JOÃO FILIPI RODRIGUES GUIMARÃES

EFEITO DA INTERAÇÃO DOMINÂNCIA X AMBIENTE NA HABILIDADE DE PREDIÇÃO GENÔMICA

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

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
G963e
2016

Guimarães, João Filipi Rodrigues, 1985-
Efeito da interação dominância x ambiente na habilidade de predição genômica / João Filipi Rodrigues Guimarães. – Viçosa, MG, 2016.
vii, 30f. : il. ; 29 cm.

Orientador: Fabyano Fonseca e Silva.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. *Pinus taeda* L - Melhoramento genético. 2. Interação genótipo-ambiente. I. Universidade Federal de Viçosa. Departamento de Biologia. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22 ed. 634.9751

JOÃO FILIPI RODRIGUES GUIMARÃES

EFEITO DA INTERAÇÃO DOMINÂNCIA X AMBIENTE NA HABILIDADE DE PREDIÇÃO GENÔMICA

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

APROVADA: 29 de Julho de 2016.

Marciane Oliveira

Rodrigo Neves Graça

Camila Ferreira Azevedo

Cosme Damião Cruz
(Coorientador)

Fabyano Fonseca e Silva
(Orientador)

Aos meus pais, Luiz Claudio e Rivane.

Às minhas irmãs Sarah e Ana.

Ao meu sobrinho Luiz Felipe.

À minha querida esposa Natálya.

Dedico

AGRADECIMENTOS

A DEUS pelo dom da vida.

Aos meus pais por nunca terem poupado esforços para que eu chegasse onde cheguei.

Às minhas irmãs e meu sobrinho por sempre serem fonte de motivação.

À minha esposa Natálya, pelo amor, compreensão, ajuda e motivação durante todo o período de doutorado.

Ao meu orientador Fabyano Fonseca pelos ensinamentos e apoio no desenvolvimento da minha tese.

Aos meus co-orientadores Professor Marcos Deon Vilela de Resende e Professor Cosme e Damião Cruz pelas sugestões e ensinamentos.

Aos conselheiros Professor Matias Kirst, Professor Patrício Muñoz e Marcio Resende Jr.

Aos membros da banca de defesa Rodrigo, Marciane e Camila pelas valiosas sugestões.

Aos companheiros do laboratório de bioinformática: Digner, Jaqueline, Vinícius, Haroldo, Daniele, Rafael, Ricardo, Renato, Dayana, Gislayne, Gabi, Isabela, Alexandre, Laís e tantos outros que tive a honra de conhecer, obrigado pela amizade de vocês.

À University of Florida e ao Forest Genomics lab por me receber durante o período de doutorado sanduíche.

Aos amigos do Forest Genomics Lab na University of Florida, Chris, Cinthia, Rodrigo Furtado, Janeo, Justyna, Annette.

Aos amigos de Gainesville Leandro, Ediene, Kelly e Felipe.

Aos amigos de república Adérico, Vinícius, Luciano e Hélcio.

À Universidade Federal de Viçosa pelo suporte para o desenvolvimento pessoal e de todos os meus estudos.

Ao Programa de Pós-Graduação em Genética e melhoramento por ter me proporcionado valiosos ensinamentos por meio de seu excelente corpo docente.

À CAPES e ao CNPq pelo suporte financeiro durante período em que fui bolsista no Brasil e no exterior.

A todos que de alguma forma contribuíram para que atingisse esse objetivo.

BIOGRAFIA

JOÃO FILIPI RODRIGUES GUIMARÃES, filho de Rivane de Fátima Rodrigues Guimarães e Luiz Claudio Dias Guimarães, nasceu em Montes Claros, Minas Gerais, no dia 04 de setembro de 1985.

Em dezembro de 2009, concluiu a graduação no curso de Agronomia da Universidade Estadual de Montes Claros (UNIMONTES), Campus Janaúba, MG.

Em agosto de 2010, ingressou no programa de Pós-Graduação em Melhoramento Genético de Plantas, em nível de Mestrado, na Universidade Federal Rural de Pernambuco, submetendo-se à defesa da dissertação em julho de 2012.

Em agosto de 2012, ingressou no programa de Pós-Graduação em Genética e Melhoramento, em nível de Doutorado, na Universidade Federal de Viçosa, submetendo-se à defesa da tese em Julho de 2016.

SUMÁRIO

RESUMO

GUIMARÃES, João Filipi Rodrigues Guimarães, D.Sc., Universidade Federal de Viçosa, julho de 2016. **Efeito da interação dominância x ambiente na habilidade de predição genômica.** Orientador: Fabyano Fonseca e Silva. Coorientadores: Cosme Damião Cruz e Marcos Deon Vilela de Resende.

A seleção genômica tornou-se ferramenta bastante útil no auxílio ao melhoramento animal e vegetal, contudo este cenário tem requerido modelos com alta capacidade de predição. Neste contexto a inclusão de efeitos não aditivos e também de interação GxA em modelos GBLUP desponta como possível fonte de melhoria de capacidade predição em estudos de seleção genômica. Em conformidade com esta demanda, o objetivo com este trabalho foi verificar se a inclusão do efeito de dominância e seu respectivo efeito de interação com o ambiente aumentaria a habilidade de predição do modelo G-BLUP. Foram comparados modelos com e sem inclusão de efeitos de dominância sob interação GxA. Para as análises foram utilizados dados reais e simulados. Os dados simulados consistiram de 923 indivíduos avaliados em três ambientes sob diferentes herdabilidades no sentido amplo, herdabilidades no sentido restrito, níveis de dominância e correlações genéticas e residuais. Os dados reais provêm de 923 indivíduos de *Pinus taeda* L. genotipados e fenotipados para característica altura em quatro localidades dos Estados Unidos da América. Os resultados demonstraram haver ligeiro incremento na habilidade de predição (em validações entre e dentro de ambientes) quando utilizado o modelo com inclusão do efeito de dominância e também o efeito de interação dominância por ambiente, contudo conclui-se que o modelo aditivo dominante teve performance estatisticamente igual ao modelo aditivo sob interação GxA.

ABSTRACT

GUIMARÃES, João Filipi Rodrigues Guimarães, D.Sc., Universidade Federal de Viçosa, july 2016. **Effect of interaction dominance x environment in genomic prediction**. Adviser: Fabyano Fonseca e Silva. Co-Advisers: Cosme Damião Cruz and Marcos Deon Vilela de Resende.

Genomic selection has become very useful tool in helping to animal and plant breeding, however this scenario has required models with high predictive ability. In this context the inclusion of non-additive effects and GxE interaction in GBLUP models is emerging as a possible source for improve the predictions of genomic selection. In line with this demand, the aim of this study was to evaluate the effect of inclusion of dominance effect and also the additive and dominance by environment interaction in genomic prediction. We compared models with and without inclusion of dominance effects in GxE interaction. For the analyzes, real and simulated data were used. The simulated data consisted in 923 individuals evaluated in three environments under different broad-sense heritability, narrow-sense heritability, dominance levels and genetic and residual correlations. The real data comes from 923 individuals of *Pinus taeda* L. genotyped and phenotyped to hight in four locations of United States. The results demonstrated a slight increase in predictive ability (in validations within and across environments) when using model with inclusion of dominance effect and the interaction effect dominance by environment, but we concluded that the additive and dominance model had a performance statistically equal to the additive model under GxE interaction.

CAPÍTULO 1 – Revisão Geral

Revisão Geral

Seleção genômica

A seleção assistida por marcadores (SAM) e a genética de associação têm como objetivo a detecção de genes de maior efeito e posteriormente a sua introgressão visando o melhoramento de características de interesse (Heffner *et al.*, 2009). No entanto, este tipo de metodologia possui limitações que inviabilizam sua utilização, tal como a necessidade de estabelecer associações marcadores-QTL para cada família em avaliação, além disso esta associação deve explicar grande parte da variação genética, já que esta técnica não é sensível a detecção de genes de menor efeito (Desta & Ortiz, 2014).

A introdução da seleção genômica (Meuwissen *et al.*, 2001) abriu o caminho para superar as limitações presentes na SAM, por meio do uso de modelos de predição com base em informações genômicas. A utilização de plataformas de genotipagem de alta densidade foi fator preponderante para tornar a seleção genômica viável, uma vez que a abrangência desse tipo de técnica aumentava a probabilidade de cada um dos *locus* da característica estarem em desequilíbrio de ligação (LD) com um número considerável de *locus* de marcadores em toda a população alvo (Shengqiang *et al.*, 2009). A seleção genômica dessa forma dispensa a necessidade de investigar associações QTL-marcadores de maneira individual que sejam significativas.

A seleção genômica usa marcadores amplamente distribuídos pelo genoma para prever o valor genético dos indivíduos (Meuwissen *et al.*, 2001). Para execução desta metodologia, uma população que foi tanto genotipada quanto fenotipada é utilizada para treinar e ajustar um modelo estatístico, esta população é denominada como população de treinamento. A partir dos efeitos de marcadores estimados na população de treinamento é então possível predizer os valores genotípicos dos indivíduos candidatos a selecção que não foram fenotipados . Este segundo conjunto de indivíduos, que foram apenas genotipados compõem a população de seleção (Habier *et al.*, 2007).

