. 2019.

OSTHUSHENRICH, T.; FRISCH, M.; ZENKE-PHILIPPI, C.; JAISER, H.; SPILLER, M.; CSELÉNYI, L.; KRUMNACKER, K.; BOXBERGER, S.; KOPAHNKE, D.; ANTJE HABEKUB, A.; ORDON, F.; HERZOG, E. Prediction of means and variance of crosses with genome-wide marker effects in barley. **Frontiers in Plant Science.** 9:1899. 2018.

SANT'ANNA, I. C.; CABRAL, F. R.; NASCIMENTO, M.; SILVA, G. N.; CARNEIRO, V. Q.; CRUZ, C. D.; OLIVEIRA, M. S.; CHAGAS, F. E. Multigenerational prediction of genetic values using genome-enabled prediction. **Plos one**, 14(1), p.e0210531. 2019.

SCHNELL F. W.; UTZ H. F. F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern, pp. 234–258 In **Bericht über die Arbeitstagung der Vereinigung Österreichischer Pflanzenzüchter**. Gumpenstein, Österreich. 1975.

SILVA, F. L., BORÉM, A.; SEDIYAMA, T.; LUDKE, W. **Soybean Breeding**. 1st ed. Springer, Gewebestrasse, Switzerland, 2017.

TIEDE, T., KUMAR, L.; MOHAMMADI, M.; SMITH, K. P. Predicting genetic variance in bi- parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. **Molecular Breeding.** 35: 1–13. 2015.

VENCOVSKY, R. Herança quantitative. In: PATERNIANI, E.; VIEGAS, G. (Ed.) **Melhoramento e produção de milho no Brasil**. 2.ed. Campinas: Fundação Cargill. p. 137-209. 1987

YAO, J.; ZHAO, D.; CHEN, X.; ZHANG, Y.; WANG, J. Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (Triticum aestivum L.). **The Crop Journal**. 6: 353-365. 2018.

ZHONG, S.; JANNINK, J. L. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. **Genetics** 177: 567–576. 2007.

PREDIÇÃO DA VARIÂNCIA GENÉTICA PARA SELEÇÃO DE CRUZAMENTOS DE SOJA E SUAS RELAÇÕES COM A DISTÂNCIA GENÉTICA

CAPÍTULO 1

PREDIÇÃO DA VARIÂNCIA GENÉTICA PARA SELEÇÃO DE CRUZAMENTOS DE SOJA E SUAS RELAÇÕES COM A DISTÂNCIA GENÉTICA

RESUMO

SILVA, Maikon Guerith Baptistella, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. Capítulo I. **Predição da variância genética para seleção de cruzamentos de soja e suas relações com a distância genética.** Orientador: Felipe Lopes da Silva.

Predizer a média e a variabilidade genética de um cruzamento antes de ser realizado, por meio da seleção genômica ampla, é uma ferramenta que visa auxiliar os melhoristas na escolha dos genitores em um programa de melhoramento de soja. Metodologias como distância genética entre os genitores apresentam baixas correlações com a variância genética de progênies predita. Portanto, o objetivo deste estudo foi de validar as relações entre as estimativas de média e variância resultantes da metodologia de seleção de cruzamentos biparentais por meio da predição genômica via progênies simuladas, e a distância genética entre os genitores com base em todas as marcas (distância total), e também, por meio da realização da partição das distâncias genéticas a partir de informações dos efeitos dos marcadores genéticos e a separação entre alelos fixados e em heterozigosidade, juntamente com as respectivas somas destes efeitos. Para isto 300 linhagens e um conjunto de 886 marcadores foram simulados e realizado todos os cruzamentos possíveis. De cada combinação biparental, foram geradas 200 linhagens RILs, tendo como base um mapa genético para as execuções das segregações. Considerando, todas as marcas, a correlação entre a distância genética e a variância das progênies RILs simuladas foi de -0,34. Porém, mediante as partições, a estimativa passou para -0,49 quando realizada a correlação com o grupo 3, que apresentou a maior soma dos efeitos em heterozigosidade. Portanto, o uso da distância total não foi totalmente informativo para encontrar populações com alta variabilidade genética. A captação da divergência nas regiões de maiores efeitos, mediante o particionamento dos marcadores, proporcionou um melhor entendimento da variância predita.

Palavras-chave: Simulação. Distância genética. Seleção genômica ampla. Seleção de genitores. Variância genética. *Glycine max.*

PREDICTION OF THE GENETIC VARIANCE FOR SELECTION SOYBEAN CROSSES AND THEIR RELATIONSHIPS WITH GENETIC DISTANCE

ABSTRACT

SILVA, Maikon Guerith Baptistella, M.Sc., Universidade Federal de Viçosa, February of 2022. Chapter I. **Prediction of the genetic variance for selection soybean crosses and their relationships with genetic distance.** Adviser: Felipe Lopes da Silva.

Predicting the mean and the genetic variability of a crossing before it is carried out, through wide genomic selection, is a tool that aims to assist breeders in choosing the parents in a soybean breeding program. Methodologies as genetic distance between the parents present low correlations with the genetic variance of predicted progenies. Therefore, the objective of this study was to validate the relationships between the mean and variance estimates resulting from the methodology of selection of biparental crosses through the genomic prediction via simulated progenies, and the genetic distance between the parents, in addition, together with the partition of genetic distances from information from the effects of genetic markers and separation between fixed and heterozygous alleles, along with their sums of these effects. For this, 300 lines and a set of 886 markers were simulated and performed all possible crosses, and each cross were generated 200 RILs lines based on a genetic map the executions of segregations. Considering, all the markers, the correlation between the genetic distance and the variance of the simulated RILs progenies was -0.34. However, through the partitions, the estimate passed to -0.49 when the correlation with Group 3, which presents the largest sum of the effects on heterozygousness. Thereforce, the use of the total distance was not fully informative to find populations with high genetic variability, but the divergence capture in the regions of greater effects, through the partitioning of the markers, where gene complementarity occurs.

Keywords: Simulation. Genetic distance. Genome wide selection. Parents selection. Genetic variance. *Glycine max.*

1. INTRODUÇÃO

A utilização de genitores com alta frequência de alelos favoráveis e variabilidade genética, proporcionando populações segregantes com alta média e maior variância genética para a característica de interesse, são os mais indicados para o desenvolvimento de genótipos superiores (Smith, 2019; Bernardo, 2010; Fehr, 1987)

Porém, devido ao número cada vez maior de linhagens candidatas a englobar os blocos de cruzamentos em programas de melhoramento genético de soja [*Glycine max* (L.) Merrill], selecionar, da forma mais eficiente possível, os melhores genitores para o desenvolvimento de populações com alto potencial, é uma das etapas mais cruciais e de maior desafio aos melhoristas, pois desta, depende o sucesso das próximas etapas (Silva et al., 2017; Bernardo, 2010).

Pode-se dizer que, uma seleção eficiente de plantas, dentro de populações promissoras, resultará em linhagens de alta performance, devido a predominância de efeitos aditivos e epistáticos de interação aditivo x aditivo para a produtividade de plantas de soja (Cooper, 1990). Logo, a predição do potencial das populações e de suas progênies endogâmicas a ser geradas permitiria que aquelas com baixo desempenho sejam eliminadas antes mesmo de o cruzamento ser realizado, permitindo a utilização de esforços somente em populações promissoras. Os trabalhos pioneiros de Schnell e Utz (1975) e posteriormente de Zhong e Jannink (2007) demonstraram a predição de médias de progênies derivadas de determinadas combinações com base em valores fenotípicos, e ambos destacaram a importância da variância genética aditiva para a identificação das melhores progênies, e consequentemente, dos melhores cruzamentos.

De acordo com a genética quantitativa clássica, a média das linhagens endogâmicas ou das *Recombinant Inbred Lines* (RILs), extraídas de um cruzamento biparental pode ser estimada pela média dos seus genitores (Bernardo, 2010). Porém, para o cálculo da variância genética, a predição se torna mais difícil, sendo necessário a avaliação das progênies, tornando a aplicação desta metodologia impraticável na rotina de um programa de melhoramento genético de soja quando há um grande número de populações segregantes.

O início dos estudos da seleção genômica ampla (GWS) após os estudos pioneiros de Meuwissen (2001) proporcionaram a predição de genótipos não fenotipados. E, juntamente com os estudos de Bernardo (2014), diversas estratégias

para a realização de predição de cruzamentos por meio de seleção das melhores progênies via simulação e validação de variâncias genéticas preditas em progênies vêm sendo desenvolvidas, e apresentando maiores ganhos genéticos quando comparado somente com a média predita dos genitores (Zhong e Jannink, 2007).

Utilizando de simulação computacional e validação em dados reais, estudos foram realizados utilizando a metodologia de predição de progênies, em que se observou uma maior acurácia de plantas autógamas em relação as alógamas (Jean et al., 2021; Sant'Anna et al., 2019; Adeyemo & Bernardo, 2019; Neyhart & Smith, 2019; Beckett et al., 2019; Osthushenrich et al., 2018; Osthushenrich et al., 2017; Lehermeier et al., 2017; Mohammadi et al., 2015; Bernardo, 2014).

Metodologias como a utilização da distância genética e ancestralidade servem como teorias para o melhor entendimento da variabilidade genética entre os genitores. Entretanto, embora haja a necessidade de mais processos de entendimento e validação, trabalhos pioneiros como de Tiede et al. (2015), demonstraram que houve resultados melhores da predição da variância aditiva de cruzamentos biparentais utilizando as simulações de progênies endogâmicas e a aplicação da GWS com base em marcadores moleculares, em comparação ao uso da distância genética.

De acordo com o estudo de Beckett et al., (2019) não houve relação entre a distância genética e a variância das progênies endogâmicas simuladas. Logo, a utilização da distância genética entre os parentais não foi um parâmetro acurado para a estimação desta variabilidade e, assim podendo acarretar um viés de predição. Porém, é importante ressaltar que a distância genética utilizada para a comparação com a variância genética, foi com base em todas as marcas. Assim, a variância genética encontrada em genitores muito similares pode estar ocorrendo devido apenas a divergência genética entre poucas marcas, porém de efeito maior, podendo ocorrer complementariedade entre os genes, e com isso, a ocorrência de segregação transgressiva.

Portanto, o objetivo deste estudo foi validar as relações entre as estimativas de média e variância resultantes da metodologia de seleção de cruzamentos biparentais por meio da predição genômica via progênies simuladas, e da distância genética entre os genitores. O presente estudo difere dos demais pelo fato de ser realizado a partição das distâncias, a partir de informações dos efeitos dos marcadores genéticos, mediante aplicação da GWS, além também, da separação entre alelos fixados e em heterozigosidade e das respectivas somas destes efeitos,

sendo capaz de captar a real divergência genética nas regiões gênicas de efeitos maior e menor para a característica de interesse.

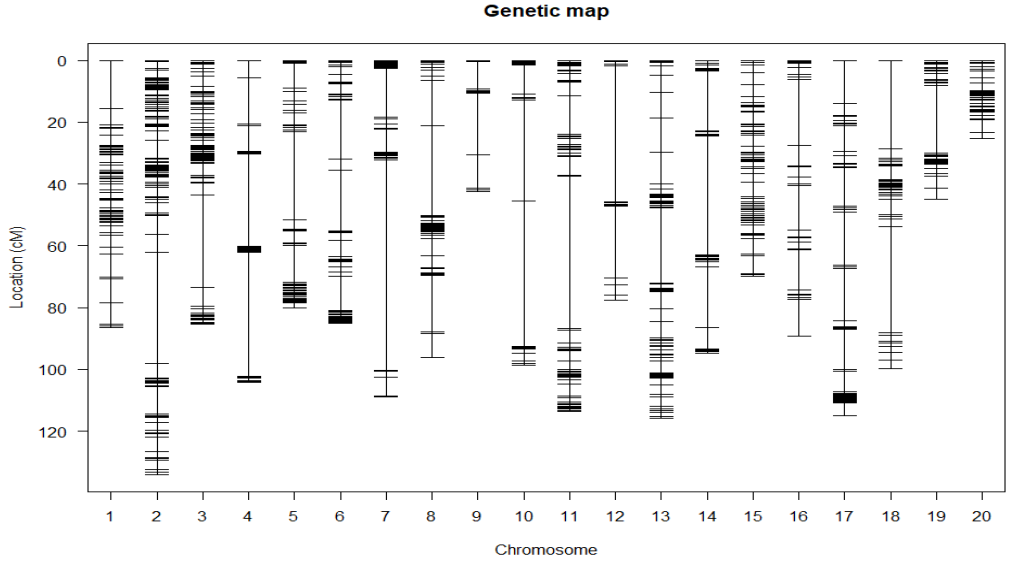
2. MATERIAL E MÉTODOS

3. Simulação genotípica e fenotípica dos genitores

Neste estudo, tomou-se como referência um programa de melhoramento genético da cultura da soja e o incremento de produtividade em kg ha^{-1} como característica quantitativa a ser avaliada. Foram simuladas 300 linhagens, denominados de L1 a L300, e 886 marcadores moleculares codominantes (SNPs) contidas ao longo dos 20 grupos de ligação (Figura 1). As informações de frequências de recombinação em *centimorgans* (cM) foram utilizadas para a inserção dos marcadores moleculares ao longo do genoma das linhagens simuladas.

A distribuição das marcas no genoma e a frequência de recombinação utilizadas foram obtidas mediante utilização das informações de posição física e mapa genético dos marcadores moleculares da população NAM 05 (IA3023xCL0J095-4-6) proveniente do projeto SoyNAM (*Nested Association Mapping (NAM) of Genes Controlling Soybean Yield and Other Key Traits*).

Figura 1 – Posição dos marcadores codominantes ao longo dos cromossomos da cultura da soja usada para simular a população de treinamento e as progênies endogâmicas.



Fonte: Silva (2022).

Foram considerados nas simulações apenas efeitos aditivos entre os alelos. Com o intuito de reproduzir uma característica quantitativa com *locus* de efeito maior e menor, dentre as 886 marcas, 50 foram classificados como *Quantitative Trait Locus* (QTLs) de efeito maior para a característica produtividade em kg ha^{-1}, distribuídos de maneira aleatória no genoma.

O valor genético de cada indivíduo foi obtido pela soma dos efeitos genotípicos de cada marca, mais a média e o desvio do erro aleatório (e), em que $e \sim N(0, \sigma_e^2)$, onde a variância residual foi obtida de acordo com o estimador da herdabilidade da característica, conforme equação abaixo:

$$Y = X + g + \sigma_e\, e \qquad \text{(Eq. 1)}$$

$$Y = X + g + \sqrt{\frac{\sigma_g^2(1-h^2)}{h^2}}\; e \qquad \text{(Eq. 2)}$$

Em que Y é o valor fenotípico simulado; X é a média de produtividade, considerado como 4300 kg ha^{-1} (Valor extraído da média de populações RILs do Projeto SoyNAM); g é a soma dos valores genéticos; σ_e é o desvio do erro aleatório; h^2 é a herdabilidade da característica, sendo considerado o valor de 0,20, σ_g^2 é a variância dos valores genotípicos e σ_e^2 é a variância do erro.

A simulação foi repetida 10 vezes para cada linhagem, e a média utilizada como valor fenotípico.

