UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E MELHORAMENTO DE PLANTAS

MODELOS DE PREDIÇÃO GENÔMICA MULTI-AMBIENTAL EM MILHO TROPICAL: PRODUTIVIDADE DE GRÃOS E *STAYGREEN*

AILTON JOSÉ CRISPIM FILHO

Orientadora:
Prof.ª Marcela Pedroso Mendes Resende

Coorientador:
Prof. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Abril – 2023

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [x] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Ailton José Crispim Filho

3. Título do trabalho

MODELOS DE PREDIÇÃO GENÔMICA MULTIAMBIENTAL EM MILHO TROPICAL: PRODUTIVIDADE DE GRÃOS E *STAYGREEN*

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Marcela Pedroso Mendes Resende**, **Professora do Magistério Superior**, em 26/05/2023, às 16:12, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ailton José Crispim Filho**, **Discente**, em 29/05/2023, às 09:15, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3781682** e o código CRC **B166A8D9**.

AILTON JOSÉ CRISPIM FILHO

MODELOS DE PREDIÇÃO GENÔMICA MULTI-AMBIENTAL EM MILHO TROPICAL: PRODUTIVIDADE DE GRÃOS E *STAYGREEN*

Tese apresentada ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Escola de Agronomia, da Universidade Federal de Goiás, como requisito parcial à obtenção do título de Doutor em Genética e Melhoramento de Plantas.

Orientadora:
Prof.ª Dr.ª Marcela Pedroso Mendes Resende

Coorientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia, GO – Brasil
2023

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

Crispim Filho, Ailton José
Modelos de predição genômica multi-ambiental em milho tropical [manuscrito] : Produtividade de grãos e staygreen / Ailton José Crispim Filho. - 2023.
LXXV, 75 f.: il.

Orientador: Profa. Dra. Marcela Pedroso Mendes-Resende; co orientador Dr. Alexandre Siqueira Guedes Coelho.
Tese (Doutorado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-graduação em Genética e Melhoramento de Plantas, Goiânia, 2023.
Bibliografia. Anexos. Apêndice.

1. seleção genômica. 2. Zea mays L.. 3. efeito de dominância. 4. habilidade preditiva. I. Mendes-Resende, Marcela Pedroso , orient. II. Título.

CDU 633

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE AGRONOMIA

ATA DE DEFESA DE TESE

Ata Nº **106** da sessão de Defesa de Tese de **Ailton José Crispim Filho** que confere o título de Doutor em Genética e Melhoramento de Plantas, na área de concentração em Genética e Melhoramento de Plantas**.**

Aos vinte e oito dias do mês de abril de dois mil e vinte e três, a partir das treze horas e trinta minutos, via videoconferencia, realizou-se a sessão pública de Defesa de Tese intitulada **"PREDIÇÃO GENÔMICA DA PRODUTIVIDADE DE GRÃOS E STAYGREEN EM MILHO TROPICAL**". Os trabalhos foram instalados pela Orientadora, Doutora Marcela Pedroso Mendes Resende (EA/UFG), com a participação dos demais membros da Banca Examinadora: Doutor Gustavo Vitti Môro (FCAV/UNESP), membro titular externo; Doutor Cláudio Lopes de Souza Júnior (ESALQ/USP), membro titular externo; Doutor Germano Martins Ferreira Costa Neto (Cornell University), membro titular externo; Doutor Fernando Henrique Ribeiro Barrozo Toledo (CIMMYT), membro titular externo e Doutor Alexandre Siqueira Guedes Coelho (EA/UFG), membro titular interno. Durante a arguição os membros da banca fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese tendo sido o candidato aprovado pelos seus membros. Proclamados os resultados pela Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora.

TÍTULO SUGERIDO PELA BANCA

MODELOS DE PREDIÇÃO GENÔMICA MULTIAMBIENTAL EM MILHO TROPICAL: PRODUTIVIDADE DE GRÃOS E *STAYGREEN*

Documento assinado eletronicamente por **Marcela Pedroso Mendes Resende**, **Professora do Magistério Superior**, em 26/05/2023, às 16:12, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Alexandre Siqueira Guedes Coelho**, **Professor do Magistério Superior**, em 30/05/2023, às 12:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3781684** e o código CRC **EE098656**.

“A verdadeira viagem da descoberta consiste não em achar novas paisagens, mas sim em ver com novos olhos.”

Marcel Proust

Aos meus pais, Iris e Ailton.
DEDICO

AGRADECIMENTOS

É difícil colocar em palavras o meu sentimento ao escrever estes agradecimentos. São tantas pessoas que fizeram e fazem parte desta conquista, momentos únicos que ficarão para sempre em minha memória. Por isso, tentarei expressar aqui, de forma singela, minha gratidão a todos que fizeram parte deste ciclo.

Agradeço minha querida orientadora, prof. ª Marcela Pedroso Mendes Resende, por ter me aceitado, novamente, como seu orientado, pelos seus preciosos ensinamentos para meu crescimento pessoal e profissional ao longo dos cursos de mestrado e doutorado. É inestimável a amizade que construímos ao longo destes últimos anos. Novamente, é uma honra ser seu orientado.

Agradeço ao prof. Alexandre Siqueira Guedes Coelho, meu coorientador, pela paciência e disponibilidade para discutirmos a construção desta tese e os possíveis caminhos que poderíamos seguir. Quaisquer reuniões ou conversas, seja *online* ou presencialmente, sempre foram uma aula para mim.

Agradeço aos professores: Patrícia Guimarães Santos Melo, Sérgio Tadeu Sibov e Bruna Mendes de Oliveira, em nome dos quais agradeço aos demais docentes do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, por toda dedicação e comprometimento com o ensino, convívio saudável e amizade durante a realização deste curso.

Aos inúmeros amigos que fiz durante meu doutoramento no PPGGMP, agradeço imensamente a todos vocês por terem me permitido compartilhar minhas histórias, minhas lutas e, acima de tudo, minhas alegrias. Não poderia deixar de citar aqui alguns nomes, Ikio Watanabe, Antônia Batista, Kleibe Bertoni, Márcio Guedes, Flávio Pereira, Érica Silva, Rodrigo Souza, Jordana de Paula, Angelina Ciappina, Lais Castro, Nayana Costa, Luís Gabriel Alvarenga, Priscila Neves e Juliana Borges, obrigado pela amizade, parceria e momentos vividos.

À Universidade Federal de Goiás, pela oportunidade de realização do curso e excelência em ensino, e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão de bolsa de estudo.

Agradeço aos professores Cláudio Lopes de Souza Júnior (ESALQ/USP) e Roberto Fritsche-Neto (*Louisiana State University*) por cederem gentilmente os dados base para realização do estudo desta tese.

Aos membros da banca examinadora, agradeço, de antemão, pelas valiosas contribuições para melhoria deste trabalho.

À minha família, agradeço aos meus pais, Iris e Ailton, meus avós, Raulina, Irames e Nerci (*in memoriam*), e à minha irmã, Polyanne, pelo amor incondicional, incentivo aos estudos e o ensino dos valores da vida. Um agradecimento super carinhoso para minha sobrinha, Cibelle, que chegou em nossas vidas para confirmar os sentidos de família e amor.

Agradeço ao meu parceiro de vida, Ricardo Nascimento, por nestes últimos tempos ter me dado todo apoio e incentivo para finalização deste trabalho.

A todos aqueles que, de alguma forma, contribuíram para a conclusão deste trabalho, minha sincera gratidão.

Ao final da escrita destes agradecimentos, uma retrospectiva emocionante para mim, agradeço a Deus, por na imensidão de seu amor, estar sempre comigo, ter me permitido mais esta conquista e por ter colocado pessoas tão valiosas e essenciais em minha vida. A Ele, toda honra e toda glória!

Muito obrigado!

SUMÁRIO

RESUMO

CRISPIM-FILHO, A. J. **Modelos de predição genômica multi-ambiental em milho tropical: produtividade de grãos e *staygreen***. 2023. 75 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2023.[1]

Staygreen e produtividade de grãos são caracteres agronômicos de interesse de serem avaliados em programas modernos de melhoramento de milho. Uma abordagem moderna para melhoramento destes caracteres pode ser a seleção genômica, cuja eficiência depende, dentre outros fatores, da escolha adequada do modelo de predição a ser utilizado, dos efeitos que serão contabilizados neste modelo e dos recursos e tempo necessários para o processo de predição dos fenótipos. Neste trabalho, três modelos paramétricos e um modelo não paramétrico foram utilizados na predição genômica multi-ambiental de híbridos simples de milho para *staygreen* e produtividade de grãos, considerando efeitos aditivos, exclusivamente, e em conjunto com efeitos de dominância. Os dados fenotípicos se referem à avaliação de 152 híbridos simples de milho, provenientes do cruzamento de 42 linhagens endogâmicas, avaliados em 13 ambientes para produtividade de grãos e 8 ambientes para *staygreen*. As linhagens foram genotipadas com 13.826 marcadores SNPs (*Single Nucleotide Polymorphism*) pelo o método GBS (*Genotyping by Sequencing*), sendo suas combinações genotípicas utilizadas para gerar os genótipos dos híbridos. As médias ajustadas para cada genótipo, em cada local, foram usadas para treinar os modelos de predição genômica. A habilidade preditiva foi mensurada por meio da correlação de Pearson, obtida por meio do sistema de *ten-fold*. As habilidades preditivas dos modelos variaram de 0,23 a 0,83 para produtividade de grãos e 0,44 a 0,72 para *staygreen*. A inclusão dos efeitos de dominância em todos os modelos paramétricos incrementou as habilidades preditivas para os dois caracteres, sendo que para produtividade de grãos o incremento médio foi de 25%. Isto confirma que a inclusão de efeitos não-aditivos no modelo de predição permite explorar melhor a heterose e ter maior precisão na seleção genômica. Os modelos não diferiram entre atributos vinculados a capacidade preditiva. Devido a menor demanda computacional do GBLUP, ele é o mais indicado para predizer o desempenho fenotípico destes caracteres neste conjunto de dados. A predição com o modelo GBLUP aditivo-dominante indica a possibilidade de seleção de melhores combinações de linhagens do que as já realizadas que, potencialmente, elevam a produtividade de grãos e *staygreen* ao selecionar os melhores 15 híbridos por predição para cada caráter separadamente.

Palavras-chave: seleção genômica, *Zea mays* L., efeito de dominância, habilidade preditiva.

[1] Orientadora: Prof.ª Dr.ª Marcela Pedroso Mendes Resende. EA – UFG.
[1] Coorientador: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA – UFG.

ABSTRACT

CRISPIM-FILHO, A. J. **Multi-environment genomic prediction models in tropical maize: grain yield and staygreen**. 2023. 75 p. Thesis (Doctor of Science in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2023.[1]

Staygreen and grain yield are agronomic traits of interest to be evaluated in modern maize breeding programs. A modern approach to improving these traits can be genomic selection, whose efficiency depends, among other factors, on the proper choice of the prediction model to be used, the effects that will be accounted for in this model and the resources and time required for the prediction process of the phenotypes. In this work, three parametric models and a non-parametric model were used in the multi-environment genomic prediction of single maize hybrids for staygreen and grain yield, considering additive effects, exclusively, and together with dominance effects. The phenotypic data refer to the evaluation of 152 single maize hybrids, from the crossing of 42 inbred lines, evaluated in 13 environments for grain yield and 8 environments for staygreen. The lines were genotyped with 13,826 SNPs (Single Nucleotide Polymorphism) markers using the GBS (Genotyping by Sequencing) method, and their genotypic combinations were used to generate the genotypes of the hybrids. Adjusted means for each genotype at each location were used to train the genomic prediction models. The predictive ability was measured using Pearson's mean correlation, obtained using the ten-fold system. The models' predictive abilities ranged from 0.23 to 0.83 for grain yield and 0.44 to 0.72 for staygreen. The inclusion of dominance effects in all parametric models increased the predictive abilities for both traits, and for grain yield the average increase was 25%. This confirms that the inclusion of non-additive effects in the prediction model allows better exploration of heterosis and greater precision in genomic selection. The models did not differ between attributes linked to predictive ability. Due to the lower computational demand of GBLUP, it is the most suitable to predict the phenotypic performance of these characters in this data set. Prediction with the additive-dominant GBLUP model indicates the possibility of selecting better combinations of inbred lines than those already performed, which potentially increase grain and staygreen productivity by selecting the best 15 hybrids per prediction for each character separately.

Keywords: genomic selection, *Zea mays* L., dominance effect, predictive ability.

[1] Advisor: Prof. Dr. Marcela Pedroso Mendes Resende. EA – UFG.
[1] Co-advisor: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA – UFG.

1 INTRODUÇÃO

O melhoramento genético de milho é direcionado para a obtenção de híbridos que proporcionem maior retorno econômico aos agricultores. O híbrido é resultado do cruzamento entre diferentes indivíduos, geralmente linhagens endogâmicas, cuja complementação genética permite explorar a heterose. Identificar combinações híbridas de alto desempenho é essencial dentro do programa de melhoramento, pois, a partir de poucas linhagens, um número muito grande de híbridos pode ser formado (Technow et al., 2014). A obtenção e avaliação de todos os híbridos possíveis em múltiplos ambientes é um processo bastante oneroso (Beyene et al., 2021; Shikha et al., 2017). Na prática, apenas um pequeno subconjunto dos possíveis híbridos é efetivamente obtido e avaliado em ensaios de campo. Prever o desempenho dos híbridos não obtidos por meio da seleção genômica é uma alternativa bastante atrativa para os melhoristas de milho (Oliveira et al., 2020).

Bernardo (1994, 1996) propôs utilizar o procedimento BLUP (*Best Linear Unbiased Prediction*) na predição do desempenho de híbridos não testados ao explorar a covariância genética entre eles e os híbridos testados a campo. Meuwissen et al. (2001) propuseram uma metodologia semelhante de seleção por meio de abordagens Bayesianas (Bayes A e Bayes B) utilizando denso número de marcadores distribuídos pelo genoma e predizendo o efeito de cada marca na expressão do fenótipo. Estas abordagens de seleção de indivíduos com base em marcadores moleculares amplamente espalhados pelo genoma foram denominadas de *Genome-Wide Selection* (GWS) ou, simplesmente, *Genomic Selection* (GS).

A GS incorpora os efeitos de marcadores moleculares a modelos de predição para obter os *Genomic Estimated Breeding Values* (GEBVs). Genótipos com valores interessantes de GEBV podem ser selecionados em fases precoces de programas de melhoramento (Cui et al., 2020; Meuwissen et al., 2001). Lorenzana & Bernardo (2009) afirmam que a GS reduz pela metade o tempo de seleção quando comparada com a seleção fenotípica para diferentes caracteres em milho. A utilização da GS torna o processo de identificação e seleção de híbridos de alto desempenho mais dinâmico, rápido e com melhor custo-benefício.

Os primeiros modelos de predição genômica foram baseados no modelo infinitesimal (Fisher, 1918), que dita a ação aditiva de muitos genes, cada um com efeitos pequenos. Neste contexto, apenas os efeitos aditivos das marcas eram considerados na análise, uma vez que, por teoria, estes são transmitidos diretamente dos genitores para as progênies (Crossa et al., 2010, 2011). Contudo, dependendo da arquitetura genética do caráter, a contabilização de efeitos não aditivos, como os efeitos de dominância e epistasia, pode aumentar a precisão das predições dos modelos genômicos ao explorar a heterose, principalmente no milho (Alves et al., 2021; Wang et al., 2018).

Vários modelos foram desenvolvidos e categorizados em regressões paramétricas e não paramétricas (Desta & Ortiz, 2014). Os modelos BLUP genômico (GBLUP) (Meuwissen et al., 2001; VanRaden, 2008), *Ridge Regression* BLUP (RR-BLUP) (Meuwissen et al. 2001; Piepho, 2009), *Least Absolute Shrinkage and Selector Operator* (LASSO) (Li & Sillanpää, 2012; Usai, 2009) são exemplos de modelos paramétricos baseados em abordagem frequentista. Bayes A e Bayes B (Meuwissen et al. 2001), *Ridge Regression* Bayesiano (de los Campos et al., 2013), LASSO Bayesiano e Bayes C (de los Campos et al., 2009), e Bayes Cπ e Bayes Dπ (Habier et al., 2011) formam outro conjunto de modelos paramétricos, porém, baseados na abordagem Bayesiana. Os modelos não paramétricos são estabelecidos com base em técnicas de *Machine Learning*, como *Random Forest* (Chen & Ishwaran, 2012; González-Recio & Forni, 2011), *Support Vector Machines* (Maenhout et al., 2007) e redes neurais (González-Camacho et al., 2012). Todo este conjunto de modelos de predição foi proposta para lidar com a diversidade de caracteres de interesse dos melhoristas, já que a arquitetura e controle genético destes caracteres são bastante distintos.

Independentemente do caráter em análise pela GS, a adoção de práticas de fenotipagem com alta qualidade experimental se faz essencial para que bons modelos possam ser obtidos. Proceder a avaliação dos genótipos em mais de um ambiente permite testar a estabilidade e adaptabilidade dos indivíduos frente a diferentes condições ambientais ou de manejo, além de permitir modelar a o efeito da interação dos genótipos com ambientes (G x E). Dependendo da magnitude desta interação, pode ocorrer redução na habilidade preditiva dos modelos de predição genômica quando seus efeitos são ignorados (Mendes & Souza Júnior, 2016).

Para caracterizar mais precisamente o desempenho fenotípico de indivíduos, recomenda-se a utilização de ensaios multi-ambientais que tentam captar e avaliar o impacto da interação genótipos x ambientes (G x E) (Jarquín et al., 2021). Na presença

de interação G x E significativa, a acurácia preditiva dos modelos de predição genômica decai significativamente quando o modelo é calibrado em um ambiente e validado em outro, mesmo utilizando a mesma população de estudo (Mendes & Souza Júnior, 2016). Informações de similaridade dos locais de teste e da presença ou não de interação G x E indicam se a melhor forma de implementar a GS para determinado conjunto de híbridos e ambientes deve ser por meio de modelos ambiente-específicos ou de um modelo geral, que tem boa performance em todo conjunto de ambientes. É importante verificar estas alternativas para uso eficiente da GS.

A produtividade de grãos e *staygreen* são caracteres essenciais de serem avaliados em programas de melhoramento de milho. A produtividade de grãos é controlada por muitos QTLs de pequeno efeito que interagem fortemente com estímulos ambientais gerando uma distribuição contínua de fenótipos e é o principal caráter de interesse dos melhoristas. *Staygreen*, apesar de ainda apresentar arquitetura genética pouco compreendida, também apresenta herança quantitativa, mas controlado por poucos QTLs de efeitos maiores (Sekhon et al., 2019) e está relacionado à manutenção do enchimento de grãos no último estágio de maturação do milho, à tolerância a estresses bióticos e abióticos, incluindo tolerância à seca e ao acamamento de plantas (Belícuas et al., 2014; Caseys, 2019; Luche et al., 2015).

A bateria de modelos existentes em GS e as pressuposições específicas de cada um foram propostas para lidar com a arquitetura, controle genético distintos de caracteres de interesse agronômico e suas possíveis interações com fatores ambientais. Segundo Desta & Ortiz (2014), caracteres controlados por poucos QTLs (*Quantitative Trait Loci*) de efeitos maiores, são mais adequados para regressões Bayesianas, enquanto caracteres com distribuições contínuas dos fenótipos, altamente poligênicos, com numerosos QTLs de pequeno efeito seriam melhor modelados pelas regressões lineares (GBLUP, RR-BLUP etc.). Contudo, com a evolução e ampla divulgação da GS, nota-se que isto não é regra. Estudos apontam modelos de regressões bayesianas e de abordagem de *machine learning* com maior robustez para predição genômica da produtividade de grãos do que os modelos de regressão lineares (Zhang et al., 2020; Kaler et al., 2022).

Por esse motivo, o esforço operacional de avaliação de modelos visando otimizar a GS para uma dada característica, e num dado germoplasma, numa dada rede experimental. Assim, objetivou-se avaliar o potencial da seleção genômica multi-ambiental para caracteres agronômicos num conjunto de híbridos simples de milho oriundo de germoplasma tropical.

2 EFEITOS DE DOMINÂNCIA EM MODELOS DE PREDIÇÃO GENÔMICA PARA *STAYGREEN* EM HÍBRIDOS DE MILHO TROPICAL

RESUMO

Staygreen é a capacidade da planta em retardar sua senescência, de forma que ela continua realizando fotossíntese mesmo quando a umidade dos grãos já está mais baixa. Com isso, há um aumento no acúmulo de biomassa e, consequentemente, na produtividade. A arquitetura genética do *staygreen* é pouco compreendida em milho, e não se sabe o quão importante é a heterose em sua expressão. Em se tratando da seleção genômica, é importante verificar o impacto da contabilização de efeitos de dominância na habilidade preditiva de modelos de predição. Neste trabalho, três modelos paramétricos e um modelo não paramétrico foram utilizados na predição genômica multi-ambiental de híbridos simples de milho para *staygreen*, considerando efeitos aditivos, exclusivamente, e em conjunto com efeitos de dominância. Os dados fenotípicos se referem à avaliação de 152 híbridos simples de milho, provenientes do cruzamento de 38 linhagens endogâmicas com 4 linhagens testadoras, avaliados em 8 ambientes (combinação de 5 locais em diferentes anos de plantio). As linhagens foram genotipadas com 13.826 marcadores SNP pelo o método GBS, sendo suas combinações genotípicas utilizadas para gerar os genótipos dos híbridos. As médias ajustadas para cada genótipo, em cada local, foram usadas para treinar os modelos de predição genômica. A habilidade preditiva foi mensurada por meio da correlação de Pearson, obtida por meio do sistema de *ten-fold*. As habilidades preditivas dos modelos variaram de 0,44 a 0,72. O aumento do número de ambientes para treinamento, melhora os valores de habilidade de predição. A inclusão dos efeitos de dominância nos modelos paramétricos incrementou sutilmente as habilidades preditivas para *staygreen*, o que sugere que a inclusão de efeitos não-aditivos no modelo de predição permite explorar certo nível de heterose e ter maior precisão na seleção genômica. Considerando o modelo de predição GBLUP aditivo-dominante, verifica-se que é possível obter novos híbridos com menores notas de *staygreen* dentro deste conjunto de linhagens, além daqueles já formados.

Palavras-chave: seleção genômica, *Zea mays* L.; habilidade preditiva; dominância; BLUE.