Existem diversas abordagens em relação aos métodos estatísticos utilizados na seleção genômica, contudo não existe metodologia unânime, um vez que cada uma delas pode assumir diferentes pressuposições (Desta & Ortiz, 2014). Para adoção do modelo, três fatores são cruciais na tomada de decisão, são eles: a arquitetura genética da característica e a distribuição dos efeitos dos genes que a compõem, a capacidade de regularização do processo de estimação sob influência de multicolinearidade e elevado

número de marcadores, e a seleção de covariáveis (marcadores) que afetam a característica em análise (Resende *et al.*, 2014).

Os principais métodos para a seleção genômica podem ser divididos em três grandes categorias: regressão explicita, regressão implícita e regressão via redução de dimensionalidade. O primeiro grupo pode ainda ser subdivido em dois: métodos de estimação penalizada, como RR-BLUP (*Ridge Regression–Best Linear Unbiased Prediction*) e LASSO (*Least Absolute Shrinkage and Selection Operator*) e métodos de estimação bayesiana, como BayesA, BayesB, entre outros. Na regressão implícita enquadram se os métodos RKHS (*Reproducing Kernel Hilbert Space*) e Redes Neurais. Dentro do grupo dos métodos de regressão com redução de dimensionalidade estão as metodologias de regressão componentes principais (PCR), componentes independentes (PCI) e quadrados mínimos parciais (PLS) (Desta & Ortiz, 2014).

Apesar das peculiaridades dos métodos estatísticos aplicados a seleção genômica anteriormente citados, a maioria deles lida de maneira satisfatória com o problema de estimação de um grande número de parâmetros com base em um número limitado de observações (Karkkainen & Sillanpaa, 2012; Wimmer *et al.*, 2013). Este problema foi durante muito tempo o principal fator limitante a seleção genômica, pois impedia a obtenção de estimativas via quadrados mínimos. Nos dias atuais este problema foi contornado pelo auxílio das técnicas computacionais e o aprimoramento das metodologias estatísticas (Fernando *et al.*, 2014).

O interesse sobre a seleção genômica no melhoramento genético se deve à possibilidade de se realizar predições do valor genético de indivíduos e assim aumentar os ganhos genéticos por unidade de tempo (Beaulieu *et al.*, 2014). A metodologia em questão viabilizaria a seleção antes mesmo da fenotipagem, o que representa acelerar o programa de melhoramento, uma vez que sob condições normais, longos períodos são necessários para a tomada de decisão com base apenas em informações fenotípicas (Resende *et al.*, 2012).

Modelos GBLUP

Na metodologia clássica de modelos mistos o numerador da matriz de parentesco é obtido com base no parentesco médio (Henderson, 1975) e é utilizado para descrever a variância e covariância entre indivíduos. O valor genômico predito por este tipo de análise é denominado BLUP e é bastante utilizado no melhoramento animal e vegetal.

VanRaden, (2008) com base na metodologia tradicional de modelos mistos propôs alterações nesta metodologia visando utilizá-la como ferramenta para predição genômica, para tanto o autor substituiu o numerador da matriz de parentesco, antes obtido pelo pedigree, por uma matriz de parentesco realizada e demonstrou que o parentesco estimado por meio de marcadores de alta densidade viabilizaria predições mais acuradas, esta metodologia foi denominado *Genomic BLUP* (GBLUP).

O modelo GBLUP têm sido amplamente utilizado em avaliações genômicas, principalmente devido a sua simplicidade e baixa exigência computacional (Xia *et al.*, 2015). O método GBLUP é computacionalmente mais simples por resultar em menor número de equações a serem resolvidas, além disso este tipo de análise fornece informações quanto a herdabilidade total explicada por todos os marcadores simultaneamente (Gao *et al.*, 2013).

O modelo GBLUP tradicional é dado pelo seguinte modelo (VanRaden, 2008):

$$y = Xb + Za + e$$

Em que y o vetor de fenótipos, b é o vetor de efeitos fixos e a o vetor de efeitos aditivos. X e Z são as matrizes de incidência de efeitos fixos e aleatórios, respectivamente. O componente e corresponde ao vetor de efeitos residuais. Os efeitos aditivos neste caso são assumidos como seguindo uma distribuição $a \sim N(0, G\sigma_a^2)$, em que G e a matriz de parentesco genômica aditiva.

Esta matriz descreve o parentesco entre indivíduos e pode ser construída a partir do conjunto de marcadores pelo qual os indivíduos foram genotipados. Dessa forma tomando A_{1_j} e A_{2_j} como dois alelos do j-ésimo marcador e p_j como a frequência do alelo A_{2_j}. A matriz G é criada da seguinte forma:

$$G = \frac{M_a M_a{}'}{\sum_{j=1}^{m} 2p_j(1-p_j)}$$

Em que M_a é uma matriz de dimensão nx$m$, sendo que n corresponde ao número de indivíduos e m o número de marcadores. O elemento M_a para o i-ésimo indivíduo no j-ésimo marcador é parametrizado da seguinte forma:

$$M_{a_{i,j}} = \begin{cases} -2p_j(A_1A_1) \\ 1 - 2p_j(A_1A_2) \\ 2 - 2p_j(A_2A_2) \end{cases}$$

Com uso da matriz de parentesco realizada espera-se obter estimativas mais acuradas da covariância entre indivíduos, pois possibilita acessar com maior precisão a informação de parentesco por utilizar informações em nível de genoma. Outro fator importante desta metodologia é o fato de corrigir a falta de informação acerca de parentes distantes.

A utilização desta matriz de parentesco está de acordo com as prerrogativas do modelo infinitesimal, assumindo que um grande número de genes distribuídos pelo genoma contribui igualitariamente para o fenótipo de uma característica (Hayes *et al.*, 2009).

Interação Genótipos por ambientes

O conjunto de peculiaridades que podem caracterizar um ambiente vão desde características edafoclimáticas até práticas de manejo, tais como época de plantio, níveis de adubação, quantidade de irrigação, etc. Os mínimos desvios em relação às condições ideais para o desenvolvimento da planta podem desencadear eventos fisiológicos que ameaçam seriamente o rendimento dos genótipos (Marjanovic-Jeromela *et al.*, 2011), estas diferentes respostas do genótipo a estes desvios são denominadas de interação genótipo por ambiente (GxA).

Diferenças entre ambientes em que um dado conjunto de genótipos é avaliado pode confundir e prejudicar a recomendação do melhorista, já que o mesmo genótipo sob influencia de diferentes ambientes tem seus respectivo desempenho alterados (Cooper *et al.*, 2014). A interação GxA pode ser classificada em dois tipos: a) simples, quando há alteração na diferença entre os genótipos, contudo a posição relativa dos genótipos não é alterada. Geralmente este tipo de interação não acarreta problemas sérios ao melhorista, uma vez que o *ranking* dos genótipos em avaliação não se altera de um ambiente para o outro. b) Complexa, caracterizada pela ausência de correlação entre os desempenhos dos genótipos entre os ambientes, de modo que estes apresentam diferentes respostas às variações ambientais, causando alteração no seu *ranking* (Cargnin *et al.*, 2006).

Sob constatação de interação genótipo por ambiente do tipo complexa, uma das alternativas para auxiliar o melhorista durante a recomendação dos genótipos é a obtenção de informações quanto à adaptabilidade e estabilidade (Scapim *et al.*, 2010). Dentro do âmbito do melhoramento de plantas o conceito adaptabilidade refere-se à capacidade dos genótipos em aproveitar vantajosamente o estímulo ambiental e a estabilidade refere-se a

um comportamento altamente previsível dos genótipos frente às alterações ambientais (Cruz et al., 2012)

Outra estratégia para minimizar os efeitos da interação GxA é dividir um ambiente amplo e heterogêneo em sub-ambientes homogêneos, possibilitando a recomendação eficiente de genótipos de acordo com as peculiaridades de cada subambiente (Taye & Makumbi, 2014). Este amplo estudo sobre a interação genótipo por ambiente é de grande interesse para o melhorista, já que em casos onde a estabilidade dos genótipos é constatada, o ganho com a seleção alcançado em um ambiente pode ser estendido para outros ambientes (Senger *et al.*, 2016).

Efeito de dominância

O efeito de dominância ocorre quando os efeitos dos alelos de um determinado *locus* não são somente aditivos, mas interagem entre si de modo que o valor do genótipo heterozigoto desvia-se da média dos valores dos genótipos homozigotos. Neste caso a e $-a$ são os valores genotípicos dos genótipos homozigotos A_1A_1 e A_2A_2, assim d é o valor genotípico do genótipo heterozigoto A_1A_2. Se $d = 0$, não há ação dominante no locus e os valores genotípicos no *locus* são puramente decorrentes de efeitos aditivos.

Os valores aditivos correspondem a $2q[a + d(q - p)]$ para o genótipo A_1A_1, $(q - p)[a + d(q - p)]$ para o genótipo A_1A_2 e $-2p[a + d(q - p)]$ para o genótipo A_2A_2, em que p é a frequência do alelo A_1 e q a frequência do alelo A_2 na população. Os desvios de dominância de um genótipo para um determinado *locus* corresponde à diferença entre o valor genotípico e o valor aditivo, e é igual a $-2q^2d$, $-2pqd$ e $2p^2d$ para os genótipos A_1A_1, A_1A_2 e A_2A_2, respectivamente.