3.1. Predição Genômica

Para a estimação dos efeitos de cada marcador, conforme etapa 1 da figura 2, utilizou-se da metodologia de seleção genômica, em que a população de treinamento foi composta pelas informações genotípicas e fenotípicas das 300 linhagens. O modelo de predição genômica utilizado foi o método estatístico Ridge-Regression Best Linear Unbiased Prediction (RR-BLUP), conforme equação do modelo linear misto abaixo:

$$y = Xb + Wm + e \qquad \text{(Eq. 3)}$$

Em que, y é o vetor dos valores fenotípicos de dimensão 300x1; b é o vetor de efeitos fixos; m se refere ao vetor dos efeitos aleatórios dos marcadores (886x1), em que $m \sim N(0, \sigma_m^2)$; e é o vetor de resíduos aleatórios, em que e ~ N (0, I$\sigma^2{}_e$), no qual I é a matriz de incidência e $\sigma^2{}_e$ é a variância residual ; X e W são as matrizes de incidência para vetores de efeitos fixos e aleatórios respectivamente. As equações de modelos mistos para a predição de m mediante o método RR-BLUP equivalem a:

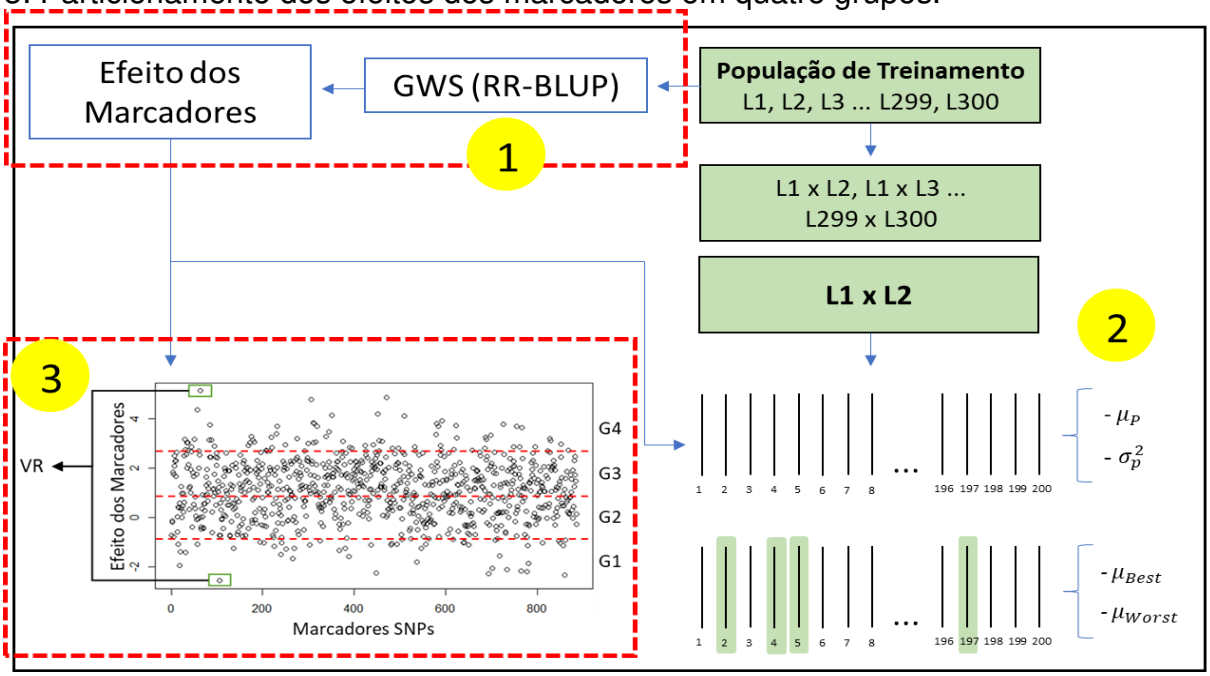
$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + I\,\frac{\sigma^2{}_e}{\sigma^2{}_m} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix} \qquad \text{(Eq. 4)}$$

Os valores genéticos genômicos (GEBV) foram preditos por meio da equação:

$$GEBV = W\hat{m} \qquad \text{(Eq. 5)}$$

A acurácia de predição foi realiza utilizando o processo de validação cruzada com 5 folds. O processo se repetiu por 50 vezes, e a capacidade preditiva foi realizada pela média das correlações de Pearson entre o valor fenotípico e o valor genético genômico estimado de todas as repetições.

Figura 2 - Esquema das etapas da metodologia emprega. 1: Estimação dos efeitos de cada marcados. 2: Simulação de 200 progênies RILs em cada cruzamento biparental. 3: Particionamento dos efeitos dos marcadores em quatro grupos.



Fonte: Silva (2022).

3.2. Simulação das progênies endogâmicas

A partir das linhagens simuladas, todas as combinações hibridas biparentais possíveis foram realizadas, totalizando 44850 cruzamentos, sem recíprocos. Considerando o modelo aditivo, a média fenotípica das combinações (μ_F) foi obtida mediante a média simples entre os valores fenotípicos simulados *per se* de cada genitor. Em seguida, utilizando do mapa genético, foram simuladas 200 progênies *Recombinant Inbred Lines* (RILs) de cada cruzamento.

Além da média dos valores genéticos genômicos (μ_p) e da variância (σ_p^2) de todas as progênies RILs simuladas de cada combinação híbrida gerada, foram obtidos a média dos valores genéticos genômicos das dez progênies superiores (μ_{Best}) e das dez piores (μ_{Worst}), conforme etapa 2 da figura 2. A simulação das 200 progênies RILs foi repetida por 20 vezes, e a média foi utilizada para cada parâmetro. Com os valores destes parâmetros preditos foi possível verificar quais cruzamentos tendem a gerar melhores populações para extração de linhagens superiores, sendo estes de interesse para os programas de melhoramento de soja.

Para fins de comparação entre a seleção de cruzamentos somente com base em valores fenotípicos e o uso de predição genômica de progênies simuladas, os 20 melhores genitores com base no valor fenotípico foram selecionados e análises de coincidência foram realizadas junto aos melhores cruzamentos preditos pela metodologia de GWS.

Análises de correlação de Pearson com os dados de produtividade simulada e índices de Coincidência (IC) utilizando os ranqueamentos das combinações hibridas foram realizados, sendo IC(%) = (Ns/Nt)*100, em que Ns é o número de combinações hibridas selecionados nos dois grupos em avaliação e Nt é o número total de combinações.

3.3. Partição das distâncias genéticas por meio dos efeitos de marcadores genéticos

Para a realização do desdobramento dos efeitos dos marcadores genéticos ao longo do genoma (etapa 3 da figura 2), eles foram ranqueados em ordem decrescente, e particionados em quatro grupos, em que o critério adotado para alocar

as marcas nos respectivos grupos foi mediante a utilização de um valor de referência, sendo a diferença entre a marca de maior e menor efeito, divido pelo número de grupos, conforme a seguinte equação.

$$VR = \frac{EMaior - E\ Menor}{Nº\ Grupos} \quad \text{(Eq. 6)}$$

Em que, VR é o Valor de Referência; EMaior se refere ao maior efeito dentre todos os marcadores; EMenor se refere ao menor efeito e Nº Grupos é a quantidade de grupos que se deseja classificar.

Tomando como exemplo a marca de maior efeito do SNP X com 6,25, e por outro lado o SNP Y com efeito -3,5, o VR estimado, considerando a formação de quatro grupos, será de 2,43. Logo, todos os SNPs que possuem seus efeitos de marcadores entre -3,5 a -1,08 estarão classificados no Grupo 1 (G1), entre -1,08 a 1,37 no Grupo 2 (G2), entre 1,37 e 3,80 no Grupo 3 (G3) e entre 3,80 a 6,25 no Grupo (G4).

Para verificar a variabilidade genética entre os genitores de todas as combinações e as suas correlações com a σ_p^2, foi realizada a distância genética via marcadores moleculares considerando todos os 886 SNPs (Ga), e, de maneira separada dentro de cada grupo gerado a partir da partição dos efeitos dos marcadores.

A metodologia utilizada para a estimação da distância genética foi a matriz de variância-covariância de parentesco proposta por VanRaden (2008), em que a maior dissimilaridade genética entre os genitores é expressa por valores inferiores, portanto, quanto mais negativo, mais divergência genética entre as linhagens. Dentro de cada grupo gerado, as marcas foram separadas com base no número de cópias do alelo dominante, e em seguida foram contabilizados os números que estavam fixados no genoma para alelos dominantes (2), recessivos (0) e os em heterozigosidade (1), e, as respectivas somas dos efeitos destes marcadores gerados pela predição genômica.

Todas as simulações e análises foram realizadas no software R (R Development Core Team 2019), utilizando a combinação de funções próprias e os pacotes rrBLUP (Endelman, 2011), PopVar (Mohammadi et al., 2015), R/qtl (Broman et al., 2003) e snpReady (Granato et al., 2018).

4. RESULTADOS

4.1. Médias fenotípicas e acurácia de predição

As simulações fenotípicas das 300 linhagens apresentaram uma média geral de 6041 kg ha^{-1}, variando de 6894 kg ha^{-1} para a L173 e de 5236 kg ha^{-1} para L3, englobando uma diferença de aproximadamente 1650 kg ha^{-1}. Dentre as 20 melhores linhagens, a média foi de 6653 kg ha^{-1}, e a diferença entre a primeira e a vigésima de apenas 410 kg ha^{-1}, o que representou aproximadamente 7% da média geral (Tabela 1).

Tabela 1 – Ranqueamento das 20 linhagens mais produtivos, em kg ha^{-1}, a partir da média de 10 repetições. E seus respectivos valores genéticos.

Rank	Genótipo	Valor Genético	Média Prod. Simulada (Kg ha^{-1})	Rank	Genótipo	Valor Genético	Média Prod. Simulada (Kg ha^{-1})
1	L173	2599	6894	10	L260	2363	6665
2	L147	2599	6866	11	L120	2363	6650
3	L48	2481	6823	12	L100	2245	6587
4	L236	2481	6815	13	L89	2245	6557
5	L60	2481	6793	14	L290	2245	6552
6	L161	2481	6772	15	L288	2245	6548
7	L108	2363	6693	16	L82	2245	6542
8	L57	2363	6684	17	L45	2245	6534
9	L78	2363	6673	18	L21	2245	6484
10	L260	2363	6665	19	L5	2245	6479
11	L120	2363	6650	20	L171	2126	6459

Foi observado um alto valor de acurácia preditiva média estimada pela validação cruzada da predição genômica, sendo de 0,74 na média das 50 repetições.

Dentre todos os 44850 cruzamentos possíveis derivados das 300 linhagens, a combinação L147xL173 apresentou a maior média fenotípica com 6880 kg ha^{-1}. Após a aplicação do modelo de predição genômica, o cruzamento entre L173xL236, terceira melhor combinação para μ_F, obteve a maior média predita das 200 linhagens RILs de 6866 kg ha^{-1} conforme observado na Tabela 2.

Tabela 2 – As 20 Combinações híbridas com base na média das melhores RILs preditas *(μ_{Best}),* em kg ha^{-1}. E seus respectivos valores de média fenotípica *(μ_F),* média da predição *(μ_p),* variância das RILs *(σ_p^2)* e a diferença entre μ_{best} e μ_p.

Rank	Cross	Média Fen. (μ_F)(N1)	Média GWS (μ_p)(N2)	Variância GWS (σ_p^2)	RILs Melhores (μ_{Best})	RILs Inferiores (μ_{Worst})	μ_{best}-μ_p
1	L48xL236	6819(10)	6822(8)	40344	7220	6423	398
2	L173xL236	6854(3)	6866(1)	32747	7216	6470	350
3	L48xL173	6858(2)	6843(3)	29975	7197	6501	354
4	L78xL173	6783(17)	6759(24)	42789	7179	6369	420
5	L147xL236	6841(6)	6816(9)	31864	7173	6477	356
6	L173xL260	6779(20)	6786(14)	36006	7168	6393	382
7	L108xL147	6780(19)	6782(17)	39696	7166	6358	384
8	L57xL173	6789(16)	6774(18)	34058	7165	6409	391
9	L147xL173	6880(1)	6838(4)	26026	7165	6555	327
10	L60xL147	6830(8)	6833(5)	26804	7162	6493	328
11	L48xL147	6845(4)	6852(2)	24998	7154	6503	302
12	L161xL173	6833(7)	6825(6)	26489	7152	6494	327
13	L173xL290	6723(45)	6731(35)	50568	7149	6296	418
14	L60xL173	6843(5)	6825(7)	21160	7141	6535	317
15	L120xL147	6758(25)	6771(20)	39973	7141	6349	370
16	L120xL236	6733(39)	6741(31)	45409	7138	6306	397
17	L60xL236	6804(12)	6803(11)	27660	7129	6453	326
18	L100xL173	6740(34)	6722(41)	38213	7126	6338	404
19	L48xL108	6758(26)	6741(30)	37475	7123	6366	382
20	L48xL60	6808(11)	6800(12)	27198	7121	6470	321

(N1) – Ranqueamanto com base em μ_F ; (N2) – Ranqueamento com base em μ_p

A combinação L48xL236 obteve produtividade média predita de 7220 kg ha^{-1} para μ_{Best}, apresentando ser o cruzamento com a maior probabilidade de encontrar os melhores genótipos superiores. Por outro lado, L3xL183 como o pior cruzamento com 5073 kg ha^{-1}. Logo, se considerar em um programa de melhoramento de soja a realização de cruzamentos entre somente as 20 melhores combinações com base na μ_F, 25% dos melhores cruzamentos preditos com base em μ_{Best} não seriam realizados, como por exemplo, a combinação L173xL290, ranqueada pela média fenotípica na posição 45, passando para a posição 13 com base no ranque das melhores RILs preditas (Tabela 2).

A linhagem L173, que apresentou a maior produtividade fenotípica, esteve presente em 50% dos 20 melhores cruzamentos considerando um ranqueamento por

μ_{Best}, demonstrando possuir alta frequência de alelos favoráveis e complementariedade com outras linhagens (Tabela 2).

4.2. Correlações entre μ_F, μ_P e μ_{Best}

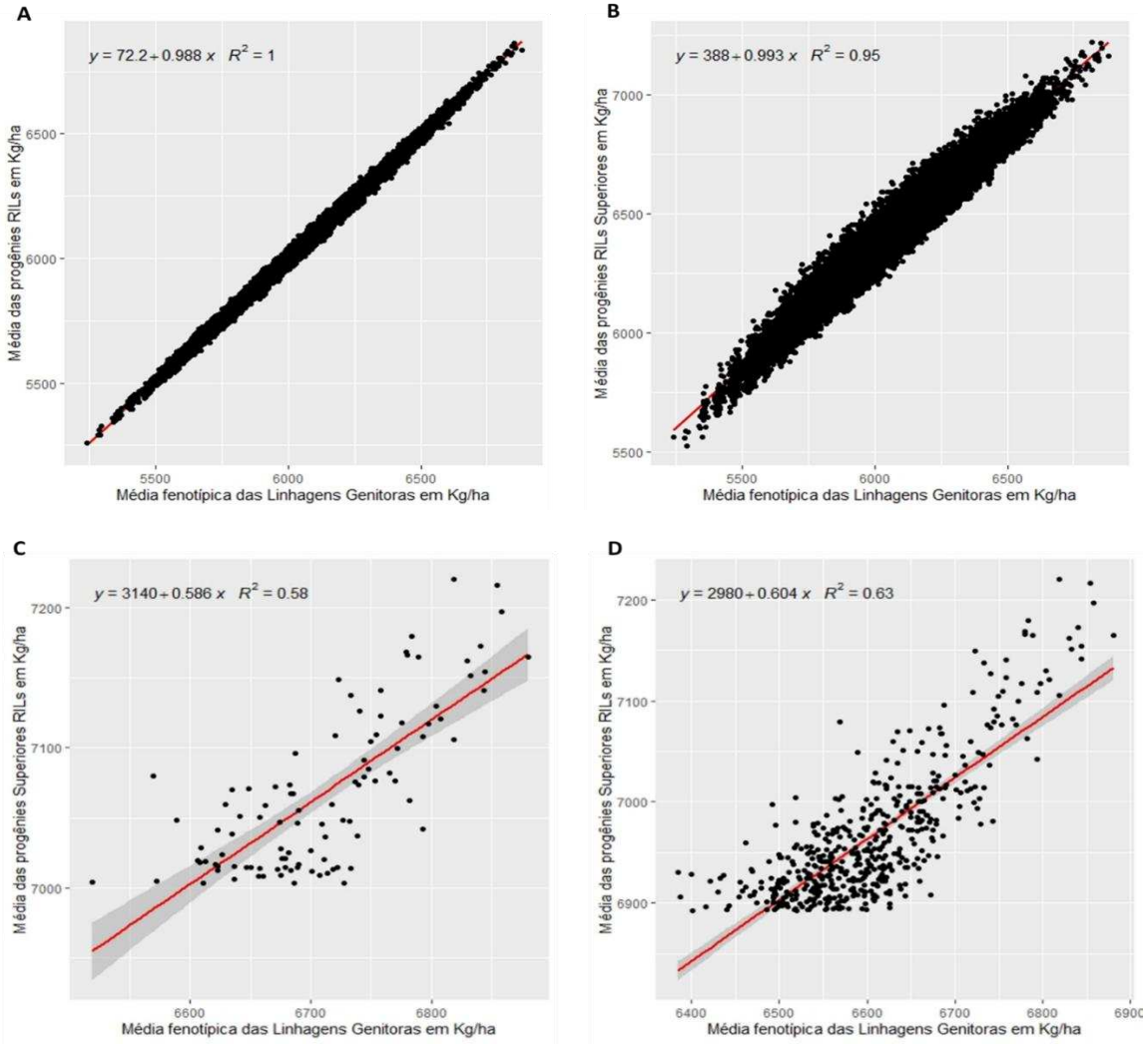
Considerando ainda o ranqueamento com base em μ_{Best}, a correlação de Pearson entre μ_F e μ_P apresentou alta magnitude, com estimativas de 0,997 (Figura 3A). Avaliando apenas os 100 primeiros cruzamentos, a estimativa da correlação foi de 0,937, mantendo os altos índices de correlações entre os dois parâmetros.