2.1 INTRODUÇÃO

Plantas com maior s*taygreen* são capazes de estender seu período fotossintético, aumentar a capacidade de acúmulo de biomassa e, consequentemente, apresentar maior produtividade (Zhang et al., 2019). No milho, este caráter está

relacionado principalmente com a manutenção do enchimento de grãos no último estágio de maturação e também à tolerância a estresses bióticos e abióticos, incluindo tolerância à seca e ao acamamento de plantas (Belícuas et al., 2014; Casey, 2019; Luche et al., 2015). Assim, maior *staygreen* é visto como um dos fatores que fazem com que os híbridos de milho modernos sejam mais produtivos que seus antecessores (Chibane et al., 2021; Zhang et al., 2019).

Apesar da imensa importância do *staygreen* em milho, ainda são poucos os estudos genômicos realizados para mapear QTLs (*Quantitative Trait Locis*) e entender a dinâmica de herança deste caráter (Kamal et al., 2019); e mais sútil ainda, estudos que envolvam seleção genômica (GS, *genomic selection*) (Kadam et al., 2016). Belícuas et al. (2014) e Sekhon et al. (2019) sugerem que *staygreen* possui herança genética quantitativa, porém, controlado por poucos QTLs de efeitos maiores que interagem com o ambiente para expressão. Neste cenário de um caráter poligênico com interações com ambiente, um *pipeline* moderno para o melhoramento de *staygreen* em milho deve ser realizado por meio da GS.

A GS é uma abordagem preditiva que usa marcadores de baixo custo e abundantes para identificar os melhores genótipos em uma população. Simulações e estudos empíricos demonstraram que a GS pode acelerar expressivamente o processo de melhoramento, manter a diversidade e aumentar o ganho genético quando comparada com a seleção fenotípica ou abordagens de mapeamento de QTL (Bernardo, 2016; Crossa et al., 2017; Heslot et al., 2012). A vantagem principal da GS é que os dados genotípicos obtidos a partir de sementes ou plântulas podem ser usados para predizer o desempenho fenotípico de indivíduos adultos sem a necessidade de fenotipagem ao longo de anos e ambientes, agregando rapidez ao processo de desenvolvimento de populações melhoradas (Bhat et al., 2016).

A bateria de modelos existentes em GS e as pressuposições específicas de cada um foram propostas para lidar com a arquitetura e controle genético distintos de caracteres de interesse agronômico. Segundo Desta & Ortiz (2014), caracteres controlados por poucos QTLs de efeitos maiores, são mais adequados para regressões Bayesianas, enquanto caracteres com distribuições contínuas dos fenótipos, altamente poligênicos, com numerosos QTLs de pequeno efeito seriam mais bem modelados pelas regressões lineares (GBLUP, RR-BLUP etc.). Contudo, as indicações de modelos considerando somente à natureza genética dos caracteres pode ser deficiente (Li et al., 2020).

Além da arquitetura e controle genético, ao se tratar da cultura do milho, um fenômeno relevante que ocorre na obtenção de híbridos é a heterose. Como o *staygreen* em milho possui arquitetura genética pouco compreendida e não se sabe o quão importante é a heterose em sua expressão, verificar o impacto da contabilização de efeitos de dominância na habilidade preditiva de modelos de GS passa a ser essencial (Kadam et al., 2016).

Neste sentido, os objetivos deste estudo foram: (1) verificar a estrutura de populações presente no conjunto de híbridos e das linhagens genitoras; (2) avaliar a *performance* de modelos paramétricos e não paramétrico sem a incorporação dos efeitos de dominância e dos paramétricos com a incorporação deste efeito para lidar com *staygreen*; (3) selecionar o melhor modelo para identificação das combinações híbridas superiores para este mesmo caráter; e (4) estimar o ganho potencial em *staygreen* com a aplicação da seleção genômica no conjunto possível de híbridos.

2.2 MATERIAL E MÉTODOS

2.2.1 Materiais vegetais e ensaios de campo

Foi utilizado um painel de 152 híbridos simples de milho obtidos do cruzamento de 38 linhagens endogâmicas elite com 4 linhagens testadoras. Os cruzamentos das linhagens foram realizados de modo que todas estivessem representadas nesta amostra de 152 híbridos simples. As linhagens são originadas de diferentes populações (IG-1, IG-2 e CMS-05) e híbridos comerciais (HS-1, XL-560 e BR-201). As quatro linhagens previamente selecionadas como testadoras também são endogâmicas, provenientes de diferentes populações e possuem germoplasma elite (Aguiar et al., 2003). Todas as linhagens e híbridos foram obtidos no Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz”, ESALQ/USP, Piracicaba-SP.

Os 152 híbridos foram avaliados juntamente com outros 104 híbridos em látice simples 16x16 em 8 ambientes, sendo que cada ambiente corresponde a uma combinação local x ano. Os locais de avaliação foram as Estações Experimentais: Estação Areão, Estação Caterpillar e Departamento de Genética da ESALQ/USP, nos anos agrícolas de 2002/2003, 2003/2004 e 2004/2005; e Estação Anhembi, nos anos agrícolas 2003/2004 e 2004/2005; todas no estado de São Paulo. O espaçamento utilizado foi de

0,80m entre linhas e 0,20m entre plantas, sendo as parcelas constituídas de uma linha de 4 m de comprimento e o estande de 20 plantas por parcela (62.500 plantas ha^{-1}).

A partir do 110º dia após a semeadura, de três em três dias, foi avaliado o desenvolvimento da "camada preta" nos grãos (*black layer*) para determinação da maturidade fisiológica dos ensaios, indicando a correta época de avaliação de *staygreen*. A retirada dos grãos para esta avaliação foi realizada em espigas da bordadura dos ensaios, cujo híbrido utilizado apresentava ciclo semelhante ao das linhagens genitoras dos híbridos dos experimentos. Assim, aos 120 dias após a semeadura dos experimentos, avaliou-se o *staygreen* em dez plantas competitivas por parcela avaliadas visualmente com uma escala de notas de 1 (maior *staygreen*) a 5 (menor *staygreen*). A nota "1" foi atribuída a plantas com todas as folhas acima da espiga, pelo menos duas folhas abaixo da espiga e os caules verdes; "2" para plantas com todas as folhas acima da espiga e os caules verdes; "3" às plantas com duas folhas acima da espiga senescentes e as demais verdes, independentemente da cor dos caules; "4" para plantas com duas folhas verdes acima da espiga e caules senescentes; e "5" as plantas com todas as folhas e caules senescentes. Além da camada preta dos grãos, o florescimento feminino dos híbridos, registrado como o número de dias desde a semeadura até 50% das plantas na parcela apresentarem estilos-estigmas, foi utilizado para ajustar o *staygreen* por análise de covariância a fim de corrigir possíveis diferenças de maturação. Mais detalhes da avaliação de *staygreen* podem ser encontradas em Belícuas et al. (2014). Todos os dados fenotípicos foram gentilmente cedidos pelo Dr. Cláudio Lopes de Souza Junior, professor titular da ESALQ/USP. Este mesmo conjunto de dados fenotípicos já foi base para estudos de predição genômica em outras abordagens (Mendes & Souza-Júnior, 2016) e *testcrosses* (Alves, 2006).

2.2.2 Análises fenotípicas

As análises foram realizadas agrupando os oito ambientes (combinação x local ano) em 4 locais de avaliação, respectivamente, utilizando o local como critério de agrupamento (Anhembi, Areão, Caterpillar e Departamento de Genética). Dois modelos mistos foram utilizados: (1) para análise individual de *staygreen* em Anhembi, já que só havia um ano de avaliação neste local e (2) para os demais locais.

$$y_{ijk} = \mu + g_i + r_k + b_{j(k)} + \varepsilon_{ijk} \quad (1)$$

em que y_{ijk} é o fenótipo do genótipo *i*, no bloco *j*, na repetição *k*; μ é o intercepto; g_i é o efeito fixo do genótipo *i*; r_k é o efeito aleatório da repetição *k*, com r_k ~ N (0, $\boldsymbol{I}\sigma_r^2$); $b_{j(k)}$ é o efeito aleatório do bloco *j*; na repetição *k*, com $b_{j(k)}$ ~ N (0, $\boldsymbol{I}\sigma_b^2$); e ε_{ijk} é o efeito aleatório não-genético, com ε_{ijk} ~ N (0, $\boldsymbol{I}\sigma_\varepsilon^2$), , em que N refere-se à distribuição normal e $\boldsymbol{I}$ é a matriz identidade.

$$y_{ijkl} = \mu + g_i + a_l + r_{k(l)} + b_{j(kl)} + (ga)_{il} + \varepsilon_{ijkl} \quad (2)$$

em que y_{ijkl} é o fenótipo do genótipo *i*, no bloco *j*, na repetição *k*, no ambiente *l*; μ é o intercepto; g_i é o efeito fixo do genótipo *i*; a_l é o efeito aleatório do ambiente *l*, com a_l ~ N (0, $\boldsymbol{I}\sigma_a^2$); $r_{k(l)}$ é o efeito aleatório da repetição *k*, no ambiente *l*, com $r_{k(l)}$ ~ N (0, $\boldsymbol{I}\sigma_r^2$); $b_{j(kl)}$ é o efeito aleatório do bloco *j*, na repetição *k*, no ambiente *l*, com $b_{j(kl)}$ ~ N (0, $\boldsymbol{I}\sigma_b^2$); $(ga)_{il}$ é o efeito aleatório da interação do genótipo *i* com o ambiente *l*, com $(ga)_{il}$ ~ N (0, $\boldsymbol{I}\sigma_{ga}^2$); e ε_{ijkl} é o efeito aleatório não-genético, com ε_{ijkl} ~ N (0, $\boldsymbol{I}\sigma_\varepsilon^2$), em que N refere-se à distribuição normal e $\boldsymbol{I}$ é a matriz identidade. Também foi realizada a análise de variância do conjunto total de ambientes, ou seja, considerando os oito ambientes para *staygreen*, seguindo o modelo misto (2).

Os valores de BLUEs (*Best Linear Unbiased Estimators*), ou médias ajustadas, dos híbridos em cada local e do conjunto de ambientes foram utilizados como os valores genéticos dos híbridos para as análises de predição genômica. A partir dos BLUEs e posterior ranqueamento em escala ordinal, os coeficientes de correlação de Spearman ($\hat{r}$) foram calculados entre os locais pela fórmula:

$$\hat{r} = \frac{\widehat{COV}_{R(X)R(Y)}}{\sqrt{\hat{\sigma}^2_{R(X)}.\hat{\sigma}^2_{R(Y)}}} \quad (3)$$

em que $\widehat{COV}_{R(X)R(Y)}$ é a covariância entre as os ranques dos BLUEs nos locais X e Y; $\hat{\sigma}^2_{R(X)}$ é a variância dos ranques dos BLUEs no local X; $\hat{\sigma}^2_{R(Y)}$ é a variância dos ranques BLUEs no local Y. A significância dos coeficientes de correlação de Spearman foi avaliada pelo teste t-Student a 5% de probabilidade.

Todas as análises foram realizadas utilizando o ambiente R versão 4.1.0 (R *Development Core Team*, 2021). Para análise de modelos mistos foi utilizado o pacote "lme4" versão 1.1-27.1 (Bates et al., 2015); para obtenção dos BLUEs o pacote "emmeans" versão 1.7.0 (Lenth, 2021) e para a análise de correlação entre os ambientes, o pacote "psych" versão 2.1.9 (Revelle, 2021).

2.2.3 Dados genotípicos

Todas as linhagens foram genotipadas utilizando o protocolo de *Genotyping-by-sequencing* (GBS) (Poland et al., 2012; Sim et al., 2012) na plataforma Illumina NextSeq 500 (Illumina Inc., San Diego, CA, EUA). Para isso, o DNA genômico das linhagens foi extraído de folhas jovens e saudáveis usando o protocolo CTAB (Doyle & Doyle, 1987). As amostras do DNA genômico das linhagens foram digeridas por duas enzimas de restrição (*Pst1-Mse1*) e incluídas em placas de sequenciamento, sendo que os fragmentos de DNA de cada amostra foram ligados a adaptadores com *barcodes* específicos para posterior genotipagem via GBS. Os dados de sequenciamento foram alinhados ao genoma de referência B73 (B73_RefGen_v4) usando o alinhador Bowtie2 (Langmead & Salzberg, 2012). Em seguida, *Single Nucleotide Polymorphisms* (SNPs) foram "chamados" por meio do pipeline GBSv2, disponível no *software* TASSEL 5.0 (Glaubitz et al., 2014). O SNP *dataset* foi filtrado, e marcadores com proporção de alelos identificados (*Call Rate*) menores do que 0,90, não-bialélicos, com frequência alélica mínima (MAF) inferior a 0,05 e com genótipo heterozigótico em pelo menos um genótipo foram removidos do conjunto de dados durante o processo de controle de qualidade. Dados de marcadores perdidos foram imputados pelo *software* Beagle versão 5.0 (Browning & Browning, 2016). Após estes procedimentos de controle de qualidade, um total de 13.826 marcadores SNP de alta qualidade foram considerados para as análises genômicas.

Os genótipos dos híbridos foram obtidos, *in silico*, pelas informações genômicas das linhagens genitoras, para os quais foram atribuídos os valores: 2 para o genótipo homozigótico alternativo, 1 para o genótipo heterozigótico e 0 para o outro genótipo homozigótico. Para exemplificar, supondo duas linhagens, L1 e L2, e quatro marcadores SNP, a construção da matriz dos genótipos dos híbridos foi realizada conforme esquema apresentado na Tabela 2.1.

Tabela 2.1. Exemplo da obtenção da matriz de genótipos dos híbridos.

Marcadores	Linhagens		Híbrido (L1 x L2)
	L1 (Genótipo)	L2 (Genótipo)	(Genótipo)
SNP1	2 (M1M1)	2 (M1M1)	2 (M1M1)
SNP2	2 (M2M2)	0 (m2m2)	1 (M2m2)
SNP3	0 (m3m3)	2 (M3M3)	1 (M3m3)
SNP4	0 (m4m4)	0 (m4m4)	0 (m4m4)

O conjunto de dados genotípicos foi cedido gentilmente pelo Dr. Roberto Fritsche-Neto (Professor Assistente na *Louisiana State University*) e está disponível com o título "*TCGA: a tropical corn germplasm assembly for genomic prediction and high-throughput phenotyping studies*" (Fritsche-Neto et al., 2020).

2.2.4 Estrutura genética de populações

A análise de estrutura de populações foi realizada no conjunto de linhagens e híbridos utilizando o pacote "LEA" versão 3.6.0 (Frichot & François, 2015) implementado no ambiente R (R *Development Core Team*, 2021), considerando os 13.826 SNP. A identificação das *K*s subpopulações foi feita com base no modelo de mistura para obtenção dos coeficientes de mistura individuais. Valores de *K* variando de 1 a 10 foram testados, com 30 repetições independentes para cada agrupamento. Cada solução numérica foi otimizada configurando simulações por Monte Carlo via Cadeias de Markov (MCMC) de 100.000 iterações. O número de grupos *K* que melhor se ajusta ao conjunto de dados foi determinado de acordo com o gráfico de entropia cruzada.

Também foi realizada a análise de distância genética entre as 42 linhagens também entre os 152 híbridos pelo método de Rogers (1972) utilizando os 13.826 SNP por meio do pacote "adegenet" versão 2.1.5 (Jombart, 2008). As distâncias genéticas foram representadas por meio de dendrogramas com o método de agrupamento hierárquico UPGMA (*Unweighted Pair-Group Method using Arithmetic Avarages*), sendo realizadas 100.000 amostragens via *bootstrap* para medir a consistência dos grupos formados. A estimação da correlação cofenética e o teste de Mantel (1967) foram realizados para medir a representatividade do dendrograma em relação as distâncias genéticas obtidas. A matriz de distância genética e o dendrograma foram ilustrados por meio da função *heatmap.2* do pacote "gplots". Os pacotes "adegenet", "gplots" (Warnes et al., 2020) também são implementados no ambiente R.

2.2.5 Predição genômica

2.2.5.1 Modelos aditivos

Três modelos de predição paramétricos (GBLUP, Bayes Cπ e Bayes Dπ) foram usados para estimar os *Genomic Estimated Breeding Values* (GEBVs) de cada

genótipo considerando apenas efeitos aditivos. Os modelos GBLUP, Bayes Cπ e Bayes Dπ foram implementados utilizando o pacote "BGLR" (Pérez & de los Campos, 2014) do *software* R (R *Development Core Team* 2021). O processo de amostragem de Gibbs para os modelos paramétricos foi feito com 10.000.000 de iterações, descartando os primeiros 10.000 resultados como *burn-in* e *thin* de 1000.

GBLUP: O método foi baseado no seguinte modelo linear:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \boldsymbol{Z_{ia}a} + \boldsymbol{\varepsilon} \quad (4)$$

em que $\boldsymbol{y_i} = (y_1, ..., y_n)$ é o vetor de BLUEs e $\boldsymbol{y_i}$ representa a observação do genótipo *i* (*i* = 1, ..., n) em cada local; **1** é um vetor com o mesmo tamanho de $\boldsymbol{y_i}$ com todas as entradas iguais a 1; μ é a média geral; $\boldsymbol{Z}_{ia}$ é a matriz de incidência que conecta os efeitos genéticos aleatórios aos fenótipos; $\boldsymbol{a}$ é o vetor de efeitos genéticos aleatórios de cada híbrido e $\boldsymbol{\varepsilon}$ é o vetor de efeitos aleatórios residuais para cada local. Este modelo assume que a distribuição do vetor $\boldsymbol{a}$ é normal multivariada com média zero e uma matriz de covariância $\sigma^2_{aj}\boldsymbol{G}$, ou seja, $\boldsymbol{a} \sim$ MVN ($\mathbf{0}$, $\sigma^2_{aj}\boldsymbol{G}$), em que σ^2_{aj} é um componente de variância genética no local *j* e $\boldsymbol{G}$ é uma matriz simétrica, semi-positivo-definida remetendo à estrutura de variância-covariância construída a partir dos marcadores SNPs. O modelo também assume que os erros em cada local são independentes com variância homogênea, σ^2_ε; e sendo $\boldsymbol{\varepsilon} \sim$ N ($\mathbf{0}$, $\sigma^2_{\varepsilon j}\boldsymbol{I}$) (onde $\boldsymbol{I}$ é a matriz identidade, e $\sigma^2_{\varepsilon j}$ é a variância residual no local *j*). A matriz $\boldsymbol{G}$ foi calculada por meio da abordagem proposta por VanRaden (2008):

$$\boldsymbol{G} = \frac{\boldsymbol{W_1 W_1'}}{2\sum_{l=1}^{m}\rho_l(1-\rho_l)} \quad (5)$$

em que $\boldsymbol{W_1}$ é a matriz de marcadores de efeito genético aditivo (SNPs), com dimensões do número de indivíduos (*n*) pelo número de locos (*m*), e ρ_l é a MAF do loco *l*.

Neste modelo a distribuição *a priori* para os efeitos dos marcadores pode ser escrita como $\rho(a) = \prod_{l=1,m}\rho(a_l)$, em que $\rho(a_l) \sim$ N (0, σ^2_{a0}), ou seja, cada efeito do marcador segue, a *priori*, uma distribuição normal com uma variância σ^2_{a0} (variância dos efeitos do marcador). O termo "0" implica que o modelo GBLUB possui variância constante entre os marcadores.

Bayes Cπ e Bayes Dπ: Para estas abordagens bayesianas, foi adotado o seguinte modelo linear:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \sum_{l=1}^{m}\boldsymbol{X_{il}g_l} + \boldsymbol{\varepsilon_i} \quad (6)$$

em que $\boldsymbol{y_i} = (y_1, ..., y_n)$ é o vetor de BLUEs e $\boldsymbol{y_i}$ representa a observação no genótipo *i* (*i* = 1, ..., n) em cada local; **1** é um vetor com o mesmo tamanho de $\boldsymbol{y_i}$ com todas as entradas iguais a 1; $\boldsymbol{\mu}$ é a média geral; g_l é o vetor de efeitos aditivos do marcador SNP *l*; $\boldsymbol{X_{il}}$ é a vetor de incidência de cada marcador (assumindo valores de 2, 1 e 0 que correspondem aos genótipos SNPs AA, Aa e aa, respectivamente) e m é o número de marcadores. Estes valores $\boldsymbol{x_{il}}$ representam diretamente o número de cópias do alelo alternativo em cada loco *l*; e $\boldsymbol{\varepsilon_i}$ é o vetor residual, tendo $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \sigma^2_{\varepsilon j}\boldsymbol{I})$ (onde $\boldsymbol{I}$ é a matriz identidade, e $\sigma^2_{\varepsilon j}$ é a variância residual no local *j*). Os modelos Bayes Cπ e Bayes Dπ assumem a mesma distribuição a *priori* para a variância residual, uma distribuição Qui-quadrado invertida escalonada, dada por $\sigma^2_{\varepsilon j}|v_{\varepsilon j},\ S_{\varepsilon j} \sim \chi^{-2}_{v_{\varepsilon j}}(v_{\varepsilon j}, S_{\varepsilon j})$ com hiperparâmetros $v_{\varepsilon j}$ (graus de liberdade) e $S_{\varepsilon j}$ (escala) ambos iguais a 2,0 a *priori* vaga. As diferenças entre estes dois modelos foram as distribuições assumidas a *priori* para os efeitos dos marcadores.

Bayes Cπ e Bayes Dπ possuem um parâmetro adicional π (probabilidade de o efeito do marcador ser igual a zero). Este parâmetro é tratado como desconhecido e atribuído a uma distribuição β a *priori*, tendo $\pi \sim \beta\ (\rho_0, \pi_0)$, com $\rho_0 > 0$ e $\pi_0 \in [0, 1]$ (Perez & de los Campos, 2014). No modelo Bayes Cπ, a distribuição a *priori* para os efeitos de marcador é dada por uma mistura de distribuições normais $g_l|\pi \sim (1 - \pi)\ N\ (0, \sigma^2_g) + \pi N\ (0, \sigma^2_g = 0)$. Este modelo assume a mesma variância genética para todos os marcadores com distribuição a *priori* dada por $\sigma^2_g|v_g, S_g \sim \chi^{-2}_{v_g}(v_g, S_g)$. O modelo Bayes Dπ também assume como *priori* para efeitos de marcador uma mistura de distribuições normais dada por $g_l|\pi \sim (1 - \pi)\ N\ (0, \sigma^2_{gl}) + \pi N\ (0, \sigma^2_{gl} = 0)$, contudo σ^2_{gl} denota que cada SNP tem sua própria variância com distribuição a *priori* dada por $\sigma^2_{gl}|v_{gl}$, $S_{gl} \sim \chi^{-2}_{v_{gl}}(v_{gl}, S_{gl})$.