Até recentemente, os estudos sobre os desvios de dominância eram escassos, já que não havia informações genômicas disponíveis. Outro fator limitante era a indisponibilidade de grandes conjuntos de dados com proporções suficientes de indivíduos com níveis de efeito dominância não nulos, tais como famílias de irmãos completos (Ertl *et al.*, 2014). Para Misztal *et al.* (1998) a explicação para existência de poucos estudos visando identificar a real contribuição do efeito de dominância na genética quantitativa está também associado ao nível de complexidade computacional utilizado neste tipo de análise e a imprecisão nas estimativas dos componentes de variância.

A variância atribuída aos desvios de dominância também é negligenciada, normalmente por não possuir aplicabilidade para prever a resposta à seleção. Embora as

estimativas sejam escassas, a variância devido aos desvios de dominância geralmente representa cerca de 10% da variação fenotípica total (Varona *et al.*, 1998).

Estudos conduzidos por (Crnokrak & Roff, 1995) demonstraram que a razão entre variância de dominância e variância aditiva (Vd/Va) em espécies de animais selvagens é de cerca de 1,17 para características comportamentais (acasalamento e territorialismo), 1,06 para características fisiológicas, demonstrando que a variância devido aos desvios de dominância nestes casos contribuiu de maneira similar a variância aditiva para a variância total destas características. Isik *et al.* (2003) avaliando as características altura, diâmetro na altura do peito, volume e incidência de ferrugem (*Cronartium quercuum* B.) em famílias de irmãos completos de *Pinus taeda* L. no sexto ano após o plantio, obtiveram valores de Vd/Va de 0,04, 0,03, 0,14 e 0 respectivamente.

Entre as estratégias disponíveis no melhoramento para estimar a influência da variância devido aos desvios de dominância, encontram-se os delineamentos genéticos, este tipo de análise é capaz de prover estimativas de parâmetros úteis na seleção de genitores para hibridação e no entendimento dos efeitos genéticos envolvidos na determinação dos caracteres (Cruz *et al.*, 2012).

Além dos delineamentos outras metodologias despontam como novas ferramentas para obtenção de resultados mais acurados acerca da contribuição dos efeitos de dominância para a variância fenotípica. A seleção genômica é uma delas e desponta como uma alternativa em complemento aos métodos clássicos no desenvolvimento de estratégias visando o melhoramento de espécies vegetais (Zhao *et al.*, 2013; Munoz *et al.*, 2014; Nishio & Satoh, 2014; de Almeida Filho *et al.*, 2016).

Neste sentido análises utilizando a seleção genômica possibilitando uma melhor compreensão da arquitetura genética das características quantitativas, facilitando o planejamento de estratégias de melhoramento que maximizem o ganho genético. Dessa forma modelos capazes de capturar diferenças decorrentes dos efeitos dominância podem ser exploradas por meio da concepção de sistemas de acasalamento que maximizam combinações alélicas favoráveis, especialmente se o programa é de espécies com propagação clonal ou com populações estruturadas em irmãos completos (Toro & Varona, 2010).

REFERÊNCIAS

de Almeida Filho JE, Guimarães JFR, e Silva FF, de Resende MD V, Muñoz P, Kirst M, Resende MFR. **2016**. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity*: 1–9.

Beaulieu J, Doerksen T, Clément S, MacKay J, Bousquet J. **2014**. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. *Heredity* **113**: 343–52.

Cargnin A, De Souza MA, Carneiro PCS, Sofiatti V. **2006**. Interação entre genótipos e ambientes e implicações em ganhos com seleção em trigo. *Pesquisa Agropecuaria Brasileira* **41**: 987–993.

Cooper M, Messina CD, Podlich D, Totir LR, Baumgarten A, Hausmann NJ, Wright D, Graham G. **2014**. Predicting the future of plant breeding: Complementing empirical evaluation with genetic prediction. *Crop and Pasture Science* **65**: 311–336.

Crnokrak P, Roff D a. **1995**. Dominance variance: associations with selection and fitness. *Heredity* **75**: 530–540.

Cruz CD, Regazzi AJ, Carneiro PCS. **2012**. *Modelos biométricos aplicados ao melhoramento genético* (CD Cruz, Ed.). Viçosa.

Desta ZA, Ortiz R. **2014**. Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science* **19**: 592–601.

Ertl J, Legarra A, Vitezica ZG, Varona L, Edel C, Emmerling R, Götz K-U. **2014**. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. *Genetics, selection, evolution : GSE* **46**: 40.

Fernando RL, Garrick DJ, Leaflet ASR. **2014**. Three Different Gibbs Samplers for BayesB Genomic Prediction Three Different Gibbs Samplers for BayesB Genomic Prediction. **660**: 1–2.

Gao H, Su G, Janss L, Zhang Y, Lund MS. **2013**. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. *Journal of dairy science* **96**: 4678–87.

Habier D, Fernando RL, Dekkers JCM. **2007**. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**: 2389–2397.

Hayes BJ, Visscher PM, Goddard ME. **2009**. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research* **91**: 47–60.

Heffner EL, Sorrells ME, Jannink J. **2009**. Genomic Selection for Crop Improvement. *Crop Science* **49**: 1–12.

Henderson CR. **1975**. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**: 423–447.

Isik F, Li B, Frampton J. **2003**. Estimates of additive, dominance and epistatic genetic variances from a clonally replicated test of loblolly pine. *Forest Science* **49**: 77–88.

Karkkainen HP, Sillanpaa MJ. **2012**. Back to basics for Bayesian model building in genomic selection. *Genetics* **191**: 969–987.

Marjanovic-Jeromela A, Nagl N, Gvozdanovic-Varga J, Hristov N, Kondic-Spika A, Vasi?? M, Marinkovi?? R. **2011**. Genotype by environment interaction for seed yield per plant in rapeseed using AMMI model. *Pesquisa Agropecuaria Brasileira* **46**: 174–181.

Meuwissen THE, Hayes BJ, Goddard ME. **2001**. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Misztal I, Varona L, Culbertson M, Bertrand JK, Mabry J, Lawlor TJ, Van Tassel CP, Gengler N. **1998**. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. *Biotechnologie, Agronomie, Société et Environnement= Biotechnology, Agronomy, Society and Environment [= BASE]* **4**: 227–233.

Munoz PR, Resende MFR, Gezan SA, Resende MD V., de los Campos G, Kirst M, Huber D, Peter GF. **2014**. Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. *Genetics* **198**: 1759–1768.

Nishio M, Satoh M. **2014**. Including dominance effects in the genomic BLUP method for genomic evaluation. *PloS one* **9**: e85792.

Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MD V, Kirst M. **2012**. Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytologist* **193**: 617–624.

Resende MDV, Silva FF e, Camila Azevedo Ferreira. **2014**. *Estatística matemática, biométrica e computacional:* (MDV Resende, Ed.). Viçosa: Suprema.

Scapim CA, Pacheco CAP, do Amaral Júnior AT, Vieira RA, Pinto RJB, Conrado TV. **2010**. Correlations between the stability and adaptability statistics of popcorn cultivars. *Euphytica* **174**: 209–218.

Senger E, Martin M, Dongmeza E, Montes JM. **2016**. Genetic variation and genotype by environment interaction in Jatropha curcas L. germplasm evaluated in different environments of Cameroon. *Biomass and Bioenergy* **91**: 10–16.

Shengqiang Z, Dekkers JCM, Fernando RL, Jannink JL. **2009**. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: A barley case study. *Genetics* **182**: 355–364.

Taye G, Makumbi D. **2014**. Studying GxE interaction under different management system and yield level using linear-bilinear models : the case of CIMMYT. *Journal of Plant Breeding and Genetics* **02**: 63–76.

Toro M a, Varona L. **2010**. A note on mate allocation for dominance handling in genomic selection. *Genetics, selection, evolution : GSE* **42**: 33.

VanRaden PM. **2008**. Efficient methods to compute genomic predictions. *Journal of dairy science* **91**: 4414–23.

Varona L, Misztal I, Bertrand JK, Lawlor TJ. **1998**. Effect of full sibs on additive breeding values under the dominance model for stature in United States Holsteins. *Journal of dairy science* **81**: 1126–1135.

Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Sch??n CC. **2013**. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics* **195**: 573–587.

Xia J, Wu Y, Fang H, Zhang W, Song Y, Zhang L, Gao X, Chen Y, Li J, Gao H. **2015**. Improving the Efficiency of Genomic Selection in Chinese Simmental beef cattle. *bioRxiv*: 022673.

Zhao Y, Zeng J, Fernando R, Reif JC. **2013**. Genomic Prediction of Hybrid Wheat Performance. *Crop Science* **53**: 802.

CAPÍTULO 2 - Efeito da interação dominância x ambiente na habilidade de predição genômica

Efeito da interação dominância x ambiente na habilidade de predição genômica

João Filipi Rodrigues Guimarães[1], Janeo Eustáquio Almeida Filho[2], Rodrigo Furtado dos Santos[3], Márcio Resende Jr.[4], Patricio Muñoz [5], Matias Kirst[6], Fabyano Fonseca e Silva[7].

