FRANCYSE EDITE DE OLIVEIRA CHAGAS DE MORAES

PREDIÇÃO GENÔMICA SOB DIFERENTES CENÁRIOS QUE INCLUEM, OU NÃO, LOCOS CONTROLADORES DE CARACTERÍSTICAS

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

Orientador: Cosme Damião Cruz

VIÇOSA - MINAS GERAIS
2022

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

T
M827p
2022

Moraes, Francyse Edite de Oliveira Chagas de, 1993-
Predição genômica sob diferentes cenários que incluem, ou não, locos controladores de características / Francyse Edite de Oliveira Chagas de Moraes. – Viçosa, MG, 2022.
1 tese eletrônica (82 f.): il. (algumas color.).

Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Biologia Geral, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.578
Modo de acesso: World Wide Web.

1. Mapeamento cromossômico - Métodos estatísticos. 2. Marcadores genéticos. I. Cruz, Cosme Damião, 1958-. II. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-Graduação em Genética e Melhoramento. III. Título.

CDD 22. ed. 572.8633

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

FRANCYSE EDITE DE OLIVEIRA CHAGAS DE MORAES

PREDIÇÃO GENÔMICA SOB DIFERENTES CENÁRIOS QUE INCLUEM, OU NÃO, LOCOS CONTROLADORES DE CARACTERÍSTICAS

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

APROVADA: 25 de julho de 2022.

Assentimento:

Francyse Edite de Oliveira Chagas de Moraes
Autora

Cosme Damião Cruz
Orientador

A Deus, aos meus pais, ao meu marido, a minha irmã e ao meu cunhado.

AGRADECIMENTOS

Agradeço a Deus por iluminar o meu caminho, sendo o Senhor da minha vida em todas as situações.

Aos meus pais, Adalberto Custódio e Maria Elesbão, meu marido, Nilander, minha irmã, Francielle e meu cunhado Sebastião Júnior, pelo amor, carinho, dedicação e incentivo para que eu conseguisse alcançar mais este objetivo.

Aos líderes espirituais que passaram pela minha vida durante a trajetória de estudos em Viçosa: Oliveira Cintra, Cirinete Simões, Anderson Cardoso e Aline Ribeiro.

À Universidade Federal de Viçosa e ao Programa de Pós-graduação em Genética e Melhoramento, pela oportunidade de cursar a graduação, o mestrado e o doutorado.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) por financiar a pesquisa.

Ao professor orientador Cosme Damião Cruz e a Michele Jorge Silva Siqueira, pela excelente orientação, pelos conhecimentos transmitidos, paciência e exemplo de profissionalismo.

Aos funcionários do BIOAGRO pela disponibilidade.

Aos secretários do Programa de Pós-graduação em Genética e Melhoramento por todo o auxílio prestado durante o mestrado e doutorado. A todos aqueles que colaboraram de alguma forma para este trabalho.

BIOGRAFIA

FRANCYSE EDITE DE OLIVEIRA CHAGAS DE MORAES, filha de Adalberto Custódio das Chagas e Maria Elesbão Neves de Oliveira Chagas, nasceu em 09 de dezembro de 1993, em Viçosa, no estado de Minas Gerais.

Em março de 2011, iniciou o curso de Licenciatura em Ciências Biológicas na Universidade Federal de Viçosa, onde obteve o título em janeiro de 2016.

Em março de 2016, iniciou o curso de Mestrado no Programa de Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em fevereiro de 2018.

Em março de 2018, iniciou o curso de Doutorado no Programa de Genética e Melhoramento na Universidade Federal de Viçosa, submetendo-se à defesa de tese em julho de 2022.

“Tudo tem o seu tempo determinado, e há tempo para todo propósito debaixo do céu. Tempo de chorar, e tempo de rir; tempo de prantear e tempo de dançar”.

(Eclesiastes 3.1 e 3.4)

RESUMO

DE MORAES, Francyse Edite de Oliveira Chagas, D.Sc., Universidade Federal de Viçosa, julho de 2022. **Predição genômica sob diferentes cenários que incluem, ou não, locos controladores de características.** Orientador: Cosme Damião Cruz.

O presente estudo avaliou o impacto do uso de diferentes conjuntos de marcadores sobre a eficiência da predição utilizando as técnicas RR-BLUP, árvore de decisão, *bagging, boosting* e *random forest*. As técnicas foram analisadas em relação a seis características. As características foram controladas pelos mesmos quarenta genes com diferentes herdabilidades (0,4, 0,6 e 0,8) acrescidos, ou não, por quatro genes com efeitos maiores de herdabilidade igual a um. O grau médio de dominância adotado foi um para todas as características. Dentro de cada gene havia um marcador. Os genes controladores de efeito menor estavam distribuídos equitativamente nos oito primeiros grupos de ligação (GL) e os quatro de efeito maior estavam nos quatro primeiros GL. Ao simular a aleatorização envolvida na formação dos gametas que originaram a população, pode segregar as marcas diferentemente do que era desejado. Além disso, os dados fenotípicos e genotípicos gerados podem ser diferentes dos pretendidos. No primeiro capítulo foi analisada a qualidade dos dados em relação a esses fatores. Foi testado se o conjunto de dados obtido por simulação expressava o padrão fenotípico e/ou genotípico das diversas características e se os marcadores segregavam corretamente. Encontrou-se que das 2010 marcas simuladas, somente cinco não segregavam como o esperado. As marcas estavam distribuídas equitativamente em dez grupos de ligação e por meio dos resultados de desequilíbrio de ligação. Mesmo com as marcas distorcidas, foi possível recuperar a ordem e a posição desses grupos. Em relação aos dados fenotípicos, encontrou-se que as características controladas por quarenta genes (x_1, x_3 e x_5) possuíam uma média de 127,97 e as características controladas por quarenta e quatro genes (x_2, x_4 e x_6) possuíam uma média de 220,21, independentemente da herdabilidade. As variâncias foram todas diferentes, mas com o padrão das que eram controladas por quarenta genes serem menores do que as controladas por quarenta e quatro genes. Ao se fazer a correlação entre os valores fenotípicos e valores genotípicos, recuperou-se o valor da herdabilidade das características próximo ao estipulado pela simulação. Observou-se que a presença de genes de efeitos maiores aumentava a herdabilidade,