2.2.5.2 Modelos aditivo-dominantes

A incorporação dos efeitos de dominância foi realizada somente nos modelos paramétricos utilizando também o pacote "BGLR" e as mesmas 10.000.000 de iterações.

Os modelos foram ajustados das seguintes formas:

GBLUP: o modelo utilizado que ajusta simultaneamente aos efeitos aditivos e de dominância dos SNPs foi:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \boldsymbol{Z_{ia}a} + \boldsymbol{Z_{iv}v} + \boldsymbol{\varepsilon_i} \quad (7)$$

em que $\boldsymbol{y_i}$, $\mathbf{1}$, $\boldsymbol{\mu}$, $\boldsymbol{Z_{ia}}$, $\boldsymbol{a}$ e $\boldsymbol{\varepsilon_i}$ representam os mesmos parâmetros definidos no modelo (4), assim como $\boldsymbol{Z}_{ia}$, $\boldsymbol{Z}_{iv}$ é a matriz de incidência que conecta os efeitos genéticos aleatórios aos fenótipos, e $\boldsymbol{v}$ se refere ao vetor dos efeitos de dominância de cada indivíduo. A matriz de covariância dos efeitos de dominância foi *Var* ($\boldsymbol{v}$) = $\boldsymbol{D}\sigma_d^2$, em que $\boldsymbol{D}$ representa a matriz de parentesco genômica-dominante (Vitezica et al., 2013), calculada pelo pacote "snpReady" (Granato et al., 2018), também do *software* R, pela equação:

$$\boldsymbol{D} = \frac{\boldsymbol{W_2}\,\boldsymbol{W_2'}}{4\sum_{l=1}^{m}\rho_l^2(1-\rho_l)^2} \quad (8)$$

em que $\boldsymbol{W_2}$ representa a matriz de marcadores de efeitos genéticos de dominância e *n*, *m* e ρ_l são os mesmos termos definidos na equação (5).

Bayes Cπ e Bayes Dπ: o modelo utilizado que ajusta aos efeitos aditivos e de dominância dos SNPs nas abordagens bayesianas foi:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \sum_{l=1}^{m}(\boldsymbol{X_{il}a_l}) + \sum_{l=1}^{m}(\boldsymbol{W_{il}d_l}) + \boldsymbol{\varepsilon_i} \quad (9)$$

em que $\boldsymbol{y_i}$, $\mathbf{1}$, $\boldsymbol{\mu}$, $\boldsymbol{X_{il}}$ e $\boldsymbol{\varepsilon_i}$ representam os mesmos parâmetros definidos no modelo aditivo (6), $\boldsymbol{W_{il}}$ é a vetor de incidência de cada marcador (assumindo valores de 1 e 0 que correspondem aos genótipos heterozigóticos e homozigóticos, respectivamente) e m é o número de marcadores, $\boldsymbol{a_l}$ é o efeito aditivo e $\boldsymbol{d_l}$ é o efeito de dominância do marcador *l*. Dada a suposição de que a epistasia está ausente, o termo residual no modelo aditivo-dominante contém apenas efeitos não genéticos, enquanto o resíduo do modelo aditivo também inclui desvios de dominância. Nos modelos Bayes Cπ e Bayes Dπ as *priori* estabelecidas para os efeitos aditivos dos marcadores seguem as mesmas do modelo (9). Nestes modelos as distribuições a *priori* para $\boldsymbol{d_l}$ também é uma mistura de normais, dados π_d e σ_d^2 para Bayes Cπ e π_d e σ_{dl}^2 para Bayes Dπ. No entanto, a fim de contabilizar a direcionalidade da dominância, o componente normal da *priori* para $\boldsymbol{d_l}$ tem uma média desconhecida diferente de zero (Almeida-Filho et al., 2019; Zeng et al., 2013), assim:

Para o modelo Bayes Cπ:

$$\boldsymbol{d_l}|\mu_d, \sigma_d^2 = \begin{cases} 0 & \text{com probabilidade } \pi_d \\ \sim N(\mu_d, \sigma_d^2) & \text{com probabilidade } 1-\pi_d \end{cases} \quad (10)$$

em que: $\sigma_d^2|v_d, S_d \sim \chi_{v_d}^{-2}(v_d, S_d)$ e $\pi_d|\rho_0, \pi_0 \sim \beta\ (\rho_0, \pi_0)$.

Para o modelo Bayes Dπ:

$$\boldsymbol{d_l}|\mu_{dl}, \sigma_{dl}^2 = \begin{cases} 0 & \text{com probabilidade } \pi_d \\ \sim N(\mu_{dl}, \sigma_{dl}^2) & \text{com probabilidade } 1-\pi_d \end{cases} \quad (11)$$

em que: $\sigma_{dl}^2|v_{dl}, S_{dl} \sim \chi_{v_{dl}}^{-2}(v_{dl}, S_{dl})$ e $\pi_d|\rho_0, \pi_0 \sim \beta\ (\rho_0, \pi_0)$.

2.2.5.3 Modelo *Random Forest*

O modelo *Random Forest* é um modelo de predição não paramétrico implementado no pacote "randomForest" (Liaw & Wiener, 2002) do *software* R (R *Development Core Team* 2021). Foi utilizado o valor de 500.000 para o número de árvores de decisão.

Random Forest é um método de aprendizagem supervisionado no qual um conjunto de dados de treinamento com grande número de preditores (por exemplo, SNPs, x_l, onde x se refere a um vetor contendo genótipos de todos os SNPs para o indivíduo *l*) é usado para predizer um dado fenótipo (y_l). É uma modificação do *bootstrap aggregating* ou *bagging* que gera uma grande coleção de árvores distribuídas de forma idênticas (Abdollahi-Arpanahi et al., 2020).

Compreende quatro parâmetros principais: *N* – número total de observações, *M* – número total de variáveis preditoras (SNPs), *mtry* – subconjunto de *M* escolhido aleatoriamente para determinar uma árvore de decisão, normalmente *mtry* << *M* e *Ntree* – número total de árvores de decisões que formam uma "floresta". Cada árvore minimiza a função de perda média nos *bootstrapped data* e indica o quão bem um modelo se encaixa em um conjunto de dados de treinamento (normalmente apresentada como um erro quadrático médio). Resumidamente, o procedimento de *Random Forest* segue as seguintes etapas:

1) selecionar aleatoriamente um subconjunto de observações (*B* amostras de *bootstrap* de treinamento);

2) selecionar aleatoriamente um subconjunto de marcadores SNP – *mtry*;

3) gerar uma única árvore T_b dividindo o subconjunto de SNPs no subconjunto das amostras para formar nós da árvore; durante a divisão de um nó em uma árvore, o SNP com a maior capacidade de diminuir o erro quadrático médio dos nós filhos é selecionado para dividir o nó;

4) usar todos os dados *out-of-bag*, ou seja, o restante dos indivíduos que não foram selecionados na etapa 1, para determinar o erro quadrático médio da predição da árvore; para cada variável (SNP) na árvore (modelo), conduzir a permutação aleatória da ordem do SNP na árvore e calcule a diferença entre o erro quadrático médio da nova árvore e o erro quadrático médio da árvore inicial;

5) gerar uma floresta de árvores $\{T_b\}_1^B$ repetindo as etapas 1–4; o valor predito do conjunto de teste individual ($\hat{y}_l$) com genótipo x_l foi calculado como:

$$\hat{y}_l = \frac{1}{B}\sum_{b=1}^{B} T_b(x_l) \quad (12)$$

6) obter valores finais de importância da variável SNP (denotados como VIM) calculando a média dos valores de erro de predição em todas as árvores na floresta que contém esse SNP. O processo de divisão do nó continua até que não haja mais alteração dos valores do erro quadrático médio em todos os nós terminais.

Para a regressão, um valor VIM de um SNP é dado pela porcentagem de incremento no erro quadrático médio após um SNP ser permutado aleatoriamente em uma nova amostra. Em *Random Forest*, todos os SNPs são classificados com base em seus valores VIM. Os valores VIM variam de valores negativos a positivos. Um grande valor positivo indica um grande aumento no erro de predição quando o SNP é permutado aleatoriamente, em comparação com o valor erro quadrático médio antes da permutação, assim mais importante é o SNP. Por outro lado, valores negativos indicam que, quando esses SNPs foram permutados aleatoriamente, os modelos de predição de novas ordens SNP tiveram um erro de previsão menor do que antes da permutação (Li et al., 2018; Abdollahi-Arpanahi et al., 2020). Para mais detalhes sobre a teoria de *Random Forest*, consultar Breiman (2001).

2.2.5.4 Validação cruzada

Foi utilizado o esquema de validação cruzada *ten-fold* para determinar a habilidade preditiva dos modelos de predição genômica. Em cada local e no conjunto geral, os 152 híbridos foram divididos aleatoriamente em 10 grupos, dois grupos com 16 híbridos e oito com 15 híbridos, o que gerou as 10 *folds*. A cada rodada do modelo, foi atribuído valores fenotípicos ausentes a um grupo, sendo este dado como o conjunto de validação. A habilidade preditiva dos modelos de predição genômica foi calculada como a correlação de Pearson entre os GEBVs e os valores de BLUEs dos híbridos. Cada modelo foi rodado 5 vezes (k=5), sendo que a habilidade preditiva apresentada se refere a média destas rodadas.

Todos os gráficos de resultados foram plotados pelo pacote “ggplot2” versão 3.4.1 (Wickham, 2016).

2.3 RESULTADOS E DISCUSSÃO

2.3.1 Análises fenotípicas

Staygreen é controlado por poucas dezenas de QTLs de grande efeito, os quais apresentam forte interação com o ambiente para regulação de sua expressão (Belícuas et al., 2014; Sekhon et al., 2019). Devido a esta arquitetura e possível controle genético, existe uma distribuição contínua de fenótipos. Este comportamento é verificado no conjunto de híbridos deste estudo. Eles apresentaram uma distribuição bastante simétrica dos BLUEs nos locais de avaliação e no conjunto geral (Figura 2.1). Os 152 híbridos tiveram média de BLUEs, por local, variando de 2,75 em Areão a 4,15 em Anhembi, e média geral dos BLUEs, entre todos os locais, de 3,56, sendo este valor geral menor que Anhembi e Caterpillar e maior que em Areão e Departamento de Genética (Figura 2.1).

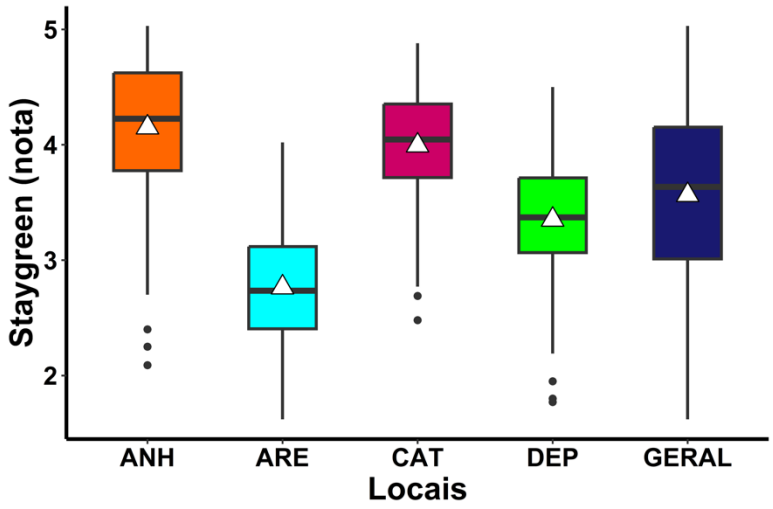

Figura 2.1 *Box plot* para os valores BLUEs de *staygreen*, com notas 1 (maior *staygeen*) a 5 (menor *staygreen*), em híbridos simples de milho avaliados em diferentes locais. Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP); e GERAL (conjunto de todos os ambientes).
[a]: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano).

O ordenamento dos BLUEs dos híbridos dentro de cada local apresentou ligeira similaridade quando estes são comparados entre os locais (Figura 2.2). Somente

quando comparados com o grupo geral é que a similaridade foi alta ($\hat{r}$ > 0,80), com exceção de Anhembi ($\hat{r}$ = 0,63).

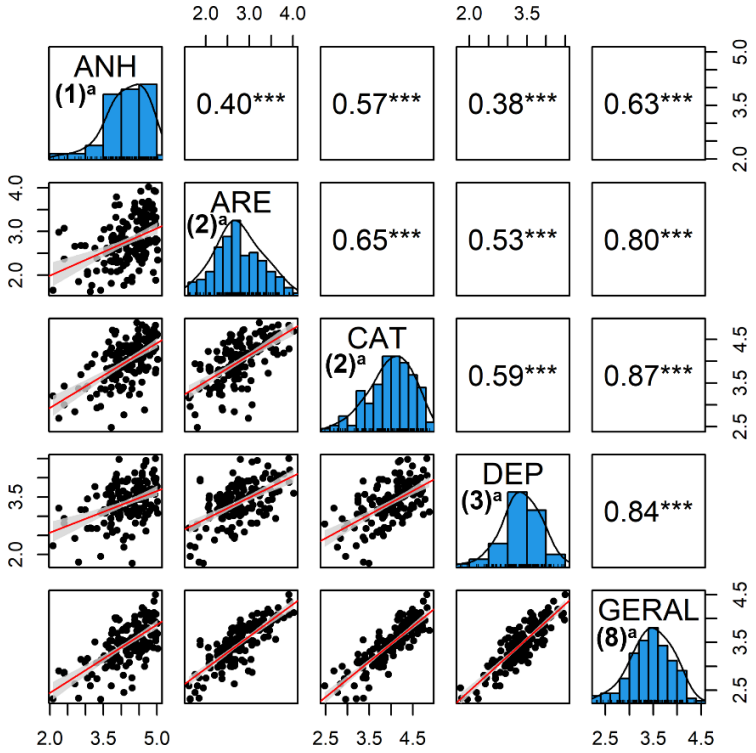

Figura 2.2 Estimativas de correlação de Spearman entre locais baseadas no ranqueamento dos BLUEs de *staygreen* (notas 1-5) em híbridos simples de milho. Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT); Departamento de Genética da ESALQ/USP (DEP) e GERAL (conjunto de todos os ambientes).
[a]: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano);
[***]: significativo a 0,1% de probabilidade de erro pelo teste t.

Contudo, é possível visualizar que todas as estimativas de correlação foram positivas e significativas (p < 0,01) (Figura 2.2). O menor valor de correlação estimado foi entre os BLUEs de Anhembi e Departamento de Genética ($\hat{r}$ = 0,38).

2.3.2 Estrutura genética de populações

Entender a dinâmica da diversidade e estrutura genética do painel de linhagens e híbridos utilizados em estudos genômicos é essencial para delinear o modo mais eficiente de explorar a variabilidade presente e o quão representativo e utilizável os resultados obtidos neste painel podem ser extrapolados para outros conjuntos de

indivíduos. Com a utilização conjunta da análise estrutura de populações realizada pelo algoritmo LEA (Frichot & François, 2015) e de distância genética pelo método de Rogers (1972), foi obtida uma visão robusta da variabilidade genética do painel de linhagens e dos híbridos e como ela está estruturada.

2.3.2.1 Linhagens

A análise de estrutura de populações agrupou o conjunto de linhagens em 5 subpopulações, as quais estão representadas por diferentes cores na Figura 2.3 (G1 a G5). As distâncias genéticas foram representadas por dendrogramas com agrupamento por UPGMA. A representação via dendrograma indicou ótimo ajuste entre as distâncias originais e as distâncias presentes no dendrograma, com valor de correlação cofenética de 0,98.

O grupo azul (G3) é composto por linhagens oriundas dos híbridos HS-1 (híbrido simples) e BR-201 (híbrido duplo), os quais possuem parentesco próximo devido o híbrido HS-1 ser o híbrido simples fêmea genitor do híbrido BR 201 (EMBRAPA, 1990; Cardoso, 1999). As linhagens originadas da população IG-2 também foram agrupadas e formaram o grupo verde (G5). Este conjunto em particular não apresenta histórico de parentesco com as demais origens, sendo obtida pelo intercruzamento das populações CMS-06 do Centro Nacional de Pesquisa do Milho e Sorgo (CNPMS – EMBRAPA) e a população ESALQ-PB4 do Departamento de Genética da ESALQ/USP (Souza Júnior et al., 1993).

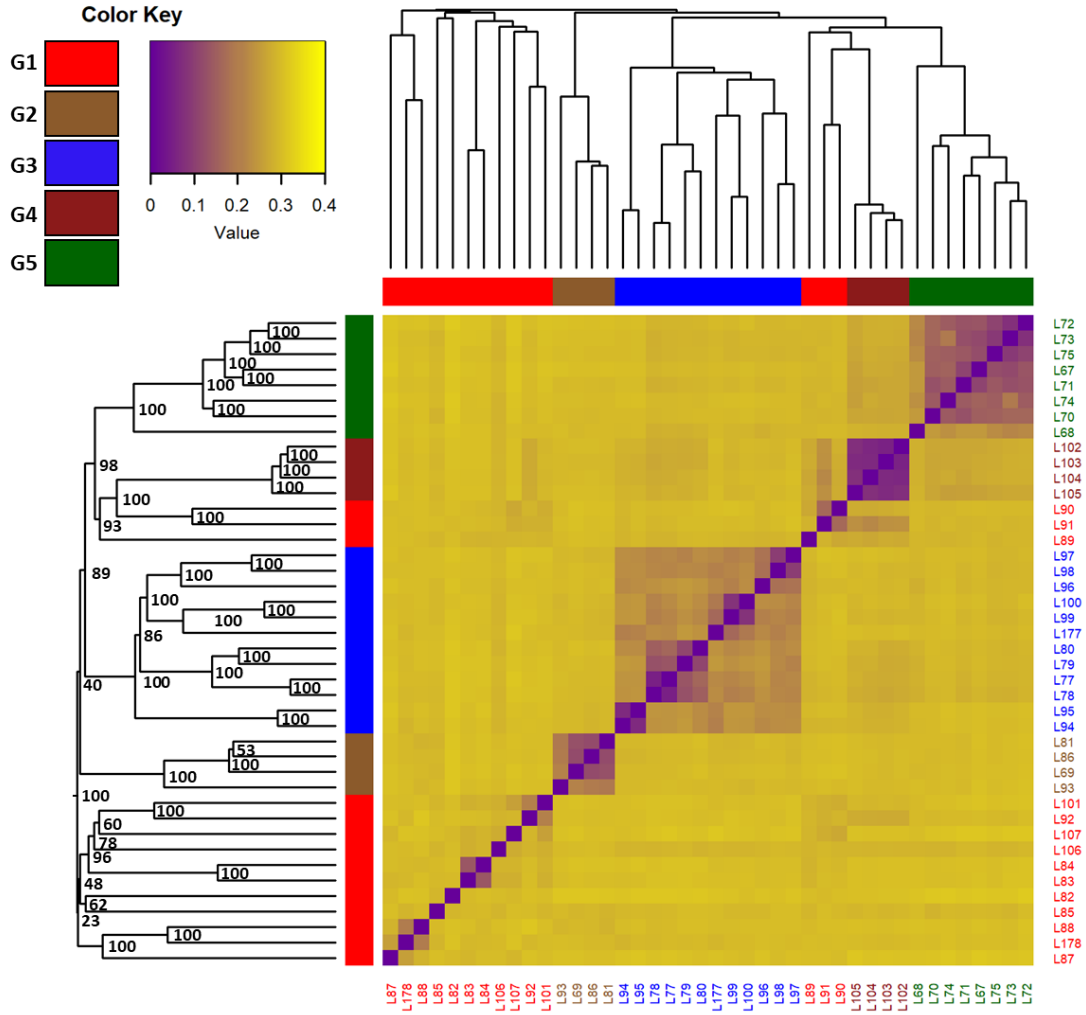

Figura 2.3 Representação das relações genéticas entre as 42 linhagens baseadas na distância de Rogers (1972) e algoritmo LEA (Frichot & François, 2015) pela combinação de dendrograma e mapa de calor com valores de consistência dos nós gerados por *Bootstrap* com 100.000 iterações e grupos gerados pela análise de estruturação (k=5) e pelo método de UPGMA.

As linhagens oriundas da população CMS-05 foram agrupadas em um único grupo (G4) e possuem baixíssima distância genética entre elas, indicada pela tonalidade escura do tom de roxo na chave de cores (Figura 2.3). Esta população é caracterizada por ser um germoplasma exótico e bastante antigo em trabalhos de melhoramento de milho no Brasil. Também denominada de Suwan DMR, a CMS-05 foi originada no Caribe e selecionada na Tailândia para resistência ao míldio-pulverulento (*Erysiphe polygoni*) (EMBRAPA, 1980).

Os grupos G1 (vermelho) e G2 (marrom) apresentam mistura das outras origens históricas. Nestes grupos estão linhagens obtidas de cruzamentos *topcrosses* entre as populações BR-105 (variedade extraída da CMS-05) e BR106 (composto do

cruzamento entre as populações Centramex, Dentado-Maya e Tuxpeño-1) (Canton, 1988; Packer et al., 1989, Souza Júnior et al., 1993), da população IG-2, que veio do intercruzamento da população CMS-05 com a ESALQ-PB5 (Souza Júnior et al., 1993), e do híbrido comercial XL-560, da extinta empresa Braskalb, que passou a se chamar Dekalb no início dos anos 2000 e é, atualmente, uma das marcas de sementes de milho da empresa Bayer AG. Os grupos G4, G1 e G2 compartilham muitos genitores em comum, o que acarreta a proximidade deles nos dendrogramas. Nesta situação, foi comum encontrar linhagens de origem histórica de um grupo sendo agrupadas em um dos demais.

Nota-se que os resultados das análises de estrutura e de divergência genética, em grande parte, estão em concordância com pedigree histórico do conjunto de linhagens, sendo este o principal fator de subdivisão das populações. Também, que apesar do painel ser constituído de poucas dezenas de linhagens, este possui uma representação relevante de linhagens de grupos tropicais (G3 e G5), subtropical (G4) e da mistura deles (G1 e G2).