[1]Futuragene Brasil Tecnologia Ltda, Itapetininga-SP, Brasil; [2]Universidade Norte Fluminense – UENF, Campos dos Goytacazes-RJ, Brasil; [3] University of Florida, Plant Molecular and Cellular Biology program, Gainesville-FL, EUA; [4]RAPID Genomics LLC,Gainesville, Gainesville-FL, EUA; [5]University of Florida, Agronomy Department, Gainesville-FL, EUA; [6]University of Florida, School of Forest Resources and Conservation, Gainesville-FL, EUA; [7]Universidade Federal de Viçosa, Departamento de Zootecnia, Viçosa-MG, Brasil.

Resumo: A seleção genômica tornou-se ferramenta bastante útil no auxílio ao melhoramento animal e vegetal, contudo este cenário tem requerido modelos com alta capacidade de predição. Neste contexto a inclusão de efeitos não aditivos e também de interação GxA em modelos GBLUP desponta como possível fonte de melhoria de capacidade predição em estudos de seleção genômica. Em conformidade com esta demanda, o objetivo com este estudo foi verificar o efeito da inclusão do efeito de dominância e também do efeitos de interação aditivos e de dominância com o ambiente na habilidade de predição genômica. Foram comparados modelos com e sem inclusão de efeitos de dominância sob interação GxA. Para as análises foram utilizados dados reais e simulados. Os dados simulados consistiram de 923 indivíduos avaliados em três ambientes sob diferentes herdabilidades no sentido amplo, herdabilidades no sentido restrito, níveis de dominância e correlações genéticas e residuais. Os dados reais provêm de 923 indivíduos de *Pinus taeda* L. genotipados e fenotipados para característica altura em quatro localidades dos Estados Unidos da América. Os resultados demonstraram haver ligeiro incremento na habilidade de predição (em validações entre e dentro de ambientes) quando utilizado o modelo com inclusão do efeito de dominância e também o efeito de interação dominância por ambiente, contudo conclui-se que o modelo aditivo dominante teve performance estatisticamente igual ao modelo aditivo sob interação GxA.

Palavras-chave: GBLUP, interação genótipo x ambiente, simulação, *Pinus taeda* L.

Effect of interaction dominance x environment in genomic prediction.

Abstract: Genomic prediction has become very useful tool in helping to animal and plant breeding, however this scenario has required models with high predictive ability. In this context the inclusion of non-additive effects and also GxE interaction in GBLUP models is emerging as a possible source for improve the predictions of genomic prediction. In line with this demand, the aim of this study was to evaluate the effect of inclusion of dominance effect and also the additive and dominance by environment interaction in genomic prediction. We compared models with and without inclusion of dominance effects in GxE interaction. For the analyzes, real and simulated data were used. The simulated data consisted in 923 individuals evaluated in three environments under different broad-sense heritability, narrow-sense heritability, dominance levels and genetic and residual correlations. The real data comes from 923 individuals of *Pinus taeda* L. genotyped and phenotyped for hight in four locations of United States. The results demonstrated a slight increase in predictive ability (in validations within and across environments) when using model with inclusion of dominance effect and also the interaction effect dominance by environment, but we concluded that the additive and dominance model had a performance statistically equal to the additive model under GxE interaction.

Keywords GBLUP, genotype by environment interaction, simulation, *Pinus taeda* L.

Introdução

A predição genômica tornou-se uma ferramenta bastante útil no auxílio ao melhoramento animal (Toro & Varona, 2010; Sun *et al.*, 2014; De Coninck *et al.*, 2014) e vegetal (Bernardo & Yu, 2007; Riedelsheimer *et al.*, 2013; Technow *et al.*, 2014; Daetwyler *et al.*, 2015) e a consequência disto tem sido a necessidade do aprimoramento dos modelos característicos para tal finalidade.

A inclusão de efeitos não aditivos em modelos de predição genômica tem se destacado como uma das possíveis alternativas para aumentar a capacidade de predição em modelos de seleção genômica. Diversos autores tem se dedicado a estudar a influência dos componentes não aditivos no modelo, seja por meio de aplicação em dados simulados ou provindos do melhoramento animal e vegetal (Toro & Varona, 2010; Sun *et al.*, 2014; Muñoz *et al.*, 2014), contudo quando há a inclusão de efeito de interação genótipo por ambiente nos modelos, apenas o efeito aditivo tem sido utilizado (Crossa, 2012; Crossa *et al.*, 2013).

Entre as diversas metodologias que podem ser utilizadas para estimar o mérito genético quando considerados os efeitos não aditivos e as interações genótipo por ambiente encontra-se o *Genomic Best Linear Unbiased Predictor* (G-BLUP), uma metodologia conhecida por sua robustez e fácil aplicabilidade, no que diz respeito à obtenção de estimativas de valores genéticos (de Los Campos *et al.*, 2013; Muñoz *et al.*, 2014). Similar ao *Pedigree Best Linear Unbiased Predictor* (P-BLUP), em que matrizes de parentesco provindas do parentesco médio são utilizadas para estimar o valor genético dos indivíduos (Henderson, 1975). Enquanto o G-BLUP utiliza matrizes de parentesco provindas das informações de marcadores com ampla cobertura do genoma (VanRaden, 2008).

A estimativa dos efeitos de dominância e seu comportamento frente aos diferentes ambientes de avaliação é uma alternativa promissora, pois a inclusão dessas informações promoveria uma melhoria na capacidade de predição dos modelos, em razão de determinar com maior segurança a contribuição genética para o fenótipo, além de minimizar o problema da herdabilidade perdida (Da *et al.*, 2014).

Em termos práticos conhecer a natureza dos efeitos de dominância em espécies vegetais perenes seria de grande utilidade, como no caso de espécies florestais, que

predominantemente são propagadas vegetativamente, neste contexto, todo o mérito genético do individuo será repassado integralmente aos clones.

A inclusão dos efeitos de interação GxA no modelos podem proporcionar melhores valores de habilidade de predição. O objetivo com este estudo foi verificar o efeito da inclusão do efeito de dominância e também do efeitos de interação aditivos e de dominância com o ambiente na habilidade de predição genômica.

Metodologia

Dados reais

Os dados fenotípicos e genotípicos reais utilizados neste estudo consistiram de uma população clonal de *Pinus taeda* L., constituída de 923 indivíduos provindos 61 famílias. Os fenótipos da população foram avaliados em quatro localidades nos Estados Unidos da América, sendo elas: Palatka, Nassau, Cuthbert e B.F. Grant Forest, os dois primeiros situados no estado da Flórida e os dois últimos no estado da Geórgia. Foram utilizadas as mensurações da característica altura no sexto ano após o plantio, avaliada em delineamento em blocos incompletos balanceados com parcelas compostas por uma árvore, tendo cada tratamento oito repetições.

Os dados genotípicos reais são compostos de 4.722 SNPs (*Single Nucleotide Polymorphism*), que foram previamente submetidos ao controle de qualidade de marcas via *Call Rate* > 90%, Análise de Equilíbrio de Hardy-Weinberg e MAF (*Minor Allele Frequency*) > 10%.

Todos os dados utilizados nesse estudo estão disponíveis na forma de material suplementar em (Resende *et al.*, 2012) ou em http://loblolly.ucdavis.edu/bipod/ftp/Genotype_Population_CCLONES.txt.

Dados simulados

Genoma

A simulação do genoma teve abordagem na teoria de coalescência, para isso foi utilizado o software Markovian Coalescent Simulator-MaCS (Chen *et al.*, 2009) e o tamanho efetivo (N_e) da população natural foi de 10000 indivíduos, além disso, foram tomadas como base informações de 1.000 gerações prévias à população base (G0), mantendo o mesmo tamanho efetivo durante todas as gerações prévias. A partir desta população foram simulados 2.000 haplótipos com 12 cromossomos, sendo que cada

cromossomo possui comprimento de 100 cM e foi composto por 2 gigas pares de base (gpb).

O parâmetro de mutação utilizado como entrada no software de simulação do genoma foi definido por meio da seguinte expressão:

$$\theta = 4.\mathrm{N_e}.\mu \qquad (1)$$

Em que μ é a taxa de mutação por geração, aqui assumida como $2{,}5\text{x}10^{-8}$ (Willyard *et al.*, 2006). O parâmetro de recombinação utilizado como entrada foi definido pela seguinte expressão:

$$\rho = 4.\mathrm{N_e}.\mathrm{r} \qquad (2)$$

Em que r é a taxa de recombinação por geração, e que será assumida como $5\text{x}10^{-10}$, esse valor representa uma probabilidade de recombinação de 1% por cM (De Coninck *et al.*, 2014).

Após a obtenção dos haplótipos foram gerados 1.000 indivíduos compondo a população base (G0). Duas gerações de recombinação por meio de estruturação em famílias de meios irmãos e posterior seleção foram conduzidas (G1 e G2) após a obtenção da população base, dando garantia da existência de desequilíbrio de ligação na população simulada. Ambas as gerações eram compostas por 1.000 indivíduos, em que a população G1 foi obtida por cruzamento aberto entre os 100 melhores indivíduos selecionados da população G0, e a geração G2 obtida por meio da seleção e recombinação dos 42 melhores indivíduos provindos dos cruzamentos em G1.