facilitando o estabelecimento de classes de discriminação genotípica. Ao se plotar os dados para análise da distribuição fenotípica, observou-se distribuição contínua em x_1, x_3 e x_5. Em x_2, x_4 e x_6 foi visto padrão contínuo com tendência a estabilização e formação de duas regiões modais. Os dados simulados remetiam ao que era esperado, podendo ser usado nas análises. No segundo capítulo, foi analisada a eficiência da predição por meio da capacidade preditiva (r^2) e da raiz do erro quadrado médio (REQM) das técnicas RR-BLUP, árvore de decisão, *bagging*, *boosting* e *random forest* em cinco diferentes conjuntos de marcadores. Como mencionado anteriormente, os genes controladores de efeito menor estavam distribuídos equitativamente nos oito primeiros grupos de ligação (GL) e os quatro de efeito maior estavam nos quatro primeiros GL. Como foram simulados dez grupos de ligação com 201 marcas codominantes em cada, havia 1608 marcas diretamente ou indiretamente relacionadas aos genes e 402 marcas desnecessárias a predição. A formação dos conjuntos de marcadores levou essas informações como critério. No grupo um, estavam todos os marcadores. No grupo dois, os 1608 marcadores diretamente ou indiretamente relacionados aos genes. No grupo três, os quarenta e quatro marcadores dentro dos genes e os 402 marcadores não relacionados. No grupo quatro, os 402 marcadores desnecessários a predição. No grupo cinco, os quarenta e quatro marcadores diretamente relacionados aos genes controladores. Ao se analisar o r^2 e REQM das técnicas, observou-se que a maioria delas promoveu resultados péssimos na situação quatro. A técnica árvore de decisão chegou a não obter os valores em algumas repetições. Como nessa situação não havia marcadores relacionados as características, era esperado que em nenhuma técnica fosse possível obter resultados. A explicação veio pelo RR-BLUP. Ele forneceu o efeito dos marcadores sobre as características. Foram encontrados efeitos falsos positivos relacionados às 402 marcas desnecessárias a predição. Continuando-se as análises, foi observado que as técnicas *bagging* e *boosting* obtiveram os maiores valores de r^2 entre todas as técnicas (0,880 e 0,815, respectivamente) e os menores valores de REQM (5,852 e 5,853). A maioria dos valores foi obtida do quinto conjunto de dados e, ou não diferiu significativamente dos outros conjuntos, ou foi diferente apenas do conjunto quatro (sem marcadores relacionados). Resultado diferente foi observado para a *random forest*. Ela foi a mais sensível, tanto aos diferentes subconjuntos de marcadores quanto as diferentes características. Para o quinto conjunto de marcadores, obteve r^2 para as características x_3, x_4, x_5 e x_6, respectivamente iguais a

0,371; 0,720; 0,514 e 0,788. Para REQM, obteve, naquele mesmo conjunto, em x_3 e x_5, respectivamente, 10,280 e 8,371. Esses valores foram os melhores e diferentes significativamente dos obtidos para as mesmas características nos outros quatro conjuntos. Os resultados obtidos mostram que o uso de diferentes técnicas exploram melhor o conjunto de dados. Também mostra que o descarte de marcadores desnecessários não prejudica o processo preditivo, algumas vezes até o melhora, sendo recomendável. Trabalhos futuros deveriam se concentrar na identificação dos marcadores diretamente envolvidos com as características.

Palavras-chave: Simulação. Capacidade preditiva. Raiz do erro quadrado médio. RR-BLUP. Aprendizado de máquina.

ABSTRACT

DE MORAES, Francyse Edite de Oliveira Chagas, D.Sc., Universidade Federal de Viçosa, July 2022. **Genomic prediction under different scenarios that include, or not, trait-controlling loci.** Advisor: Cosme Damião Cruz.

The present study evaluated the impact of the use of different sets of markers on the prediction efficiency using the RR-BLUP, decision tree, bagging, boosting and random forest techniques. The techniques were analyzed in relation to six characteristics. The traits were controlled by the same forty genes with different heritability (0.4, 0.6 and 0.8) plus, or not, by four genes with greater heritability effects equal to one. The average degree of dominance adopted was one for all characteristics. Within each gene was a marker. The minor-effect controller genes were evenly distributed in the first eight linkage groups (GL) and the four major-effect genes were in the first four GL. By simulating the randomization involved in the formation of gametes that gave rise to the population, it can segregate the marks differently from what was desired. In addition, the phenotypic and genotypic data generated may differ from those intended. In the first chapter, the quality of the data in relation to these factors was analyzed. It was tested whether the dataset obtained by simulation expressed the phenotypic and/or genotypic pattern of the different traits and whether the markers segregated correctly. It was found that of the 2010 simulated brands, only five did not segregate as expected. The marks were evenly distributed across ten linkage groups and across linkage disequilibrium results. Even with the distorted marks, it was possible to recover the order and position of these groups. Regarding the phenotypic data, it was found that the traits controlled by forty genes (x_1, x_3 and x_5) had an average of 127.97 and the traits controlled by forty-four genes (x_2, x_4 and x_6) had an average of 220.21, regardless of heritability. The variances were all different, but with the pattern of those controlled by forty genes being smaller than those controlled by forty-four genes. By making the correlation between the phenotypic and genotypic values, the heritability value of the traits close to that stipulated by the simulation was recovered. It was observed that the presence of genes with greater effects increased heritability, facilitating the establishment of genotypic discrimination classes. When plotting the data for analysis of the phenotypic distribution, a continuous distribution was observed in x_1, x_3 and x_5. In x_2, x_4 and x_6 a continuous pattern was seen with a tendency to

stabilization and formation of two modal regions. The simulated data referred to what was expected and could be used in the analyses. In the second chapter, the prediction efficiency was analyzed through the predictive capacity (r^2) and the root mean square error (REQM) of the RR-BLUP, decision tree, bagging, boosting and random forest techniques in five different sets of markers. . As mentioned earlier, the minor-effect controller genes were evenly distributed in the first eight linkage groups (GL) and the four major-effect genes were in the first four GL. As ten linkage groups were simulated with 201 codominant markers in each, there were 1608 markers directly or indirectly related to genes and 402 markers unnecessary for prediction. The formation of the marker sets took this information as a criterion. In group one, there were all the markers. In group two, the 1608 markers directly or indirectly related to the genes. In group three, the forty-four markers within genes and the 402 unrelated markers. In group four, the 402 markers unnecessary the prediction. In group five, the forty-four markers were directly related to the controlling genes. When analyzing the r^2 and REQM of the techniques, it was observed that most of them promoted poor results in situation four. The decision tree technique did not obtain the values in some repetitions. As in this situation there were no markers related to the characteristics, it was expected that in no technique it would be possible to obtain results. The explanation came from RR-BLUP. It provided the effect of markers on traits. False positive effects were found related to the 402 unnecessary marks for prediction. Continuing the analysis, it was observed that the bagging and boosting techniques obtained the highest values of r^2 among all the techniques (0.880 and 0.815, respectively) and the lowest values of REQM (5.852 and 5.853). Most values were obtained from the fifth dataset and either did not differ significantly from the other sets or differed only from set four (no related markers). Different result was observed for random forest. She was the most sensitive, both to different subsets of markers and to different characteristics. For the fifth set of markers, r^2 was obtained for the characteristics x_3, x_4, x_5 and x_6, respectively equal to 0.371; 0.720; 0.514 and 0.788. For REQM, he obtained, in that same set, at x_3 and x_5, respectively, 10.280 and 8.371. These values were the best and significantly different from those obtained for the same characteristics in the other four sets. The results obtained show that the use of different techniques better explore the dataset. It also shows that discarding unnecessary markers does not harm the predictive process, sometimes even improves it, which is recommended. Future work should focus on identifying the markers directly involved with the traits.

Keywords: Simulation. Predictive capability. Root mean square error. RR-BLUP. Machine learning.