Estimativas de alta magnitude também foram encontrados quando foi comparado μ_F e μ_{Best}, com 0,973 (Figura 3B). Porém, à medida que foram realizadas as correlações dentro de grupos de ranqueamentos, as magnitudes foram decaindo, como observado entre as 100 e 500 melhores combinações com valores de 0,76 e 0,79 respectivamente (Figuras 3C e 3D). Os IC entre as 100 e 500 primeiras combinações foram muito semelhantes, com 73% e 74% respectivamente, portanto, aproximadamente um quarto das combinações híbridas apresentaram um alto potencial de desenvolver linhagens superiores e se utilizado o ranqueamento de μ_F para a escolha dos cruzamentos, estes não seriam realizados.

4.3. Relação entre a variância genética e a distância genética total e particionada

Considerando todas os cruzamentos possíveis, a média predita de σ_p^2 foi de 31130 kg^2 ha^{-2}, variando de 7706 kg^2 ha^{-2} a 88648 kg^2 ha^{-2}, respectivamente para o cruzamento entre as linhagens L236xL262 e a combinação L99xL137, sendo este o cruzamento que apresentou o maior ganho genético quando comparado a diferença entre μ_{best}-μ_p, com incremento em produtividade de aproximadamente 10%. A correlação entre o ganho genético e σ_p^2 foi de 0,91, portanto, para a obtenção de incremento genético em futuras cultivares comerciais, além da utilização de genitores com alta média fenotípica, ou seja, com alta frequência de alelos favoráveis, a divergência genética entre os genitores também é fundamental pois poderá ser explorada a complementariedade de alelos favoráveis.

Figura 3 – A: Scatterplot entre a média fenotípica entre as linhagens genitoras (μ_F) e a média de todas as progênies RILs (μ_p), para todas as 44850 combinações. B: Scatterplot entre a média fenotípica entre as linhagens genitoras (μ_F) e a média das melhores progênies RILs (μ_{best}), para todas as 44850 combinações. C: Scatterplot entre a média fenotípica entre as linhagens genitoras (μ_F) e a média das melhores progênies RILs (μ_{best}), para as 100 melhores combinações. D: Scatterplot entre a média fenotípica entre as linhagens genitoras (μ_F) e a média das melhores progênies RILs (μ_{best}), para as 500 melhores combinações.



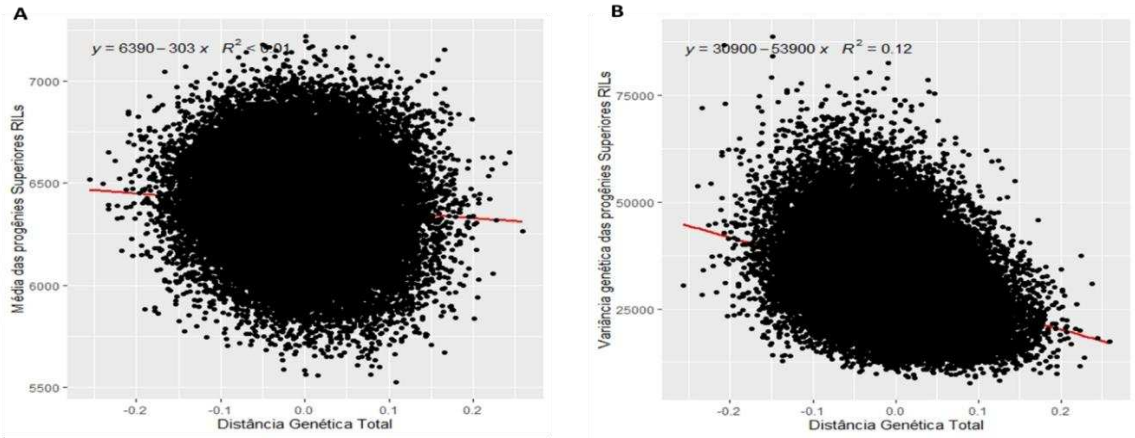
Fonte: Silva (2022).

A correlação entre Ga e σ_p^2 foi de magnitude baixa com estimativa de -0,340 (Figura 4B), e de 0,25 entre σ_p^2 e μ_{Best}. Não foi encontrado correlação entre Ga e μ_{Best}, conforme observado na Figura 4A.

A amplitude entre o maior e menor efeito obtidos pelo método RR-BLUP mediante a GWS foi de 41,02, logo, o VR encontrado entre os efeitos dos marcadores foi de 10,25. Conforme Tabela 3, aproximadamente 73% de todos os marcadores estão concentrados dentro de G2, o que já era esperado, visto que a característica

simulada neste estudo foi de caráter quantitativo, e, portanto, grande número de genes de pequeno efeito.

Figura 4 – A: Scatterplot entre a distância genética total (Ga) e a média das melhores progênies RILs (μ_{best}), para todas as 44850 combinações. B: Scatterplot entre a distância genética total (Ga) e a variância genética das progênies RILs (σ_p^2), para todas as 44850 combinações.



Fonte: Silva (2022).

A combinação que apresentou a maior divergência genética com base no total de marcas foi L27xL192 (30560 $kg^2\ ha^{-2}$) com Ga de -0,256, ficando na posição 12433 dentre todos os cruzamentos e apresentando maior probabilidade de obter linhagens superiores. Por outro lado, o cruzamento L186xL263 (17358 $kg^2\ ha^{-2}$) apresentou a maior similaridade, com estimativa Ga de 0,259, na posição 32462.

Tabela 3 – Número de SNPs e estimativas de maior e menor efeito das marcas em cada Grupo particionado

Grupo	Nª. SNPs	Maior Efeito	Menor Efeito
G1	87	-4,68	-14,93
G2	650	5,57	-4,64
G3	123	15,74	5,62
G4	26	26,09	15,85

Os cruzamentos L57xL147 e L137xL236 apresentaram Ga praticamente similares, de -0,079 e -0,073 respectivamente, estando ranqueados entre as 35 mais promissoras combinações, com destaque para L137xL236 que subiu mais de 300 posições no ranqueamento quando comparado à utilização da média fenotípica. Avanço este de posições devido a sua alta média, e principalmente à elevada σ_p^2 das suas progênies simuladas, com aproximadamente 68000 $kg^2\ ha^{-2}$, sendo mais que o

dobro de L57xL147. Isso também foi evidenciado entre as combinações ranqueadas em primeiro e terceiro lugares, conforme Tabela 4, em que L48xL236 apresentou uma σ_p^2 34% maior que L48xL173, e com a mesma Ga. A explicação para esta diferença foi observada nas partições, pois, nos grupos de efeitos de marcadores de G3 e G4, o número de marcadores em heterozigosidade (S1) e suas respectivas somas de efeitos foi superior para L48xL236. Para G2 houve similaridade entre ambos, enquanto para G1 houve maior divergência para L48xL173, mas são efeitos de menor intensidade quando comparado a G4 por exemplo (Tabela 3).

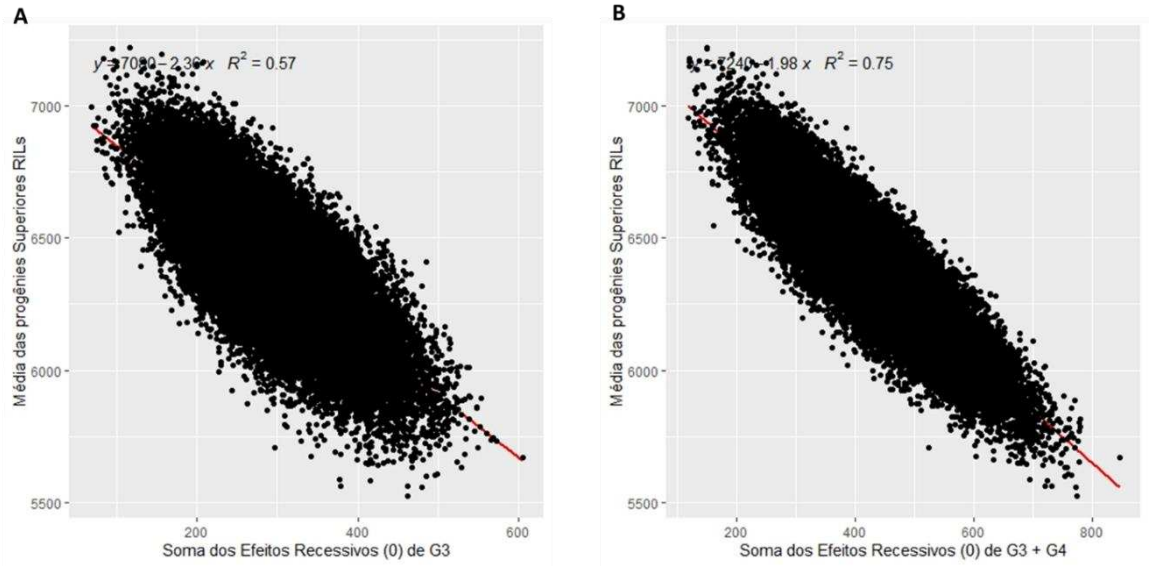
As correlações entre Ga e as distância genéticas de cada grupo foram respectivamente de 0,384, 0,930, 0,558 e 0,264, para G1, G2, G3 e G4. Porém, embora encontrado alta correlação entre Ga e G2, quando G2 foi comparado com σ_p^2, a correlação diminui drasticamente a magnitude para -0,19, enquanto entre G3 e σ_p^2 foi de -0,49, apresentando a maior estimativa dentre os quatro grupos.

Dentre as 20 melhores combinações, o cruzamento que apresentou a maior porcentagem de alelos fixados, considerando tanto dominantes (S2) e recessivos (S0) foi L48xL147 com aproximadamente 62% de todas as marcas em homozigose, com maior porcentagem de alelos fixados dominantes. Ainda dentro dos 20 melhores cruzamentos e considerando a soma dos efeitos, os alelos fixados dominantes do G3 apresentam um maior peso com estimativa média de 473 kg ha^{-1}, sendo também o grupo de maior importância para os alelos em heterozigosidade (S1).

Para G1, que apresenta os efeitos de marcadores de alta magnitude negativos, é fundamental uma maior presença de alelos recessivos, para que não haja perda de produtividade, logo, o cruzamento que apresentou a maior soma destes alelos foi L48xL60. Por outro lado, L173xL236 obteve 24 marcas em homozigose dominante, acarretando um efeito negativo de -167 (Tabela 4).

A maior correlação entre a soma dos efeitos de alelos fixados e μ_{best} foi em G3 para o alelo recessivo com estimativa de -0,75, ou seja, os cruzamentos com menor soma dos efeitos dos alelos recessivos presentes no genoma, apresentam maiores probabilidades de encontrar genótipos superiores. (Figura 5A). Realizando a soma dos efeitos entre 2 Grupos, a correlação de maior magnitude foi encontrada entre os grupos de efeitos de marcadores recessivos positivos (G3 + G4), com -0,86 (Figura 5B).

Figura 5 – A: Scatterplot entre a soma dos efeitos recessivos do Grupo 3 e a média das melhores progênies RILs (μ_{best}), para todas as 44850 combinações. B: Scatterplot entre a soma dos efeitos recessivos dos Grupos 3 e 4 em conjunto e a média das melhores progênies RILs (μ_{best}), para todas as 44850 combinações.



Fonte: Silva (2022).

Tabela 4 – As 20 combinações híbridas ranqueadas com base na média das melhores RILs preditas (μ_{Best}), em kg ha^{-1}. E seus respectivos valores de variância das RILs (σ_p^2), distância total (Ga), distância dentro de cada agrupamento, além do número de SNPs e estimativas da soma dos efeitos das marcas em cada Grupo particionado.

Rank	Cross	σ_p^2	Ga	G1				G2				G3				G4			
				Dist G1	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$	Dist G2	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$	Dist G3	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$	Dist G4	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$
1	L48xL236	40344	0,001	-0,049	-224$^{(28)}$	-280$^{(39)}$	-146$^{(20)}$	-0,005	31$^{(163)}$	139$^{(305)}$	157$^{(182)}$	0,114	117$^{(14)}$	456$^{(54)}$	516$^{(55)}$	-0,227	34$^{(2)}$	282$^{(15)}$	180$^{(9)}$
2	L173xL236	32747	0,057	0,211	-216$^{(29)}$	-267$^{(34)}$	-167$^{(24)}$	0,020	-11$^{(165)}$	212$^{(306)}$	125$^{(179)}$	0,146	94$^{(13)}$	458$^{(53)}$	537$^{(57)}$	0,068	57$^{(3)}$	225$^{(12)}$	214$^{(11)}$
3	L48xL173	29975	0,001	-0,265	-181$^{(24)}$	-358$^{(49)}$	-110$^{(14)}$	0,000	11$^{(152)}$	145$^{(303)}$	171$^{(195)}$	0,164	156$^{(19)}$	435$^{(49)}$	498$^{(55)}$	0,096	37$^{(2)}$	204$^{(11)}$	255$^{(13)}$
4	L78xL173	42789	-0,050	-0,171	-163$^{(22)}$	-340$^{(46)}$	-146$^{(19)}$	-0,029	10$^{(147)}$	183$^{(306)}$	133$^{(197)}$	-0,067	84$^{(11)}$	546$^{(62)}$	459$^{(50)}$	-0,083	37$^{(2)}$	223$^{(12)}$	236$^{(12)}$
5	L147xL236	31864	0,009	-0,167	-185$^{(24)}$	-336$^{(44)}$	-128$^{(19)}$	0,029	3$^{(173)}$	161$^{(295)}$	162$^{(182)}$	0,026	97$^{(14)}$	541$^{(61)}$	451$^{(48)}$	0,015	40$^{(2)}$	220$^{(12)}$	236$^{(12)}$
6	L173xL260	36006	0,048	0,058	-198$^{(26)}$	-303$^{(39)}$	-148$^{(22)}$	0,055	10$^{(174)}$	161$^{(290)}$	155$^{(186)}$	0,019	100$^{(13)}$	526$^{(59)}$	462$^{(51)}$	-0,013	20$^{(1)}$	296$^{(15)}$	179$^{(10)}$
7	L108xL147	39696	-0,042	-0,013	-204$^{(27)}$	-287$^{(38)}$	-159$^{(22)}$	-0,051	15$^{(143)}$	146$^{(312)}$	165$^{(195)}$	-0,016	133$^{(15)}$	523$^{(60)}$	432$^{(48)}$	-0,032	40$^{(2)}$	209$^{(11)}$	246$^{(13)}$
8	L57xL173	34058	-0,046	-0,167	-184$^{(24)}$	-356$^{(48)}$	-109$^{(15)}$	-0,042	20$^{(167)}$	160$^{(316)}$	147$^{(167)}$	0,039	134$^{(17)}$	501$^{(55)}$	453$^{(51)}$	-0,151	16$^{(1)}$	262$^{(14)}$	217$^{(11)}$
9	L147xL173	26026	0,039	-0,121	-181$^{(25)}$	-337$^{(44)}$	-131$^{(18)}$	0,007	8$^{(158)}$	116$^{(301)}$	202$^{(191)}$	0,285	207$^{(25)}$	377$^{(44)}$	504$^{(54)}$	0,171	18$^{(1)}$	191$^{(10)}$	286$^{(15)}$
10	L60xL147	26804	0,005	-0,189	-210$^{(27)}$	-323$^{(44)}$	-116$^{(16)}$	-0,019	9$^{(146)}$	157$^{(302)}$	160$^{(202)}$	0,189	207$^{(25)}$	419$^{(49)}$	463$^{(49)}$	0,328	43$^{(2)}$	158$^{(8)}$	295$^{(16)}$
11	L48xL147	24998	0,167	0,038	-241$^{(32)}$	-246$^{(33)}$	-162$^{(22)}$	0,178	23$^{(185)}$	98$^{(242)}$	206$^{(223)}$	0,183	199$^{(24)}$	437$^{(49)}$	453$^{(50)}$	0,211	34$^{(2)}$	171$^{(9)}$	290$^{(15)}$
12	L161xL173	26489	0,023	-0,083	-176$^{(24)}$	-338$^{(44)}$	-135$^{(19)}$	-0,012	33$^{(156)}$	156$^{(312)}$	137$^{(182)}$	0,302	144$^{(18)}$	362$^{(44)}$	583$^{(61)}$	-0,122	16$^{(1)}$	277$^{(14)}$	203$^{(11)}$
13	L173xL290	50568	-0,034	0,016	-174$^{(23)}$	-315$^{(43)}$	-161$^{(21)}$	-0,004	18$^{(160)}$	135$^{(309)}$	173$^{(181)}$	-0,260	100$^{(13)}$	639$^{(73)}$	349$^{(37)}$	0,154	20$^{(1)}$	217$^{(11)}$	259$^{(14)}$
14	L60xL173	21160	-0,016	-0,230	-183$^{(24)}$	-370$^{(50)}$	-96$^{(13)}$	-0,048	16$^{(135)}$	165$^{(319)}$	145$^{(196)}$	0,239	177$^{(22)}$	390$^{(45)}$	522$^{(56)}$	0,213	41$^{(2)}$	199$^{(10)}$	256$^{(14)}$
15	L120xL147	39973	-0,070	-0,001	-223$^{(30)}$	-288$^{(37)}$	-138$^{(20)}$	-0,068	20$^{(174)}$	126$^{(307)}$	181$^{(169)}$	-0,189	134$^{(17)}$	633$^{(71)}$	321$^{(35)}$	0,228	43$^{(2)}$	163$^{(9)}$	291$^{(15)}$
16	L120xL236	45409	-0,022	-0,035	-219$^{(27)}$	-296$^{(41)}$	-134$^{(19)}$	-0,007	16$^{(188)}$	191$^{(298)}$	120$^{(164)}$	-0,049	95$^{(13)}$	565$^{(64)}$	428$^{(46)}$	-0,210	38$^{(2)}$	282$^{(15)}$	176$^{(9)}$
17	L60xL236	27660	0,023	-0,066	-216$^{(27)}$	-309$^{(42)}$	-124$^{(18)}$	0,001	34$^{(154)}$	163$^{(305)}$	129$^{(191)}$	0,189	130$^{(17)}$	427$^{(50)}$	532$^{(56)}$	0,057	66$^{(3)}$	222$^{(12)}$	208$^{(11)}$
18	L100xL173	38213	0,077	0,037	-163$^{(22)}$	-325$^{(41)}$	-161$^{(24)}$	0,075	18$^{(157)}$	146$^{(293)}$	163$^{(200)}$	0,126	151$^{(18)}$	459$^{(53)}$	479$^{(52)}$	0,028	37$^{(2)}$	221$^{(12)}$	238$^{(12)}$
19	L48xL108	37475	-0,035	-0,104	-211$^{(27)}$	-293$^{(41)}$	-145$^{(19)}$	-0,099	47$^{(131)}$	116$^{(326)}$	163$^{(193)}$	0,316	195$^{(22)}$	355$^{(39)}$	539$^{(62)}$	0,061	72$^{(4)}$	195$^{(10)}$	229$^{(12)}$
20	L48xL60	27198	0,012	0,087	-260$^{(34)}$	-245$^{(33)}$	-145$^{(20)}$	-0,039	8$^{(138)}$	193$^{(308)}$	126$^{(204)}$	0,173	183$^{(22)}$	422$^{(48)}$	484$^{(53)}$	0,253	61$^{(3)}$	169$^{(9)}$	265$^{(14)}$