2.3.2.2 Híbridos

As análises de populações agruparam os híbridos em 4 subpopulações, os quais remetem os cruzamentos com os 4 testadores utilizados como genitores masculinos deste painel (Figura 2.4). Assim como nas análises das linhagens, as distâncias genéticas dos híbridos foram representadas por dendrograma com agrupamento por UPGMA. Neste caso, a correlação cofenética para a representação do dendrograma foi de 0,91. Como o padrão da diversidade genética neste painel de híbridos é dado pela divergência dos testadores, suas combinações com as demais linhagens são separadas por cores na Figura 3.4 (T1 a T4).

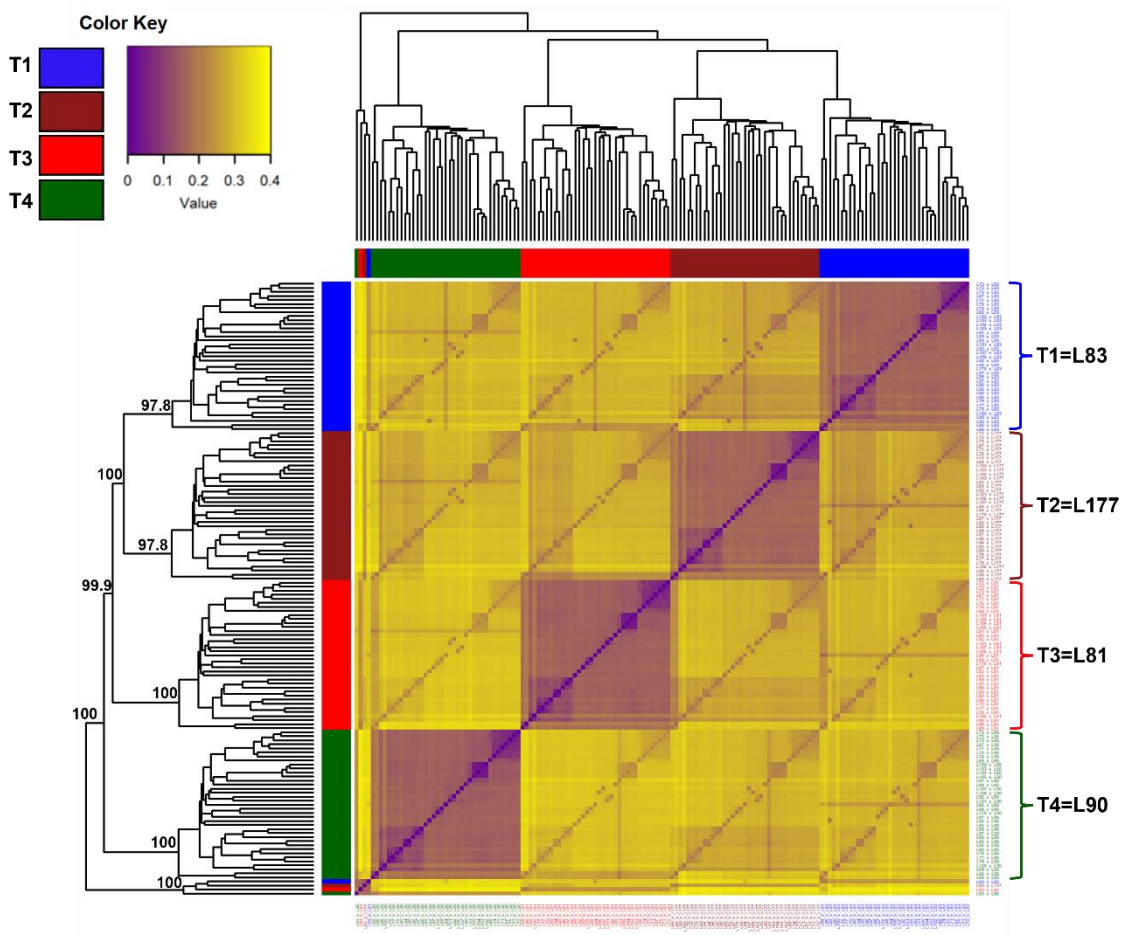

Figura 2.4 Representação das relações genéticas entre os 152 híbridos baseadas na distância de Rogers (1972) e algoritmo LEA (Frichot & François, 2015) pela combinação de dendrograma e mapa de calor com valores de consistência dos nós gerados por *Bootstrap* com 10.000 iterações e grupos gerados pela análise de estruturação (k=4) e pelo método de UPGMA. T1 a T4 representam os testadores com os quais as demais 38 linhagens foram cruzadas.

No dendrograma deste painel de híbridos os cruzamentos dos testadores com a linhagem L93 não se agruparam em qualquer outro grupo, gerando um pequeno grupo a parte. Vale ressaltar que na análise de estrutura genética pelo algoritmo LEA, foram indicadas apenas 4 subpopulações para os híbridos. Este ocorrido pode ser em função de que algumas linhagens e combinações apresentaram valores abaixo de 0,7 de probabilidade para pertencerem a determinado grupo, ou seja, baixa consistência em apontar a origem daquele indivíduo.

2.3.3 Habilidades preditivas dos modelos de predição genômica

As habilidades preditivas dos modelos tiveram grande amplitude de variação para *staygreen* (0,44 a 0,70) (Figura 2.5). De modos geral, os menores valores de habilidade preditiva foram detectados em Caterpillar e Departamento de Genética, cerca de 0,50. Estes baixos valores de habilidade preditiva podem ser causados por uma menor precisão na obtenção dos dados fenotípicos associada a baixa variância genética presente no painel avaliado, influenciando diretamente a magnitude da herdabilidade (Pandey et al., 2020; Zhang et al., 2017).

Caracteres com maior variância genética apresentam também maior herdabilidade, e são mais previsíveis via GS (Morais-Júnior et al., 2018). Destaca-se que *staygreen* é mensurado via avaliação subjetiva com pequeno intervalo de notas, o que pode afetar a herdabilidade, e consequentemente, os valores da habilidade preditiva. Maiores valores de habilidade preditiva foram obtidos quando se considerou todos os ambientes (GERAL), ultrapassando 0,70 (Figuras 2.5). Possivelmente, isto ocorre devido os BLUEs neste conjunto apresentarem uma precisão muito maior que quando elas são estimadas nos locais separados.

De modo geral, com a incorporação dos efeitos de dominância nos modelos de predição paramétricos, houve uma sutil melhora das habilidades preditivas para os modelos calibrados em Anhembi, Areão, Departamento de Genética e no conjunto GERAL, com média de 0,08 de acréscimo (Figura 2.5). No caso dos modelos calibrados em Caterpillar, os valores de habilidade preditiva foram praticamente iguais. Kadam et al. (2016) e Lyra et al. (2019) não verificaram incrementos na capacidade preditiva para este caráter ao utilizarem modelos aditivos-dominantes em detrimento aos modelos exclusivamente aditivos.

Gentinetta et al. (1986), em estudo pioneiro, relataram que *staygreen* é controlado por apenas um loco com dois alelos, apresentando dominância completa. Contudo, estudos publicados durante a década de 2000 indicaram que *staygreen* é um caráter de herança quantitativa, com presença, em algum grau, de efeitos não aditivos (Banziger et al., 2000).

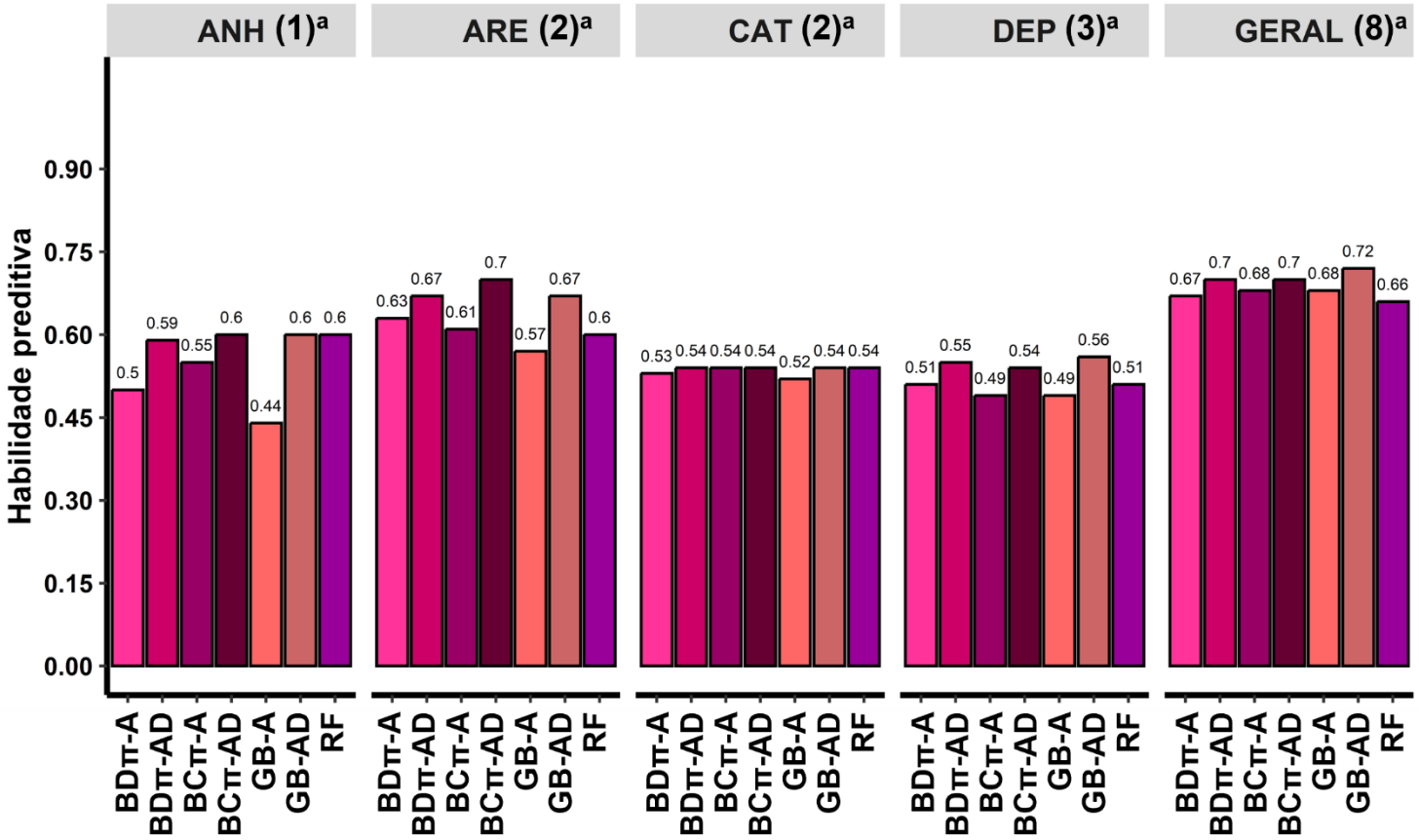

Figura 2.5 Habilidades preditivas de modelos de predição genômica para *staygreen* (notas 1-5) em híbridos simples de milho avaliados em diferentes locais. Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP) e GERAL (conjunto de todos os ambientes). Modelos paramétricos aditivos Bayes Dπ (BDπ-A), Bayes Cπ (BCπ-A), GBLUP (GB-A); Modelos paramétricos aditivo-dominantes: Bayes Dπ (BDπ-AD), Bayes Cπ (BCπ-AD) e GBLUP (GBP-AD); e Modelo não-paramétrico: e *Random Forest* (RF).
[a]: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano).

Lançando mão de estudos de associação genômica ampla (*Genome Wide Association Studies* – GWAS) em milho, Sekhon et al. (2019) determinaram que *staygreen* é um caráter de herança quantitativa controlado por poucos genes de maiores efeitos. Belícuas et al. (2014) identificaram cerca de 17 QTLs responsáveis por 73,08% da variância genética de *stay green* em milho tropical, e que os efeitos aditivos são maiores que os efeitos de dominância no controle genético do caráter, de modo que existe baixa expressão de heterose em *staygreen*. Isto possivelmente ocorre porque a heterose resulta indiretamente em maior *staygreen* ao ocasionar o maior acúmulo de matéria seca durante o período de enchimento de grãos em milho (Araus et al., 2010; Wang et al., 2019). Com base nos resultados deste estudo, o *staygreen* é controlado por efeitos aditivos com baixa expressão de efeitos de dominância, corroborando as conclusões de Bañziger et al. (2000) e Belícuas et al. (2014), e em termos de GS, com Kadam et al. (2016) e Lyra et al. (2019).

O modelo escolhido para predizer os desempenhos fenotípicos do conjunto total de híbridos poderia ser qualquer um dos modelos paramétricos com a incorporação dos efeitos de dominância. Levando em conta a demanda computacional, determinou-se que o modelo GBLUP é o mais indicado para exercer a GS para *staygreen* nesta situação.

Utilizou-se o modelo GBLUP aditivo-dominante no conjunto GERAL com 152 híbridos fenotipados para predizer o desempenho de todos os 861 híbridos simples possíveis das 42 linhagens do painel para *staygreen*. Com a predição genômica, nota-se o quanto se pode ganhar em tempo, economia de recursos e obtenção de híbridos mais promissores em programas de melhoramento de milho. Os ganhos e a disposição das melhores combinações preditas pelo modelo escolhido estão representados na Figura 2.6.

Ao selecionar com base nos desempenhos preditos as 15 melhores combinações (*Top15*) dos 152 híbridos testados (primeiro conjunto) e *Top15* dos 861 possíveis (segundo conjunto), percebe-se uma redução média de 0,45 na nota de *staygreen* (Figura 2.6) entre os *Top15* oriundos do primeiro para o segundo conjunto. Não houve coincidência entre os *Top15* dos dois conjuntos (Figura 2.6), ou seja, as melhores combinações para este caráter ainda não foram geradas.

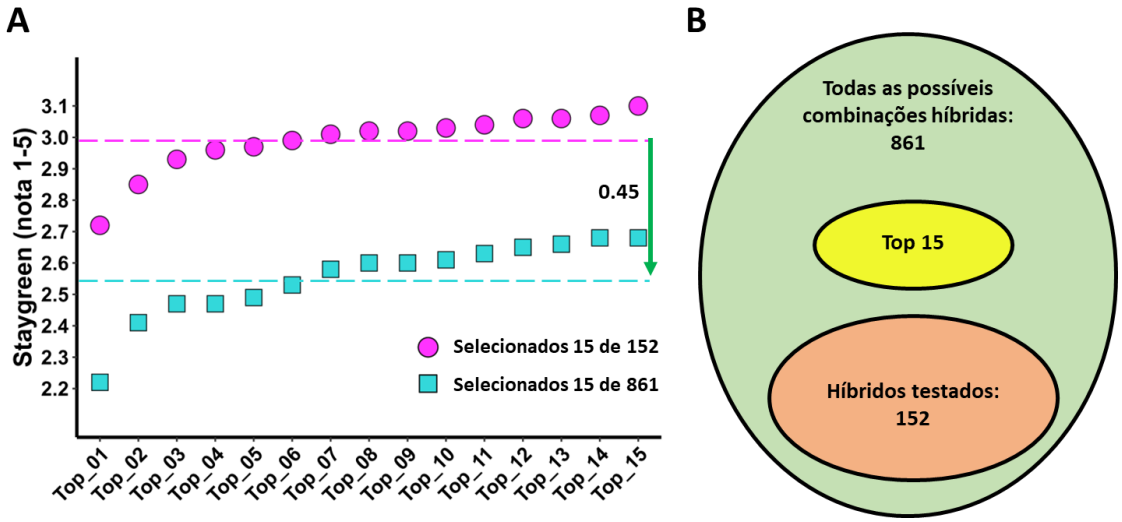

Figura 2.6 (A) distribuição das 15 melhores combinações de cruzamentos selecionadas a partir dos valores preditos dos 152 híbridos testados e dos valores preditos de todas as 861 combinações híbridas possíveis para *staygreen*. Linhas pontilhadas em rosa e azul representam o desempenho médio predito das 15 melhores combinações de cruzamentos selecionadas de 152 híbridos testados e dos 861 híbridos possíveis, respectivamente. Seta verde indica a diferença entre a médias dos dois conjuntos de híbridos. (B): disposição das combinações híbridas promissoras (*Top 15*) no conjunto de híbridos testados e não testados para *staygreen*.

2.4 CONCLUSÕES

Há um aumento da habilidade preditiva em modelos de seleção genômica para *staygreen* em milho quando são considerados os efeitos de dominância no modelo e quando se utiliza um maior número de ambientes para obtenção dos dados fenotípicos.

O modelo mais indicado para predizer o desempenho fenotípico para *staygreen* neste conjunto de dados é o GBLUP com a contabilização dos efeitos aditivos e de dominância.

2.5 REFERÊNCIAS

ABDOLLAHI-ARPANAHI, R.; GIANOLA, D.; PEÑAGARICANO, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. **Genetics Selection Evolution**, v. 52, n. 1, p. 1–15, 2020.

AGUIAR, A.M.; et al. Combining ability of inbred lines of maize and stability of their respective single-crosses. **Scientia Agricola**, Piracicaba, v.60, n.1, p.83-89, 2003.

ALMEIDA FILHO, J. E. et al. Genomic prediction of additive and non-additive effects using genetic markers and pedigrees. **G3: Genes, Genomes, Genetics**, v. 9, n. 8, p. 2739–2748, 2019.

ARAUS, J. L.; SÁNCHEZ, C.; CABRERA-BOSQUET, L. Is heterosis in maize mediated through better water use? **New Phytologist**, v. 187, n. 2, p. 392–406, 2010.

BÄNZIGER, M.; EDMEADES, G. O.; BECK, D. & BELLON, M. **Breeding for drought and nitrogen stress tolerance in maize from theory to practice**. 1. ed. Mexico: CIMMYT, 2000. 69 p.

BATES, D. et al. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1-48, 2015.

BELÍCUAS, P. R. et al. Inheritance of the stay-green trait in tropical maize. **Euphytica**, v. 198, n. 2, p. 163–173, 2014.

BERNARDO, R. Bandwagons I, too, have known. **Theoretical and Applied Genetics**, v. 129, n. 12, p. 2323–2332, 2016.

BHAT, J. A. et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. **Frontiers in Genetics**, v. 7, 2016.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BROWNING, B. L.; BROWNING, S. R. Genotype imputation with millions of reference samples. **American Journal of Human Genetics**, v. 98, n. 1, p. 116–126, 2016.

CANTON, T. **Avaliação de oito ciclos de seleção recorrente na população de milho (*Zea mays* L.) Swan DMR**. 1988. 112 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1988.

CARDOSO, R. G. **Depressão por endogamia dos componentes da produção em populações e híbridos de milho (*Zea mays* L.)**. 1999. 134 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1999.

CASEYS, C. Senescence: The genetics behind stay-green corn. **Plant Cell**, v. 31, n. 9, p. 1934–1935, 2019.

CHIBANE, N. et al. Relationship between delayed leaf senescence (stay-green) and agronomic and physiological characters in maize (*Zea mays* L.). **Agronomy**, v. 11, n. 2, p. 276, 2021.

CROSSA, J. et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. **Trends in Plant Science**, v. 22, n. 11, p. 961–975, 2017.

DESTA, Z. A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in Plant Science**, v. 19, n. 9, p. 592–601, 2014.

DIAS, K. O. D. G. et al. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. **Heredity**, v. 121, n. 1, p. 24–37, 2018.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, n. 19, v. 1, p. 11-15, 1987.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **BR 201: Tolerância à acidez e alta produtividade**.: Sete Lagoas: Centro Nacional de Pesquisa de Milho e Sorgo. 1990. Disponível em: <https://www.embrapa.br/busca-de-publicacoes/-/publicacao/487318/br-201-hibrido-duplo-de-milho>. Acesso em: 08 dez. 2021.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Melhoramento de Milho**.: Sete Lagoas: Centro Nacional de Pesquisa de Milho e Sorgo. 1980. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/81491/1/Melhoramento-populacoes.pdf&&clen=3547497>. Acesso em: 08 dez. 2021.

FRICHOT, E.; FRANÇOIS, O. LEA: An R package for landscape and ecological association studies. **Methods in Ecology and Evolution**, v. 6, n. 8, p. 925–929, 2015.

FRITSCHE-NETO, R. et al. TCGA: a tropical corn germplasm assembly for genomic prediction and high-throughput phenotyping studies. **Mendeley Data**, V3, 2020.

GENTINETTA, E. et al. A major gene for delayed senescence in maize - Pattern of photosynthates accumulation and inheritance. **Plant Breeding**, v. 97, n. 3, p.193-203, 1986.

GLAUBITZ, J. C. et al. TASSEL-GBS: A High-Capacity Genotyping by Sequencing Analysis Pipeline. **PLOS ONE**, v. 9, n. 2, p. e90346, 2014.

GRANATO, I. S. C. et al. snpReady: a tool to assist breeders in genomic analysis. **Molecular Breeding**, v. 38, n. 8, 2018.

HESLOT, N. et al. Genomic selection in plant breeding: A comparison of models. **Crop Science**, v. 52, n. 1, p. 146–160, 2012.

JOMBART, T. adegenet: a R package for the multivariate analysis of genetic markers. **Bioinformatics**, v. 24, n. 11, p. 1403–1405, 2008.

KADAM, D. C. et al. Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. **G3: Genes, Genomes, Genetics**, v. 6, n. 11, p. 3443–3453, 2016.

KAMAL, N. M. Stay-green trait: a prospective approach for yield potential, and drought and heat stress adaptation in globally important cereals. **International Journal of Molecular Sciences**, v. 20, n. 23, p. 5837, 2019.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 2012.

LENTH, R. V. (2021). **emmeans: estimated marginal means, aka least-squares means**. R package version 1.7.0. Disponível em <https://CRAN.R-project.org/package=emmeans>

LI, B. et al. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics**, v. 9, p. 1–20, 2018.

LI, G. et al. Genome-wide prediction in a hybrid maize population adapted to Northwest China. **Crop Journal**, v. 8, n. 5, p. 830–842, 2020.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002.

LUCHE, H. S. et al. Stay-green: A potentiality in plant breeding. **Ciência Rural**, v. 45, n. 10, p. 1755–1760, 2015.

LYRA, D. H. et al. Modeling copy number variation in the genomic prediction of maize hybrids. **Theoretical and Applied Genetics**, v. 132, n. 1, p. 273–288, 2019.

MANTEL, N. The detection of disease clustering and a generalized regression approach. **Cancer Research**, v. 27, p. 209-220, 1967.