Simulação dos fenótipos

Para a simulação dos fenótipos considerou-se que Q_{1j} e Q_{2j} são dois alelos do j-ésimo QTL (*Quantitative Trait Loci*), possibilitando a ocorrência dos genótipos $Q_{1j}Q_{1j}$, $Q_{1j}Q_{2j}$e $Q_{2j}Q_{2j}$.

Para a simulação dos fenótipos, nas gerações G0, G1 e G2 foram tomados 10.999 *loci* do genoma simulado, dos quais 10.000 foram utilizados como marcadores e outros 999 como QTLs.

Os fenótipos das gerações G0 e G1 foram simulados considerando apenas um ambiente com apenas a ocorrência de efeitos aditivos, dessa forma os efeitos de cada QTL para essas gerações foram amostrados de uma distribuição normal com as seguintes especificações:

$$\alpha_j \sim N(0, \sigma^2_{\alpha_j}).$$

Em que α_j são os efeitos aditivos dos QTLs. Os efeitos residuais para as gerações G0 e G1 assim, como os efeitos aditivos foram amostrados de uma distribuição normal com as seguintes especificações:

$$e_i \sim N(0, \sigma^2_e).$$

Em que e_i é o valor residual nos genótipos e σ^2_eé dado por:

$$\sigma^2_e = \frac{(1-h^2)\sigma^2_a}{h^2} \quad (3)$$

Levando-se em consideração herdabilidade no sentido restrito (h^2) e a variância genética aditiva (σ^2_a) na geração. Apenas na geração G2 os genótipos selecionados foram simulados em mais de um ambiente, já que em programas de melhoramento vegetal apenas nas fases de recomendação os genótipos são avaliados em mais de um local.

Os efeitos aditivos (α_j) dos QTL's para G2 foram simulados para três ambientes (A, B e C) e em cada ambiente foram amostrados valores utilizando uma distribuição normal multivariada de acordo com a seguinte descrição:

$$\alpha_j \sim MVN\left(0, \begin{bmatrix} \sigma^2_{\alpha_A} & r_{AB}\sigma_{a_A}\sigma_{\alpha_B} & r_{AC}\sigma_{a_A}\sigma_{\alpha_C} \\ r_{BA}\sigma_{a_B}\sigma_{\alpha_A} & \sigma^2_{a_B} & r_{BC}\sigma_{a_B}\sigma_{\alpha_C} \\ r_{CA}\sigma_{a_C}\sigma_{\alpha_A} & r_{CB}\sigma_{a_C}\sigma_{\alpha_B} & \sigma^2_{a_C} \end{bmatrix}\right).$$

Já os valores atribuídos aos efeitos de dominância (d_j) foram determinados como o produto entre o α_j e o grau médio de dominância (φ) (Nishio & Satoh, 2014), que por sua vez foi amostrado de uma distribuição normal multivariada como descrito a seguir:

$$\varphi_j \sim MVN\left(0, \begin{bmatrix} \sigma^2_{\varphi_A} & r_{AB}\sigma_{\varphi_A}\sigma_{\varphi_B} & r_{AC}\sigma_{\varphi_A}\sigma_{\varphi_C} \\ r_{BA}\sigma_{\varphi_B}\sigma_{\varphi_A} & \sigma^2_{\varphi_B} & r_{BC}\sigma_{\varphi_B}\sigma_{\varphi_C} \\ r_{CA}\sigma_{\varphi_C}\sigma_{\varphi_A} & r_{CB}\sigma_{\varphi_C}\sigma_{\varphi_B} & \sigma^2_{\varphi_C} \end{bmatrix}\right).$$

O valor aditivo dos indivíduos (u) e os desvios de dominância no caso do modelo aditivo dominante (δ), foram obtidos utilizando as seguintes expressões de acordo com a teoria proposta (Falconer & Mackay, 1996).

$$u_i = \sum_i \left[I(x_{ij} = 1)2q_j + I(x_{ij} = 0)(q_j - p_j) - I(x_{ij} = -1)2p_j\right]\hat{\alpha}_j \quad (4)$$

$$\delta_i = \sum_i \left[-I(x_{ij} = 1)2q_j^2 + I(x_{ij} = 0)2p_jq_j - I(x_{ij} = -1)2p_j^2\right]d_j \quad (5)$$

Em que p_j é a frequência alélica de Q_{1j}e $q_j = 1 - p_j$, $\hat{\alpha}_j = a_j + d_j(q_j - p_j)$, é o efeito médio de substituição alélica e I é uma função indicadora dos marcadores.

Para obtenção dos valores fenotípicos, serão adicionados os efeitos ambientais aos valores de efeito genotípico total, tais efeitos residuais foram amostrados da seguinte distribuição:

$$e_i \sim MVN\left(0, \begin{bmatrix} \frac{(1-h_A^2)\sigma_A^2}{h_A^2} & r_{AB}\sqrt{\frac{(1-h_A^2)\sigma_A^2}{h_A^2}}\sqrt{\frac{(1-h_B^2)\sigma_B^2}{h_B^2}} & r_{AC}\sqrt{\frac{(1-h_A^2)\sigma_A^2}{h_A^2}}\sqrt{\frac{(1-h_C^2)\sigma_C^2}{h_C^2}} \\ r_{BA}\sqrt{\frac{(1-h_B^2)\sigma_B^2}{h_B^2}}\sqrt{\frac{(1-h_A^2)\sigma_A^2}{h_A^2}} & \frac{(1-h_B^2)\sigma_B^2}{h_B^2} & r_{BC}\sqrt{\frac{(1-h_B^2)\sigma_B^2}{h_B^2}}\sqrt{\frac{(1-h_C^2)\sigma_C^2}{h_C^2}} \\ r_{CA}\sqrt{\frac{(1-h_C^2)\sigma_C^2}{h_C^2}}\sqrt{\frac{(1-h_A^2)\sigma_A^2}{h_A^2}} & r_{CB}\sqrt{\frac{(1-h_C^2)\sigma_C^2}{h_C^2}}\rho_{23}\sqrt{\frac{(1-h_B^2)\sigma_B^2}{h_B^2}} & \frac{(1-h_C^2)\sigma_C^2}{h_C^2} \end{bmatrix}\right)$$

.

As herdabilidades no sentido amplo foram fixados para cada ambiente buscando-se obter um cenário similar àqueles avaliados por Resende *et al.*(2012). Dessa forma foram atribuídos aos ambientes A, B e C valores de herdabilidade no sentido amplo de ≈0,10, ≈0,20 e ≈0,30, herdabilidades no sentido restrito de ≈0,10, ≈0,15 e ≈0,20 e níveis de dominância de ≈0, ≈0,05 e ≈0,10, respectivamente.

Para os efeitos aditivos, o grau médio de dominância e o efeito residual foram considerados correlações para a amostragem pela distribuição normal multivariada nos valores de 0,8, 0,2 e 0,2 entre os ambientes A e B, A e C, e B e C respectivamente.

Visando aumentar o grau de confiabilidade das estimativas, foram simulados dez replicatas do conjunto de dados simulados, para isso foi utilizado o software R (RCoreTeam, 2013), onde foram empregados os pacotes Basic e MASS (Venables & Ripley, 2002).

Modelos GBLUP

Para a predição dos valores genômicos foram utilizados os seguintes modelos em suas formas matriciais, para tanto foi utilizado o software ASReml (Gilmour *et al.*, 2009).

$$y = Xb + Z_1 i_{bl} + Z_2 a + Z_4 i_a + e \quad (6)$$

$$y = Xb + Z_1 i_{bl} + Z_2 a + Z_3 d + Z_4 i_a + e \quad (7)$$

$$y = Xb + Z_1 i_{bl} + Z_2 a + Z_3 d + Z_4 i_a + Z_5 i_d + e \quad (8)$$

Em que y é um vetor de fenótipos, b e o vetor de efeitos fixos, X corresponde a matrix de incidência dos efeitos fixos. Os vetores i_{bl}, a, d, i_a, i_d, e e se referem aos efeitos aleatórios de bloco incompleto dentro de repetição, aditivo, dominante, interação aditivo-

ambiente, interação dominante-ambiente e resíduo tendo respectivamente como matrizes de incidência Z_1, Z_2, Z_3, Z_4, Z_5.

Para os efeitos aleatórios foram adotadas as seguintes distribuições: $a \sim N(0, A\sigma^2{}_a)$; $d \sim N(0, D\sigma^2{}_d)$; $i_{bl} \sim N(0, I\sigma^2{}_{i_{bl}})$; $i_a \sim N(0, A \otimes I\sigma^2{}_{i_a})$; $i_d \sim N(0, D \otimes I\sigma^2{}_{i_d})$; $e \sim N(0, I\sigma^2)$, Em que A e D representam as matrizes de parentesco aditivo e de dominância entre os indivíduos e que foram obtidas das informações oriunda dos SNPs, I é uma matriz identidade e $\otimes$ representam o produto de Kronecker.