$^{(N)}$ – Número de marcadores

5. DISCUSSÃO

Reunir em uma nova cultivar, alelos favoráveis de genitores distintos é o objetivo de grande parte de programas de melhoramento genético de soja visando características quantitativas. Porém, para isso, é necessário a escolha correta dos genitores que são complementares, e que desenvolvam populações segregantes capazes de extrair progênies superiores. Quando o objetivo é o incremento de produtividade de grãos, esta escolha se torna mais difícil, devido ao fato de ser uma característica herdada por muitos genes de pequeno efeito. Logo, os melhores cruzamentos são aqueles que possibilitam a formação de populações base com alta média e de grande variabilidade genética (Nass, 2001).

De acordo com Baenziger e Peterson (2001) há duas categorias para a definição dos parentais. A primeira com base apenas nas informações de performance dos genitores, em que a metodologia mais empregada é mediante a utilização de suas médias fenotípicas. Porém, com uso deste método não é possível antever a variabilidade genética da combinação. Ferramentas que empregam o uso da distância genética via marcadores moleculares, coeficiente de parentesco e técnicas de análise multivariada podem ser utilizadas no auxílio ao melhorista para a predição da variância do cruzamento.

A segunda categoria é com base no desempenho das progênies, em que as metodologias mais conhecidas são as análises dialélicas, Jinks e Pooni (1976) e a estimativa de m+a'. Estas são ferramentas em que podemos obter estimativas de genitores com maior capacidade de combinação e quais as populações segregantes com maiores probabilidades de se extrair linhagens superiores. Entretanto, são alternativas que apresentam limitações quando o número de genitores envolvidos é elevado, pois necessitam de avaliações das progênies de maneira criteriosa, podendo não ser viável economicamente em um programa de melhoramento genético de soja.

Neste contexto, sem a necessidade da avaliação a campo, a simulação genética auxilia os programas de melhoramento a obter informações prévias sobre as progênies que poderão ser geradas a partir da escolha de determinados genitores, ou até de milhares, e realizar os cruzamentos somente daqueles de alto potencial.

Diversos trabalhos com simulação vêm sendo desenvolvidos nesta área, onde são capazes de promover os primeiros resultados de maneira mais rápida e gerando novas informações. Neste estudo, realizamos a simulação da maneira mais próxima

da realidade possível, com a criação de linhagens de soja e as progênies endogâmicas respeitando um mapa genético real, tornando as segregações genéticas ainda mais acurados. Um dos exemplos mais relevantes da predição genômica, que guiou diversos outros trabalhos, foi proporcionado por meio de simulação computacional, em que Meuwissen (2001) demonstrou toda a metodologia da GWS ao mundo científico.

A acurácia preditiva média encontrada na validação cruzada da predição foi similar a encontrada por Meuwissen (2001) utilizando a mesma metodologia estatística. Com dados reais, Beckett et al. (2019), obtiveram capacidades preditivas semelhantes, de 0,72 para produtividade de grãos de milho, gerando resultados satisfatórios na identificação de combinações híbridas promissoras. Para outras culturas autógamas, Akdemir e Sanches (2019) trabalhando com trigo e Smallwood et al. (2019) com soja, obtiveram acurácia de 0,65 e 0,48, respectivamente.

A utilização de apenas efeitos aditivos entre os alelos contribuiu para alta acurácia preditiva e consequentemente para a alta correlação encontrada entre μ_F e μ_p quando se considerou todas as combinações. Embora seja considerada uma correlação de alta magnitude, como também encontrado no trabalho de Beckett et al. (2019), as variações de ranqueamento entre as primeiras posições podem levar à exclusão de futuros cruzamentos promissores, pois, caso um programa de melhoramento optasse por escolher gerar apenas 20 populações segregantes com alto potencial, e a seleção for realizada com base em μ_F , as combinações L78xL173 e L60xL161 não seriam efetuadas, ou seja, 10%.

Mantendo o número de populações a serem escolhidas, quando comparamos a metodologia mais popular para seleção de genitores, a μ_F , contra a método por predição genômica, μ_{Best}, houve 75% de coincidência entre os cruzamentos selecionados, em que se destaca a combinação entre os genitores L173 e L290 que passou do *rank* 45 em μ_F , para a décima terceira em μ_{Best}. Uma das razões para este avanço foi devido a sua alta σ_p^2 (50568 kg^2 ha^{-2}), explicada principalmente pela alta quantidade de SNPs em heterozigosidade no G3, sendo possível a exploração de mais de 630 kg ha^{-1} mediante a segregação genética, conforme observado na soma dos efeitos destes marcadores (Tabela 4).

Considerando, todas as marcas, a correlação entre a distância genética e a σ_p^2 das progênies simuladas foi baixa, assim como encontrado no trabalho de Beckett

et al., (2019). Porém para o melhor entendimento desta predição, e diferentemente do trabalho citado anteriormente, a realização da partição das distâncias e o agrupamento com base nos valores dos efeitos dos marcadores, fez com que a correlação aumentasse de -0,34 entre σ_p^2 e Ga, para -0,49 quando comparado σ_p^2 somente com a distância do G3. O aumento desta correlação está relacionada com a maior soma dos efeitos dos marcadores SNPs em heterozigosidade dentro de G3, com média de 512 kg, ocasionando assim variações expressivos nos GEBV preditos das progênies RILs simuladas de acordo com as segregações existentes.

Estes resultados evidenciaram que, a variância genética das progênies preditas do cruzamento apresentou uma maior correlação com a distância genética naquele grupo em que possui as marcas em heterozigosidade com maior soma do efeito genético, ocasionando maior impacto no GEBV predito. Portanto, o número de marcadores heterozigotos totais não é o fator mais importante para a classificar um cruzamento com maior probabilidade de gerar genótipos superiores, e sim ocorrer divergência genética naquela região que apresenta maior magnitude dos efeitos destas marcas relacionadas a característica de interesse, havendo assim complementariedade gênica.

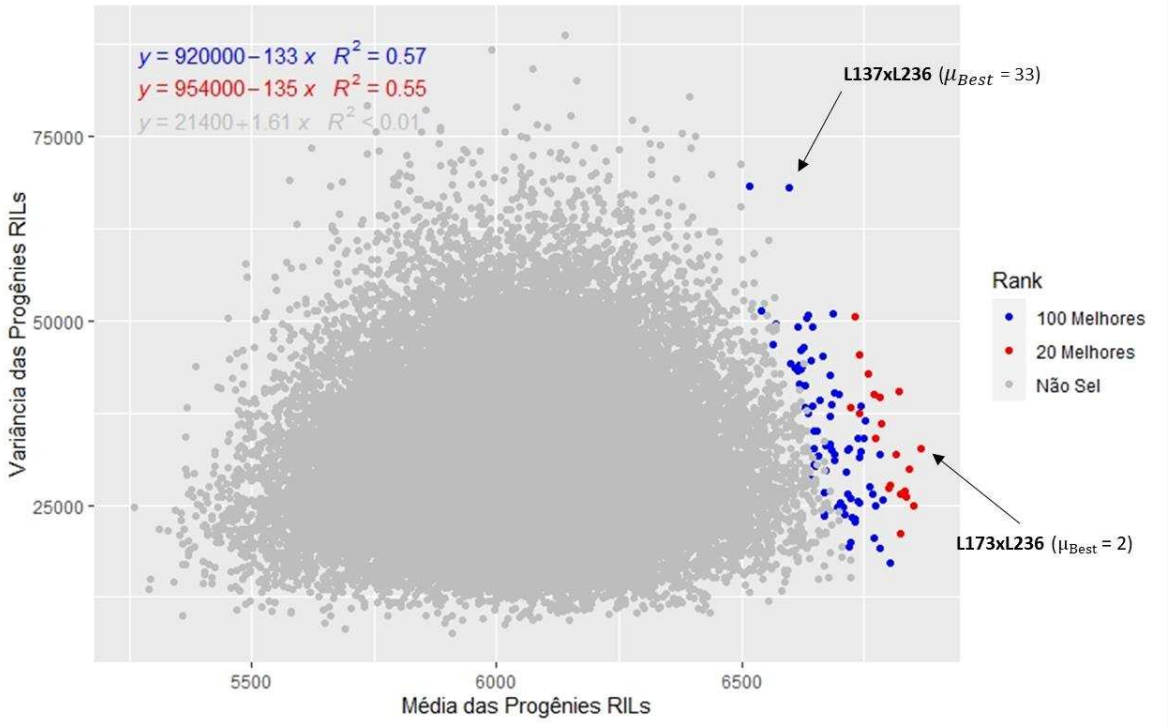
A distância genética total, Ga, não foi informativa a respeito da geração das melhores linhagens superiores, portanto o seu uso pode causar um viés, e até indicação errônea na escolha de blocos de cruzamentos. Se ainda usado, o melhorista deve estar ciente que dois genitores podem apresentar alta similaridade genética total, porém uma divergência em regiões especificas e que causam alto efeito na variância, ou também alta dissimilaridade, porém em regiões que possuem efeitos pequenos ou até nulos na expressão da característica.

Além da busca pela variabilidade entre os genitores da maneira mais eficiente possível, não se deve esquecer do outro fator importante para uma boa população segregante, que é alta média, causada pelo acúmulo de alelos favoráveis e que apresentam fortes correlações com as probabilidades de extração de genótipos superiores, principalmente nas regiões com maiores somas de efeitos de marcadores (Figuras 5A e 5B). Para G1 e G2, as combinações com genes fixados recessivos merecem destaque, pois nestes grupos há predominância de efeitos de marcadores negativos, causando declínio na produtividade. Enquanto para G3 e G4, os genes fixados em dominância são os de maior relevância para o incremento da média, pois os efeitos foram todos positivos.

De acordo os trabalhos de Osthushenrich et al. (2018) e Zhong e Jannink (2007) cultivares elite em programas de melhoramento tendem a formar populações segregantes de alta média, porém com baixa variância, apresentando uma correlação negativa, enquanto as maiores variâncias são ocasionadas por doadores de baixa média, geralmente de pool gênicos diferentes. Neste trabalho, observou-se uma tendencia de similaridade com os resultados de Osthushenrich et al. (2018) e Zhong e Jannink (2007), principalmente para a combinação entre L137xL236, em que L137 não está classificado entre as melhores linhagens, e por outro lado L236 é a quarta mais produtiva.

Porém, conforme Figura 6, foram obtidas correlações de aproximadamente -0,74 entre μ_P e σ_p^2 tanto para as 20 melhores combinações, como para as 100. Portanto, embora ocorra uma correlação de média a alta magnitude, ainda há combinações entre cultivares elites que podem ocorrer uma variância entre eles a ser explorada, principalmente em regiões especificas de alto valor de efeito alélico, como é o caso da combinação L173xL236.

Figura 6 – Scatterplot entre a média das progênies RILs (μ_P) e a variância das progênies RILs (σ_p^2) para todas as 44850 combinações. Em vermelho, estão as 20 melhores combinações ranqueadas com base em μ_{Best}, e em azul as demais até a centésima melhor combinação.



Fonte: Silva (2022).

Como sugestão para futuros estudos, esta metodologia pode ser validada utilizando dados reais de um programa de melhoramento genético, aplicando modelo de predição genômicos junto aos genitores, e validação após as seleções a campo das progênies avançadas. E outros métodos estatísticos para a predição podem ser utilizados como modelos Bayesianos, conforme utilizados por Lehermeier et al. (2017), com a finalidade de principalmente aumentar a acurácia preditiva da variância genética para caracteres quantitativos.

6. CONCLUSÃO

Em conclusão, as metodologias de simulação computacional tanto dos genitores, como das progênies, se mostraram eficiente para o presente trabalho, sendo relevante para uma melhor definição da seleção dos genitores e das progênies. As partições das distâncias e as somas dos seus efeitos dos marcadores em cada grupo foi uma estratégia fundamental para elucidar as baixas correlações entre distância genética e variância das progênies. Portanto, conclui-se que a dissimilaridade entre dois genitores com base em todos os marcadores moleculares por si só não deve ser utilizada como parâmetro definitivo para a escolha potenciais cruzamentos.

Finalmente, a abordagem proposta neste estudo, mediante o uso da GWS, é mais uma nova ferramenta que permite aos melhoristas de soja tomar decisões mais eficientes para a escolha dos genitores a serem utilizados nos blocos de cruzamentos, antes mesmos de serem realizados, principalmente para caracteres de baixa herdabilidade, como a produtividade de grãos, ou outros caracteres conforme seus objetivos preestabelecidos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ADEYEMO, E.; BERNARDO, R. Predicting genetic variance from genomewide marker effects estimated from a diverse panel of maize inbreds. **Crop Science**. 59:583–590. 2019.

AKDEMIR, D.; J. I. SÁNCHEZ. Design of training populations for selective phenotyping in genomic prediction. **Scientific Reports**, 9:1446. 2019.

BAENZIGER, P. S.; PETERSON, C. Genetic variation: Its origin and use for breeding self-pollinated species. In Plant Breeding in the 1990's, edited by H. Stalker and J. Murphy, pp. 69–100, **North Carolina State University**, North Carolina. 1991.

BECKETT, T. J.; ROCHEFORD, T. R.; MOHAMMADI, M. Re-imagining maize inbred potential: identifying breeding crosses using genetic variance of simulated progeny. Crop Science, vol. 59: 1-12. 2019.

BERNARDO, R. **Breeding for quantitative traits in plants**. Stemma Press, Woodbury, second edition, 2010.

BERNARDO, R. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. **Crop Science**. 54:2586–2595. 2014.

BROMAN K. W., WU H., SEN Ś., CHURCHILL G. A. R/qtl: QTL mapping in experimental crosses. **Bioinformatics** 19:889-890. 2003.

COOPER, R. L. Modified early generation testing procedure for yield selection in soybean. Crop Science, v. 30: 659-670, 1990.

ENDELMAN J. B. "Ridge regression and other kernels for genomic selection with R package rrBLUP." *Plant Genome*, **4**, 250-255. 2011.

FEHR, W. Principles of cultivar development. **Theory and technique**. Macmillan, New York. 1987.

GRANATO, I. S. C., GALLI, G., COUTO, E. G. O. snpReady: a tool to assist breeders in genomic analysis. **Molecular Breeding**, Heidelberg, v. 38, n. 8, p. 1-7, 2018.

JEAN, M.; COBER, E.; O'DONOUGHUE, L; RAJCAN, I.; BELZILE, F. Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. **Crop Science**. 61. 10.1002/csc2.20583. 2021.

JINKS, J. L., POONI, H. S. Predicting the properties of recombinant inbreed lines derived by single seed descent. **Heredity**, Oxford, v. 36, n. 2, p. 243-266. 1976.