SUMÁRIO

INTRODUÇÃO GERAL

Em 1952, com o experimento de Hershey e Chase, a humanidade reconheceu o ácido desoxirribonucleico (DNA) como controlador dos processos celulares e das características apresentadas pelos seres vivos. Desde então, os estudos sobre essa molécula foram sendo aprofundados, de tal forma que hoje ela é usada para gerar e selecionar a variabilidade existente (HERSHEY & CHASE, 1952).

O ácido desoxirribonucleico é formado por nucleotídeos, que possuem açúcar pentose, base nitrogenada e grupamento fosfato. Esses nucleotídeos podem ter quatro tipos de bases nitrogenadas: adenina, timina, citosina ou guanina (LEVENE, 1920).

A sequência de bases em cada molécula determina a espécie trabalhada. Quando há nucleotídeos diferentes da sequência padrão, tem-se marcadores moleculares denominados polimorfismos de nucleotídeos únicos (SNPs).

Esses SNPs podem ser usados no melhoramento genético em técnicas que selecionam com base, tanto nas informações fenotípicas, quanto nas moleculares. Uma dessas técnicas é a seleção genômica ampla (GWS).

Na GWS ocorre o reconhecimento simultâneo de todos os polimorfismos, estimando-se o valor genético genômico (VGG) dos indivíduos (MASSMAN *et al.*, 2012). Por fazer essa análise simultânea, muitas vezes a importância dos polimorfismos durante a predição é negligenciada.

A GWS pode ser aplicada em metodologias como o Ridge Regression-Best Linear Unbiased Prediction (RR-BLUP), árvore de decisão e seus refinamentos (*random forest, bagging* e *boosting*). As técnicas baseadas no aprendizado de máquinas têm sido utilizadas com sucesso em trabalhos de predição (PARMLEY *et al.* ,2019).

Na árvore de decisão, a amostra a ser trabalhada é dividida em conjuntos homogêneos com base nas variáveis explicativas mais diferenciadores da variável resposta. Essa estratégia de divisão segue a abordagem de cima para baixo, conhecida como divisão binária recursiva (ESTRADA, 2015).

O *bagging* é uma técnica usada para reduzir a variância das previsões ou a taxa de erro de classificação. Essa técnica combina o resultado de vários classificadores modelados em diferentes sub-amostras do mesmo conjunto de dados.

No *random forest* é utilizada uma fração do número de variáveis para tentar eliminar a correlação entre as árvores geradas, melhorando ainda mais a precisão das previsões (JAMES *et al.*, 2013). No *boosting* são criadas árvores sequencialmente, buscando minimizar a função de perda (FREUND e SCHAPIRE, 1998).

Quando as empresas de genotipagem identificam os SNPs, muitos são descartados baseando-se na baixa frequência, presença em poucos exemplares, entre outros critérios (LAURIE *et al.*, 2010).

Os SNPs restantes são analisados pelos melhoristas, e podem estar associados às características de interesse, ou não. As características analisadas, por sua vez, podem ser controladas por genes de efeitos maiores e/ou possuir herdabilidades diferentes.

O presente trabalho visa considerar se a GWS, apesar de fazer o reconhecimento simultâneo das marcas, é afetada pela presença de polimorfismos desnecessários a predição. Analisou-se a capacidade preditiva e a raiz do erro quadrado médio de diferentes modelos preditivos. Esses modelos foram construídos utilizando-se diferentes conjuntos de marcadores, agrupados pela sua influência sobre as características. As informações fenotípicas foram provenientes de seis características, controladas por genes com diferentes efeitos e herdabilidades. A GWS foi aplicada nas metodologias RR-BLUP, árvore de decisão e seus refinamentos.

REVISÃO DE LITERATURA

1. Material genético dos seres vivos: o ácido desoxirribonucleico (DNA)

Desde que Morgan e seus colaboradores perceberam que os fatores de Mendel estavam associados aos cromossomos, desejou-se descobrir qual substância estava ali presente e se, de fato, era determinante na manifestação das características. As proteínas presentes nos cromossomos eram consideradas as responsáveis, sendo o DNA visto apenas como um material repetitivo associado a elas (MORGAN, 1911).

Griffith, ao tentar elaborar uma vacina contra pneumonia, percebeu que bactérias não patogênicas, quando em contato com bactérias patogênicas mortas por aquecimento, tornavam-se canais de doença. Ele propôs que as bactérias antes inofensivas haviam adquirido um “princípio transformante” das cepas maléficas, tornando-se patogênicas (GRIFFITH, 1928).

Avery, McCarty e MacLeod dispuseram-se a identificar o “princípio transformante” de Griffith, por meio de sua purificação. Os resultados apontaram para o DNA, mas eles foram criteriosos nas suas constatações. Sugeriram que talvez um contaminante em doses pequenas fosse o verdadeiro princípio (AVERY, MCCARTY E MACLEOC, 1943).

Somente com o experimento de Hershey e Chase em 1952 com fagos, que, finalmente, foi descoberto que o ácido desoxirribonucleico é o responsável pelas informações hereditárias. Sabia-se que as proteínas tinham enxofre em sua composição, enquanto o DNA tinha fosfato em seus nucleotídeos, além de açúcar desoxirribose e base nitrogenada (adenina, timina, guanina ou citosina) (LEVENE, 1920).

Hershey e Chase sabiam que os fagos se prendiam à superfície de uma célula bacteriana e injetavam alguma substância em seu interior. Esta substância coordenava a produção de muitos fagos pelas células hospedeiras (HERSHEY E CHASE, 1952).

Para saber qual era a substância injetada, marcaram o DNA dos fagos e as proteínas usando, respectivamente, isótopos radioativos de fósforo e enxofre. Como as células infectadas tinham somente fósforo radioativo, evidenciou-se que o DNA era a substância injetada (HERSHEY E CHASE, 1952).

2. Características quantitativas no melhoramento de plantas

Ao se fazer a análise estatística das características quantitativas, é percebida principalmente a variância genética aditiva. Essa variância correspondente ao dobro da variabilidade gamética, cujos efeitos são expressos pelos efeitos alélicos aditivos. Essa percepção deve-se mais aos efeitos da distribuição da frequência alélica do que a ação gênica em si, pois os efeitos não aditivos (dominância e epistasia) têm papel crucial na regulação genética dessas características (CRUZ, 2012; LONG *et al.*, 2011).

A variância genética de dominância deve-se a relação entre os alelos de um mesmo gene e a variância genética epistática entre alelos de genes diferentes. Essa última ocorre porque as características quantitativas são controladas por vários genes. Eles geralmente têm pequenos efeitos sobre a característica, sendo a distribuição dos valores genotípicos das diferentes classes contínuas. A genética quantitativa surgiu para modelar a ação desses genes (CRUZ, 2012).

Algumas técnicas usuais de melhoramento das espécies consistem na análise fenotípica dos dados. No estudo da herança de caracteres quantitativos, a média de um conjunto de indivíduos será uma medida mais confiável. Os efeitos do ambiente tendem a se cancelar entre as repetições. Outra medida que torna os dados mais confiáveis é a variância (CRUZ, 2012).