A matriz A utilizada corresponde a matriz de parentesco genômica aditiva proposta por VanRaden (2008) e foi obtida de acordo com a seguinte expressão:

$$A = \frac{W_a W_{a'}}{\sum 2p_j q_j} \qquad (9)$$

Em que w_a é uma matriz de dimensão n x m composta por n indivíduos e m marcadores, assumindo que A_1 e A_2 são dois alelos do locus j. Foram adotadas as codificações 0, 1 e 2 para as configurações alélicas A_2A_2, A_2A_1 e A_1A_1. Considerando que p_j e q_j são as respectivas frequências do lócus i, realizou-se a padronização da matriz w_a por meio da seguinte reparametrização:

$$W_a \begin{cases} (2-2p) \\ (1-2p) \\ (-2p) \end{cases} \text{para os genótipos} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

Para a construção da matriz de parentesco genômica de dominância (D) foi adotada a metodologia proposta por Vitezica et al. (2013), a qual é descrita pela seguinte expressão:

$$D = \frac{W_d W_{d'}}{\sum (2p_j q_j)^2} \qquad (10)$$

w_d tal como w_a é uma matriz de dimensão n x m composta por n indivíduos e m marcadores, entretanto neste caso foram assumidas as codificações 0,1 e 0 para A_2A_2, A_2A_1 e A_1A_1, em que a padronização da matriz w_d foi realizada de acordo com a seguinte parametrização:

$$w_d \begin{cases} (-2q^2) \\ (2pq) \\ (-2p^2) \end{cases} \text{para os genótipos} \begin{cases} A_1A_1 \\ A_1A_2 \\ A_2A_2 \end{cases}$$

Validação cruzada dos resultados

Para a validação cruzada foi utilizado esquema de validação em *ten-fold*, neste esquema os genótipos foram divididos aleatoriamente em 10 subamostras (*fold*), nove dessas subamostra foram utilizadas como população de treinamento. A décima subamostra por sua vez foi utilizada como população de validação (Tabela 1). O final da validação se deu após constatação de que todas as subamostras foram utilizadas como população de validação.

Os estudos de validação consistiram no calculo da correlação média entre os ciclos/fold entre valores genotípicos estimados e simulados e também a correlação média entre valores genotípicos e fenotípicos (habilidade de predição), para tanto foram utilizadas as seguintes expressões, respectivamente.

$$r_{(\hat{g}g)} = \frac{COV(\hat{g}g)}{\sqrt{\sigma^2_{\hat{g}}\sigma^2_g}} \qquad (11)$$

Em que $r_{(\hat{g}g)}$ é a correlação ente o valor genotípico predito ($\hat{g}$) e o valor genotípico simulados (g), $COV(\hat{g}g)$ é a covariância entre valores genotípicos preditos e simulados, $\sigma^2_{\hat{g}}$ é a variância genotípica calculada com base nos valores genotípicos estimados e σ^2_g é a variância genotípica calculada com base nos valores genotípicos simulados.

$$r_{(\tilde{g}y)} = \frac{COV(\tilde{g}y)}{\sqrt{\sigma^2_{\tilde{g}}\sigma^2_y}} \qquad (12)$$

Em que $r_{\hat{g}y}$ é a correlação ente o valor genotípico ($\hat{g}$) e o valor fenotípico (y) dos genótipos. $COV(\hat{g}y)$ é a covariância entre valores genotípicos preditos e os valores fenotípicos, $\sigma^2_{\hat{g}}$ é a variância genotípica calculada com base nos valores genotípicos preditos e σ^2_y é a variância fenotípica calculada com base nos valores fenotípicos.

Na validação entre ambientes foram utilizadas as mesmas expressões demonstradas em (11) e (12), a correlação foi obtida a partir de valores genotípicos (no conjunto de dados simulados) e fenotípicos dos demais ambientes avaliados (y').

Qualidade do ajuste dos modelos

Visando à avaliação da qualidade do ajuste dos modelos utilizados neste trabalho, os mesmos foram avaliados e comparados quanto ao critério de informação de Akaike (AIC) (Akaike 1974), dado pela seguinte expressão:

$$AIC = -2l(\theta) + 2p \qquad (13)$$

Em que $l(\theta)$ é a função de log-verossimilhança obtida em cada *fold* e p é o número de parâmetros do modelo avaliado.

Resultados

Comparação entre modelos

De maneira geral todos modelos mostraram desempenho similar, contudo os modelos #2 e #3 tiveram desempenho levemente destacado em relação ao modelo#1 sob cenários simulados com efeitos de dominância e também no conjunto de dados reais (figuras 2, 3 e 4).

A análise de ajuste do modelo com base no AIC demonstrou que todos os modelos avaliados possuem capacidade de ajuste similar (Figura 1), contudo os valores médios obtidos para o modelo#2 e modelo#3 foram inferiores ao modelo#1.

Dados simulados

Correlação entre valores genotípicos preditos e simulados

O ambiente A se destacou por apresentar o melhor resultado tanto para $r_{(\hat{g}g)}$ (0,472) quanto para $r_{(\hat{g}g')}$ (0,401), neste último caso se deu quando a validação foi realizada em B, estes resultados foram obtidos com os modelos modelo#3 e #2, respectivamente (Figura 1).

Avaliando-se o ambiente C é possível constatar que este é o ambiente que fornece as piores estimativas de valores genotípicos, no entanto seu melhor desempenho para $r_{(\hat{g}g)}$ se dá com o modelo#1 (0,386). Quando a abordagem é direcionada para $r_{(\hat{g}g')}$ o melhor resultado de C se dá quando a validação é feita em B, chegando neste caso ao valor de 0,258 e tal como em $r_{(\hat{g}g)}$ a estimativa foi obtida com o modelo#1 (Figura 1). Ainda tratando de $r_{(\hat{g}g')}$, os valores genotípicos preditos do ambiente C utilizando o modelo#1 propiciaram performance ligeiramente superiores aos demais modelos avaliados.

Correlação entre valor genotípico predito e o fenotípico

De maneira similar à validação $r_{(\hat{g}g)}$, os resultados para $r_{(\hat{g}y)}$ foram superiores a $r_{(\hat{g}y')}$, demonstrando que a validação dentro do ambiente para qual foram preditos os valores genotípicos proporcionaram melhores resultados do que validação entre ambientes.

Os valores genotípicos preditos para o ambiente A proporcionaram de maneira geral valores mais elevados $r_{(\hat{g}y)}$ entre os ambientes simulados, onde o melhor resultado foi 0,252 utilizando o modelo#3, contudo quando a abordagem é referente a $r_{(\hat{g}y')}$ o melhor resultado se deu na validação em B (0,171) por meio do modelo#1 (Figura 2).

Apesar dos valores genotípicos preditos no ambiente A demonstrarem bons resultados quando a validação foi realizada dentro do mesmo ambiente, o melhor resultado obtido para $r_{(\hat{g}y')}$ se deu com o uso de valores genotípicos preditos para o ambiente B com o modelo#2 e que foram validados em A (0,201) (Figura 2).

Correlações provenientes dos valores genotípicos preditos no ambiente C demonstraram desempenho inferior em relação àqueles obtidos dos demais ambientes simulados, seu melhor resultado para $r_{(\hat{g}y)}$ foi 0,131, já para $r_{(\hat{g}y')}$ o melhor resultado foi de 0,141, nesta situação a validação foi realizada com valores fenotípicos obtidos em A, ambos os resultados utilizando o modelo#1(Figura 2). Em tendência similar à $r_{(\hat{g}g')}$, os resultados de $r_{(\hat{g}y')}$ utilizando valores genotípicos preditos para C por meio do modelo#1 desempenham melhor performance do que o modelo#2 e o modelo#3.

Dados reais

Validação no valor fenotípico

Considerando a validação dentro dos ambientes, os modelos #2 e #3 atingiram de forma geral os melhores desempenhos para habilidade de predição (Figura 4). Palatka foi o ambiente que obteve o melhor resultado para $r_{(\hat{g}y)}$ 0,459, neste caso o modelo utilizado foi o modelo#2. Os resultados obtidos para Cuthbert em contrapartida indicaram que este é o ambiente com menor desempenho para $r_{(\hat{g}y)}$, sua melhor performance foi atingida utilizando o modelo#3, obtendo-se o valor de 0,266 (Figura 3).

Quando a validação é direcionada para $r_{(\hat{g}y')}$ os valores genotípicos preditos para Nassau foram de maneira geral os que proporcionaram o melhor desempenho, com destaque para a validação em fenotípicos mensurados em Palatka, onde o modelo#3 propiciou o valor de 0,363, sendo este o melhor resultado para validação entre ambientes (Figura 4).

Valores genotípicos preditos para genótipos alocados em Cuthbert proporcionaram resultados de $r_{(\hat{g}y')}$ que variaram entre 0.209 a 0.281, estes valores correspondem a

habilidade de predição destes valores genotípicos quando a validação foi realizada em Palatka utilizando o modelo#2 e em Nassau utilizando o modelo#3.