LEHERMEIER, C., S. TEYSS`EDRE, AND C.-C. SCH¨ON. Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. **Genetics Early Onli**: 1–10. 2017.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157: 1819–1829. 2001.

MOHAMMADI, M.; TIEDE, T., SMITH, K. P. Popvar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. **Crop Science**. 55: 2068–2077. 2015.

NASS, L. L.: VALOIS, A. C. C.: MELO, I. S.: VALADARES-INGLIS, M. C. Recursos genéticos e melhoramento – plantas. **Fundação MT**. 2001.

NEYHART, J. L.; SMITH, K. P. Validating genomewide predictions of genetic variance in a contemporary breeding program. **Crop Science**, vol. 59:1062-1072. 2019.

OSTHUSHENRICH, T., M. FRISCH, E. HERZOG. Genomic selection of crossing partners on basis of the expected mean and variance of their derived lines. **PloS ONE** 12: 1–12. 2017.

OSTHUSHENRICH, T.; FRISCH, M.; ZENKE-PHILIPPI, C.; JAISER, H.; SPILLER, M.; CSELÉNYI, L.; KRUMNACKER, K.; BOXBERGER, S.; KOPAHNKE, D.; ANTJE HABEKUB, A.; F. ORDON, F.; HERZOG, E. Prediction of means and variance of crosses with genome-wide marker effects in barley. **Frontiers in Plant Science.** 9:1899. 2018.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna. 2019.

SANT'ANNA, I.C.; CABRAL, F.R.; NASCIMENTO, M.; SILVA, G.N.; CARNEIRO, V.Q.; CRUZ, C.D.; OLIVEIRA, M.S.; CHAGAS, F.E. Multigenerational prediction of genetic values using genome-enabled prediction. **Plos one**, 14(1), p.e0210531. 2019.

SCHNELL F. W.; UTZ H. F. F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern, pp. 234–258 In **Bericht über die Arbeitstagung der Vereinigung Österreichischer Pflanzenzüchter**. Gumpenstein, Österreich, 1975.

SILVA, F. L., A. BORÉM, T. SEDIYAMA, AND W. LUDKE. **Soybean Breeding**. 1st ed. Springer, Gewebestrasse, Switzerland, 2017.

SMALLWOOD, C. J.; SAXTON, A. M.; GILLMAN, J.D.; BHANDARI, H.S.; WADL, P. A.; FALLEN, B. D.; HYTEN, D. L.; SONG, Q.; PANTALONE, V. R. Context-Specific Genomic Selection Strategies Outperform Phenotypic Selection for Soybean Quantitative Traits in the Progeny Row Stage. **Crop Science**, vol. 59: 54-67. 2019.

TIEDE, T., KUMAR, L.; MOHAMMADI, M.; SMITH, K. P. Predicting genetic variance in bi- parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. **Molecular Breeding.** 35: 1–13. 2015.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science** 91: 4414–4423. 2008.

ZHONG S.; JANNINK J. L. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. **Genetics** 177: 567–576. 2007.

ÍNDICE DE SELEÇÃO PARA À IDENTIFICAÇÃO DE POTENCIAIS CRUZAMENTOS DE SOJA ATRAVÉS DA PREDIÇÃO GENÔMICA

CAPÍTULO 2

ÍNDICE DE SELEÇÃO PARA À IDENTIFICAÇÃO DE POTENCIAIS CRUZAMENTOS DE SOJA ATRAVÉS DA PREDIÇÃO GENÔMICA

RESUMO

SILVA, Maikon Guerith Baptistella, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. Capítulo II. **Índice de seleção para à identificação de potenciais cruzamentos de soja através da predição genômica.** Orientador: Felipe Lopes da Silva.

Selecionar os melhores cruzamentos é crucial para o sucesso na obtenção de cultivares de soja adaptadas as diferentes regiões do Brasil. Devido ao número elevado de genitores elites candidatos a englobar os blocos de cruzamentos, metodologias que visam a avaliações de progênies são impraticáveis. Ferramentas como simulação computacional, juntamente com a seleção genômica vem sendo importantes alternativas para a predição. Porém, baixas correlações de variância predita e observada podem causar um viés na seleção de cruzamentos. O objetivo deste estudo foi avaliar uma nova proposta de metodologia de seleção de cruzamentos mediante a predição genômica e um índice de seleção desenvolvido após o particionamento dos marcadores conforme seus efeitos genéticos. Para isto foram utilizadas 102 linhagens do programa de melhoramento genético de soja da empresa GDM Seeds do Brasil S.A, totalizando 5151 combinações possíveis. Deste total, 10 foram realizados, se tornando a população de validação. Foram utilizados a metodologia de simulação de progênies simuladas e um índice de seleção por soma de postos com base nos efeitos dos marcadores genético, divido estes, em grupos conforme a magnitude. Os resultados indicaram uma melhor acurácia preditiva da seleção de cruzamentos com base no índice de seleção, pois caracterizou melhor a complementariedade de genes. A correlação entre a distância genética e a classificação com base da distância dentro de cada grupo particionado foi capaz de explicar a variância genética existente em cada cruzamento.

Palavras-chave: Índice de seleção. Distância genética, Seleção genômica ampla. Seleção de genitores. Variância genética. *Glycine max.*

SELECTION INDEX FOR THE IDENTIFICATION OF POTENTIAL SOYBEAN CROSSES THROUGH GENOMIC PREDICTION

ABSTRACT

SILVA, Maikon Guerith Baptistella, M.Sc., Universidade Federal de Viçosa, February of 2022. Chapter I. **Selection index for the identification of potential soybean crosses through genomic prediction.** Adviser: Felipe Lopes da Silva.

Selecting the best crosses is crucial for success in obtaining soybean cultivars adapted to different regions of Brazil. Due to the high number of candidate elite parents to encompass the crossing blocks, methodologies aimed at progeny evaluations are impractical. Tools such as computer simulation, along with genomic selection have been important alternatives for prediction. However, low correlations of predicted and observed variance can cause a bias in the selection of crosses. The goal of this study was to evaluate a new proposal for a cross selection methodology through genomic prediction and a selection index developed after partitioning the markers according to their genetic effects. For this, 102 lines of the soybean genetic improvement program of the company GDM Seeds do Brasil S.A were used, totaling 5151 possible combinations. Of this total, 10 were performed, becoming the validation population. The simulation methodology of simulated progenies and a selection index by sum of ranks based on the effects of genetic markers were used, divided into groups according to magnitude. The results indicated a better predictive accuracy of the selection of crosses based on the selection index, as it better characterized the complementarity of genes. The correlation between the genetic distance and the classification based on the distance within each partitioned group was able to explain the genetic variance existing in each cross.

Keywords: Index selection. Genetic distance. Accuracy. Genome wide selection. Parents selection. Genetic variance. *Glycine max.*

1. INTRODUÇÃO

Os avanços em tecnologias e a ocupação em áreas antes inadequadas ao cultivo, proporcionaram ao Brasil se tornar o maior produtor e exportador mundial de soja [*Glycine max* (L.) Merr.], alcançando na última safra 2021/2022 mais de 40,3 milhões de hectares semeados em todo o país. Dentre os fatores que impulsionaram a cultura ao longo dos anos, como práticas de manejo de solo e culturais, o melhoramento genético possui um papel fundamental com o desenvolvimento de cultivares cada vez mais produtivas e adaptadas a todas as regiões produtoras. (CONAB, 2022; Nass et al., 2001).

Um programa de melhoramento genético para a obtenção de cultivares com performances superiores, se inicia com a escolha correta dos parentais utilizados nas obtenções das populações base. O sucesso depende da definição clara dos objetivos a serem alcançados, juntamente com a característica a ser melhorada e, respectivamente, o seu caráter genético. Grande parte dos caracteres de importância na cultura da soja são quantitativos, sendo controlados por vários genes de pequeno efeito e muito influenciados pelo ambiente, ocorrendo assim, baixos valores de herdabilidade (Nass et al., 2001).

As populações base desejadas são aquelas que apresentam alta média fenotípica e maior variância genética para a característica de interesse (Neyhart e Smith, 2019; Bernardo, 2010; Fehr, 1987). Metodologias como o uso da média fenotípica dos pais juntamente com a informação de pedigree são as mais usuais em programas de melhoramento de soja para a definição dos blocos de cruzamentos (Lado, 2017). O trabalho pioneiro de Schnell e Utz (1975) e em seguida de Zhong e Jannink (2007) demonstraram uma alternativa para a predição do potencial de cruzamentos para a extração de genótipos superiores, e ambos destacaram a importância da variância genética aditiva como um fator primordial para a identificação das melhores progênies. Porém, está variância só é possível ser obtida com a avaliação fenotípica das progênies das populações geradas, tornando esta metodologia vagarosa quando há um número grande de cruzamentos gerados, e consequente um número alto de populações desenvolvidas.

Logo, com o advento da seleção genômica, proposta por Meuwissen, et al. (2001), e o avanço em ferramentas computacionais, diversos trabalhos vêm sendo desenvolvidos com o intuito de se obter as melhores alternativas para a seleção de

genitores, principalmente após o trabalho de Bernando (2014). Atualmente, uma das metodologias utilizando a predição genômica é mediante à simulação computacional de progênies endogâmicas, sendo possível a predição de milhares de cruzamentos e a obtenção das estimativas das médias e variâncias de cada combinação, conforme trabalhos de Tiede et al., (2015), Mohammadi et al., (2015), Yao et al., (2018), Adeyemo e Bernardo (2019), Beckett et al., (2019), Neyhart e Smith, (2019) e Jean et al., (2021).

Para a aplicação desta metodologia, a população de treinamento é composta por informações genotípicas e fenotípicas dos parentais, de acordo com a característica de interesse. Em seguida, linhagens endogâmicas recombinantes (RILs) são geradas via simulação computacional tendo como referência um mapa genético, e o valor dos efeitos das marcas aplicados nestas progênies para gerar as estimativas dos valores genéticos genômicos (GEBVs). Sendo possível obter a predição da média e variância genética dentro de cada população segregante.

Porém, nestes trabalhos, além da baixa correlação entre a distância genética dos parentais e a variância genética predita, também foi observado uma baixa correlação na validação da predição da variância genética das populações, o que pode acarretar uma seleção equivocada de determinados cruzamentos, ou a não seleção de combinação com alto potencial para ocorrência de segregação transgressiva. Fato este comprovado mediante os estudos de Osthushenrich et al., (2018), em que obtiveram excelentes resultados para a seleção de cruzamentos de cevada utilizando da metodologia de simulação computacional, onde os 50% melhores cruzamentos foram todos corretamente preditos. Uma das razões para este resultado encontrado se de principalmente a uma alta correlação entra a variância genética observada e predita entre as populações desenvolvidas.

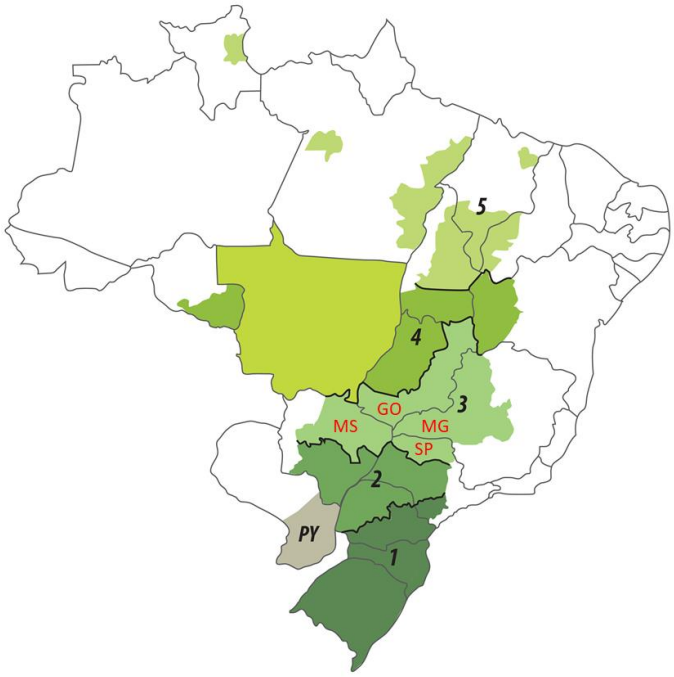
Portanto, o objetivo deste trabalho foi avaliar uma nova proposta de metodologia de seleção de cruzamentos mediante a predição genômica e um índice de seleção desenvolvido após o particionamento dos marcadores conforme seus efeitos genéticos. Além disso, foi realizado uma comparação com a metodologia de simulação de progênies.

2. MATERIAL E MÉTODOS

2.1. Material genético

Para incorporar a população de treinamento da seleção genômica, foram utilizadas 102 linhagens do programa de melhoramento genético de soja da empresa GDM Seeds do Brasil S.A., pertencentes a Macrorregião sojícola 3 (M3) na região central do Brasil, compreendo principalmente o estado de Goiás (GO), região do Triangulo Mineiro em Minas Gerais (MG), norte do Mato Grosso do Sul (MS) e norte de São Paulo (SP), conforme Figura 7. O Grupo de Maturação Relativa (GMR) das cultivares comerciais usualmente utilizadas pelos agricultores da região, variam de 6.5 a 8.0. Todas as informações fenotípicas e genotípicas foram gentilmente cedidas pela empresa GDM Seeds do Brasil.

Figura 7 – As cinco principais regiões produtoras de soja do Brasil, com acréscimo da região Sudeste do Paraguai, que possui similaridade com a Macrorregião 2 do Brasil. Em destaque os estados que compreendem a Macrorregiao 3 (GO, MS, MG e SP).



Fonte: Silva (2022)

A genotipagem por sequenciamento (genotyping-by-sequencing – GBS) das linhagens foi realizada pela “Genomic Diversity Facility” na Universidade de Cornell obtendo um total de 76143 marcadores do tipo *Single Nucleotide Polymorphisms* (SNPs). Após o processo de controle de qualidade, 1752 marcas foram obtidas mediante uma proporção máxima de dados perdidos (CR) de 80% e uma frequência

mínima do alelo menor (maf) de 5%. O processo de imputação foi realizado com base na média dos valores das marcas presentes. Tendo como base as informações das posições físicas dos marcadores, foi gerado um mapa genético mediante comparação junto a população NAM 05 (IA3023xCL0J095-4-6) proveniente do projeto SoyNAM (*Nested Association Mapping (NAM) of Genes Controlling Soybean Yield and Other Key Traits*). A análise de divergência genética com base em análise de componentes principais das 102 linhagens utilizando as 1752 marcas se encontra no Apêndice 1.

2.2. Dados fenotípicos

As 102 linhagens foram avaliadas para a característica produtividade de plantas em 18 localidades da M3 na safra 2015/2016, em que 70 linhagens estavam inseridas em um desenho experimental em blocos aumentados com 5 testemunhas em comum e 45 progênies em avaliação, obtendo o melhor estimador linear não-viesado (BLUE) via modelos mistos conforme equação:

$$y = Xp + Wn_{(l)} + Zn + Mt + e \qquad \text{(Eq. 7)}$$

Em que, y é o vetor dos valores fenotípicos; p é o vetor de efeito fixo das progênies; $n(l)$ se refere ao vetor dos efeitos aleatórios ensaio (n) dentro de cada localidade (l), em que $n(l) \sim N(0, \sigma^2_{n(l)})$; n se refere ao vetor dos efeitos aleatórios de localidade, em que $l \sim N(0, \sigma^2_l)$; t se refere ao vetor dos efeitos aleatórios da interação progênies x localidade, em que $t \sim N(0, \sigma^2_t)$; e e é o vetor de resíduos aleatórios, em que e ~ N (0, I$\sigma^2{}_e$), no qual I é a matriz de incidência e $\sigma^2{}_e$ é a variância residual ; X, W, Z e M são as matrizes de incidência para vetores de efeitos fixos e aleatórios respectivamente.

Para as demais linhagens, o melhor estimador linear não-viesado via modelos foi obtido mediante o delineamento em blocos casualizados:

$$y = Xp + Wr_{(l)} + Zn + Mt + e \qquad \text{(Eq. 8)}$$

Em que, y é o vetor dos valores fenotípicos; p é o vetor de efeito fixo das progênies; $r(l)$ se refere ao vetor dos efeitos aleatórios de repetição (r) dentro de cada

localidade (l), em que $r(l) \sim N(0, \sigma^2_{r(l)})$; n se refere ao vetor dos efeitos aleatórios de localidade, em que $l \sim N(0, \sigma^2_l)$; t se refere ao vetor dos efeitos aleatórios da interação progênies x localidade, em que $t \sim N(0, \sigma^2_t)$; e e é o vetor de resíduos aleatórios, em que e ~ N (0, I$\sigma^2{}_e$), no qual I é a matriz de incidência e $\sigma^2{}_e$ é a variância residual ; X, W, Z e M são as matrizes de incidência para vetores de efeitos fixos e aleatórios respectivamente.