MORAIS JÚNIOR, O. P. et al. Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. **Crop Science**, v. 58, n. 2, p. 592–607, 2018.

PACKER, D. et al. Caracterização das populações de milho (*Zea mays* L.) do Centro Nacional de Pesquisa de Milho e Sorgo. **Departamento de Genética**, ESALQ/USP, Piracicaba, p. 1-37, 1989.

PANDEY, M. K. et al. Genome-based trait prediction in multi- environment breeding trials in groundnut. **Theoretical and Applied Genetics**, v. 133, n. 11, p. 3101–3117, 2020.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483–495, 2014.

POLAND, J. A. et al. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **PLOS ONE**, v. 7, n. 2, p. e32253, 2012.

R Core Team (2021). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAMBAUT, A. (2018). **FigTree v1.4.4**. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. Disponível em: <http://tree.bio.ed.ac.uk/software/figtree/>

REVELLE, W. (2021). **psych: procedures for personality and psychological research**. Northwestern University, Evanston, Illinois, USA. Disponível em: <https://CRAN.R-project.org/package=psych Version = 2.1.9,.>

ROGERS, J. S. Measures of genetic similarity and genetic distance: studies in genetics. **University of Texas**, v. 7213, p.145-153, 1972.

SEKHON, R. S. et al. Integrated genome-scale analysis identifies novel genes and networks underlying senescence in maize. **Plant Cell**, v. 31, n. 9, p. 1968–1989, 2019.

SIM, S. C. et al. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. **PLOS ONE**, v. 7, n. 7, p. e40563, 2012.

SOUZA JÚNIOR et al. Estimativas de parâmetros genéticos na interpopulação de milho BR-105 x BR-106 e suas implicações no melhoramento. **Pesquisa Agropecuária Brasileira**, v. 28, n. 4, p. 473-479, 1993.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VENCOVSKY, R.; BARRIGA, P. **Genética biométrica no fitomelhoramento**. 1. ed. Ribeirão Preto: Revista Brasileira de Genética, 1992. 496 p.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, v. 195, n. 4, p. 1223–1230, 2013.

WANG, Z. et al. Physiological basis of heterosis for nitrogen use efficiency of maize. **Scientific Reports**, v. 9, n. 1, p. 1–11, 2019.

WARNES, G. R. et al. 2020. **gplots: various r programming tools for plotting data**. R package version 3.1.1. Disponível em: <https://CRAN.R-project.org/package=gplots>

WICKHAM, H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York. Disponível em: < https://ggplot2.tidyverse.org.>

ZENG, J. et al. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, v. 45, n. 1, p. 1–17, 2013.

ZHANG, A. et al. Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. **Frontiers in Plant Science**, v. 8, 2017.

ZHANG, J. et al. Identification and characterization of a novel stay-green QTL that increases yield in maize. **Plant Biotechnology Journal**, v. 17, p. 2272-2285.

3 PREDIÇÃO GENÔMICA MULTI-AMBIENTAL PARA PRODUTIVIDADE DE GRÃOS EM MILHO TROPICAL

RESUMO

A eficiência seleção genômica (GS) depende, dentre outros fatores, da escolha adequada do modelo de predição a ser utilizado, dos efeitos que serão contabilizados neste modelo e dos recursos e tempo necessários para o processo de predição dos fenótipos. Na cultura do milho, em que há expressão de heterose para diversos caracteres, modelos de GS que incorporam efeitos não-aditivos podem ser superiores àqueles puramente aditivos. Neste trabalho, três modelos paramétricos e um modelo não paramétrico foram utilizados na predição genômica multi-ambiental de híbridos simples de milho para produtividade de grãos, considerando efeitos aditivos, exclusivamente, e em conjunto com efeitos de dominância. Os dados fenotípicos se referem à avaliação de 152 híbridos simples de milho, provenientes do cruzamento de 42 linhagens endogâmicas, avaliados para produtividade de grãos em 13 ambientes (combinação de 6 locais em diferentes épocas e anos de plantio). As linhagens foram genotipadas com 13.826 marcadores SNPs pelo o método GBS, sendo suas combinações genotípicas utilizadas para gerar os genótipos dos híbridos. As médias ajustadas para cada genótipo, em cada local, foram usadas para treinar os modelos de predição genômica. A habilidade preditiva foi mensurada por meio da correlação de Pearson, obtida por meio do sistema de *ten-fold*. Os locais apresentaram similaridade para o comportamento dos híbridos, indicando ser vantajoso utilizar um modelo de predição geral para este conjunto de ambientes. Com isso, se reduz drasticamente o número de modelos ambiente-específicos a serem utilizados para predição nos diferentes locais. A inclusão dos efeitos de dominância em todos os modelos paramétricos incrementou as habilidades preditivas em cerca de 25%. Isto confirma que a inclusão de efeitos não-aditivos no modelo de predição permite explorar melhor a heterose e ter maior precisão na seleção genômica. Os modelos não diferiram entre atributos vinculados a capacidade preditiva quando ambos contabilizaram efeitos de dominância. Devido a menor demanda computacional do GBLUP, ele é o mais indicado para predizer a produtividade de grãos neste conjunto de dados. A predição com o modelo GBLUP aditivo-dominante indica a possibilidade de seleção de melhores combinações de linhagens do que as já realizadas que, potencialmente, elevam a produtividade de grãos em cerca de 0,21 t ha^{-1} ao selecionar os melhores 15 híbridos por predição.

Palavras-chave: seleção genômica, *Zea mays* L.; habilidade preditiva; efeitos não-aditivos.

3.1 INTRODUÇÃO

A escolha adequada do modelo de predição que melhor estime os efeitos dos marcadores ao longo do genoma faz parte da rotina nos *pipelines* da seleção genômica (GS, *genomic selection*) (Li et al., 2020). Existem diferentes modelos que são propostos e categorizados em regressões paramétricas (baseados em abordagens frequentistas e bayesianas) e não paramétricas (baseados em *machine learning*). Os modelos respondem de modo diferente na obtenção das predições por variarem nas pressuposições em como tratar a variância dos efeitos dos marcadores (Meuwissen et al., 2001; Schrauf et al., 2021).

A bateria de modelos existentes em GS e as pressuposições específicas de cada um foram propostas para lidar com a arquitetura e controle genético distintos de caracteres de interesse agronômico. A produtividade de grãos no milho é o principal caráter alvo do melhoramento e, apesar de ser o mais estudado, compreender sua arquitetura genética e interações com fatores ambientais, ainda é complexo. Sabe-se, contudo, que a produtividade de grãos em milho é governada por muitos QTLs de pequeno efeito, que interagem entre si e com o ambiente gerando uma distribuição contínua de fenótipos.

Segundo Desta & Ortiz (2014), caracteres altamente poligênicos, com distribuições contínuas dos fenótipos e com numerosos QTLs de pequeno efeito seriam melhor modelados pelas regressões lineares (GBLUP, RR-BLUP etc.), enquanto que caracteres controlados por poucos QTLs (*Quantitative Trait Loci*), de efeitos maiores, seriam mais adequados para regressões Bayesianas. Contudo, com a evolução e ampla divulgação da GS, nota-se que isto não é regra. Estudos apontam modelos de regressões bayesianas e de abordagem de *machine learning* com maior robustez para predição genômica da produtividade de grãos do que os modelos de regressão lineares (Zhang et al., 2020; Kaler et al., 2022). Percebe-se, então, a necessidade de testar diferentes modelos a cada novo conjunto de dados.

Para caracterizar mais precisamente o desempenho fenotípico de indivíduos, recomenda-se a utilização de ensaios multi-ambientais que tentam captar e avaliar o impacto da interação genótipos x ambientes (G x E) (Jarquín et al., 2021). Na presença de interação G x E significativa, a acurácia preditiva dos modelos de predição genômica decai significativamente quando o modelo é calibrado em um ambiente e validado em outro, mesmo para a mesma população de estudo (Mendes & Souza Júnior, 2016).

Informações de similaridade dos locais de teste e da presença ou não de interação G x E indicam se a melhor forma de implementar a GS para determinado conjunto de híbridos e ambientes deve ser por meio de modelos ambiente-específicos ou de um modelo geral, que tem boa performance em todo conjunto de ambientes. É importante verificar estas alternativas para uso eficiente da GS.

Por se tratar da cultura do milho, a incorporação de efeitos de dominância nos modelos pode ser benéfica para aumentar a precisão das predições para caracteres que manifestam o efeito da heterose, como a produtividade de grãos (Alves et al., 2021; Kadam et al., 2016; Technow et al., 2012). Alves et al. (2021) e Dias et al. (2018) obtiveram aumentos expressivos em capacidades preditivas para produtividade de grãos em híbridos simples de milho ao contabilizar o efeito de dominância.

Portanto, o modelo de predição mais adequado, juntamente com os efeitos a serem contabilizados, deve ser determinado para cada população (Li et al., 2020; Robertsen et al., 2019). Neste sentido, objetivou-se: (1) verificar os efeitos da similaridade dos locais de teste e treinamento ao recomendar o uso de modelos ambiente-específicos ou de um modelo geral baseados em abordagens paramétricas (e não paramétricas; (2) avaliar a *performance* dos modelos sem a incorporação dos efeitos de dominância e dos paramétricos com a incorporação deste efeito; e (3) selecionar o melhor modelo, com efeitos de dominância ou não, para identificação das combinações híbridas superiores em milho tropical para produtividade de grãos.

3.2 MATERIAL E MÉTODOS

3.2.1 Materiais vegetais e ensaios de campo

Utilizou-se um painel de 152 híbridos simples de milho obtidos do cruzamento de 38 linhagens endogâmicas elite com 4 linhagens testadoras. O cruzamento das linhagens foi realizado de modo que todas estivessem representadas nesta amostra de 152 híbridos simples. As linhagens elite são originadas de diferentes populações (IG-1, IG-2 e CMS-05) e híbridos comerciais (HS-1, XL-560 e BR-201). As quatro linhagens previamente selecionadas como testadoras também são endogâmicas, provenientes de diferentes populações e possuem germoplasma elite (Aguiar et al., 2003). Todas as linhagens e híbridos foram obtidos no Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz”, ESALQ/USP, Piracicaba-SP.

Os 152 híbridos foram avaliados juntamente com outros 104 híbridos em látice simples 16x16 em 13 ambientes, sendo que cada ambiente corresponde a uma combinação local x ano ou local x nº de experimentos. Os locais de avaliação foram as Estações Experimentais: Estação Areão, Estação Caterpillar e Departamento de Genética da ESALQ/USP, nos anos agrícolas de 2002/2003, 2003/2004 e 2004/2005; Estação Anhembi, nos anos agrícolas 2003/2004 e 2004/2005; próximas ao município de Piracicaba, no estado de São Paulo e; e Estação Experimental da Empresa Biomatrix no município de Patos de Minas, no estado de Minas Gerais, no ano agrícola de 2004/2005 (dois experimentos). O espaçamento utilizado foi de 0,80m entre linhas e 0,20m entre plantas, sendo as parcelas constituídas de uma linha de 4 m de comprimento e o estande de 20 plantas por parcela (62.500 plantas ha^{-1}).

A produtividade de grãos foi avaliada determinando o peso de grãos de toda a parcela em kg $parcela^{-1}$, corrigido para o teor de umidade de 15% e estande de 20 plantas. Posteriormente, os valores obtidos foram convertidos para t ha^{-1}. Todos os dados fenotípicos foram gentilmente cedidos pelo Dr. Cláudio Lopes de Souza Junior, professor titular da ESALQ/USP. Este mesmo conjunto de dados fenotípicos já foi base para estudos de predição genômica em outras abordagens (Mendes & Souza-Júnior, 2016) e *testcrosses* (Alves, 2006).

3.2.2 Análises fenotípicas

As análises de variância foram realizadas agrupando os 13 ambientes (combinação local x ano ou local x nº experimento) em 5 locais de avaliação, respectivamente, utilizando o local como critério de agrupamento. Estabeleceu-se o seguinte modelo linear misto para as análises fenotípicas:

$$y_{ijkl} = \mu + g_i + a_l + r_{k(l)} + b_{j(kl)} + (ga)_{il} + \varepsilon_{ijkl} \quad (1)$$

em que y_{ijkl} é o fenótipo do genótipo *i*, no bloco *j*, na repetição *k*, no ambiente *l*; μ é o intercepto; g_i é o efeito fixo do genótipo *i*; a_l é o efeito aleatório do ambiente *l*, com a_l ~ N (0, $\boldsymbol{I}\sigma_a^2$); $r_{k(l)}$ é o efeito aleatório da repetição *k*, no ambiente *l*, com $r_{k(l)}$ ~ N (0, $\boldsymbol{I}\sigma_r^2$); $b_{j(kl)}$ é o efeito aleatório do bloco *j*, na repetição *k*, no ambiente *l*, com $b_{j(kl)}$ ~ N (0, $\boldsymbol{I}\sigma_b^2$); $(ga)_{il}$ é o efeito aleatório da interação do genótipo *i* com o ambiente *l*, com $(ga)_{il}$ ~ N (0, $\boldsymbol{I}\sigma_{ga}^2$); e ε_{ijkl} é o efeito aleatório não-genético, com ε_{ijkl} ~ N (0, $\boldsymbol{I}\sigma_\varepsilon^2$), em que N refere-se à distribuição normal e $\boldsymbol{I}$ é a matriz identidade. Também foi realizada a

análise de variância do conjunto total de ambientes, ou seja, considerando os 13 ambientes seguindo o mesmo modelo misto.

Os valores de BLUEs (*Best Linear Unbiased Estimators*), ou médias ajustadas, dos híbridos em cada local e do conjunto de ambientes foram utilizados como os valores genéticos dos híbridos para as análises de predição genômica. A partir dos BLUEs e posterior ranqueamento em escala ordinal, os coeficientes de correlação de Spearman ($\hat{r}$) foram calculados entre os locais pela fórmula:

$$\hat{r} = \frac{\widehat{COV}_{R(X)R(Y)}}{\sqrt{\hat{\sigma}^2_{R(X)} . \hat{\sigma}^2_{R(Y)}}} \quad (2)$$

em que $\widehat{COV}_{R(X)R(Y)}$ é a covariância entre as os ranques dos BLUEs nos locais X e Y; $\hat{\sigma}^2_{R(X)}$ é a variância dos ranques dos BLUEs no local X; $\hat{\sigma}^2_{R(Y)}$ é a variância dos ranques BLUEs no local Y. A significância dos coeficientes de correlação de Spearman foi avaliada pelo teste t a 5% de probabilidade.

Para fins de explorar possíveis efeitos da interação de genótipos com os locais, foram estimadas as herdabilidades em nível de média de parcelas ($\hat{h}^2$) para cada local e no conjunto de ambientes pela equação abaixo, tomando, estritamente neste cenário, os efeitos dos genótipos como aleatórios:

$$\hat{h}^2 = \frac{\hat{\sigma}^2_g}{\hat{\sigma}^2_g + \left(\frac{\hat{\sigma}^2_{ga}}{l}\right) + \left(\frac{\hat{\sigma}^2_e}{rl}\right)} \quad (3)$$

em que $\hat{\sigma}^2_g$ é a variância genética total dos híbridos; $\hat{\sigma}^2_{ga}$ é a variância da interação de genótipos com ambientes; $\hat{\sigma}^2_e$ é a variância residual, *r* o número de repetições *l* o número de locais.

Os coeficientes de variação ambiental ($CV_e\%$) e genético ($CV_g\%$), e relativo (β) foram determinados utilizando os estimadores:

$$CV_e\% = \frac{\sqrt{\hat{\sigma}^2_E}}{\bar{X}}; \quad (4)$$

$$CV_g\% = \frac{\sqrt{\hat{\sigma}^2_g}}{\bar{X}}; \quad (5)$$

$$\beta = CV_g / CV_e \quad (6)$$

sendo "$\bar{X}$" a média geral do caráter.

Todas as análises foram realizadas utilizando o ambiente R versão 4.1.0 (R *Development Core Team*, 2021). Para análise de modelos mistos e estimação das herdabilidades foi utilizado o pacote "lme4" versão 1.1-27.1 (Bates et al., 2015); para obtenção dos BLUEs, o pacote "emmeans" versão 1.7.0 (Lenth, 2021), e para a análise de correlação entre os ambientes, o pacote "psych" versão 2.1.9 (Revelle, 2021).

3.2.3 Dados genotípicos

Todas as linhagens foram genotipadas utilizando o protocolo de *Genotyping-by-sequencing* (GBS) (Poland et al., 2012; Sim et al., 2012) na plataforma Illumina NextSeq 500 (Illumina Inc., San Diego, CA, EUA). Para isso, o DNA genômico das linhagens foi extraído de folhas jovens e saudáveis usando o protocolo CTAB (Doyle & Doyle, 1987). As amostras do DNA genômico das linhagens foram digeridas por duas enzimas de restrição (*Pst1-Mse1*) e incluídas em placas de sequenciamento, sendo que os fragmentos de DNA de cada amostra foram ligados a adaptadores com *barcodes* específicos para posterior genotipagem via GBS. Os dados de sequenciamento foram alinhados ao genoma de referência B73 (B73_RefGen_v4) usando o alinhador Bowtie2 (Langmead & Salzberg, 2012). Em seguida, *Single Nucleotide Polymorphisms* (SNPs) foram "chamados" por meio do pipeline GBSv2, disponível no *software* TASSEL 5.0 (Glaubitz et al., 2014). O SNP *dataset* foi filtrado, e marcadores com proporção de alelos identificados (*Call Rate*) menores do que 0,90, não-bialélicos e com frequência alélica mínima (MAF) inferior a 0,05 foram removidos do conjunto de dados durante o processo de controle de qualidade. Dados de marcadores perdidos foram imputados pelo *software* Beagle versão 5.0 (Browning & Browning, 2016). Após estes procedimentos de controle de qualidade, um total de 13.826 marcadores SNP de alta qualidade foram considerados para as análises genômicas.

Tabela 3.1. Exemplo da obtenção da matriz de genótipos dos híbridos.

Marcadores	Linhagens		Híbrido (L1 x L2)
	L1 (Genótipo)	L2 (Genótipo)	(Genótipo)
SNP1	2 (M1M1)	2 (M1M1)	2 (M1M1)
SNP2	2 (M2M2)	0 (m2m2)	1 (M2m2)
SNP3	0 (m3m3)	2 (M3M3)	1 (M3m3)
SNP4	0 (m4m4)	0 (m4m4)	0 (m4m4)

O conjunto de dados genotípicos foi cedido gentilmente pelo Dr. Roberto Fritsche-Neto (Professor Assistente na *Louisiana State University*) e está disponível com o título "*TCGA: a tropical corn germplasm assembly for genomic prediction and high-throughput phenotyping studies*" (Fritsche-Neto et al., 2020).

3.2.4 Predição genômica

3.2.4.1 Modelos aditivos

Três modelos de predição paramétricos (GBLUP, Bayes Cπ e Bayes Dπ) foram usados para estimar os *Genomic Estimated Breeding Values* (GEBVs) de cada genótipo considerando apenas efeitos aditivos. Os modelos GBLUP, Bayes Cπ e Bayes Dπ foram implementados utilizando o pacote "BGLR" (Pérez & de los Campos, 2014) do *software* R (R *Development Core Team* 2021). O processo de amostragem de Gibbs para os modelos paramétricos foi feito com 10.000.000 de iterações, descartando os primeiros 10.000 resultados como *burn-in* e *thin* de 1000.

GBLUP: O método foi baseado no seguinte modelo linear:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \boldsymbol{Z_{ia}a} + \boldsymbol{\varepsilon} \quad (7)$$

em que $\boldsymbol{y_i} = (y_1, ..., y_n)$ é o vetor de BLUEs e $\boldsymbol{y_i}$ representa a observação do genótipo *i* (*i* = 1, ..., n) em cada local; **1** é um vetor com o mesmo tamanho de $\boldsymbol{y_i}$ com todas as entradas iguais a 1; μ é a média geral; $\boldsymbol{Z}_{ia}$ é a matriz de incidência que conecta os efeitos genéticos aleatórios aos fenótipos; $\boldsymbol{a}$ é o vetor de efeitos genéticos aleatórios de cada híbrido e $\boldsymbol{\varepsilon}$ é o vetor de efeitos aleatórios residuais para cada local. Este modelo assume que a distribuição do vetor $\boldsymbol{a}$ é normal multivariada com média zero e uma matriz de covariância $\sigma^2_{aj}\boldsymbol{G}$, ou seja, $\boldsymbol{a} \sim$ MVN ($\mathbf{0}$, $\sigma^2_{aj}\boldsymbol{G}$), em que σ^2_{aj} é um componente de variância genética no local *j* e $\boldsymbol{G}$ é uma matriz simétrica, semi-positivo-definida remetendo à estrutura de variância-covariância construída a partir dos marcadores SNPs. O modelo

também assume que os erros em cada local são independentes com variância homogênea, σ_ε^2; e sendo $\boldsymbol{\varepsilon} \sim$ N ($\mathbf{0}$, $\sigma_{\varepsilon j}^2 \boldsymbol{I}$) (onde $\boldsymbol{I}$ é a matriz identidade, e $\sigma_{\varepsilon j}^2$ é a variância residual no local *j*). A matriz $\boldsymbol{G}$ foi calculada por meio da abordagem proposta por VanRaden (2008):

$$\boldsymbol{G} = \frac{\boldsymbol{W_1}\,\boldsymbol{W_1'}}{2\sum_{l=1}^{m}\rho_l(1-\rho_l)} \quad (8)$$

em que $\boldsymbol{W_1}$ é a matriz de marcadores de efeito genético aditivo (SNPs), com dimensões do número de indivíduos (*n*) pelo número de locos (*m*), e ρ_l é a MAF do loco *l*.

Neste modelo a distribuição *a priori* para os efeitos dos marcadores pode ser escrita como $\rho(a) = \prod_{l=1,m}\rho(a_l)$, em que $\rho(a_l) \sim$ N (0, σ_{a0}^2), ou seja, cada efeito do marcador segue, a *priori*, uma distribuição normal com uma variância σ_{a0}^2 (variância dos efeitos do marcador). O termo "0" implica que o modelo GBLUB possui variância constante entre os marcadores.