Nota-se que apesar dos valores genotípicos preditos em Cuthbert não proporcionarem os piores resultados para $r_{(\hat{g}y\prime)}$, a validação em fenótipos avaliados neste ambiente resultam em baixos valores de $r_{(\hat{g}y\prime)}$, se considerarmos apenas os resultados $r_{(\hat{g}y\prime)}$ em que a validação ocorre neste ambiente, os valores variam de 0,107 até 0,175, mantendo a superioridade do modelo#3 em comparação ao modelo#1(Figura 4).

Discussão

Segundo El-Dien *et al.* (2015), para que a implementação da seleção genômica no melhoramento de espécies florestais seja bem sucedida, é crucial a adoção de modelos capazes de fornecer informações precisas sobre a interação genótipo por ambiente. Uma das metodologias utilizadas comparar o ajuste dos modelos de seleção genômica é critério de informação de Akaike, em que quanto menor o valor de AIC melhor o ajuste do modelo (Schulz-Streeck & Piepho, 2010), com base nessa afirmação verifica-se que o modelo#1 obteve melhor ajuste do que os modelos #2 e #3 influenciados pelo maior número de parâmetros.

Os modelos avaliados utilizando dados reais e simulados não apresentaram diferenças acentuadas e a explicação para isto segundo Muñoz *et al.* (2014), seria decorrente do componente aditivo do modelo ser capaz de capturar parte do efeito de dominância, dessa forma o modelo apenas com componentes aditivo não seria penalizado pela omissão do efeito de dominância. Estendendo este conceito para os efeitos de interação, pode-se deduzir que parte dos efeitos de interação dominância por ambiente pode ser capturado pelo efeito de interação aditivo por ambiente, prova disto seria o desempenho similar entre os modelos #2 e #3, já que no primeiro houve apenas a inclusão dos efeitos de interação aditiva por ambiente.

Verifica-se que o modelo#3 apresenta resultados superiores aos demais modelos sob em ambientes onde o nível de dominância é elevado. Resultados como este foram reportados por Nishio & Satoh (2014), de Almeida Filho *et al.*, (2016) quando avaliadas características poligênicas simuladas por meio de modelo aditivo dominante. De acordo com estes resultados, pode-se deduzir que o desempenho de modelos que incluem efeitos de dominância e seu respectivo efeito de interação com o ambiente teriam seu

desempenho otimizado sob cenários em características com alto grau de contribuição do efeito de dominância para a variância genética.

Validação cruzada entre ambientes simulados

Na validação entre ambientes simulados o que se viu foi que os ambientes A e B simulados com alta correlação genética e residual entre si, apresentaram potencial elevado para predição indireta com base nos resultados de $r_{(\hat{g}g)\prime}$ e $r_{(\hat{g}y\prime)}$, em contrapartida o ambiente C, que foi simulado com baixas correlações genéticas e residuais em relação aos demais ambientes apresentou baixo potencial de uso de valores genotípicos preditos para a validação e consequente seleção indireta entre A e B.

Validações realizadas no ambiente C demonstraram a superioridade do modelo aditivo em todos cenários avaliados ($r_{(\hat{g}g)}$, $r_{(\hat{g}g\prime)}$, $r_{(\hat{g}y)}$, $r_{(\hat{g}y\prime)}$), este resultado corrobora com a hipótese de que o modelo incluindo apenas o efeito aditivo e a interação aditivo por ambiente demonstraria melhores resultados, já que os genótipos do ambiente C foram simulados com nível quase nulo para o nível de dominância.

Validação entre ambientes oriundos de dados reais

A validação entre ambientes utilizando o conjunto de dados reais demonstrou que os ambientes Palatka e Nassau possuem potencial para utilização de seus respectivos valores genotípicos preditos via GBLUP estáveis, possibilitando validações e seleção indireta entre si (Figura 3). Validações entre Palatka e Nassau provavelmente atendem o mesmo pressuposto verificado para os ambientes simulados A e B, ou seja a alta correlação genética e residual entre os ambientes. Resende *et al.*(2012) utilizando este mesmo conjunto de dados, concluíram que a proximidade geográfica entre estes ambientes foi crucial para justificar a utilização da validação cruzada entre eles.

Para Spindel *et al.* (2016) o sucesso da aplicação da seleção genômica em programas de melhoramento tal como em metodologias clássicas é dependente da estratificação de ambientes para que se alcance bons resultados na habilidade de predição. Segundo Heffner *et al.* (2011) ampliar a avaliação de genótipos em vários ambientes possibilitaria ainda minimizar riscos de associados a eventos climáticos atípicos que podem comprometer a qualidade da fenotipagem.

A adoção de modelos com a inclusão de efeitos de dominância por si só possibilitaria um ganho extra na habilidade de predição, além disso este tipo de modelo

poderia ser utilizado para fornecer informações uteis na alocação de cruzamentos visando maximizar ganhos com efeitos de dominância (Toro & Varona, 2010; Nishio & Satoh, 2014). Além disso com a inclusão do efeito de interação dominância por ambiente, seria possível realizar a predição de genótipos não avaliados em um determinado ambiente, bem como obter informações sobre alocação de cruzamentos específicos para cada ambiente.

Influência da herdabilidade na validação cruzada

A partir dos resultados observou-se que os ambientes em que a característica simulada e a característica altura (ambiente C e Cuthbert) apresentaram os menores valores de herdabilidade no sentido amplo houve uma baixa habilidade de predição de todos os modelos avaliados. Combs & Bernardo (2013) utilizando modelo aditivo em ambiente único encontraram esta mesma relação entre habilidade de predição e a herdabilidade, contudo os autores destacam que esta relação não é totalmente conhecida e que exceções podem acontecer.

Conclusão

Conclui-se que apesar de haver ligeiro incremento na habilidade de predição (entre e dentro de ambientes) quando utilizado o modelo com inclusão do efeito de dominância e também o efeito de interação dominância por ambiente, sua performance não é estatisticamente diferente do modelo que inclui apenas os componentes aditivos.

Agradecimentos:

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao United States Department of Agriculture (USDA).

REFERÊNCIAS

Akaike H. **1974**. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19**: 716–723.

de Almeida Filho JE, Guimarães JFR, e Silva FF, de Resende MD V, Muñoz P, Kirst M, Resende MFR. **2016**. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity*: 1–9.

Bernardo R, Yu J. **2007**. Prospects for genomewide selection for quantitative traits in maize. *Crop Science* **47**: 1082–1090.

Chen G, Marjoram P, Wall J. **2009**. Fast and flexible simulation of DNA sequence data. *Genome research*: 136–142.

Combs E, Bernardo R. **2013**. Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. *The Plant Genome* **6**: 1–7.

De Coninck A, Fostier J, Maenhout S, De Baets B. **2014**. DAIRRy-BLUP: a high-performance computing approach to genomic prediction. *Genetics* **197**: 813–822.

Crossa J. **2012**. From genotype× environment interaction to gene× environment interaction. *Current genomics*: 225–244.

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, *et al.* 2013. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 (Bethesda, Md.)* **3**: 1903–26.

Da Y, Wang C, Wang S, Hu G. **2014**. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PloS one* **9**: e87666.

Daetwyler HD, Hayden MJ, Spangenberg GC, Hayes BJ. **2015**. Selection on Optimal Haploid Value Increases Genetic Gain and Preserves More Genetic Diversity Relative to Genomic Selection. *Genetics* **200**: 1341–1348.

El-Dien OG, Ratcliffe B, Klapste J, Chen C, Porth I, El-Kassaby YA, Gamal El-Dien O, Ratcliffe B, Klapste J, Chen C, *et al.* 2015. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. *Bmc*

Genomics **16**: 370.

Falconer D, Mackay TFC. **1996**. *Introduction to Quantitative Genetics*. Harlow, UK: Longman.

Gilmour A, Gogel B, Cullis B, Thompson R. **2009**. ASReml user guide release 3.0. *VSN International Ltd, …*.

Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME. **2011**. Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. *Crop Science* **51**: 2597.

Henderson CR. **1975**. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**: 423–447.

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. **2013**. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**: 327–45.

Muñoz PR, Resende MFR, Gezan S a, Resende MDV, de Los Campos G, Kirst M, Huber D, Peter GF. **2014**. Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics* **198**: 1759–68.

Nishio M, Satoh M. **2014**. Including dominance effects in the genomic BLUP method for genomic evaluation. *PloS one* **9**: e85792.

RCoreTeam. **2013**. R: A Language and Environment for Statistical Computing.

Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MD V, Kirst M. **2012**. Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. *New Phytologist* **193**: 617–624.

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE. **2013**. Genomic predictability of interconnected biparental maize populations. *Genetics* **194**: 493–503.

Schulz-Streeck T, Piepho H-P. **2010**. Genome-wide selection by mixed model ridge regression and extensions based on geostatistical models. *BMC Proceedings* **4**: S8.

Spindel JE, Begum H, Akdemir D, Collard B, Redoña E, Jannink J, Mccouch S.

2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**: 395–408.

Sun C, VanRaden PM, Cole JB, O'Connell JR. **2014**. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PloS one* **9**: e103934.

Technow F, Schrag T a., Schipprack W, Bauer E, Simianer H, Melchinger AE. **2014**. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* **197**: 1343–1355.