A razão pelo qual há está separação das linhagens em dois grupos experimentais, se deve a etapa do programa de melhoramento genético da GDM em que elas se encontram. As primeiras 70 se encontram em uma última etapa de fase preliminar de seleção, enquanto as demais foram avaliadas em ensaios de Valor de Cultivo e Uso (VCU). Entretanto ambos os ensaios foram semeados nas mesmas datas e localidades. Incrementar as linhagens de VCU proporcionou enriquecer o modelo de predição genômica, e aumentar a população de treinamento.

2.3. Predição genômica

Para a estimação dos efeitos de cada marcador, conforme etapa 1 da figura 8, utilizou-se da metodologia de seleção genômica, em que a população de treinamento utilizada foi composta pelas informações genotípicas e fenotípicas das 102 linhagens. O modelo de predição genômica utilizado foi o método estatístico Ridge-Regression Best Linear Unbiased Prediction (RR-BLUP), conforme equação do modelo linear misto abaixo:

$$y = Xb + Wm + e \qquad \text{(Eq. 9)}$$

Em que, y é o vetor dos melhores estimadores linear não-viesado; b é o vetor de efeitos fixos; m se refere ao vetor dos efeitos aleatórios dos marcadores, em que $m \sim N(0, \sigma^2_m)$; e é o vetor de resíduos aleatórios, em que e ~ N (0, I$\sigma^2{}_e$), no qual I é a matriz de incidência e $\sigma^2{}_e$ é a variância residual ; X e W são as matrizes de incidência para vetores de efeitos fixos e aleatórios respectivamente. As equações de modelos mistos para a predição de m mediante o método RR-BLUP equivalem a:

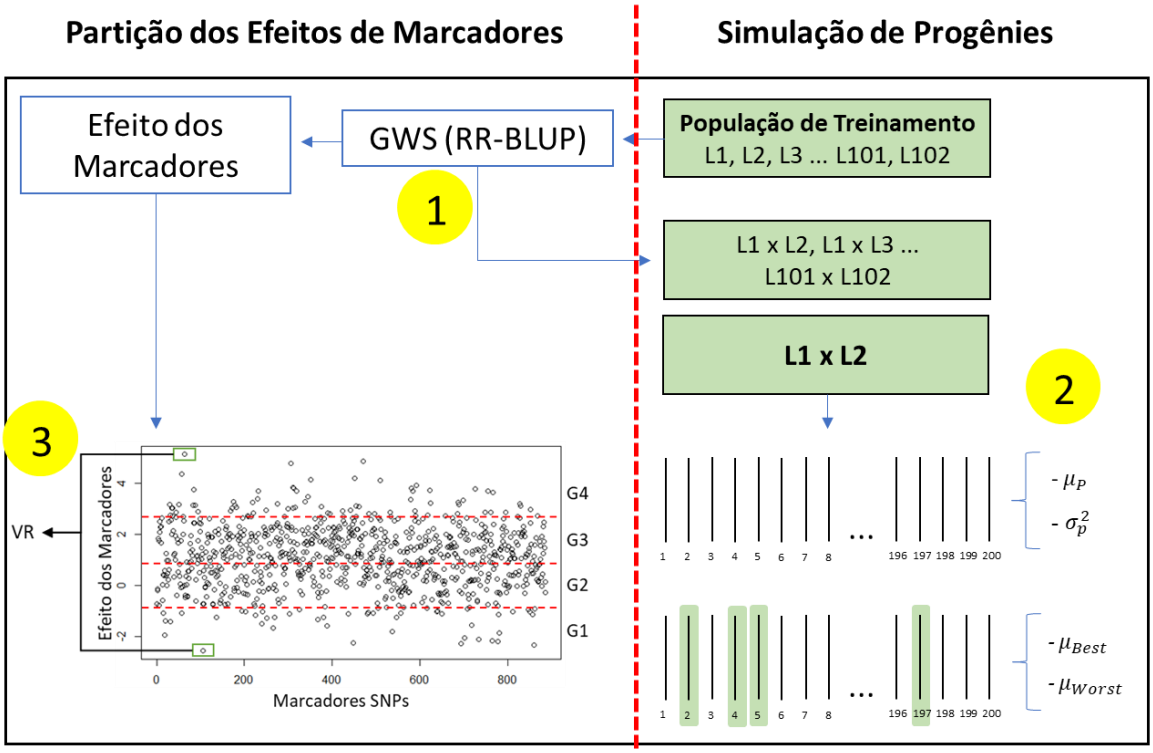
$$\begin{bmatrix} X'X & X'W \\ W'X & W'W + I\,\frac{\sigma^2 e}{\sigma^2 m} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ W'y \end{bmatrix} \quad \text{(Eq. 10)}$$

Os valores genéticos genômicos (GEBV) foram preditos por meio da equação:

$$GEBV = W\hat{m} \quad \text{(Eq. 11)}$$

A acurácia de predição foi realiza utilizando o processo de validação cruzada com 5 folds. O processo se repetiu por 50 vezes, e a capacidade preditiva foi realizada pela média das correlações de Pearson entre o valor fenotípico e o valor genético genômico estimado de todas as repetições.

Figura 8 - Esquema de todas as etapas da metodologia emprega. Etapa 1: Estimação dos efeitos de cada marcados. Etapa 2: Simulação de 200 progênies RILs em cada cruzamento biparental. Etapa 3: Particionamento dos efeitos dos marcadores em quatro grupos.



Fonte: Silva (2022).

2.4. Metodologia via simulação de progênies endogâmicas

A partir das informações genotípicas dos genitores, todas as combinações hibridas biparentais possíveis foram realizadas, totalizando 5151 cruzamentos, sem recíprocos. Considerando um modelo aditivo, a média fenotípica das combinações (μ_F) foi obtida mediante a média simples entre os valores fenotípicos de cada genitor, ou seja, entre as estimativas BLUE. Em seguida, utilizando do mapa genético, foram simuladas 200 progênies *Recombinant Inbred Lines* (RILs) de cada cruzamento.

Além da média dos valores genéticos genômicos (μ_p) e da variância (σ_p^2) de todas as progênies RILs simuladas de cada combinação híbrida gerada, foram obtidos a média dos valores genéticos genômicos das dez progênies superiores (μ_{Best}) e das dez piores (μ_{Worst}), conforme etapa 2 da figura 2.

A simulação das 200 progênies RILs foi repetida por 20 vezes, e a média foi utilizada para cada parâmetro. Com os valores destes parâmetros preditos foi possível verificar quais cruzamentos tendem a gerar melhores populações para extração de linhagens superiores, sendo estes de interesse para os programas de melhoramento de soja.

2.5. Nova abordagem por meio de partição

Para a realização do desdobramento dos efeitos dos marcadores genéticos ao longo do genoma, conforme etapa 3 da figura 2, eles foram ranqueados em ordem decrescente, e particionados em quatro grupos, em que o critério adotado para alocar as marcas nos respectivos grupos foi mediante a utilização de um valor de referência, sendo a diferença entre a marca de maior e menor efeito, divido pelo número de grupos, conforme a seguinte equação:

$$VR = \frac{EMaior - EMenor}{Nº\ Grupos} \quad \text{(Eq. 12)}$$

Em que, VR é o Valor de Referência; EMaior se refere ao maior efeito dentre todos os marcadores; EMenor se refere ao menor efeito e Nº Grupos é a quantidade de grupos que se deseja classificar.

Tomando como exemplo a marca de maior efeito do SNP X com 6,25, e por outro lado o SNP Y com efeito -3,5, o VR estimado, considerando a formação de quatro grupos, será de 2,43. Logo, todos os SNPs que possuem seus efeitos de

marcadores entre -3,5 a -1,08 estarão classificados no Grupo 1 (G1), entre -1,08 a 1,37 no Grupo 2 (G2), entre 1,37 a 3,80 no Grupo 3 (G3) e entre 3,80 a 6,25 no Grupo (G4).

Para verificar a variabilidade genética entre os genitores de todas as combinações e a as suas correlações com a σ_p^2, foi realizada a distância genética via marcadores moleculares considerando todos os 1752 SNPs (Ga), e, de maneira separada dentro de cada grupo gerado a partir da partição dos efeitos dos marcadores.

A metodologia utilizada para a estimação da distância genética foi a matriz de variância-covariância de parentesco proposta por VanRaden (2008), em que a maior dissimilaridade genética entre os genitores é expressa por valores inferiores, portanto, quanto mais negativo, mais divergência genética entre as linhagens. Dentro de cada grupo gerado, as marcas foram separadas com base no número cópias do alelo dominante, e em seguida foram contabilizados a quantidade que estavam fixados no genoma para alelos dominantes (2), recessivos (0) e os em heterozigosidade (1), e, as respectivas somas dos efeitos destes marcadores gerados pela predição genômica.

2.6. Índice de seleção Mulamba e Mock

Com a finalidade de desenvolver uma nova metodologia para seleção de cruzamentos potenciais à geração de novas cultivares, foi proposto um índice de ranqueamento com base na soma dos efeitos de alelos em homozigose e em heterozigosidade, de cada partição, sendo possível encontrar aquelas combinações que apresentam os melhores alelos para cada região do genoma.

A metodologia do índice de ranqueamento de postos de Mulamba e Mock (1978) foi aplicado na soma dos efeitos dos marcadores. A classificação do ranqueamento dos cruzamentos para a soma dos efeitos dos marcadores fixados dominantes (2) e os em heterozigosidade (1) foram realizados na ordem decrescente. Por outro lado, a soma dos efeitos dos marcadores fixados recessivos (0) se ordenou de maneira crescente. Os cruzamentos selecionados foram aqueles que apresentaram as menores soma de postos. Os pesos utilizados foram iguais para todas as variáveis.

2.7. População de validação

Para a validação da metodologia, dentre as 5151 combinações possíveis, 10 foram realizadas e avançadas até as suas respectivas populações $F_{2:4}$, e avaliadas em teste de progênies na safra 2017/2018, gerando assim uma população de validação (PV) para a metodologia (dados também cedidos GDM Seeds do Brasil S.A.). A definição destes genitores que pertenceram aos cruzamentos da PV foram definidos pelo melhorista com base nas suas características fenotípicas de produtividade per se, além do GMR e aspecto de planta. O número de genótipos avaliadas dentro de cada população variou devido a disponibilidade de sementes e a demanda do melhorista responsável, conforme Tabela 5. O número de progênies selecionadas em cada cruzamento foi realizado com base na produtividade e no grupo de maturidade relativa, não sendo selecionado progênies com GMR abaixo de 6.5 e superior a 7.8.

Tabela 5 – Cruzamentos realizados para o processo de validação, grupo de maturidade relativa dos genitores, número de progênies avaliadas e selecionadas com diferentes taxas de seleção.

Cross	GMR* Genitores	Nº Progênies $F_{2:4}$	Nº Progênies Selecionadas	% Seleção
L636 x L453	69 x 78	102	23	23%
L173 x L955	66 x 71	62	19	31%
L528 x L085	75 x SI**	34	7	21%
L085 x L141	SI x 74	82	7	9%
L908 x L530	72 x 63	198	5	3%
L161 x L708	65 x 68	237	5	2%
L161 x L141	65 x 74	60	4	7%
L908 x L141	72 x 74	72	2	3%
L879 x L908	65 x 72	29	1	3%
L453 x L421	78 x 68	90	1	1%

*GMR: Grupo de maturidade relativa; **SI: Sem informação.

2.8. Softwares e scripts

Todas as simulações e análises foram realizadas no software R (R Core Team 2019, utilizando combinações de funções próprias e pacotes como rrBLUP (Endelman, 2011), PopVar (Mohammadi et al., 2015), R/qtl (Broman et al., 2003) e snpReady (Granato et al., 2018).

3. RESULTADOS

A média dos melhores estimadores não viesados (BLUE) para produtividade de plantas obtida nas 102 linhagens foi de 3972 kg ha^{-1}, com variação de 4729 kg ha^{-1} para a L1367543 e de 3149 kg ha^{-1} para L2489400, englobando uma amplitude de aproximadamente 1650 kg ha^{-1}. A acurácia preditiva média estimada pela validação da predição genômica, mediante a utilização do RR-BLUP, foi de 0,60.

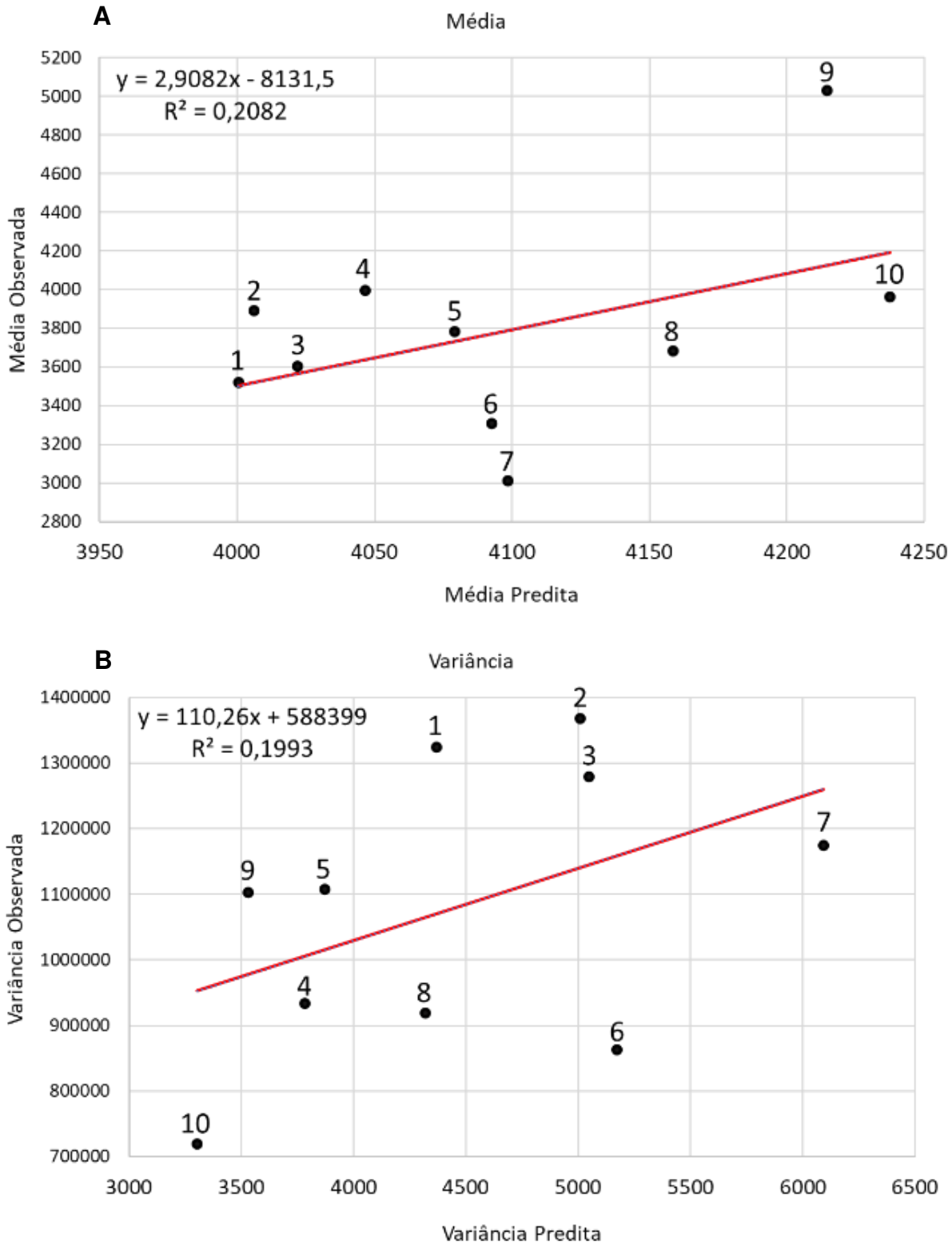
Considerando as 5151 combinações possíveis, foi observado a presença da linhagem L1367543 em 14 dos 20 melhores cruzamentos para extração de linhagens superiores tendo como base o ranqueamento por μ_{Best}, apresentando uma alta frequência de alelos favoráveis em seu genoma.

Mantendo o mesmo ranqueamento (por μ_{Best}), dentre as combinações biparentais da PV, com exceção do L636 x L453, os demais foram classificados dentro dos 50% melhores dentro do grupo dos 5151 possíveis. Com destaque para a combinação L173 x L955 e L528 x L085 que englobam no grupo dos 10% melhores cruzamentos preditos, ranqueados respectivamente nas posições 434 e 506.

O comportamento dos 10 cruzamentos do grupo de validação entre os dados observados pelas progênies $F_{2:4}$ e as estimativas de predição foram diferentes, conforme observado nas Figuras 9A e 9B. A correlação entre observado e predito para os cruzamentos foi similar tanto para a média, como para a variância, com estimativas próximas a 0,44.