3. Modelo preditivo a partir da herdabilidade

Para melhorar o processo de escolha dos indivíduos utiliza-se a herdabilidade. Ela representa a proporção da variância fenotípica com origem genética (BORÉM & MIRANDA, 2013). A herdabilidade pode ser mensurada por:

$$h^2 = \frac{{\sigma_g}^2}{{\sigma_f}^2}$$

Em que:

h^2: herdabilidade;

${\sigma_g}^2$: variância genotípica;

${\sigma_f}^2$: variância fenotípica.

A correlação entre o valor fenotípico e genotípico depende diretamente desse valor. Quando se utiliza a variância aditiva ao invés da genotípica, obtêm-se a herdabilidade no sentido restrito (CAMARGO *et al.*, 1998).

4. Marcadores genéticos

Os marcadores moleculares podem ser definidos como segmentos de DNA que estão fisicamente ligados aos locos. Eles podem ser evidenciados por métodos que combinam o uso de enzimas de restrição a hibridização entre sequências complementares de DNA. Alguns métodos são o *Restriction Fragment Lenght Polymorphisms* (RFLP) e a técnica de *Polymerase Chain Reaction* (PCR) (MARIN *et al*., 2005).

O grande potencial do uso de marcadores moleculares no melhoramento, reside no fato de eles serem praticamente ilimitados em número, de fácil detecção e se comportarem como “caracteres” de herança simples e previsível, não sendo afetados pelo meio (MARIN *et al*., 2005).

Existem vários tipos de marcadores, como os *Simple Sequente Repeats* (SSR), o *Polymerase Chain Reaction – Random Amplified Polymorphism* DNA (PCR – RAPD), *Restriction Fragment Lenght Polymorphisms* (RFLP) e os SNPs (CRICK E WATSON, 1956).

Os SNPs consistem em nucleotídeos substituídos em pontos específicos do DNA, como nos éxons, íntrons ou em regiões intergênicas. Devido à pressão seletiva, que limita as alterações de aminoácidos nas sequências proteicas, a ocorrência de SNPs em regiões gênicas é limitada (MOTTA, 2016).

5. Genotipagem dos organismos e seleção de marcadores

As informações moleculares são obtidas pela genotipagem dos organismos. Nas plantas são retiradas amostras de tecido foliar e, em seguida, são realizados os procedimentos para extração do material genético (RAFFAN E SEMPLE, 2011).

O DNA é encaminhado para empresas que realizam a genotipagem. Com o genoma da espécie sequenciado, são produzidos adaptadores universais de DNA ou RNA, que se complementam com as moléculas enviadas (RAFFAN E SEMPLE, 2011).

Os adaptadores são chamados de SNPchips. As bases nitrogenadas dos nucleotídeos da amostra são coradas diferentemente por meio de técnicas modernas de fluorescência. Essas cores são detectadas por comprimentos de ondas diferentes (RAFFAN E SEMPLE, 2011).

Os fragmentos de sequência do DNA são alinhados com a sequência do DNA de referência, para observar a discordância entre as sequências e determinar se a variante é homozigótica ou heterozigótica. Bancos de dados de referência são utilizados para determinar se as variantes são novas ou previamente reconhecidas como SNPs (RAFFAN E SEMPLE, 2011).

Muitos SNPs são descartados seguindo critérios como valor de GCScore, *Minor Allele Frequency* (MAF), *Call Rate* e, em alguns casos, a análise do equilíbrio de Hardy-Weinberg (EHW) (LAURIE *et al.*, 2010).

O GCScore é uma medida de confiança associada a cada marcador, que varia de zero a um. Geralmente, os marcadores com valores inferiores a 0,5 são excluídos pelas empresas (LAURIE *et al.*, 2010).

A MAF é uma medida relacionada com a variação dos alelos na população. Alelos pouco frequentes são pouco informativos e não apresentam relevância genética na população. Considerando-se a forma alélica menos frequente, caso sua frequência seja inferior a 1% ou 5%, descarta-se aquele SNP (LAURIE *et al.*, 2010).

O *Call Rate* corresponde a taxa de atendimento na genotipagem de um grupo de organismos. Essa é uma medida usada para eliminar marcadores com grande quantidade de genótipos perdidos. Geralmente são excluídos marcadores com *Call Rate* inferior a 95% ou 90% (LAURIE *et al.*, 2010).

Para o teste do EHW, os marcadores fora do EHW são descartados (LAURIE *et al.*, 2010).

Os marcadores que são mantidos após os descartes podem estar associados às características de interesse ou não. Esse poderia ser mais um dos critérios de descarte, antes do uso dos SNPs nos programas de melhoramento genético. SILVA (2021) observou que a redução do número de marcadores preservou as mesmas conclusões biológicas da situação em que todos os marcadores foram usados.

6. Uso da informação molecular para estabelecer critério de seleção de indivíduos

A seleção genômica ampla (GWS) surgiu no início dos anos 2000 como uma nova abordagem para a seleção assistida por marcadores (MAS). Ela seleciona indivíduos favoráveis com base em seus valores genéticos genômicos preditos, fundamentada na ocorrência de desequilíbrio de ligação entre os marcadores e os

locos que governam as características quantitativas (QTL) da população estudada (MEUWISSEN *et al*., 2001; RESENDE, 2008; CROSSA, 2017).

No primeiro trabalho proposto para uso dessa tecnologia, não havia a possibilidade de identificação de marcadores por todo o genoma, tendo sido o trabalho feito com dados simulados (MEUWISSEN *et al*., 2001).

O objetivo da GWS é obter um modelo que seja capaz de predizer o valor genético do indivíduo, mas que não necessariamente determine genes específicos envolvidos no controle do caráter. São necessárias três populações de trabalho nessa metodologia: população de treinamento, de validação e de seleção (CRUZ, SALGADO E BHERING, 2013).

A população de treinamento é usada para a genotipagem dos marcadores moleculares e fenotipagem das características de interesse, obtendo-se equações de predição dos valores genético genômicos (VGG) (CRUZ, SALGADO E BHERING, 2013).

A predição é realizada usando-se os efeitos estimados com base na população de treinamento e submetidos à análise de correlação com os valores genéticos obtidos via análise dos dados fenotípicos, conseguidos por metodologia padrão (BLUP) (CRUZ, SALGADO E BHERING, 2013).

A população de validação é usada para testar as equações de predição e avaliar a acurácia do modelo previamente estabelecido. Sendo a acurácia a proximidade entre os resultados obtidos pelos modelos preditivos e os valores reais (CRUZ, SALGADO E BHERING, 2013).

Para ser possível analisar a acurácia das equações de predição, é necessário que a população de validação e de treinamento sejam independentes. Assim sendo, todos os erros associados aos valores preditos e observados são também independentes, e toda correlação entre esses valores é de natureza genética e indica a capacidade preditiva da metodologia (CRUZ, SALGADO E BHERING, 2013).

Na população de seleção tem-se apenas os marcadores avaliados nos candidatos à seleção. Essa população não necessita ter seus fenótipos avaliados. As equações de predição derivadas da população de descoberta são, então, usadas na predição dos VGGs ou fenótipos futuros dos candidatos à seleção (CAVALCANTI *et al*., 2012).