Bayes Cπ e Bayes Dπ: Para estas abordagens bayesianas, foi adotado o seguinte modelo linear:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \sum_{l=1}^{m}\boldsymbol{X_{il}}\boldsymbol{g_l} + \boldsymbol{\varepsilon_i} \quad (9)$$

em que $\boldsymbol{y_i} = (y_1, ..., y_n)$ é o vetor de BLUEs e $\boldsymbol{y_i}$ representa a observação no genótipo *i* (*i* = 1, ..., n) em cada local; **1** é um vetor com o mesmo tamanho de $\boldsymbol{y_i}$ com todas as entradas iguais a 1; $\boldsymbol{\mu}$ é a média geral; g_l é o vetor de efeitos aditivos do marcador SNP *l*; $\boldsymbol{X_{il}}$ é a vetor de incidência de cada marcador (assumindo valores de 2, 1 e 0 que correspondem aos genótipos SNPs AA, Aa e aa, respectivamente) e m é o número de marcadores. Estes valores $\boldsymbol{x_{il}}$ representam diretamente o número de cópias do alelo alternativo em cada loco *l*; e $\boldsymbol{\varepsilon_i}$ é o vetor residual, tendo $\boldsymbol{\varepsilon_i} \sim$ N ($\mathbf{0}$, $\sigma_{\varepsilon j}^2 \boldsymbol{I}$) (onde $\boldsymbol{I}$ é a matriz identidade, e $\sigma_{\varepsilon j}^2$ é a variância residual no local *j*). Os modelos Bayes Cπ e Bayes Dπ assumem a mesma distribuição a *priori* para a variância residual, uma distribuição Qui-quadrado invertida escalonada, dada por $\sigma_{\varepsilon j}^2 | v_{\varepsilon j},\ S_{\varepsilon j} \sim \chi_{v_{\varepsilon j}}^{-2}(v_{\varepsilon j}, S_{\varepsilon j})$ com hiperparâmetros $v_{\varepsilon j}$ (graus de liberdade) e $S_{\varepsilon j}$ (escala) ambos iguais a 2,0 a *priori* vaga. As diferenças entre estes dois modelos foram as distribuições assumidas a *priori* para os efeitos dos marcadores.

Bayes Cπ e Bayes Dπ possuem um parâmetro adicional π (probabilidade de o efeito do marcador ser igual a zero). Este parâmetro é tratado como desconhecido e atribuído a uma distribuição β a *priori*, tendo $\pi \sim \beta$ (ρ_0, π_0), com $\rho_0 > 0$ e $\pi_0 \in [0, 1]$ (Perez & de los Campos, 2014). No modelo Bayes Cπ, a distribuição a *priori* para os efeitos de marcador é dada por uma mistura de distribuições normais $g_l|\pi \sim (1 - \pi)$ N (0,

σ_g^2) + πN (0, σ_g^2 = 0). Este modelo assume a mesma variância genética para todos os marcadores com distribuição a *priori* dada por $\sigma_g^2|v_g, S_g \sim \chi_{v_g}^{-2}(v_g, S_g)$. O modelo Bayes Dπ também assume como *priori* para efeitos de marcador uma mistura de distribuições normais dada por $g_l|\pi \sim (1 - \pi)$ N (0, σ_{gl}^2) + πN (0, σ_{gl}^2 = 0), contudo σ_{gl}^2 denota que cada SNP tem sua própria variância com distribuição a *priori* dada por $\sigma_{gl}^2|v_{gl}$, $S_{gl} \sim \chi_{v_{gl}}^{-2}(v_{gl,}S_{gl})$.

3.2.4.2 Modelos aditivo-dominantes

A incorporação dos efeitos de dominância foi realizada somente nos modelos paramétricos utilizando também o pacote "BGLR" e as mesmas 10.000.000 de iterações.

Os modelos foram ajustados das seguintes formas:

GBLUP: o modelo utilizado que ajusta simultaneamente aos efeitos aditivos e de dominância dos SNPs foi:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \boldsymbol{Z_{ia}a} + \boldsymbol{Z_{iv}v} + \boldsymbol{\varepsilon_i} \quad (10)$$

em que $\boldsymbol{y_i}$, $\mathbf{1}$, $\boldsymbol{\mu}$, $\boldsymbol{Z_{ia}}$, $\boldsymbol{a}$ e $\boldsymbol{\varepsilon_i}$ representam os mesmos parâmetros definidos no modelo (7), assim como $\boldsymbol{Z}_{ia}$, $\boldsymbol{Z}_{iv}$ é a matriz de incidência que conecta os efeitos genéticos aleatórios aos fenótipos, e $\boldsymbol{v}$ se refere ao vetor dos efeitos de dominância de cada indivíduo. A matriz de covariância dos efeitos de dominância foi *Var* ($\boldsymbol{v}$) = $\boldsymbol{D}\sigma_d^2$, em que $\boldsymbol{D}$ representa a matriz de parentesco genômica-dominante (Vitezica et al., 2013), calculada pelo pacote "snpReady" (Granato et al., 2018), também do *software* R, pela equação:

$$\boldsymbol{D} = \frac{\boldsymbol{W_2}\,\boldsymbol{W_2'}}{4\sum_{l=1}^{m}\rho_l^2(1 - \rho_l)^2} \quad (11)$$

em que $\boldsymbol{W_2}$ representa a matriz de marcadores de efeitos genéticos de dominância e *n*, *m* e ρ_l são os mesmos termos definidos na equação (8).

Bayes Cπ e Bayes Dπ: o modelo utilizado que ajusta aos efeitos aditivos e de dominância dos SNPs nas abordagens bayesianas foi:

$$\boldsymbol{y_i} = \boldsymbol{\mu 1} + \sum_{l=1}^{m}(\boldsymbol{X_{il}a_l}) + \sum_{l=1}^{m}(\boldsymbol{W_{il}d_l}) + \boldsymbol{\varepsilon_i} \quad (12)$$

em que $\boldsymbol{y_i}$, $\mathbf{1}$, $\boldsymbol{\mu}$, $\boldsymbol{X_{il}}$ e $\boldsymbol{\varepsilon_i}$ representam os mesmos parâmetros definidos no modelo aditivo (9), $\boldsymbol{W_{il}}$ é a vetor de incidência de cada marcador (assumindo valores de 1 e 0 que correspondem aos genótipos heterozigóticos e homozigóticos, respectivamente) e m é o número de marcadores, $\boldsymbol{a_l}$ é o efeito aditivo e $\boldsymbol{d_l}$ é o efeito de dominância do marcador *l*.

Dada a suposição de que a epistasia está ausente, o termo residual no modelo aditivo-dominante contém apenas efeitos não genéticos, enquanto o resíduo do modelo aditivo também inclui desvios de dominância. Nos modelos Bayes Cπ e Bayes Dπ as *priori* estabelecidas para os efeitos aditivos dos marcadores seguem as mesmas do modelo (9). Nestes modelos as distribuições a *priori* para $\boldsymbol{d_l}$ também é uma mistura de normais, dados π_d e σ_d^2 para Bayes Cπ e π_d e σ_{dl}^2 para Bayes Dπ. No entanto, a fim de contabilizar a direcionalidade da dominância, o componente normal da *priori* para $\boldsymbol{d_l}$ tem uma média desconhecida diferente de zero (Almeida-Filho et al., 2019; Zeng et al., 2013), assim:
Para o modelo Bayes Cπ:

$$\boldsymbol{d_l}|\mu_d, \sigma_d^2 = \begin{cases} 0 & \text{com probabilidade } \pi_d \\ \sim N(\mu_d, \sigma_d^2) & \text{com probabilidade } 1-\pi_d \end{cases} \quad (13)$$

em que: $\sigma_d^2|v_d, S_d \sim \chi_{v_d}^{-2}(v_d, S_d)$ e $\pi_d|\rho_0, \pi_0 \sim \beta\ (\rho_0, \pi_0)$.
Para o modelo Bayes Dπ:

$$\boldsymbol{d_l}|\mu_{dl}, \sigma_{dl}^2 = \begin{cases} 0 & \text{com probabilidade } \pi_d \\ \sim N(\mu_{dl}, \sigma_{dl}^2) & \text{com probabilidade } 1-\pi_d \end{cases} \quad (14)$$

em que: $\sigma_{dl}^2|v_{dl}, S_{dl} \sim \chi_{v_{dl}}^{-2}(v_{dl}, S_{dl})$ e $\pi_d|\rho_0, \pi_0 \sim \beta\ (\rho_0, \pi_0)$.

3.2.4.3 Modelo *Random Forest*

O modelo *Random Forest* é um modelo de predição não paramétrico implementado no pacote "randomForest" (Liaw & Wiener, 2002) do *software* R (R *Development Core Team* 2021). Foi utilizado o valor de 500.000 para o número de árvores de decisão.

Random Forest é um método de aprendizagem supervisionado no qual um conjunto de dados de treinamento com grande número de preditores (por exemplo, SNPs, x_l, onde x se refere a um vetor contendo genótipos de todos os SNPs para o indivíduo *l*) é usado para predizer um dado fenótipo (y_l). É uma modificação do *bootstrap aggregating* ou *bagging* que gera uma grande coleção de árvores distribuídas de forma idênticas (Abdollahi-Arpanahi et al., 2020).

Compreende quatro parâmetros principais: *N* – número total de observações, *M* – número total de variáveis preditoras (SNPs), *mtry* – subconjunto de *M* escolhido aleatoriamente para determinar uma árvore de decisão, normalmente *mtry* << *M* e *Ntree* – número total de árvores de decisões que formam uma "floresta". Cada árvore minimiza a função de perda média nos *bootstrapped data* e indica o quão bem um modelo se

encaixa em um conjunto de dados de treinamento (normalmente apresentada como um erro quadrático médio). Resumidamente, o procedimento de *Random Forest* segue as seguintes etapas:

1) selecionar aleatoriamente um subconjunto de observações (*B* amostras de *bootstrap* de treinamento);

2) selecionar aleatoriamente um subconjunto de marcadores SNP – *mtry*;

3) gerar uma única árvore T_b dividindo o subconjunto de SNPs no subconjunto das amostras para formar nós da árvore; durante a divisão de um nó em uma árvore, o SNP com a maior capacidade de diminuir o erro quadrático médio dos nós filhos é selecionado para dividir o nó;

4) usar todos os dados *out-of-bag*, ou seja, o restante dos indivíduos que não foram selecionados na etapa 1, para determinar o erro quadrático médio da predição da árvore; para cada variável (SNP) na árvore (modelo), conduzir a permutação aleatória da ordem do SNP na árvore e calcule a diferença entre o erro quadrático médio da nova árvore e o erro quadrático médio da árvore inicial;

5) gerar uma floresta de árvores $\{T_b\}_1^B$ repetindo as etapas 1–4; o valor predito do conjunto de teste individual ($\hat{y}_l$) com genótipo x_l foi calculado como:

$$\hat{y}_l = \frac{1}{B}\sum_{b=1}^{B} T_b(x_l) \quad (15)$$

6) obter valores finais de importância da variável SNP (denotados como VIM) calculando a média dos valores de erro de predição em todas as árvores na floresta que contém esse SNP. O processo de divisão do nó continua até que não haja mais alteração dos valores do erro quadrático médio em todos os nós terminais.

Para a regressão, um valor VIM de um SNP é dado pela porcentagem de incremento no erro quadrático médio após um SNP ser permutado aleatoriamente em uma nova amostra. Em *Random Forest*, todos os SNPs são classificados com base em seus valores VIM. Os valores VIM variam de valores negativos a positivos. Um grande valor positivo indica um grande aumento no erro de predição quando o SNP é permutado aleatoriamente, em comparação com o valor erro quadrático médio antes da permutação, assim mais importante é o SNP. Por outro lado, valores negativos indicam que, quando esses SNPs foram permutados aleatoriamente, os modelos de predição de novas ordens SNP tiveram um erro de previsão menor do que antes da permutação (Li et al., 2018; Abdollahi-Arpanahi et al., 2020). Para mais detalhes sobre a teoria de *Random Forest*, consultar Breiman (2001).

3.2.4.4 Validação cruzada

Dois esquemas de validação cruzada *ten-fold* foram utilizados para determinar a habilidade preditiva dos modelos de predição genômica. Nos dois esquemas, os 152 híbridos foram divididos aleatoriamente em 10 grupos, dois grupos com 16 híbridos e oito com 15 híbridos, o que gerou as *ten-folds*. A cada rodada do modelo, foi atribuído valores fenotípicos ausentes a um grupo, sendo este dado como o conjunto de validação. A habilidade preditiva dos modelos de predição genômica foi calculada como a correlação de Pearson entre os GEBVs e os valores de BLUEs dos híbridos de acordo com o esquema de validação.

Os esquemas de validação foram da seguinte forma: (1) predição dentro de locais – os modelos foram treinados em um local, em que o processo de *ten-fold* foi aplicado, e os GEBVs da população de validação foram correlacionados com os BLUEs dos mesmos híbridos no mesmo local para obtenção das habilidades preditivas; (2) predição entre locais – os modelos foram treinados em um local, em que o *ten-fold* foi aplicado, e os GEBVs da população de validação foram correlacionados com BLUEs dos mesmos híbridos, porém, de outro local para obtenção das habilidades preditivas. A segunda abordagem de validação cruzada serve para destacar a possível influência da interação de híbridos com os locais nas habilidades preditivas dos modelos. Cada modelo foi rodado 5 vezes (k=5), sendo que a habilidade preditiva apresentada se refere a média destas rodadas. Uma representação gráfica dos esquemas de validação cruzada *ten-fold* utilizados neste estudo está apresentada na Figura 2.1.

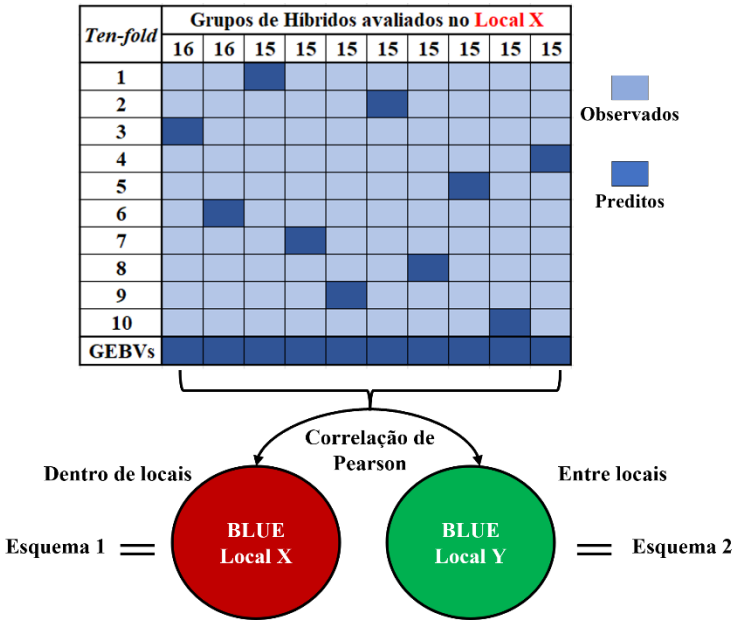

Figura 3.1. Representação dos esquemas de validação cruzada *ten-fold*, em que GEBVs refere-se aos *Genomic Estimated Breeding Values* e BLUE aos *Best Linear Unbiased Estimador*.

Todos os gráficos de resultados foram plotados pelo pacote "ggplot2" versão 3.4.1 (Wickham, 2016).

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Análises fenotípicas

A distribuição dos BLUEs nos diferentes locais foi relativamente simétrica, apresentando alguns *outliers* em todos os conjuntos, principalmente abaixo do limite inferior do primeiro quartil (Figura 3.2). Em todos os locais, a média de produtividade de grãos ficou acima de 7,00 t ha^{-1}, com variação de 7,31 t ha^{-1} para Anhembi a 8,85 t ha^{-1} para o Departamento de Genética.

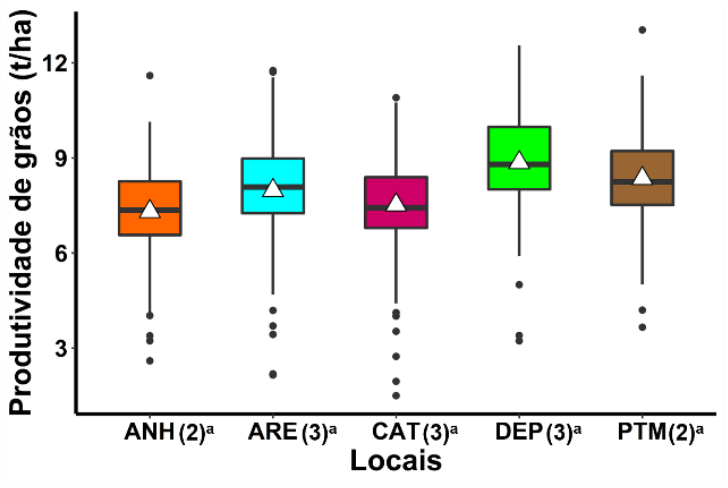

Figura 3.2. *Box plot* com os valores de BLUEs de produtividade de grãos em híbridos simples de milho avaliados em diferentes locais. Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP); Estação Experimental da Empresa Biomatrix em Patos de Minas (PTM).
ª: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano ou local x nº de experimentos).

As estimativas dos coeficientes de herdabilidade ($\hat{h}^2$) para produtividade de grãos foram de alta magnitude nos locais, variando de 0,82 (Anhembi) a 0,89 (Departamento de Genética). Na $\hat{h}^2$ para o conjunto de ambientes (GERAL), como esperado, é notado aumento expressivo de seu valor em relação aos locais individuais devido ao maior número de ambientes utilizados para a estimação. Nesta situação, a $\hat{h}^2$ foi de 0,95 para produtividade de grãos (Tabela 3.2).

Tabela 3.2. Estimativas de herdabilidade ($\hat{h}^2$), coeficientes de variação genético (CV_g), residual (CV_e) e relativo (β) para produtividade de grãos (t ha^{-1}) em híbridos simples de milho avaliados em diferentes locais.

Parâmetro	Locais					
	ANH (2)ª	ARE (3)ª	CAT (3)ª	DEP (3)ª	PTM (2)ª	GERAL (13)ª
$\hat{h}^2$	0,82	0,84	0,84	0,89	0,87	0,95
CV_g	16,86	17,61	18,75	16,15	17,83	16,60
CV_e	14,77	15,20	16,78	12,15	13,89	14,63
β	1,14	1,15	1,12	1,33	1,28	1,13

Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP); Estação Experimental da Empresa Biomatrix em Patos de Minas (PTM) e GERAL (conjunto de todos os ambientes);
ª: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano ou local x nº de experimentos).

Os valores de coeficiente de variação residual (CV_e), também denominado coeficiente de variação experimental, foram de baixa a intermediária magnitudes (< 20%) (Tabela 2.2), indicando boa precisão experimental e consistência na estimação dos parâmetros genéticos (Fritsche-Neto et al., 2012). Também foi obtido o coeficiente de variação relativo (β), o qual constitui uma medida de influência do ambiente sobre a expressão dos fenótipos já que relaciona informações da variação genética (CV_g) com a ambiental (Vencovski & Barriga, 1992). Estes mesmos autores estabelecem que valores de β acima de 1 indicam situações favoráveis à estimação precisa de parâmetros genéticos e à seleção, o que foi observado em todos os locais para produtividade de grãos.

As estimativas da variância para a interação dos híbridos com anos dentro de cada local não foram significativas (Tabela 3.3), indicando que o agrupamento dos ambientes por local de avaliação e posterior estimação dos BLUEs dos híbridos não sofreram efeito relevante da interação de híbridos com ambientes. Quando se leva em consideração o conjunto GERAL, o componente da interação é significativo para produtividade de grãos (Tabela 3.3). Isto porque, a manifestação da interação é esperada com o aumento de ambientes de avaliação, uma vez que as condições edafoclimáticas específicas de cada ambiente permite que este fenômeno biológico seja de alta magnitude e possa ter seus efeitos captados em análises estatísticas. Estas estimativas servem como medidas de similaridade ou não do comportamento dos híbridos entre os diferentes locais e o efeito diferencial destes locais sobre a expressão fenotípica dos híbridos.

Tabela 3.3. Estimativas do componente de variância para a interação de híbridos x locais, em porcentagem (%) da variação fenotípica total, para produtividade de grãos em híbridos simples de milho avaliados em diferentes locais.

Locais	ANH (2)[a]	ARE (3)[a]	CAT (3)[a]	DEP (3)[a]	PTM (2)[a]	GERAL (13)[a]
ANH	4,67	11,05**	11,64	12,57**	12,06	-
ARE	-	11,76	10,58**	11,31**	13,51**	-
CAT	-	-	11,28	11,59**	12,19*	-
DEP	-	-	-	5,95	10,64*	-
PTM	-	-	-	-	0,00	-
GERAL	-	-	-	-	-	12,96**

Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP); Estação Experimental da Empresa Biomatrix em Patos de Minas (PTM) e GERAL (conjunto de todos os ambientes);

[a]: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano ou local x nº de experimentos);

**, *: significativo a 1 e 5% de probabilidade de erro, respectivamente, pelo teste de Chi-quadrado (χ^2, LTR).

Entre os locais dois a dois, Anhembi pode ser visto como um local de condições ambientais e comportamental dos híbridos similares à Caterpillar e a Patos de Minas. As estimativas de variância para a interação dos híbridos com anos nos outros locais dois a dois foram significativas e de maior magnitude, principalmente para Areão e Departamento de Genética, que apresentaram interação significativa com todos os demais (Tabela 3.3).

A similaridade entre o comportamento dos híbridos nos diferentes locais foi verificada, também, a partir do gráfico da correlação de Spearman entre BLUEs dos híbridos nos locais dois a dois e com o conjunto de todos eles (Figura 3.3). Sob uma perspectiva de melhoramento de plantas, este parâmetro pode ser considerado um indicador da magnitude da interação G x E, no qual baixos valores indicam diferenças substanciais na classificação dos genótipos devido a mudança do ambiente (Alves et al., 2021).