O cruzamento L528 x L085 apresentou a maior média observada de suas progênies $F_{2:4}$ com 5030 kg ha^{-1}, enquanto L161 x L708 com 3011 kg ha^{-1} obteve a menor média observada. Com relação as médias preditas, L173 x L955 e L636 x L453 apresentaram respectivamente as maiores e menores predições, com estimativas variando de 4237 kg ha^{-1} e 4000 kg ha^{-1}. Para a variância observada, a média obtida foi de 1079066 kg^2 ha^{-2}, variando de 1367784 kg^2 ha^{-2} para L161 x L141 e 719396 kg^2 ha^{-2} para L173 x L955. O cruzamento L161 x L708 apresentou a maior variância predita com 6091 kg^2 ha^{-2}, e L173 x L955 com a menor predição de 3303 kg^2 ha^{-2} (Figuras 9A, 9B e Tabela 6).

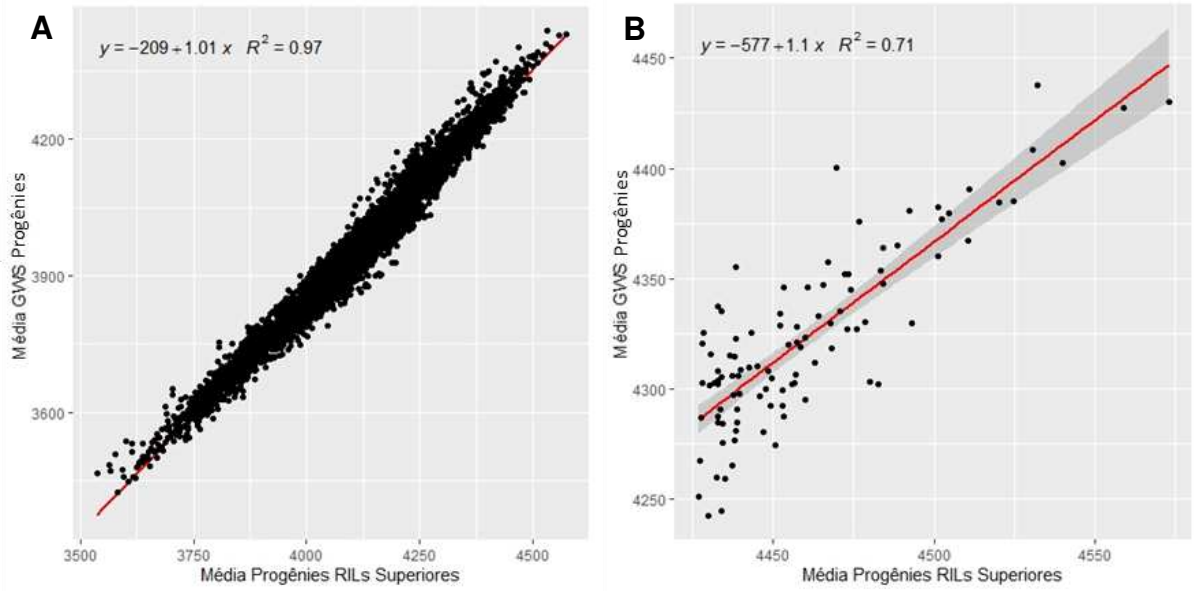
Figura 9 – A: Scatterplot entre a média observada das progênies avaliadas e a média predita de todas as progênies RILs (μ_p). B: Scatterplot entre a variância observada das progênies avaliadas e a variância predita de todas as progênies RILs. 1: L636 x L453; 2: L161 x L141; 3: L879 x L908; 4: L908 x L530; 5: L908 x L141; 6: L453 x L421; 7: L161 x L708; 8: L085 x L141; 9: L528 x L085; 10: L173 x L955.



Fonte: Silva (2022).

Foi observado uma correlação de alta magnitude entre μ_P e μ_{Best}, com estimativa de 0,98 considerando todos os 5151 cruzamentos possíveis e de 0,84 entre os 100 cruzamentos mais promissores com base em μ_{Best} (Figuras 10A e 10B). Devido a estas correlações de alta magnitude foi observado uma não alteração no ranqueamento entre os 10 cruzamentos da PV, quando comparados os dois parâmetros preditos.

Figura 10 – A: Scatterplot entre a média observada das progênies RILs Superiores preditas e a média predita das 200 progênies, para todos os 5151 cruzamentos. B: Scatterplot entre a média observada das progênies RILs Superiores preditas e a média predita das 200 progênies, considerando apenas os 100 melhores cruzamentos com base em progênies superiores.



Fonte: Silva (2022).

Destacando apenas as combinações da PV, e, mediante a utilização da metodologia por progênies simuladas e realizando um ranqueamento com base nas médias das melhores progênies dentro de cada cruzamento, o cruzamento L173 x L955 apresentou a maior probabilidade de encontrar genótipos superiores, com média de 4352 kg ha^{-1}. Considerando uma seleção de apenas as cinco melhores combinações, 80% dos melhores cruzamentos preditos apresentaram um maior número de progênies selecionadas. O único cruzamento não selecionado por esta metodologia foi o L636 x L453, sendo este o que obteve o maior número de progênies selecionados (Tabela 6).

Tabela 6 – Ranqueamento dos 10 cruzamentos biparentais da população de validação nas diferentes metodologias aplicadas, tanto fenotípicas, quanto de predição.

Cross(R)	Rank			
	Fenotípico		Predito	
	Número total de progênies Selecionadas	Porcentagem de progênies Selecionadas	μ_{Best}	Mulamba
L636 x L453 (2669)	1	2	10	3
L173 x L955 (434)	2	1	1	1
L528 x L085 (506)	3	3	2	2
L085 x L141 (875)	4	4	3	5
L161 x L708 (1283)	5	9	4	9
L908 x L530 (2224)	6	8	7	7
L161 x L141 (2456)	7	5	9	10
L908 x L141 (1794)	8	7	6	4
L879 x L908 (2336)	9	6	8	8
L453 x L421 (1310)	10	10	5	6

(R) – Ranqueamento da predição via progênies simuladas e classificadas segundo as melhores progênies superiores, considerando todas as 5151 combinações possíveis.

Diferentemente da estratégia anterior, de comparar com base no número de progênies avançadas de cada cruzamento, também foi realizado uma comparação com base na porcentagem de progênies selecionadas conforme Tabela 5. A coincidência entre predito e validação reduziu para 60%, onde os cruzamentos L636 x L453 e L161 x L141 não foram selecionados (Tabela 6).

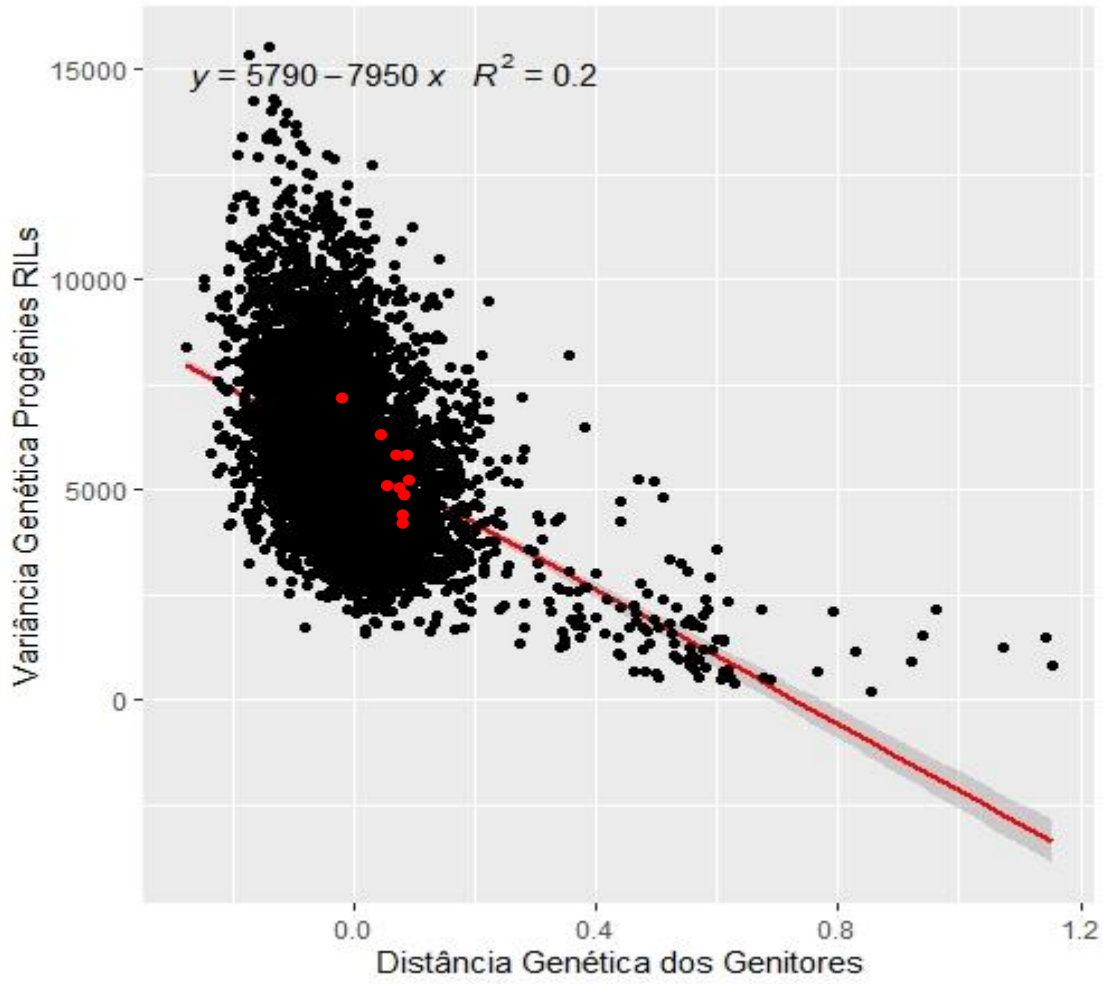
Os marcadores de maior e menor efeito genético apresentaram respectivamente valores de 5,75 e -6,73 respectivamente. Assim, o VR encontrado para a constituição dos grupos particionados foi de aproximadamente 3,12. O número de marcadores fixados e não fixados e as suas respectivas somas dos efeitos em cada cruzamento da PV estão apresentados na Tabela 7.

Dentre as combinações da PV, aquele com a maior dissimilaridade genética considerados todos os 1752 marcadores foi L161 x L708 (-0,139), ocasionada principalmente pela alta divergência genética no G1 devido 24 marcas em heterozigose, o que gerou uma soma de mais de 108 kg ha^{-1}, sendo este um dos motivos para a alta variabilidade genética encontrada na predição. Os dois melhores cruzamentos, L173xL955 e L528xL085, apresentaram, principalmente nos grupos 3 e 4, alta frequência de alelos favoráveis fixados, acarretando uma maior média fenotípica às progênies derivadas destas combinações (Tabela 7).

Considerando a seleção dos cinco melhores cruzamentos, o ranqueamento pela metodologia de índice de seleção de Mulamba e Mock também apresentou uma coincidência de 80% entre o predito e com base nas combinações com maiores progênies avançadas, e com base na porcentagem de seleção. A correlação com os cruzamentos com maior número de progênies avançadas foi superior perante a metodologia de progênies simuladas, aumentando de 0,24 para 0,62. O cruzamento L636 x L453 seria realizado se aplicado este método, saindo da última posição para a terceira. O fator decisivo para a seleção desta combinação foi a melhor captação das marcas heterozigotas destes cruzamentos, e consequentemente a sua variância genética predita. As demais combinações selecionadas estão apresentadas na Tabela 6.

A correlação entre a distância genética total e a variância genética predita foi -0,44 considerando todos os 5151 cruzamentos. Para a população de validação, a correlação observada foi de -0,76 (Figura 11). Por outro lado, a correlação entre a distância genética total e a variância observada foi de baixa magnitude, com estimativa de -0,18. A correlação entre a variância observada e a soma dos efeitos dos marcadores em heterozigosidade do G4 foi de 0,57, demonstrando a relevância deste grupo para a captação da real variabilidade das futuras progênies dos cruzamentos.

Figura 11 - Scatterplot entre a distância genética total (Ga) e a variância genética das progênies RILs (σ_p^2), para todas as 5151 combinações. Em vermelho os 10 cruzamentos da população de validação.



Fonte: Silva (2022).

Tabela 7 – As 10 combinações híbridas ranqueadas com base no índice de Mulamba e Mock. E seus respectivos valores de variância das RILs (σ_p^2), distância total (Ga), distância dentro de cada agrupamento, além do número de SNPs e estimativas dos efeitos das marcas para os alelos em homozigose e heterozigosidade em cada Grupo particionado.

Rank	Cross	σ_p^2	Ga	G1				G2			
				Dist	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$	Dist	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$
1	L173 x L955	3304	0,048	0,151	-28,0$^{(6)}$	-56,1$^{(12)}$	-40,7$^{(10)}$	0,106	-64,1$^{(37)}$	-346,2$^{(209)}$	-438,1$^{(324)}$
2	L528 x L085	3532	0,062	0,182	-25,9$^{(6)}$	-67,4$^{(14)}$	-31,6$^{(8)}$	0,115	-68,7$^{(37)}$	-353,0$^{(212)}$	-426,7$^{(321)}$
3	L636 x L453	4370	-0,002	0,087	-13,3$^{(3)}$	-62,5$^{(14)}$	-49,1$^{(11)}$	0,105	-66,3$^{(37)}$	-353,5$^{(225)}$	-428,5$^{(308)}$
4	L908 x L141	3873	-0,018	-0,031	-29,0$^{(6)}$	-60,8$^{(14)}$	-35,1$^{(8)}$	-0,127	-62,0$^{(34)}$	-344,3$^{(226)}$	-442,0$^{(310)}$
5	L085 x L141	4321	-0,043	-0,091	-32,2$^{(7)}$	-76,8$^{(17)}$	-15,9$^{(4)}$	0,010	-70,6$^{(38)}$	-362,5$^{(227)}$	-415,3$^{(305)}$
6	L453 x L421	5172	-0,011	0,149	-16,7$^{(4)}$	-62,9$^{(14)}$	-45,3$^{(10)}$	0,033	-72,9$^{(40)}$	-315,4$^{(186)}$	-460,1$^{(344)}$
7	L908 x L530	3783	0,040	0,217	-17,7$^{(4)}$	-55,4$^{(12)}$	-51,7$^{(12)}$	-0,038	-53,2$^{(30)}$	-335,8$^{(213)}$	-459,3$^{(327)}$
8	L879 x L908	5047	-0,027	-0,426	-4,4$^{(1)}$	-77,4$^{(17)}$	-43,1$^{(10)}$	-0,018	-67,8$^{(37)}$	-327,9$^{(213)}$	-452,6$^{(320)}$
9	L161 x L708	6091	-0,139	-0,717	0,0$^{(0)}$	-108,7$^{(24)}$	-16,1$^{(4)}$	-0,176	-48,1$^{(29)}$	-360,4$^{(223)}$	-439,9$^{(318)}$
10	L161 x L141	5010	0,040	-0,213	-13,0$^{(3)}$	-90,2$^{(20)}$	-21,6$^{(5)}$	-0,020	-49,5$^{(36)}$	-330,8$^{(201)}$	-468,0$^{(333)}$
Rank	Cross	σ_p^2	Ga	G3				G4			
				Dist	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$	Dist	S0$^{(N)}$	S1$^{(N)}$	S2$^{(N)}$
1	L173 x L955	3304	0,048	0,002	10,0$^{(14)}$	161,3$^{(248)}$	603,9$^{(813)}$	0,118	7,8$^{(2)}$	64,7$^{(18)}$	198,8$^{(59)}$
2	L528 x L085	3532	0,062	0,012	24,8$^{(35)}$	194,0$^{(312)}$	556,4$^{(728)}$	0,198	2,7$^{(1)}$	63,9$^{(17)}$	204,7$^{(61)}$
3	L636 x L453	4370	-0,002	-0,047	19,8$^{(31)}$	277,3$^{(387)}$	478,1$^{(657)}$	-0,222	6,7$^{(2)}$	171,0$^{(48)}$	93,6$^{(29)}$
4	L908 x L141	3873	-0,018	0,064	52,4$^{(58)}$	190,5$^{(290)}$	532,3$^{(727)}$	-0,171	6,8$^{(2)}$	117,1$^{(32)}$	147,4$^{(45)}$
5	L085 x L141	4321	-0,043	-0,076	28,9$^{(34)}$	190,9$^{(299)}$	555,3$^{(742)}$	-0,012	17,5$^{(5)}$	93,8$^{(25)}$	160,0$^{(49)}$
6	L453 x L421	5172	-0,011	-0,040	20,5$^{(30)}$	189,4$^{(269)}$	565,2$^{(776)}$	-0,054	21,7$^{(6)}$	96,2$^{(26)}$	153,4$^{(47)}$
7	L908 x L530	3783	0,040	0,078	30,9$^{(34)}$	240,7$^{(349)}$	503,6$^{(692)}$	0,059	6,3$^{(2)}$	100,3$^{(28)}$	164,7$^{(49)}$
8	L879 x L908	5047	-0,027	-0,009	39,1$^{(44)}$	241,5$^{(338)}$	494,6$^{(693)}$	-0,097	3,0$^{(1)}$	108,0$^{(30)}$	160,4$^{(48)}$
9	L161 x L708	6091	-0,139	-0,102	22,3$^{(24)}$	211,2$^{(302)}$	541,7$^{(749)}$	-0,048	7,5$^{(2)}$	100,8$^{(28)}$	163,1$^{(49)}$
10	L161 x L141	5010	0,040	0,077	46,2$^{(54)}$	172,8$^{(252)}$	556,2$^{(769)}$	0,127	41,6$^{(11)}$	99,2$^{(27)}$	130,5$^{(41)}$

$^{(N)}$ – Número de marcadores

4. DISCUSSÃO

Há duas categorias para a escolha de genitores propostas por Baezinger e Peterson (2001), a primeira relacionada apenas nas informações de performance dos genitores, tendo a predição da média a mais utilizada. Porém, a variabilidade predita do cruzamento não é possível realizar. Para isto, a segunda categoria engloba a análise do desempenho das progênies dos cruzamentos, em que é possível a identificação daquelas combinações com altas médias e variâncias genética para a extração de genótipos superiores. Entretanto, em programas de melhoramento genético de soja, o número de cultivares e linhagens elite que podem ser utilizados como genitores são elevados, gerando milhares de combinações possíveis, e excedendo a capacidade de testes a campo quando se faz uso de metodologias de predição com base em avaliação fenotípica de progênies. Portanto, ferramentas de predições genômicas permitem aos melhoristas explorar de uma maneira sem custo a campo, milhares de combinações antes de serem realizadas, onde é possível a exclusão daquelas que apresentarem uma baixa probabilidade de gerar novas cultivares comerciais, de acordo com os objetivos iniciais do programa de melhoramento, e focar somente nos cruzamentos mais promissores.