A GWS tem recebido cada vez mais informações, como a interação entre genótipo e ambiente, consideração dos efeitos não aditivos e análise combinada de várias características em vários ambientes (BURGUEÑO *et al.*, 2012).

A metodologia apontada é capaz de aumentar os ganhos genéticos e também diminuir o tempo para obtenção destes ganhos, e encurtar o tempo dos ciclos com o aumento da eficiência no uso do germoplasma, em relação as medidas convencionais (CROSSA, 2017).

Por meio da GWS é possível realizar a escolha dos pais para formação de uma população a ser melhorada. Essa população pode chegar a ter ganho mesmo em características negativamente relacionadas, mostrando que o desafio da seleção simultânea de variáveis agronômicas pode ser futuramente superado com a utilização dessa técnica (HAO, 2019).

7. Técnicas biométricas de predição

Existem várias abordagens de predição de valores genéticos a partir das informações de marcadores moleculares. De acordo com o paradigma estatístico, a predição é fundamentada em princípios estatísticos que envolvem operações sumarizadas em médias, variâncias e covariâncias. Pressuposições sobre distribuições também são fundamentais. No paradigma do aprendizado de máquinas, as estratégias buscam a obtenção de soluções por algum critério de otimização que, no caso da seleção genômica, consiste na busca de subdivisões de espaços preditores que conduzem à maior eficiência de predição.

7.1 Abordagem por métodos estatísticos

Os principais métodos estatísticos para estabelecer critérios de seleção e gerar valores preditos com informações de variáveis auxiliares podem ser divididos em três classes: regressão explícita, implícita e com redução dimensional (RESENDE 2002, 2008).

Os métodos da classe de regressão explícita podem ser divididos em dois grupos: (i) métodos de estimação penalizada, como RR-BLUP (MEUWISSEN *et al.*, 2001) e LASSO (TIBSHIRANI, 1996); (ii) métodos de estimação bayesiana, tais como Bayes A e Bayes B (MEUWISSEN *et al.*, 2001).

Os métodos de regressão com redução dimensional, por sua vez, compreendem os componentes independentes, quadrados mínimos parciais e de componentes principais (SOLBERG *et al.*, 2009). Na classe de regressão implícita, destacam-se os métodos semi-paramétricos RKHS (Reproducing Kernel Hilbert Spaces) (GIANOLA E DE LOS CAMPOS, 2008).

As metodologias baseadas em penalizações são eficientes na predição de valores genéticos em características com controle gênico aditivo (SILVA, 2018; SANT'ANNA, 2018). A técnica, quando aplicada em estudos que envolvem marcadores moleculares, consiste em estabelecer uma regressão linear simples para estimar o efeito individual de cada marcador, que pode ser descrita pela equação:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Onde:

y é o valor fenotípico e

x representa o valor genotípico de aa (-1), Aa (0) ou AA (1).

Essa expressão pode ser generalizada para o cálculo do efeito referente a todos os marcadores em uma equação multivariada com os β efeitos sendo estimados simultaneamente, assim tem-se a expressão:

$$y = \beta_0 + x_1\beta_1 + \cdots + x_m\beta_m + \varepsilon$$

Onde β_m representa o efeito para cada marcador.

A partir dos efeitos estimados, é possível calcular o valor genético genômico estimado – *Estimated Genomic Breeding Values* (EGBV), que de forma matricial é calculado conforme a expressão abaixo:

$$EGBV = X\hat{\beta}$$

Em que:

X é matriz de incidência e

$\hat{\beta}$ é o vetor do efeito de marcadores.

Contudo, a implementação da GWS impõe desafios estatísticos e computacionais, tais como a dimensionalidade do modelo, colinearidade entre marcas e a complexidade das características quantitativas (SANT'ANNA, 2018).

Para contornar os desafios da GWS, vários métodos têm sido propostos, no qual se diferem pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo (SANT'ANNA, 2018). Entre eles, os métodos de redução de dimensionalidade, tais como: o *Ridge Regression Best Linear Unbiased Prediction*

(RR-BLUP), redução via combinações lineares independentes, mínimos quadrados parciais e seleção de um subconjunto de marcadores por meio de procedimentos específicos como sondagens ou regressão *Stepwise*, sob os princípios bayseanos (Bayes A, Bayes B, Bayes C, LASSO Bayesiano, etc).

7.2 Abordagem por técnicas de aprendizado de máquinas

O aprendizado de máquina corresponde ao estudo de algoritmos possibilitando previsão e inferência. Ele visa escolher conjunto de modelos que podem melhor prever dados não observados. As técnicas baseadas em aprendizado de máquinas podem ser divididas em árvore de decisão e seus refinamentos: *bagging*, *random forest* e *boosting* (SILVA JUNIOR *et al.*, 2021).

i) Árvore de decisão

A árvore de decisão é uma metodologia que divide o conjunto de dados trabalhados em dois, de acordo com algum critério, de forma que os indivíduos alocados em cada grupo recebem o mesmo valor predito. Caso a variável analisada ao dividir o conjunto de dados seja contínua, a árvore de decisão formada é do tipo regressão. Caso seja discreta, fala-se em árvore de classificação (SILVA JUNIOR *et al.*, 2021).

A árvore de decisão é composta por nós internos, ramos e nós externos, também chamados de folhas. Quando os indivíduos de um grupo ainda serão reagrupados em grupos menores, fala-se em um nó interno, que se liga ao novo grupo pelos ramos. Quando não há formação de novos grupos, fala-se em nó externo (folha) (SILVA JUNIOR *et al.*, 2021).

Na árvore de regressão, as duas regiões formadas são obtidas por divisões binárias recursivas. As divisões visam obter a variável X_p e o ponto s que dividam o espaço em duas regiões, de forma que se obtenha o menor erro quadrático médio.

Em uma árvore de decisão muito grande, pode ocorrer *overfitting* (super-ajuste) dos dados, enquanto em uma árvore pequena, pode não haver uma boa captura das informações. Para melhorar a taxa de acerto do modelo na aplicação em outros conjuntos de dados de treinamento, são utilizados métodos de poda (*pruning*) da árvore (SILVA JUNIOR *et al.*, 2021).

Inicialmente se constrói a árvore até que os grupos formados tenham no máximo cinco indivíduos. Depois, é feita a poda usando o custo complexidade da poda, tornando, assim, a árvore de regressão menor e menos complexa (HASTIE *et al.*, 2009).

ii) ***Bagging***

O *Bootstrap Aggregation* (*Bagging*) aplica a técnica de *bootstrap*. São obtidas várias amostras com reposição do conjunto de dados e, para cada amostra, um modelo (BREIMAN, 2001).

Os dados originais são substituídos e novos conjuntos são formados a cada amostragem. Os modelos criados são utilizados para reduzir a variabilidade obtida nas árvores de decisão e para obter uma média, que será o modelo final (BREIMAN, 2001).