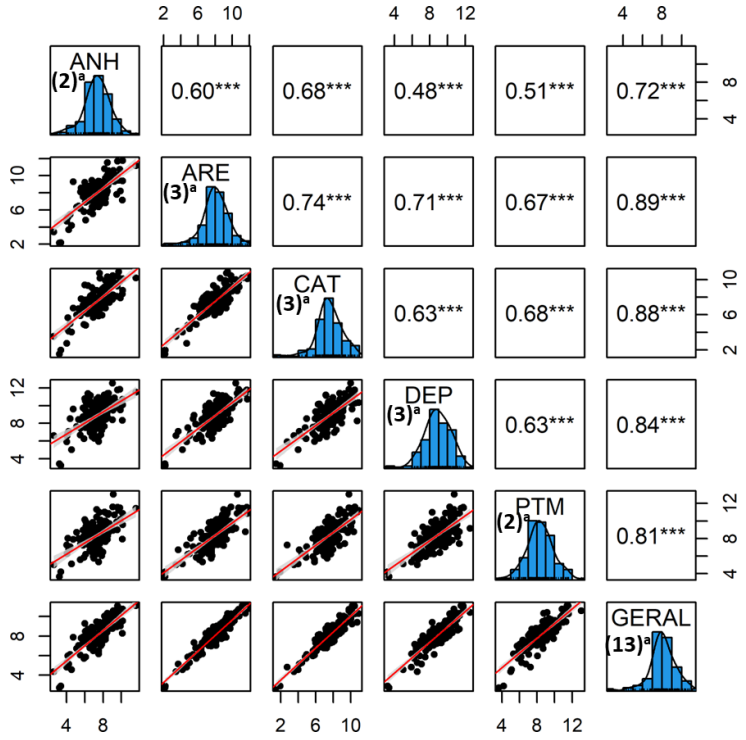

Figura 3.3. Estimativas de correlação de Spearman entre locais baseadas no ranqueamento de BLUEs de produtividade de grãos (t ha^{-1}) em híbridos simples de milho. Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP); Estação Experimental da Empresa Biomatrix em Patos de Minas (PTM) e GERAL (conjunto de todos os ambientes).
ª: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano ou local x nº de experimentos);
***: significativo a 0,1% de probabilidade de erro pelo teste t.

Todas as estimativas de correlações foram positivas e significativas ($p < 0,01$), variando em magnitude entre os pares de locais (Figura 3.3). O menor valor de correlação estimado foi entre Anhembi e Departamento de Genética ($\hat{r} = 0,48$), justamente o par de locais que apresentou maiores valores de interação entre híbridos e locais (Tabela 3.3). Nesta análise, foi incluído o conjunto GERAL para verificar a similaridade de cada local individual com o conjunto total de ambientes. A Estação Anhembi foi o único local com estimativa de correlação menor que 0,80 com o conjunto GERAL. O Departamento de Genética, de modo geral, apresentou menor correlação com os demais locais, com exceção do conjunto GERAL, o que corrobora os resultados da Tabela 3.3, a qual indica presença de interação significativa deste local com todos os demais.

O interesse em verificar a similaridade ou não entre os locais se justifica para sustentar os resultados da análise de predição genômica dos híbridos que serão discutidos nos próximos tópicos. A falta de similaridade do comportamento fenotípico de uma mesma população avaliada em diferentes locais está ligada principalmente a variações edafoclimáticas e, consequentemente da manifestação da interação G x E (Bandeira & Sousa et al., 2017; Costa-Neto et al., 2021; Jarquín et al., 2021). Na presença de interação G x E significativa, a habilidade preditiva dos modelos de predição genômica pode decair significativamente quando o modelo é calibrado em um ambiente e validado em outro (Mendes & Souza Júnior, 2016).

3.3.2 Habilidades preditivas dos modelos de predição genômica

A GS é uma ferramenta genômica promissora para predizer o desempenho fenotípico de indivíduos apenas genotipados, sem necessidade de fenotipagem (Kaler et al., 2022). Busca-se, aqui, entender como a similaridade dos locais de treinamento e teste influenciam as habilidades preditivas dos modelos de predição e indicar se o uso de um modelo geral pode ser mais vantajoso que utilizar modelos ambiente-específicos; além de verificar o impacto da incorporação dos efeitos de dominância nos modelos paramétricos nos valores de habilidade de predição, devido o milho apresentar alta heterose. Com isso, selecionar o melhor modelo de predição para identificação das combinações híbridas superiores neste conjunto de linhagens para produtividade de grãos. Por fim, estabelecer algumas implicações da GS no melhoramento de milho visando híbridos.

3.3.2.1 Modelos ambiente-específicos *versus* modelo geral

As habilidades preditivas dos modelos tiveram grande amplitude de variação entre os diferentes esquemas de treinamento e validação (0,23 a 0,83) (Figura 3.4). Como esperado, independentemente do modelo, valores mais altos de habilidade preditiva foram encontrados nas situações em que os modelos foram treinados e validados no mesmo local, com exceção de situações que o conjunto GERAL foi o local de validação. Na validação dentro de locais, os locais isolados tiverem valores máximos de habilidade preditiva semelhantes, entre 0,75 (Areão) e 0,81(Caterpillar) (Figura 3.4).

O comportamento diferenciado dos valores de habilidade preditiva dentro dos locais está relacionado, além do modelo utilizado, principalmente à combinação da precisão de obtenção dos dados fenotípicos e a variância genética presente no painel avaliado, ou seja, a magnitude da herdabilidade (Pandey et al., 2020; Zhang et al., 2017). Maiores valores de habilidade preditiva foram obtidos quando se considerou todos os ambientes (GERAL), ultrapassando 0,80 para produtividade de grãos (Figura 3.4). O valor de $\hat{h}^2$ nesta situação foi quase de uma unidade para produtividade de grãos (0,95). Morais Júnior et al. (2018) destacam que caracteres com maior variância genética apresentam também maior herdabilidade e, portanto, são mais previsíveis por meio de uma abordagem genômica, tal como a GS.

A validação dos modelos de predição em locais diferentes daqueles que ocorreram o treinamento, de modo geral, levou a decréscimos relevantes nas habilidades preditivas para produtividade de grãos, principalmente quando os locais de validação foram Departamento de Genética e Patos de Minas (Figuras 3.4). O comportamento de decréscimo na *performance* dos modelos quando a validação ocorre entre locais, mesmo quando se utiliza o mesmo conjunto híbridos e marcadores, está pautado no efeito da condição ambiental inerente a cada local (Costa-Neto et al., 2021; Jarquín et al., 2021). Este efeito dos locais sobre a expressão fenotípica dos caracteres é a interação G x E (Chaves, 2001).

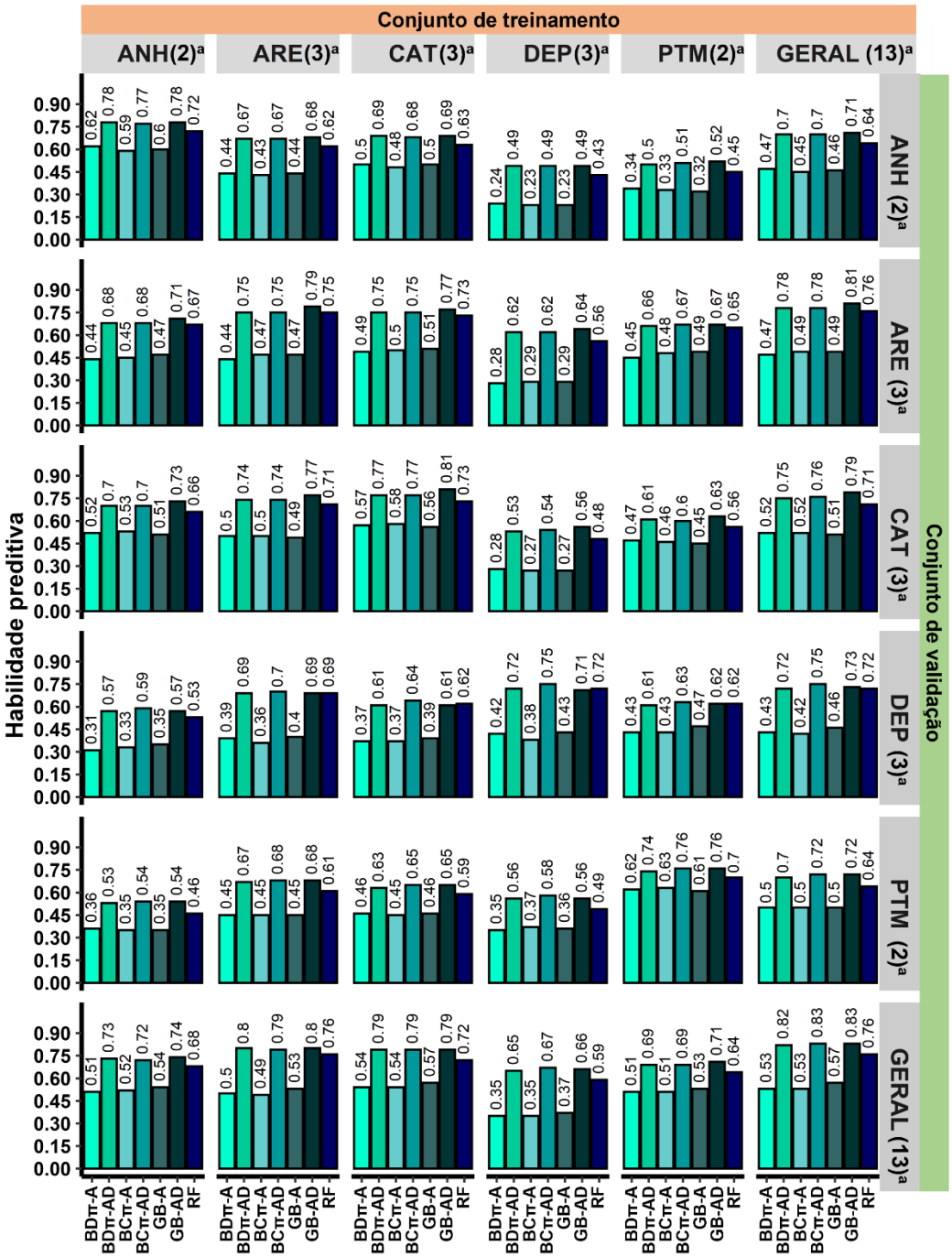

Figura 3.4. Habilidades preditivas de modelos de predição genômica para produtividade de grãos (t ha^{-1}) em híbridos simples de milho avaliados em diferentes locais. Locais: Estação Anhembi (ANH); Estação Areão (ARE), Estação Caterpillar (CAT), Departamento de Genética da ESALQ/USP (DEP); Estação Experimental da Empresa Biomatrix em Patos de Minas (PTM) e GERAL (conjunto de todos os ambientes). Modelos aditivos: Bayes Dπ (BDπ-A), Bayes Cπ (BCπ-A), GBLUP (GB-A) e *Random Forest* (RF). Modelos aditivo-dominantes: Bayes Dπ (BDπ-AD), Bayes Cπ (BCπ-AD) e GBLUP (GB-AD).

ª: número de ambientes em que o caráter foi avaliado por local (ambientes refere-se à combinação local x ano ou local x nº de experimentos).

Otimizar uma rede de ensaios que capture os efeitos da interação G xE para maximizar a acurácias dos GEBVs representa um desafio para os melhorista de plantas (Huang et al., 2018). É esperado que, com o aumento do número de ambientes testados, aumente também a diversidade ambiental e a variação deste fator seja significativa devido às diversas condições climáticas e de solo específicas, o que torna muito difícil otimizar os modelos de GS (Pandey et al., 2020). A interação G x E significativa leva à diminuição da habilidade preditiva dos modelos, principalmente na tentativa de expandir um modelo treinado e calibrado em um local para predizer o desempenho de indivíduos em outros ambientes (Mendes & Souza Júnior, 2016; Huang et al., 2018).

Os modelos treinados em Anhembi para produtividade de grãos, tiveram altos valores de habilidade preditiva quando validados em Areão e Caterpillar, atingindo 0,68 e 0,69, respectivamente (Figura 3.4). Na Tabela 3.3, quando se analise a significância do componente de interação G x E entre os locais dois a dois, nota-se que Anhembi e Areão tiveram este componente significativo. Porém, este efeito não foi relevante o suficiente para acarretar quedas consideráveis de habilidade preditiva neste par de locais. De modo contrário, verificou-se que habilidade preditiva dos modelos calibrados em Anhembi tiveram baixa eficiência ao serem validados no Departamento de Genética e em Patos de Minas. Principalmente no Departamento de Genética, os valores de habilidade preditiva ficaram abaixo de 0,50, justificando o efeito significativo do componente da interação G x E entre os dois locais (Tabela 3.3). Também foi detectada baixa correlação entre o ranqueamento dos híbridos em Anhembi e Departamento de Genética (0,48) e Anhembi e Patos de Minas (0,51), o que favorece uma menor habilidade preditiva nestes pares de locais (Figura 3.3).

Os modelos treinados em Areão foram validados com valores moderados a altos de habilidade preditiva nos demais locais, todos com valores máximos acima de 0,60, mesmo este local apresentando componente de variação da interação G x E significativo com todos os demais locais. Para os modelos treinados em Caterpillar, a validação apresentou valores notáveis de habilidade preditiva em Anhembi (0,44 a 0,71), Areão (0,49 a 0,77) e Patos de Minas (0,45 a 0,63). Também, Caterpillar apresentou o componente de interação G x E significativo com quase todos os outros locais (Tabela 3.3), contudo, a similaridade entre o ranqueamento dos híbridos deste local com os demais foi maior que 0,63 (Figura 3.3), o que pode justificar a boa predição dos híbridos nos locais distintos do qual o modelo foi treinado.

Nos modelos calibrados no Departamento de Genética, a validação ocorreu satisfatoriamente em Areão, Caterpillar e Patos de Minas (Figura 3.4). Como pode ser visto na Tabela 2.3, os pares de locais que envolveram o Departamento de Genética foram os que apresentaram os maiores valores significativos do componente de interação G x E. Possivelmente, por isso, as habilidades preditivas nestes pares de locais foram menores que as demais, como indica na literatura o efeito negativo da interação G x E na habilidade de predição de modelos genômicos (Bandeira & Souza et al., 2017; Morais Júnior et al., 2018; Costa-Neto et al., 2021; Jarquín et al., 2021). Da mesma forma, Areão e Caterpillar foram os melhores locais de validação para os modelos treinados em Patos de Minas e apresentaram altos valores de similaridade do comportamento dos híbridos ($\hat{r}$ = 0,67 e $\hat{r}$ = 0,68, respectivamente) (Figura 3.4).

Independentemente do local de treinamento dos modelos, quando a validação ocorre no conjunto GERAL, os valores de habilidade preditiva são os mais altos (Figura 3.4). Possivelmente, devido as médias obtidas neste conjunto, BLUEs, apresentarem uma precisão muito maior que quando elas são estimadas em locais separados. Assim, tanto os modelos treinados em GERAL e validados nos locais isolados ou, de forma contrária, quanto quando os modelos dos locais isolados são validados em GERAL, suas *performances* são bastante semelhantes e atrativas.

Portanto, neste conjunto de híbridos e locais é possível a utilização de um único modelo geral baseado no conjunto de todos os ambientes (GERAL) para predição do desempenho dos indivíduos, tanto num contexto de todos os ambientes, quanto neles isoladamente. Esta recomendação é baseada na comparação dos valores de habilidade preditiva obtidos dos modelos quando treinados e validados no mesmo local e quanto o conjunto GERAL é utilizado como local de validação, cuja similaridades das habilidades são altas. Além disso, esta abordagem economiza tempo e recursos computacionais por estabelecer apenas um modelo e não ter que assumir modelos específicos para cada local.

3.3.2.2 Impacto da incorporação dos efeitos de dominância nos modelos de predição

A dominância desempenha papel essencial na base genética da heterose (Tang et al., 2010). Contabilizar seus efeitos em modelos de predição genômica pode aumentar significativamente as habilidades preditivas em culturas como o milho, cujos caracteres apresentam algum nível de expressão deste fenômeno (Technow et al., 2012; Almeida-Filho et al., 2019; Dias et al., 2018; Li et al., 2021; Wang et al., 2019). A produtividade

de grãos em milho é um dos caracteres que mais expressa heterose. Mesmo assim, alguns estudos de GS não encontraram melhora nas habilidades preditivas dos modelos ao contabilizaram os efeitos de dominância neste (Li et al., 2020).

Neste conjunto de híbridos, foi observado um impacto positivo ao se incorporar os efeitos de dominância nos modelos de predição para produtividade de grãos em todos os locais e no conjunto GERAL, com um aumento médio de cerca de 25% nos valores de habilidade preditiva dos modelos (Figura 3.4). Quanto aos ensaios situados no Departamento de Genética o incremento de habilidade preditiva atingiu, em alguns casos, mais de 30%. Finalmente, no conjunto GERAL, as habilidades preditivas ultrapassaram 0,80 com os modelos aditivo-dominantes para produtividade de grãos (Figura 3.4). Estes resultados corroboram os encontrados por vários autores que utilizaram modelos de predição genômica com incorporação dos efeitos de dominância (Technow et al., 2012; Almeida-Filho et al., 2019; Dias et al., 2018; Wang et al., 2019, Alves et al., 2021). Em simulação, Santos et al. (2015) notaram que as acurácias preditivas de diferentes modelos de predição saltaram de 0,70, 0,76 e 0,77 para 0,83, 0,90 e 0,94 para produtividade de grãos, com herdabilidades variando de 0,30, 0,50 e 0,70 respectivamente, com a incorporação dos efeitos de dominância nos modelos.

Com base nos resultados obtidos neste estudo e em confronto com demais similares na literatura, é possível inferir que os efeitos a serem contabilizados nos modelos de predição devem ser determinados para cada caráter e população. Para produtividade de grãos, é indispensável testar modelos aditivo-dominantes para obtenção de melhores performances da GS, dada a natureza poligênica desse caráter e a alta influência de efeitos de heterose.

3.3.2.3 Seleção do modelo de predição com melhor performance para produtividade de grãos

Foram testados três modelos paramétricos com e sem incorporação de efeitos de dominância (Bayes Dπ, Bayes Cπ e GBLUP) e um não paramétrico (*Random Forest*) para selecionar o que mais precisamente consegue predizer o desempenho fenotípico do conjunto de híbridos para produtividade de grãos. A seleção do modelo foi direcionada pelos resultados de GS considerando todos os ambientes (GERAL). Isto porque, este conjunto apresenta BLUEs mais precisos para predizer o desempenho dos híbridos para qualquer ambiente deste conjunto.

As habilidades preditivas dos modelos para o desempenho dos híbridos foram de moderadas a altas, variando de 0,53 a 0,83 (Figura 3.4). Quando se observa apenas os três modelos exclusivamente aditivos com o modelo de *Random Forest*, nota-se superioridade deste último (habilidade preditiva = 0,76), com aumento médio de habilidade preditiva maior que 0,20 em relação aos demais modelos para produtividade de grãos (Bayes Dπ e Bayes Cπ = 0,53, GBLUP = 0,57). Contrário aos resultados deste estudo, Zhang et al. (2020) encontraram melhores habilidades preditivas para o GBLUP (0,83) em comparação ao *Random Forest* (0,62) também utilizando exclusivamente efeitos aditivos no modelo paramétrico.

No cenário em que é incluído o efeito de dominância nos modelos paramétricos, verifica-se que todos superam o *Random Forest* (Figura 3.4). Também, apesar dos modelos baseados em regressões paramétricas, como os Bayesianos e o GBLUP, apresentarem suposições distintas a respeito das variâncias dos efeitos dos marcadores para a explicação da variação fenotípica do caráter e no modo que a informação dos marcadores é adicionada, suas estimativas de habilidade preditiva foram bastante semelhantes. Os valores de habilidade preditiva foram de 0,82, 0,83 e 0,83 nos modelos Bayes Dπ, Bayes Cπ e GBLUP, respectivamente.

Os modelos oriundos do "alfabeto bayesiano" consistem em regressões dos fenótipos em marcadores, ao passo que os modelos frequentistas, como GBLUP e RR-BLUP, usam os marcadores para construir matrizes de relacionamento genômico, usadas, por sua vez, para modelar a covariância entre os efeitos genéticos (Schrauf et al., 2021). Em termos de suposições dos efeitos dos marcadores nos modelos, o GBLUP assume que todos os marcadores têm variâncias iguais com pequenos efeitos na variação genética, mas diferente de zero. O modelo Bayes Dπ assume como *priori* para efeitos de marcador uma mistura de distribuições normais e denota que cada SNP tem sua própria variância. No modelo Bayes Cπ, a distribuição a *priori* para os efeitos de marcador também é dada por uma mistura de distribuições normais, no entanto, este modelo assume a mesma variância genética para todos os marcadores.

Todavia, os modelos paramétricos, sejam bayesianos ou frequentistas, não apresentaram diferenças relevantes nas habilidades preditivas neste estudo, assim como foi verificado por Li et al. (2020). Contudo, isto não é regra, como já foi discutido, a magnitude e variação das habilidades preditivas entre estes modelos pode oscilar por e estes dependerem fortemente, dentre outros fatores, do tipo de caráter alvo da GS (Riedelsheimer et al. 2012; Li et al. 2020).

O modelo escolhido para predizer os desempenhos fenotípicos de produtividade de grãos neste conjunto de híbridos poderia ser qualquer um dos modelos paramétricos com a incorporação dos efeitos de dominância. Mas, levando em conta o esforço computacional, foi optado pelo modelo GBLUP.

3.3.2.4 Implicações da seleção genômica para o melhoramento de milho híbrido

Utilizou-se o modelo GBLUP aditivo-dominante, treinado e validado no conjunto de ambientes GERAL com 152 híbridos fenotipados, para predizer o desempenho de todos os 861 híbridos simples possíveis das 42 linhagens do painel para produtividade de grãos. Os resultados corroboram aos relatados na literatura e comprovam o poder da seleção genômica para a identificação de combinações híbridas superiores (Figura 3.5).

O direcionamento de cruzamentos entre linhagens, a seleção em estágios iniciais, e redução do número de ensaios e locais de testagem são os principais fatores responsáveis por aumentar os ganhos genéticos com a implementação da GS (Beyene et al., 2021; Montesinos-López et al., 2021). Os ganhos e a disposição das melhores combinações preditas pelo modelo escolhido estão representados na Figura 3.5.

Ao selecionar com base nos desempenhos preditos as 15 melhores combinações (*Top15*) dos 152 híbridos testados (primeiro conjunto) e *Top15* dos 861 possíveis (segundo conjunto), verifica-se um incremento médio de 0,21 t/ha para produtividade de grãos (Figura 3.5A) entre os *Top15* oriundos do primeiro para o segundo conjunto. Para produtividade de grãos, houve coincidência de 7 combinações selecionadas nos dois conjuntos (Figura 3.5B), isto é, 8 combinações ainda precisam ser realizadas para obtenção aproximada do diferencial de ganho de 0,21 t/ha.