Neste estudo, a acurácia de predição encontrada ficou na média perante a outros trabalhos na literatura. Também com a cultura da soja, Smallwood et al (2019) obtiveram acurácia de 0,48 e Jean et al (2021) com excelentes resultados de acurácia com estimativas de 0,70. Considerando uma outra cultura autógama, Akdemir e Sanches (2019) encontraram uma acurácia semelhante à deste trabalho com 0,65. Em culturas de espécies alógamas, como o milho, Beckett et al (2019) obtiveram acurácia de 0,72. O método de imputação de marcadores utilizado foi realizado conforme resultado encontrado por Rutkoski et al. (2013) em que utilizando a média das outras marcas a acurácia preditiva não sofreu grandes alterações perante outras metodologias como regressão de árvore aleatória e decomposição em valores singulares.

Um dos motivos para não se ter encontrado melhores acuraria de predição se deve ao fato de ter algumas linhagens da macrorregião 2 dentre as 102 linhagens avaliadas. Por ser uma cultura de dias curtos, o fotoperíodo é responsável pela indução do florescimento na soja (Silva et al., 2017). Logo, devido a este intercâmbio entre cultivares adaptadas a diferentes regiões sojícolas, e com comportamentos

diferentes ao fotoperíodo, um viés na geração do modelo preditivo e tanto na validade podem ocorrer. Utilizando como exemplo, um genitor da macrorregião 2 com grupo de maturação 6.2, tende a florescer mais rápido quando inserido na macrorregião 3, gerando uma redução no potencial produtivo devido a esta antecipação floral. Porém, em seu genoma, ele apresenta genes de alto potencial produtivo.

O segundo viés está relacionado as progênies oriundas de cruzamento onde um dos parentais é de outra macrorregião, em que se espera uma alta segregação para GMR, e assim, algumas progênies não são avançadas pelo fato de não apresentar o grupo de maturação adaptado ao ambiente avaliado, e não pela razão de não possuir em seu genoma genes para altas produtividades.

Neste presente trabalho, investigamos a seleção de cruzamento com base no desempenho das predições genômicas via simulação de progênies, e por meio de uma nova abordagem mediante a criação de um índice de seleção levando em consideração o particionamento e agrupamento do efeito dos marcadores, além também dos somatórios destes efeitos dentro de cada agrupamento, para cada cruzamento predito. Este particionamento permitiu uma melhor predição da variabilidade genética de cada cruzamento em comparação com a metodologia de progênies simuladas.

Considerando características quantitativas e o número de cruzamentos para a validação semelhante ao deste trabalho, Osthushenrich (2018), e Adeyemo e Bernardo (2019) encontraram correlações entre variâncias observadas e preditas respectivamente de 0,70 para produtividade de cevada e de 0,03 para altura de plantas de milho, sendo este último praticamente sem correlação. Neyhart e Smith (2019) realizaram um trabalho de predição de cruzamentos com base em progênies simuladas, e desenvolveram 27 populações obtidas por cruzamentos biparentais de cevada para a validação, estimando média e variância de progênies obtidas destes cruzamentos para característica quantitativas como altura de plantas e severidade de doença. Houve correlação entre as médias observadas e preditas, mas não para a variância. Neste estudo, a correlação das variâncias observadas e preditas por meio da simulação de progênies foi de 0,44, sendo superior aos trabalhos citados, porém ainda com correlação de média magnitude, sendo necessário um acréscimo para melhores estimativas e assertividade na seleção de cruzamentos.

A predição da média, utilizando métodos clássicos, é possível de se realizar, porém, a predição da variância é mais complexa (Bernardo, 2010). Segundo Adeyemo

e Bernardo (2019) uma das possíveis causas para as baixas correlações na predição da variância utilizando a metodologia de progênies simuladas, se deve ao fato de que cada combinação biparental pode apresentar taxas de recombinações diferentes, alterando assim a variância observada, enquanto para as simulações, se faz uso de apenas uma taxa para todas os cruzamentos. Outro fator que pode contribuir para este viés, está relacionado ao efeito *shrinkage* do modelo RR-BLUP, podendo a variância ser subestimada (Lehermeier, 2017).

Por outro lado, mediante a utilização da metodologia de particionamento das marcas, quando a correlação foi realizada entre a variância observada e a soma dos efeitos dos marcadores em heterozigosidade do G4, que representa os efeitos de maior magnitude positivos, a estimativa foi de 0,57. Este resultado evidenciou a importância de se quantificar a divergência genética entre os cruzamentos em regiões divergentes e de alta relevância (G1 e G4) para a determinação da variabilidade genética.

Nos grupos particionados G2 e G3, devido ao maior número de marcas, principalmente devido ao efeito *shrinkage* de aproximar a estimativa a zero, muitos marcadores podem ir anulando outros, e com isso, podemos não captar da melhor maneira a variabilidade do cruzamento. Está é a razão pela qual muitos cruzamentos genéticos com alta dissimilaridade total não expressam uma alta variabilidade genética, conforme evidenciado pelo cruzamento L085 x L141, que apresentou uma das maiores distancias genéticas, porém uma variância menor que outros cruzamentos, não apresentando em nenhum grupo altos somatórios de marcas em heterozigosidade.

Por outro lado, a combinação L636 x L453, embora não apresentasse uma distância genética total elevada, os seus marcadores heterozigotos dos grupos de marcas de efeito positivos G3 e G4 possuem alto somatório, resultando em uma maior variabilidade ao cruzamento, e com isso o aparecimento de genótipos transgressivos de alto potencial produtivo, conforme a seleção de aproximadamente 23% das progênies avaliadas. Logo, com o uso do índice de Mulamba, foi possível a identificação destes cruzamentos que possuem, além de altos somatórios de alelos fixados positivos, as maiores taxas de alelos heterozigotos em regiões de alto interesse.

Os cruzamentos L173 x L955 e L161 x L141 apresentam estimativas de distância genética total semelhantes, porém tanto as variâncias preditas como as

observadas foram divergentes, pois tanto para o G1, como também para G4 houve uma divergência entre as marcas em heterozigosidade (Tabela 6). Isto evidencia a importância do particionamento e que o uso da distância total pode induzir o melhorista a uma escolha equivocada de bons genitores.

Escolher corretamente os genitores pode representar progresso genético, economia de tempo e de recursos financeiros. O baixo desempenho dos cruzamentos L908 x L530 e L161 x L708 demonstra o alto gasto com recursos que poderiam ter sido evitados se aplicado as metodologias de predições de cruzamentos ou inserido menos progênies para avaliação a campo, pois foram avaliadas respectivamente 198 e 237 progênies $F_{2:4}$ e para ambos os cruzamentos foram avançadas apenas 5 linhagens. Podendo utilizar os recursos disponíveis em outras populações, ocasionando uma melhor distribuição de progênies avaliadas a campo.

Embora o número de coincidência tenha sido similar, a correlação usando a metodologia do índice se mostrou mais adequada quando utilizado os dados de validação. Logo, estudos com um número maior de genitores em validação podem ser realizado para a confirmação de que o uso do índice de Mulamba e Mock se ajusta melhor a uma predição de cruzamentos de cultivares de soja.

5. CONCLUSÃO

Os resultados obtidos neste estudo demostraram que a predição genômica associada com o particionamento dos efeitos dos marcadores proporcionou uma maior assertividade na escolha dos blocos de cruzamentos dentro de um programa de melhoramento genético de soja.

As correlações dos efeitos dos marcadores dos grupos com as maiores magnitudes apresentaram uma maior acurácia com a variância observada, tornando uma nova alternativa para o conhecimento prévio da variabilidade daquele cruzamento.

O índice de seleção por Mulamba e Mock se mostrou eficiente para a seleção de genitores, podendo ser utilizada em conjunto com metodologia de predição via simulação de progênies, como também de maneira individual, ficando a cargo de escolha do melhorista.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADEYEMO, E.; BERNARDO, R. Predicting genetic variance from genomewide marker effects estimated from a diverse panel of maize inbreds. **Crop Science**. 59:583–590. 2019.

AKDEMIR, D.; J. I. SÁNCHEZ. Design of training populations for selective phenotyping in genomic prediction. **Scientific Reports**, 9:1446. 2019.

BAENZIGER, P. S.; PETERSON, C. Genetic variation: Its origin and use for breeding self-pollinated species. In Plant Breeding in the 1990's, edited by H. Stalker and J. Murphy, pp. 69–100, **North Carolina State University**, North Carolina. 1991.

BECKETT, T. J.; ROCHEFORD, T. R.; MOHAMMADI, M. Re-imagining maize inbred potential: identifying breeding crosses using genetic variance of simulated progeny. Crop Science, vol. 59: 1-12. 2019.

BERNARDO, R. **Breeding for quantitative traits in plants**. Stemma Press, Woodbury, second edition, 2010.

BERNARDO, R. Genomewide selection of parental inbreds: Classes of loci and virtual biparental populations. **Crop Science**. 54:2586–2595. 2014.

BROMAN K. W., WU H., SEN Ś., CHURCHILL G. A. R/qtl: QTL mapping in experimental crosses. **Bioinformatics** 19:889-890. 2003.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. Acompanhamento da Safra Brasileira de Grãos, Brasília, DF, v. 9, safra 2021/22, n. 4 quarto levantamento, janeiro. 2022.

ENDELMAN J. B. "Ridge regression and other kernels for genomic selection with R package rrBLUP." *Plant Genome*, **4**, 250-255. 2011.

FEHR, W. Principles of cultivar development. **Theory and technique**. Macmillan, New York. 1987.

GRANATO, I. S. C., GALLI, G., COUTO, E. G. O. snpReady: a tool to assist breeders in genomic analysis. **Molecular Breeding**, Heidelberg, v. 38, n. 8, p. 1-7, 2018.

JEAN, M., COBER, E., O'DONOUGHUE, L, RAJCAN, I., BELZILE, F. Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. **Crop Science**. 61. 10.1002/csc2.20583. 2021.

LADO, B., BATTENFIELD, S., GUZMÁN, C., QUINCKE, M., SINGH, R.P., DREISIGACKER, S., PEÑA, R. J., FRITZ, A., SILVA, P., POLAND, J., GUTIÉRREZ, L. Strategies for Selecting Crosses Using Genomic Prediction in Two Wheat Breeding Programs. **Plant Genome**. 2017.

LEHERMEIER, C., S. TEYSS`EDRE, AND C.-C. SCH¨ON. Genetic gain increases by applying the usefulness criterion with improved variance prediction in selection of crosses. **Genetics Early Onli**: 1–10. 2017.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157: 1819–1829. 2001.

MOHAMMADI, M.; TIEDE, T., SMITH, K. P. Popvar: A genome-wide procedure for predicting genetic variance and correlated response in biparental breeding populations. **Crop Science**. 55: 2068–2077. 2015.

MULAMBA, N.N.; MOCK, J.J. Improvement of yield potential of the Eto Blanco maize (Zea mays L.) population by breeding for plant traits. **Egyptian Journal of Genetics and Cytology**, Alexandria, v.7, p.40-51, 1978.

NASS, L. L.: VALOIS, A. C. C.: MELO, I. S.: VALADARES-INGLIS, M. C. Recursos genéticos e melhoramento – plantas. **Fundação MT**. 2001.

NEYHART, J. L.; SMITH, K. P. Validating genomewide predictions of genetic variance in a contemporary breeding program. **Crop Science**, vol. 59:1062-1072. 2019.

OSTHUSHENRICH, T.; FRISCH, M.; ZENKE-PHILIPPI, C.; JAISER, H.; SPILLER, M.; CSELÉNYI, L.; KRUMNACKER, K.; BOXBERGER, S.; KOPAHNKE, D.; ANTJE HABEKUB, A.; F. ORDON, F.; HERZOG, E. Prediction of means and variance of crosses with genome-wide marker effects in barley. **Frontiers in Plant Science.** 9:1899. 2018.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna. 2019.

RUTKOSKI, J. E., POLAND, J., JANNINK, J. L., SORRELLS, M. E. Imputation of unordered markers and the impact on genomic selection accuracy. **G3 (Bethesda)**. 2013.

SCHNELL F. W.; UTZ H. F. F1-Leistung und Elternwahl in der Züchtung von Selbstbefruchtern, pp. 234–258 In **Bericht über die Arbeitstagung der Vereinigung Österreichischer Pflanzenzüchter**. Gumpenstein, Österreich, 1975.

SILVA, F. L., A. BORÉM, T. SEDIYAMA, AND W. LUDKE. **Soybean Breeding**. 1st ed. Springer, Gewebestrasse, Switzerland, 2017.

SMALLWOOD, C. J.; SAXTON, A. M.; GILLMAN, J.D.; BHANDARI, H.S.; WADL, P. A.; FALLEN, B. D.; HYTEN, D. L.; SONG, Q.; PANTALONE, V. R. Context-Specific Genomic Selection Strategies Outperform Phenotypic Selection for Soybean Quantitative Traits in the Progeny Row Stage. **Crop Science**, vol. 59: 54-67. 2019.

TIEDE, T., KUMAR, L.; MOHAMMADI, M.; SMITH, K. P. Predicting genetic variance in bi- parental breeding populations is more accurate when explicitly modeling the segregation of informative genomewide markers. **Molecular Breeding.** 35: 1–13. 2015.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science** 91: 4414–4423. 2008.

YAO, J.; ZHAO, D.; CHEN, X.; ZHANG, Y.; WANG, J. Use of genomic selection and breeding simulation in cross prediction for improvement of yield and quality in wheat (Triticum aestivum L.). **The Crop Journal**. 6: 353-365. 2018.

ZHONG S.; JANNINK J. L. Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. **Genetics** 177: 567–576. 2007.

CONCLUSÕES GERAIS

A simulação computacional é uma ferramenta extremamente poderosa em programas de melhoramento de soja, pois é capaz de encontrar soluções prévias aos diversos temas relacionados a genética, podendo gerar economia de tempo e recursos as companhias de melhoramento.

A seleção genômica apresentou ser uma metodologia eficaz para a definição das melhores combinações híbridas a serem incorporadas em programas de melhoramento genético de soja.

O emprego do particionamento dos efeitos dos marcadores genéticos após o uso da seleção genômica se mostrou eficiente para a seleção de combinações híbridas complementares, gerando melhores acurarias de predição para as variâncias genéticas de cada cruzamento.

O uso da distância genética com base em todos os marcadores foi pouco informativo para a definição de complementariedade entre os genitores, causando viés em correlações com variância genética predita.

O índice de seleção de Mulamba e Mock mediante a soma de postos do ranqueamento da soma dos efeitos dos marcadores foi eficiente para a escolha de genitores, e seus respectivos cruzamentos. Ocasionando na seleção de combinações que não seriam realizadas tanto pela metodologia clássica de média fenotípica dos pais, como também pela simulação de progênies simuladas.