A técnica de *Bagging* monta modelos robustos, menos propensos a *overfitting*. Utiliza-se uma quantidade de árvores que proporcione uma estabilização do erro (JAMES *et al.*, 2013).

iii) Random Forest

No *Bagging*, as árvores obtidas estarão altamente correlacionadas, uma vez que está sujeito a quase sempre a mesma variável estar no topo da árvore (HASTIE *et al.*, 2009; JAMES *et al.*, 2013). Além disso, a média de valores que são altamente correlacionados, não resulta em uma grande redução da variância, como ocorre quando a média é feita com valores não correlacionados (JAMES *et al.*, 2013).

O *Random Forest* apresenta o mesmo princípio de *Bagging*, mas amostrando a quantidade de variáveis preditoras (m<p) utilizadas em cada partição. Ele obtém os valores preditos mais independentes, uma vez que proporciona redução da variabilidade encontrada nas árvores de decisão.

É sugerido que o número de variáveis preditoras utilizadas em cada partição seja $m = \sqrt{p}$ para árvore de classificação e $m = p/3$ para árvores de regressão (HASTIE *et al.*, 2009). Dessa forma, as predições das árvores se tornam menos correlacionadas e a mesma variável não estará sempre no topo da árvore.

O conjunto de treinamento é cerca de $\frac{2}{3}$ do tamanho do conjunto de dados original, ou seja, cerca de $\frac{1}{3}$ não serão utilizados no processo de treinamento. Essa parte dos dados forma o erro fora da cesta (OOB), que pode ser utilizado no processo de validação do modelo.

iv) Boosting

No *Boosting* são criadas várias árvores sequencialmente, utilizando, para isso, de informação prévia da árvore anterior (SOUZA *et al.*, 2020). Nele, uma única árvore é aperfeiçoada.

Necessita-se de vários modelos, sendo um processo de aprendizagem lento. É utilizada a validação cruzada para se escolher o número de árvores que será construído, reduzindo dessa forma a possibilidade de *overfitting*, dado que todos os indivíduos participarão do conjunto de validação (BENGIO E GRANDVALET, 2004).

REFERÊNCIAS

Avery, OT. MacLeod, CM. McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. 1943.

Bengio,Y. Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation.2004.

Borém, A. Miranda, GV. Melhoramento de Plantas. 2013.

Breiman, L. Bagging Predictors. Machine Learning 24: 123–140.2001.

Burgueño, J. De Los Campos, G. Weigel, K. Crossa, J. Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers.2012.

Camargo, CE de O. Filho, AWPF. Felício, JC. Herdabilidade e correlações entre características agronômicas em populações híbridas de trigo. 1998.

Cavalcanti, JJV. Resende, MDV. Santos, FHC dos. Pinheiro, CR. Simultaneous prediction of the effects of molecular markers and genome wide selection in cashew. 2012.

Crick, FH. Watson, JD. Structure of Small Viruses. Nature. Vol.177 pp.473-5 ref.19.1956.

Crossa, J. Perez-Rodríguez, P. Cuevas, J. Montesinos-Lopez, O. Jarquín, D. de Los Campos, G. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22, 961–975.2017.

Cruz, CD. Princípios de Genética Quantitativa. 2ª edição, capítulo 1. 2012.

Cruz, CD. Salgado, CC. Bhering, LL. Genômica Aplicada. Capítulo X: Seleção Genômica Ampla. Pag. 210. 2013.

Estrada, G Del CC. Árvore de decisão aplicada à análise de risco da severidade da ferrugem do cafeeiro na Guatemala. 2015.

Freund, Y. Schapire, RE. Large Margin Classification Using the Perceptron Algorithm. 1998.

Gianola, D. De Los Campos, G..Inferring genetic values for quantitative traitsnon-parametrically. Genetics Research, 90(6), pp.525-540.2008.

Griffith. Epidemiology & Infection. Volume 27, Issu 2, pp. 113 – 159.1928.

Hastie, T. Tibshirani, R. Friedman, J. The Elements of Statiscal Learning Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 745p.2009.

Hao, Y. Wang, H. Yang, X. Zhang, HC. Li, D.Li, H. Wang, G. Wang, J Fu. Genomic prediction using existing historical data contributing to selection in biparental populations: a study of kernel oil in maize. Plant Genome. 2019.

Hershey, a.d..Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. 1952.

James, G. Witten, D. Hastie, T. Tibshirani, R. An Introduction to Statistical Learning. Springer, New York. 2013.

Laurie, CC. Doheny, K. F. Mirel, DB. Pugh, EW. Bierut, LJ. Bhangale, T. Boehm, F. Caporaso, NE. Cornelis, MC. Edenberg, HJ. Gabriel, SB. Harris, EL. Hu, FB. Jacobs, KB. Kraft, P. Landi, MT. Lumley, T. Manolio, TA. McHugh, C. Painter, I. Paschall, J. Rice, JP. Rice, KM. Zheng, X. Weirl, BS. Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies. 2010.

Levene, PA. Properties of the nucleotides obtained from yeast nucleic acid.1920.

Long, N. Gianola, D. Rosa, GJM. Marker-assisted prediction of non-additive genetic values. 2011.

Marin, ALA. Cervigni, GDL. Moreira, MA. Barros, EG. Seleção Assistida por Marcadores Moleculares Visando ao Desenvolvimento de Plantas Resistentes a Doenças, com Ênfase em Feijoeiro e Soja**.** 2005.

Massman, JM. Jung, H-J G. Bernardo, R. Genome wide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize. 2012.

Meuwissen, THE. Hayes, BJ. Goddard, ME. Prediction of total genetic value using genome-wide dense marker maps. 2001.

Morgan, TH. Random segregation versus coupling in Mendelian inheritance. 1911.

Motta, LB. Genotipagem por sequenciamento para identificação de SNPs e associação com características agronômicas em *Coffea canéfora*. 2016.

Parmley, KA. Higgins, RH. Ganapathysubramanian, B. Machine Learning Approach for Prescriptive Plant Breeding. Sci Rep 9, 17132. 2019.

Raffan, E. Semple, RK. Next generation sequencing—implications for clinical practice. 2011.

Resende, MDV. Genética biométrica e estatística no melhoramento de plantas perenes. Brasília, 2002.

Resende, MDV. Genômica quantitativa e seleção no melhoramento de plantas perenes e animais. Colombo: Embrapa Florestas, p. 330.2008.

Sant'anna, IC. Redes Neurais Artificiais para Predição Genômica na Presença de Interações Epistáticas. Viçosa: Universidade Federal de Viçosa. 93 p. 2018.

Silva, GN. Predição de valores genéticos por abordagens de Seleção Genômica Ampla e de Inteligência Computacional. Viçosa: Universidade Federal de Viçosa. 108 p. 2018.

Silva Júnior, AC da. Silva, MJ da. Cruz, CD. Sant'anna, I de C.Silva, GN. Nascimento, M. Azevedo, CF. Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. 2021.

Silva, MJ da. Efficiency of genomic prediction according to decreased the SNP´s markers and different degrees of dominance, heritability, and epistatic interactions. 2021.

Solberg, TR. Sonesson, AK. Wooliams, JA. Meuwissen, THE. Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution, 41:299. 2009.