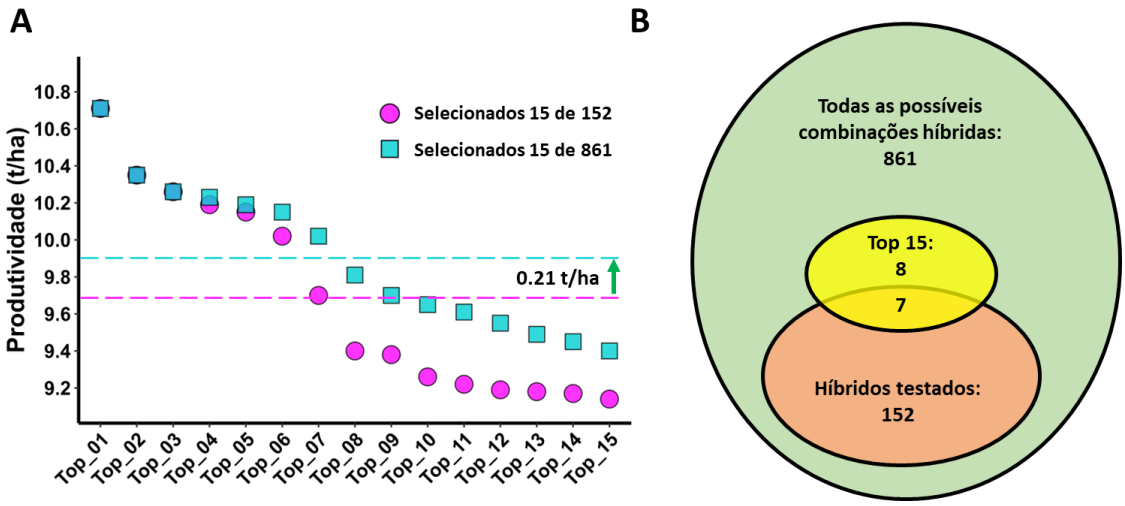

Figura 3.5. (A): distribuição das 15 melhores combinações de cruzamentos selecionadas a partir dos valores preditos dos 152 híbridos testados e dos valores preditos de todas as 861 combinações híbridas possíveis para produtividade de grãos (t/ha). Linhas pontilhadas em rosa e azul representam o desempenho médio predito das 15 melhores combinações de cruzamentos selecionadas de 152 híbridos testados e dos 861 híbridos possíveis, respectivamente. Seta verde indica a diferença entre a médias dos dois conjuntos de híbridos. (B): disposição das combinações híbridas promissoras (*Top 15*) no conjunto de híbridos testados e não testados para produtividade de grãos.

Esta abordagem está passando por um período de intensa pesquisa científica e aplicação em programas de melhoramento. Com foi visto, o incremento dos ganhos genéticos está atrelado a redução de custo e tempo gasto na obtenção de genótipos superiores para caracteres de interesse.

3.4 CONCLUSÕES

Estabelecer um modelo geral baseado no conjunto de ambientes (GERAL) é a alternativa mais atrativa para predizer o desempenho dos híbridos para produtividade de grãos, tanto num contexto de combinação de todos os ambientes quanto neles isoladamente.

Incorporar os efeitos de dominância nos modelos de predição implica fortemente no aumento da habilidade preditiva para caracteres que possuem alta expressão de heterose, como a produtividade de grãos.

O modelo mais indicado para predizer o desempenho fenotípico da produtividade de grãos, neste conjunto de dados, é o GBLUP com a contabilização dos efeitos aditivos e de dominância.

Com o GBLUP aditivo-dominante é possível realizar seleção de melhores combinações de linhagens do que as já realizadas que, potencialmente, elevam a produtividade de grãos em cerca de 0,21 t ha^{-1} ao selecionar os melhores 15 híbridos pela GS neste caráter.

3.5 REFERÊNCIAS

ABDOLLAHI-ARPANAHI, R.; GIANOLA, D.; PEÑAGARICANO, F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. **Genetics Selection Evolution**, v. 52, n. 1, p. 1–15, 2020.

AGUIAR, A.M.; et al. Combining ability of inbred lines of maize and stability of their respective single-crosses. **Scientia Agricola**, Piracicaba, v.60, n.1, p.83-89, 2003.

ALMEIDA FILHO, J. E. et al. Genomic prediction of additive and non-additive effects using genetic markers and pedigrees. **G3: Genes, Genomes, Genetics**, v. 9, n. 8, p. 2739–2748, 2019.

ALVES, F. C. et al. Impact of the complexity of genotype by environment and dominance modeling on the predictive accuracy of maize hybrids in multi-environment prediction models. **Euphytica**, v. 217, n. 3, 2021.

BANDEIRA & SOUSA, M. et al. Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. **G3: Genes, Genomes, Genetics**, v. 7, n. 6, p. 1995–2014, 2017.

BATES, D. et al. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1-48, 2015.

BEYENE, Y. et al. Application of genomic selection at the early stage of breeding pipeline in tropical maize. **Frontiers in Plant Science**, v. 12, p. 1–11, 2021.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BROWNING, B. L.; BROWNING, S. R. Genotype imputation with millions of reference samples. **American Journal of Human Genetics**, v. 98, n. 1, p. 116–126, 2016.

CHAVES, L. J. Interação de genótipos com ambientes. In: NASS, L. L.; VALOIS, A. C. C.; MELO, I. S. de; VALADARES-INGLIS, M. C. (Ed.). **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação MT, p. 673-714, 2001. 1183 p.

COSTA-NETO, G.; FRITSCHE-NETO, R.; CROSSA, J. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. **Heredity**, v. 126, n. 1, p. 92–106, 2021.

DESTA, Z. A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in Plant Science**, v. 19, n. 9, p. 592–601, 2014.

DIAS, K. O. D. G. et al. Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. **Heredity**, v. 121, n. 1, p. 24–37, 2018.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, n. 19, v. 1, p. 11-15, 1987.

FRITSCHE-NETO, R. et al. TCGA: a tropical corn germplasm assembly for genomic prediction and high-throughput phenotyping studies. **Mendeley Data**, V3, 2020.

FRITSCHE-NETO, R. et al. Atualização da proposta de classificação dos coeficientes de variação para a cultura do milho. **Acta Scientiarum - Agronomy**, v. 34, n. 1, p. 99–101, 2012.

GLAUBITZ, J. C. et al. TASSEL-GBS: A High-Capacity Genotyping by Sequencing Analysis Pipeline. **PLOS ONE**, v. 9, n. 2, p. e90346, 2014.

GRANATO, I. S. C. et al. snpReady: a tool to assist breeders in genomic analysis. **Molecular Breeding**, v. 38, n. 8, 2018.

HUANG, M. et al. The accuracy of genomic prediction between environments and populations for soft wheat traits. **Crop Science**, v. 58, n. 6, p. 2274–2288, 2018.

JARQUIN, D. et al. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. **Frontiers in Genetics**, v. 11, p. 592–769, 2021.

KADAM, D. C. et al. Genomic prediction of single crosses in the early stages of a maize hybrid breeding pipeline. **G3: Genes, Genomes, Genetics**, v. 6, n. 11, p. 3443–3453, 2016.

KALER, A. S. et al. Genomic prediction models for traits differing in heritability for soybean, rice, and maize. **BMC Plant Biology**, v. 22, n. 1, p. 87, 2022.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 2012.

LENTH, R. V. (2021). **emmeans: estimated marginal means, aka least-squares means**. R package version 1.7.0. Disponível em <https://CRAN.R-project.org/package=emmeans>

LI, B. et al. Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. **Frontiers in Genetics**, v. 9, p. 1–20, 2018.

LI, D. et al. Genetic dissection of hybrid performance and heterosis for yield-related traits in maize. **Frontiers in Plant Science**, v. 12, p. 1–19, 2021.

LI, G. et al. Genome-wide prediction in a hybrid maize population adapted to Northwest China. **Crop Journal**, v. 8, n. 5, p. 830–842, 2020.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18–22, 2002.

MENDES, M. P.; SOUZA JÚNIOR, C. L. Genomewide prediction of tropical maize single-crosses. **Euphytica**, v. 209, n. 3, p. 651–663, 2016.

MEUWISSEN, T. H. E. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

MONTESINOS-LÓPEZ, O. A. et al. A new deep learning calibration method enhances genome-based prediction of continuous crop traits. **Frontiers in Genetics**, v. 12, p. 1–12, 2021.

MORAIS JÚNIOR, O. P. et al. Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials. **Crop Science**, v. 58, n. 2, p. 592–607, 2018.

PANDEY, M. K. et al. Genome-based trait prediction in multi- environment breeding trials in groundnut. **Theoretical and Applied Genetics**, v. 133, n. 11, p. 3101–3117, 2020.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression and prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483–495, 2014.

POLAND, J. A. et al. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **PLOS ONE**, v. 7, n. 2, p. e32253, 2012.

R Core Team (2021). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

REVELLE, W. (2021). **psych: procedures for personality and psychological research**. Northwestern University, Evanston, Illinois, USA. Disponível em: <https://CRAN.R-project.org/package=psych Version = 2.1.9,.>

RIEDELSHEIMER, C.; TECHNOW, F.; MELCHINGER, A. E. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. **BMC Genomics**, v. 13, n. 1, 2012.

ROBERTSEN, C.; HJORTSHØJ, R.; JANSS, L. Genomic Selection in Cereal Breeding. **Agronomy**, v. 9, n. 2, p. 95, 2019.

SANTOS, J. P. R. et al. Genome-wide prediction of maize single-cross performance, considering non-additive genetic effects. **Genetics and Molecular Research**, v. 14, n. 4, p. 18471–18484, 2015.

SCHRAUF, M. F.; DE LOS CAMPOS, G.; MUNILLA, S. Comparing Genomic Prediction Models by Means of Cross Validation. **Frontiers in Plant Science**, v. 12, p. 1–11, 2021.

SIM, S. C. et al. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. **PLOS ONE**, v. 7, n. 7, p. e40563, 2012.

TECHNOW, F. et al. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. **Theoretical and Applied Genetics**, v. 125, n. 6, p. 1181–1194, 2012.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VENCOVSKY, R.; BARRIGA, P. **Genética biométrica no fitomelhoramento**. 1. ed. Ribeirão Preto: Revista Brasileira de Genética, 1992. 496 p.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, v. 195, n. 4, p. 1223–1230, 2013.

WANG, Z. et al. Physiological basis of heterosis for nitrogen use efficiency of maize. **Scientific Reports**, v. 9, n. 1, p. 1–11, 2019.

WICKHAM, H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York. Disponível em: < https://ggplot2.tidyverse.org.>

ZENG, J. et al. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genetics Selection Evolution**, v. 45, n. 1, p. 1–17, 2013.

ZHANG, A. et al. Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. **Frontiers in Plant Science**, v. 8, 2017.

ZHANG, M. et al. Accurate prediction of maize grain yield using its contributing genes for gene-based breeding. **Genomics**, v. 112, n. 1, p. 225–236, 2020.

4 CONSIDERAÇÕES FINAIS

Considerando um programa que possua 1000 linhagens elite passíveis de serem combinadas, podemos obter 499.500 combinações híbridas destas linhagens duas a duas, além disso, ainda existe a possibilidade da formação de híbridos duplos e triplos, como no caso do milho. É praticamente impossível um programa de melhoramento conseguir testar todas as combinações possíveis entre suas linhagens em vários ambientes para ranquear as de melhor desempenho fenotípico Nesse sentido, o uso da predição genômica multi-ambiental é uma ferramenta interessante para redução de custos e tempo na triagem de genótipos sob múltiplos ambientes em fases mais avançadas de programas de melhoramento, já que permite a predição do comportamento fenotípico das combinações com base nos genótipos dos genitores.

Pensando na "equação do melhorista", a implicação maior da GS é na possibilidade de explorar uma maior leva da variância genética presente no conjunto de genótipos trabalhados em menor intervalo de tempo. Isto é, o ganho genético potencial para determinado caráter se torna maior. Além de permitir que o melhorista concentre recursos financeiros e humanos em genótipos ou combinações que sejam potencialmente superiores com base nos resultados de GS.

Portanto, é essencial a implementação da GS em programas de melhoramento de plantas visando maior eficiência de seleção, redução de custo e tempo gasto na obtenção de genótipos superiores.

5 REFERÊNCIAS

ALVES, F. C. et al. Impact of the complexity of genotype by environment and dominance modeling on the predictive accuracy of maize hybrids in multi-environment prediction models. **Euphytica**, v. 217, n. 3, 2021.

BELÍCUAS, P. R. et al. Inheritance of the stay-green trait in tropical maize. **Euphytica**, v. 198, n. 2, p. 163–173, 2014.

BERNARDO, R. Prediction of maize single-cross performance using rflps and information from related hybrids. **Crop Science**, v. 34, n. 1, p. 20–25, 1994.

BERNARDO, R. Testcross additive and dominance effects in best linear unbiased prediction of maize single-cross performance. **Theoretical and Applied Genetics**, v. 93, n. 7, p. 1098–1102, 1996.

BEYENE, Y. et al. Application of genomic selection at the early stage of breeding pipeline in tropical maize. **Frontiers in Plant Science**, v. 12, p. 1–11, 2021.

CASEYS, C. Senescence: The genetics behind stay-green corn. **Plant Cell**, v. 31, n. 9, p. 1934–1935, 2019.

CHEN, X.; ISHWARAN, H. Random forests for genomic data analysis. **Genomics**, v. 99, n. 6, p. 323–329, 2012.

CROSSA, J. et al. Genomic selection and prediction in plant breeding. **Journal of Crop Improvement**, v. 25, n. 3, p. 239–261, 2011.

CROSSA, J. et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v. 186, n. 2, p. 713–724, 2010.

CUI, Z. et al. Assessment of the potential for genomic selection to improve husk traits in maize. **G3 Genes|Genomes|Genetics**, v. 10, n. 10, p. 3741–3749, 2020.

DE LOS CAMPOS, G. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, n. 1, p. 375–385, 2009.

DE LOS CAMPOS, G. et al. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v. 193, n. 2, p. 327–345, 2013.

DESTA, Z. A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in Plant Science**, v. 19, n. 9, p. 592–601, 2014.

FISHER, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Transactions of the Royal Society of Edinburgh**, v. 52, n. 2, p. 399–433, 1919.

GONZÁLEZ-CAMACHO, J. M. et al. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics, v. 125, n. 4, p. 759–771, 2012.

GONZÁLEZ-RECIO, O.; FORNI, S. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. **Genetics Selection Evolution**, v. 43, n. 1, p. 7, 2011.

HABIER, D. et al. Extension of the bayesian alphabet for genomic selection. **BMC bioinformatics**, v. 12, p. 186, 2011.

JARQUIN, D. et al. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. **Frontiers in Genetics**, v. 11, p. 592–769, 2021.

KALER, A. S. et al. Genomic prediction models for traits differing in heritability for soybean, rice, and maize. **BMC Plant Biology**, v. 22, n. 1, p. 87, 2022.

LI, Z.; SILLANPÄÄ, M. J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. **Theoretical and applied genetics**, v. 125, n. 3, p. 419–435, 2012.

LORENZANA, R. E.; BERNARDO, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. **Theoretical and Applied Genetics**, v. 120, n. 1, p. 151–161, 2009.

LUCHE, H. S. et al. Stay-green: A potentiality in plant breeding. **Ciência Rural**, v. 45, n. 10, p. 1755–1760, 2015.

MAENHOUT, S. et al. Support vector machine regression for the prediction of maize hybrid performance. **Theoretical and Applied Genetics**, v. 115, n. 7, p. 1003–1013, 2007.

MENDES, M. P.; SOUZA JÚNIOR, C. L. Genomewide prediction of tropical maize single-crosses. **Euphytica**, v. 209, n. 3, p. 651–663, 2016.

MEUWISSEN, T. H. E. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819–1829, 2001.

OLIVEIRA, A. A. et al. Genomic prediction applied to multiple traits and environments in second season maize hybrids. **Heredity**, v. 125, n. 1–2, p. 60–72, 2020.

PIEPHO, H. Ridge regression and extensions for genomewide selection in maize. **Crop Science**, v. 49, p. 1165–1176, 2009.

SEKHON, R. S. et al. Integrated genome-scale analysis identifies novel genes and networks underlying senescence in maize. **Plant Cell**, v. 31, n. 9, p. 1968–1989, 2019.

SHIKHA, M. et al. Genomic selection for drought tolerance using genome-wide SNPs in Maize. **Frontiers in Plant Science**, v. 8, p. 1–12, 2017.

TECHNOW, F. et al. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. **Genetics**, v. 197, n. 4, p. 1343–1355, 2014.

USAI, M. G.; GODDARD, M. E.; HAYES, B. E. N. J. LASSO with cross-validation for genomic selection. **Genetics Research**, v. 91, n. 6, p. 427–436, 2009.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

WANG, X. et al. Genomic selection methods for crop improvement: Current status and prospects. **Crop Journal**, v. 6, n. 4, p. 330–340, 2018.

ZHANG, M. et al. Accurate prediction of maize grain yield using its contributing genes for gene-based breeding. **Genomics**, v. 112, n. 1, p. 225–236, 2020.

APÊNDICE

Apêndice A. Valores preditos para as melhores 15 combinações híbridas testados a campo (Pred152) e combinações possíveis no conjunto de linhagens (Pred861) para produtividade de grãos (t/ha) e *staygreen* (nota 1-5) em milho tropical.

Produtividade de grãos (t/ha)				*Staygreen* (nota 1-5)			
Comb.	Pred152	Comb.	Pred861	Comb.	Pred152	Comb.	Pred861
73x177*	10,71	73x177*	10,71	80x83	2,72	80x89	2,22
70x177*	10,35	70x177*	10,35	89x81	2,85	79x89	2,41
75x177*	10,26	75x177*	10,26	101x177	2,93	70x89	2,47
67x177*	10,19	69x70	10,23	89x177	2,96	89x101	2,47
72x177*	10,15	67x177*	10,19	70x83	2,97	89x107	2,49
74x177*	10,02	72x177*	10,15	79x83	2,99	84x89	2,53
71x177*	9,70	74x177*	10,02	70x177	3,01	70x80	2,58
105x177	9,40	69x73	9,81	89x83	3,02	80x107	2,60
84x177	9,38	71x177*	9,70	79x177	3,02	72x89	2,60
106x177	9,26	69x107	9,65	107x83	3,03	89x98	2,61
75x81	9,22	70x75	9,61	67x177	3,04	89x102	2,63
87x177	9,19	70x86	9,55	72x177	3,06	80x84	2,65
89x177	9,18	104x177	9,49	101x83	3,06	70x107	2,66
101x177	9,17	73x84	9,45	103x177	3,07	73x89	2,68
91x177	9,14	70x106	9,40	72x83	3,10	89x100	2,68
Média	9,69	Média	9,90	Média	2,99	Média	2,55

*: Mesmas combinações dentro das 15 melhores em cada conjunto.

ANEXO

Anexo A. Origem e caracterização de grãos das linhagens genitoras dos híbridos deste estudo.

Seq.	ID_M360[1]	Nome	Origem	Tipo de grão[1]	Cor de grão[1]	Seq.	ID_M360[1]	Nome	Origem	Tipo de grão[1]	Cor de grão[1]
L1	id_0067	16-02D	IG-2	Duro	Laranja	L22	id_0092	82-01D	TOPCROSS	Duro	Laranja
L2	id_0068	25-04D	IG-2	Duro	Laranja	L23	id_0093	131-01F	TOPCROSS	Duro	Laranja
L3	id_0069	39-05D	IG-2	Dentado	Laranja	L24	id_0094	16-04R	BR-201	Duro	Vermelho
L4	id_0070	55-02D	IG-2	Duro	Laranja	L25	id_0095	16-07R	BR-201	Duro	Laranja
L5	id_0071	66-08D	IG-2	Semi-duro	Laranja	L26	id_0096	22-02D	BR-201	Duro	Amarelo
L6	id_0072	94-02D	IG-2	Semi-dentado	Amarelo	L27	id_0097	23-05D	BR-201	Duro	Amarelo claro
L7	id_0073	102-2D	IG-2	Duro	Laranja	L28	id_0098	24-03D	BR-201	Duro	Amarelo
L8	id_0074	120-04F	IG-2	Duro	Amarelo claro	L29	id_0099	29-03D	BR-201	Duro	Laranja
L9	id_0075	149-05D	IG-2	Semi-duro	Laranja	L30	id_0100	35-04F	BR-201	Semi-duro	Laranja
L10	id_0077	20-02R	HS-1	Duro	Laranja	L31	id_0101	18-08AF	CMS-05	Duro	Laranja
L11	id_0078	30-07F	HS-1	Duro	Laranja	L32	id_0102	37-02BD	CMS-05	Duro	Laranja
L12	id_0079	31-01F	HS-1	Duro	Laranja	L33	id_0103	37-03BD	CMS-05	Duro	Laranja
L13	id_0080	33-04D	HS-1	Semi-duro	Laranja	L34	id_0104	37-04BF	CMS-05	Duro	Laranja
L14	id_0082	04-05F	IG-1	Duro	Laranja	L35	id_0105	37-07BD	CMS-05	Duro	Laranja
L15	id_0084	88-05F	IG-1	Duro	Laranja	L36	id_0178	61-02F	TOPCROSS	Semi-dentado	Laranja amarelado
L16	id_0085	8F	XL-560	Semi-duro	Laranja	L37	id_0106	84-03F	TOPCROSS	Duro	Laranja
L17	id_0086	14D	XL-560	Duro	Laranja	L38	id_0107	08-04BD	CMS-05	Duro	Laranja
L18	id_0087	56D	XL-560	Duro	Laranja	L39 (T1)	id_0083	L-08-05F	IG-1	Semi-duro	Laranja
L19	id_0088	128D	XL-560	Semi-dentado	Laranja	T40 (T2)	id_0081	L-36-07F	HS-1	Dentado	Amarelo
L20	id_0089	45-03D	TOPCROSS	Duro	Amarelo	T41 (T3)	id_0090	L-49-02D	TOPCROSS	Duro	Laranja
L21	id_0091	53-01F	TOPCROSS	Duro	Laranja	T42 (T4)	id_0177	L-46-10D	BR-201	Duro	Laranja

[1]: Informações retiradas de anexos do data base "*TCGA: a tropical corn germplasm assembly for genomic prediction and high-throughput phenotyping studies"* (Fritsche-Neto et al., 2020).