Toro M a, Varona L. **2010**. A note on mate allocation for dominance handling in genomic selection. *Genetics, selection, evolution : GSE* **42**: 33.

VanRaden PM. **2008**. Efficient methods to compute genomic predictions. *Journal of dairy science* **91**: 4414–23.

Venables WN, Ripley BD. **2002**. *Modern Applied Statistics with S*. New York: Springer.

Vitezica ZG, Varona L, Legarra A. **2013**. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics* **195**: 1223–30.

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R. **2006**. Fossil Calibration of Molecular Divergence Infers a Moderate Mutation Rate and Recent Radiations for Pinus. *Molecular Biology and Evolution* **24**: 90–101.

Tabela 1 : Esquema hipotético de obtenção de valores genotípicos por fold, utilizando modelo GBLUP.

Ambiente	**Fold**									
	Fold1	Fold2	Fold 3	Fold 4	Fold 5	Fold6	Fold7	Fold8	Fold 9	Fold10
Amb1	y_{A1}	y_{A2}	y_{A3}	y_{A4}	y_{A5}	y_{A6}	y_{A7}	y_{A8}	y_{A9}	NA
Amb2	y_{B1}	y_{B2}	y_{B3}	y_{B4}	y_{B5}	y_{B6}	y_{B7}	y_{B8}	y_{B9}	NA
Amb3	y_{C1}	y_{C2}	y_{C3}	y_{C4}	y_{C5}	y_{C6}	y_{C7}	y_{C8}	y_{C9}	NA

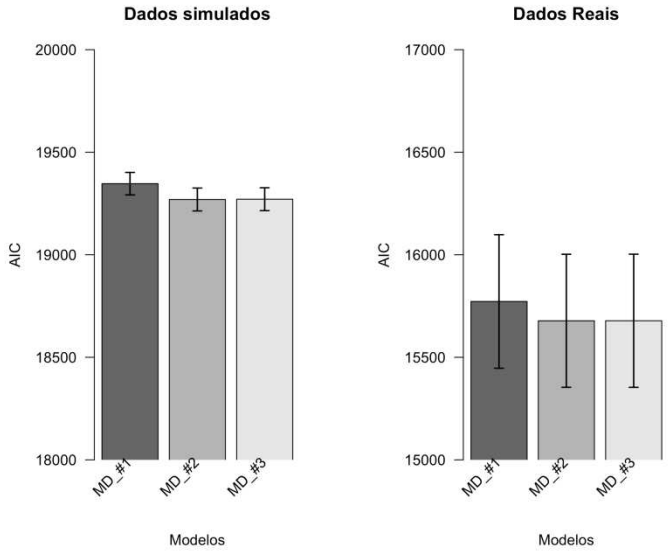

Figura 1: Qualidade de ajuste dos modelos GBLUP aplicados a análise de dados reais e simulados utilizando o Critério de informação de Akaike (AIC).

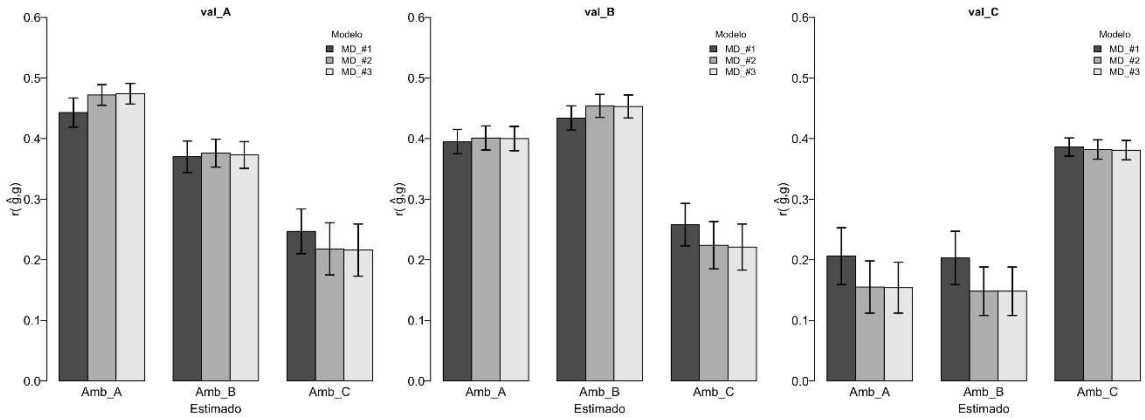

Figura 2: Correlação entre valores genotípicos preditos e simulados, validados entre e dentro de três ambientes com diferentes herdabilidades e níveis de dominância.

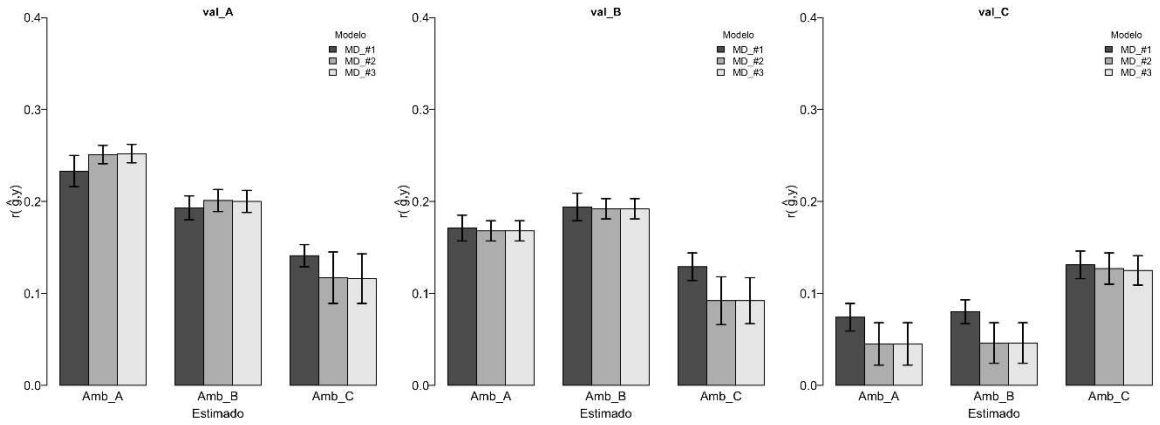

Figura 3: Correlação entre valores genotípicos preditos e fenotípicos simulados, validados entre e dentro de três ambientes com diferentes herdabilidades e níveis de dominância.

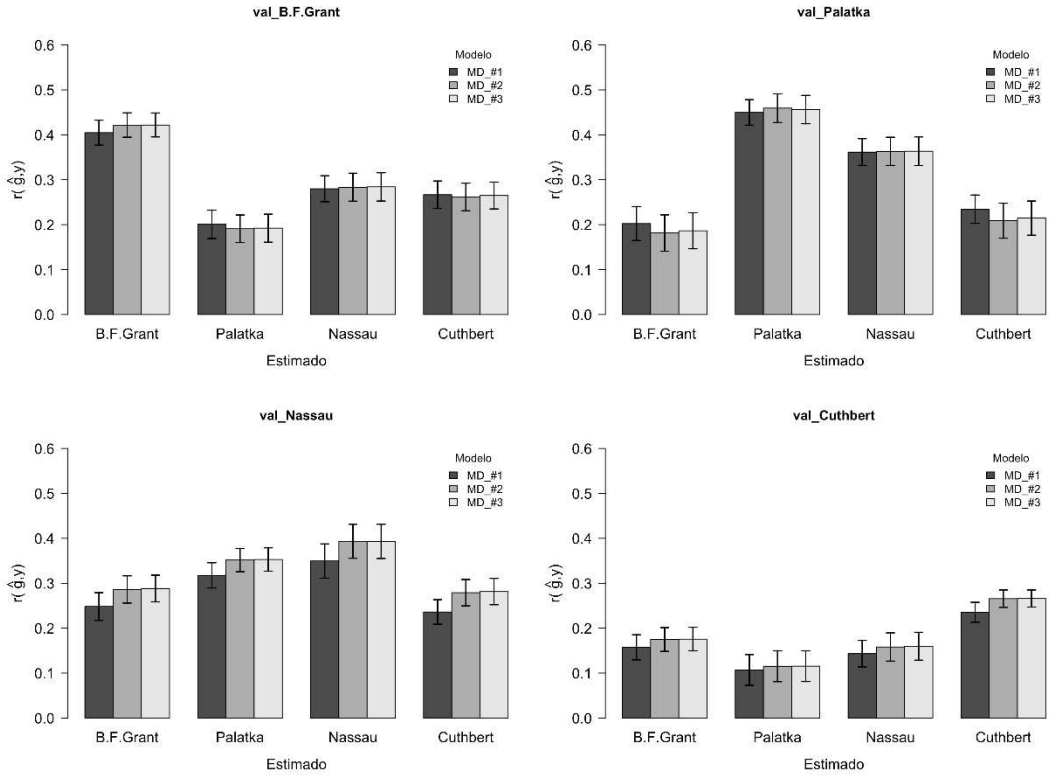

Figura 4: Correlação entre valores genotípicos preditos e valores fenotípicos, entre e dentro de quatro localidades dos Estados Unidos da América (Palatka, Nassau, Cuthbert e B.F. Grant Forest).