Sousa, IC. Nascimento, M. Silva, GN. Nascimento, ACC. Cruz, CD.Fonseca, F. Almeida, DP. Pestana, KN. Azevedo CF. Zambolim, L.Caixeita, ET. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola. 2020.

Tibshirani, R.Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. B 58: 267–288.1996.

CAPÍTULO 1

QUALIDADE DOS DADOS SIMULADOS PARA FINS DE ESTUDO DE PREDIÇÃO NO MELHORAMENTO GENÉTICO

RESUMO

DE MORAES, Francyse Edite de Oliveira Chagas, D.Sc., Universidade Federal de Viçosa, julho de 2022. **Qualidade de dados simulados para fins de estudo de predição no melhoramento genético.** Orientador: Cosme Damião Cruz.

O presente capítulo avaliou a eficiência da simulação em gerar dados que se equiparam aos dados reais. Ao simular a aleatorização envolvida na formação dos gametas que originaram a população, pode segregar as marcas diferentemente do que era desejado. Além disso, os dados fenotípicos e genotípicos gerados podem ser diferentes dos pretendidos. Para analisar a qualidade dos dados em relação a esses fatores, foi testado se o conjunto de dados obtido por simulação expressava o padrão fenotípico e/ou genotípico das diversas características e se os marcadores segregavam corretamente. O conjunto de dados representa uma população F_2 derivada do cruzamento entre pais homozigotos contrastantes. Foram simuladas seis características para fins de predição. As características foram controladas pelos mesmos quarenta genes com diferentes herdabilidades (0,4, 0,6 e 0,8) acrescidos, ou não, por quatro genes com efeitos maiores de herdabilidade igual a um. O grau médio de dominância adotado foi um para todas as características. Dentro de cada gene havia um marcador. Os genes controladores de efeito menor estavam distribuídos equitativamente nos oito primeiros grupos de ligação (GL) e os quatro de efeito maior estavam nos quatro primeiros GL. Encontrou-se que das 2010 marcas simuladas, somente cinco não segregavam como o esperado. As marcas estavam distribuídas equitativamente em dez grupos de ligação e por meio dos resultados de desequílibrio de ligação. Mesmo com as marcas distorcidas, foi possível recuperar a ordem e a posição desses grupos. Em relação aos dados fenotípicos, encontrou-se que as características controladas por quarenta genes (x_1, x_3 e x_5) possuíam uma média de 127,97 e as características controladas por quarenta e quatro genes (x_2, x_4 e x_6) possuíam uma média de 220,21, independentemente da herdabilidade. As variâncias foram todas diferentes, mas com o padrão das que eram controladas por quarenta genes serem menores do que as controladas por quarenta e quatro genes. Ao se fazer a correlação entre os valores fenotípicos e valores genotípicos, recuperou-se o valor da herdabilidade das características próximo ao estipulado pela simulação. Observou-se que a presença de genes de efeitos maiores aumentava a herdabilidade, facilitando

o estabelecimento de classes de discriminação genotípica. Ao se plotar os dados para análise da distribuição fenotípica, observou-se distribuição contínua em x_1, x_3 e x_5. Em x_2, x_4 e x_6 foi visto um padrão contínuo com tendência a estabilização e formação de duas regiões modais. Os resultados obtidos demonstram a eficiência da simulação na criação de dados que condizem com a realidade que se deseja imitar.

Palavras-chave: Simulação. Características. Grau médio de dominância. Média. Variância. Desequilíbrio de ligação.

ABSTRACT

DE MORAES, Francyse Edite de Oliveira Chagas, D.Sc., Universidade Federal de Viçosa, July 2022. **Quality of simulated data for purposes of prediction study in genetic improvement.** Advisor: Cosme Damião Cruz.

This chapter evaluated the efficiency of the simulation in generating data that match real data. By simulating the randomization involved in the formation of gametes that gave rise to the population, it can segregate the marks differently from what was desired. In addition, the phenotypic and genotypic data generated may differ from those intended. To analyze the quality of the data in relation to these factors, we tested whether the dataset obtained by simulation expressed the phenotypic and/or genotypic pattern of the various traits and whether the markers segregated correctly. The dataset represents an F_2 population derived from the cross between contrasting homozygous parents. Six features were simulated for prediction purposes. The traits were controlled by the same forty genes with different heritability (0.4, 0.6 and 0.8) plus, or not, by four genes with greater heritability effects equal to one. The average degree of dominance adopted was one for all characteristics. Within each gene was a marker. The minor-effect controller genes were evenly distributed in the first eight linkage groups (GL) and the four major-effect genes were in the first four GL. It was found that of the 2010 simulated brands, only five did not segregate as expected. The tags were evenly distributed across ten linkage groups and through linkage disequilibrium results. Even with the distorted marks, it was possible to recover the order and position of these groups. Regarding the phenotypic data, it was found that the traits controlled by forty genes (x_1, x_3 and x_5) had an average of 127.97 and the traits controlled by forty-four genes (x_2, x_4 and x_6) had an average of 220.21, regardless of heritability. The variances were all different, but with the pattern of those controlled by forty genes being smaller than those controlled by forty-four genes. By making the correlation between the phenotypic and genotypic values, the heritability value of the traits close to that stipulated by the simulation was recovered. It was observed that the presence of genes with greater effects increased heritability, facilitating the establishment of genotypic discrimination classes. When plotting the data for analysis of the phenotypic distribution, a continuous distribution was observed in x_1, x_3 and x_5. In x_2, x_4 and x_6 a continuous pattern was seen with a tendency to stabilization and formation of two

modal regions. The results obtained demonstrate the efficiency of the simulation in creating data that match the reality that you want to imitate.

Keywords: Simulation. Characteristics. Mean dominance degree. Mean. Variance. Linkage disequilibrium.

INTRODUÇÃO

A predição do valor genético é de grande importância no melhoramento genético que pratica rotineiramente o processo de eliminar indivíduos ou de selecioná-los para fins de recombinação e formação de nova progênie.

Para realizar a predição, pode-se contar com dados obtidos de ensaios experimentais planejados ou contar com dados gerados por simulação.

As simulações permitem obter grande volume de informações, em que se conhece valores paramétricos e as hipóteses podem ser convenientemente testadas. Os dados são obtidos em curto período de tempo e não há custos de implantação e condução de experimentos (BHERING, 2008).

Esses dados obtidos por simulação também podem ser usados para analisar a eficiência da seleção genômica (MEUWISSEN *et al.*, 2001), avaliar a importância de caracteres auxiliares de uma característica principal (SILVA JUNIOR *et al.*, 2021), avaliar a eficiência da seleção genômica com base na regressão quantílica regularizada para a seleção de genótipos (OLIVEIRA *et al.*, 2021), entre outros trabalhos.

Quando o assunto é a predição do valor fenotípico via marcadores moleculares, tem-se que preocupar com a qualidade tanto das informações genotípicas quanto das fenotípicas.

No contexto genotípico é necessário que se defina apropriadamente o tipo de população para fins de inferência, podendo ser famílias (irmãos completos ou meio irmão), F_2, RIL, retrocruzamento, dentre outras. A observância quanto ao tipo de segregação mendeliana, genoma da espécie trabalhada, densidade de marcadores, grupos de ligação e desequilíbrio de ligação é fundamental.

No contexto fenotípico, tem-se fatores perturbadores relevantes, pois o fenótipo é o resultado da expressão do genótipo, com efeitos de dominância e epistasia, e do ambiente.

Nos estudos de simulação, fazemos distinção entre as características monogênicas, oligogênicas, poligênicas e aquelas que são controladas por grupos de genes de efeitos maiores com efeitos adicionais de poligenes de pequeno efeito.

Os genes de efeitos maiores seriam aqueles que acrescentam muitas unidades de medida ao valor genotípico da característica, enquanto os poligenes de pequeno efeito acrescentam poucas unidades de medida.

Para estudos envolvendo caraterísticas monogênicas, nos apoiamos nos trabalhos de Mendel que associou a cada fator uma característica (1ª Lei de Mendel) e afirmou que os fatores determinantes de características diferentes segregavam de maneira independente (2ª Lei de Mendel). Esses fatores são os alelos, formas alternativas de um gene (MENDEL, 1965).

A medida que os estudos genéticos foram avançando, descobriu-se que havia fatores responsáveis por mais de uma característica (pleiotropia) e características controladas por dois ou mais fatores, podendo esses segregar independentemente ou não.

A relação entre alelos do mesmo gene é determinada pelo grau médio de dominância (g.m.d.) da característica, que expressa a posição relativa do heterozigoto em relação à média dos homozigotos (CRUZ, 2012).

Essa posição relativa é dada pela divisão “d/a”, em que “d” representa a distância do valor genotípico do heterozigoto em relação ao ponto médio e “a” representa a distância do valor genotípico do homozigoto dominante em relação ao ponto médio. O ponto médio (μ) corresponde a metade da distância entre o valor genotípico do homozigoto dominante e o valor genotípico do homozigoto recessivo (CRUZ, 2012).

O g.m.d. afeta o número de classes fenotípicas observadas para uma característica. Ocorrendo dominância completa, haverá duas classes e, nas demais situações, haverá três classes.

Para estudos envolvendo caraterísticas oligogênicas, temos que considerar também a existência de interações entre alelos de diferentes genes, que determina o efeito epistático.

Para entender a epistasia, pode-se usar o contexto molecular, em que cada gene está associado a uma enzima. Caso a forma alélica presente em um gene esteja relacionada à produção de uma enzima defeituosa, a expressão de outros genes que dependeriam daquela enzima será afetada (SOUZA *et al.*, 2015).

Para simular uma característica controlada por apenas dois genes (por exemplo numa população F_2, com interação do tipo genes duplos recessivos sem efeito cumulativo e segregação 9:7), podemos nos fundamentar em dois modelos.

O primeiro modelo é definido por princípios biológicos envolvendo a expressão gênica, por meio de vias biosintéticas. Assim, na simulação de uma característica que apresenta dois padrões fenótipos, por exemplo, de alto e baixo teor com segregação 9:7, tem-se o modelo de predição descrito na Figura 1.

Por este modelo, podemos predizer o fenótipo (alto ou baixo) de qualquer constituição genotípica de predição (SOUZA *et al.*, 2015).

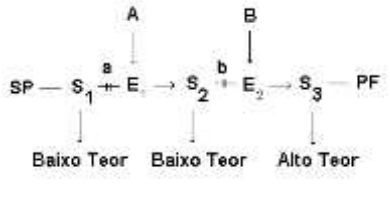


Figura 1. Modelo de predição de característica oligogência mostrando a ação de dois genes A/a e B/b numa via biossintética destacando subtância precursora (SP), substâncias intermediárias (S_1, S_2 e S_3) e um produto final (ou também intermediário) e as enzimas determinadas pelos alelos dos genes estudados.

O segundo modelo é obtido recorrendo aos princípios da Genética Quantitativa em que o valor fenotípico de determinado genótipo é determinado pela ação do efeito aditivo (a), da dominância (d) e epistasia, além do efeito ambiental. Assim, considerando dois genes A/a e B/b podemos associar os efeitos genotípicos ao modelo: (CRUZ, 2012)

$$Y_{ij} = B_0 + B_1a_{A/a} + B_2a_{B/b} + B_3d_{A/a} + B_4d_{B/b} + B_6aa + B_7ad + B_8da + B_9dd$$

Neste modelo, deve-se atribuir valores 1, 0, -1 para o efeito a ($a_{A/a} + a_{B/b}$) para AA, Aa e aa; valores 0, 1, 0 para o efeito d ($d_{A/a} + d_{B/b}$) e os produtos destes coeficientes para os efeitos epistáticos (aa, ad, da e dd). Pode ser demonstrado que, para o padrão 9:7, tem-se em B_0 uma constante e os demais coeficientes B_i (i=1,2...9) todos iguais a um (CRUZ, 2012).

Por fim, temos que fazer as considerações sobre as duas últimas situações que envolvem a simulação de característica controlada por poligenes com, ou não, a ação adicional de alguns poucos genes de efeitos maiores. Para estes casos, precisamos considerar uma série de fatores no processo de simulação, para que o valor estabelecido tenha uma boa aproximação da realidade.

Deve ser considerado que quanto mais genes, geralmente menor é o efeito de cada um, e as diferenças entre os genótipos tornam-se mínimas, ocorrendo padrão contínuo de distribuição das classes. Assim, apesar da dominância e a epistasia

provocar certa assimetria nos dados genotípicos, o elevado número de classes sujeitas à ação modificadora do ambiente conduz a um padrão relativamente simétrico na distribuição dos dados (CRUZ, 2012).

Antes de se iniciar o processo preditivo com base nos dados simulados, é interessante verificar se esses dados condizem com a realidade que desejam imitar. Por exemplo, analisar se a segregação dos marcadores usados condiz com o esperado (CHAGAS, 2018).

Ao se simular, a aleatorização envolvida na formação dos gametas que originarão aquela população, pode-se segregar as marcas diferentemente do que se é desejado. Essa segregação incorreta pode impedir a recuperação da posição das marcas e do número de grupos de ligação (GL). Além disso, erros do próprio programa podem gerar dados com valores genotípicos e fenotípicos diferentes do desejado, não se obtendo a herdabilidade planejada.

Com base no exposto, neste estudo será considerado a geração de dados simulados para fins de estudo de predições no melhoramento genético, avaliando-se a qualidade dos dados gerados, submetidos aos diferentes graus de herdabilidade e efeito diferencial de genes dentro do contexto genotípico e fenotípico.

MATERIAIS E MÉTODOS

1- Conjunto de dados

Foi simulada uma população F_2 composta por 1000 indivíduos, derivados do cruzamento entre pais homozigotos contrastantes.

2- Genotipagem

Foram simulados 2100 marcadores moleculares codominantes do tipo SNP, com dois alelos por marcador. Os marcadores foram distribuídos em um genoma estabelecido por 10 grupos de ligação (GL), refletindo uma espécie diploide com 2n = 2x = 20. Cada GL tem 200 centimorgans, sendo a distribuição dos marcadores equidistantes.

3- Fenotipagem

Seis características foram simuladas. Nelas, a importância dos efeitos “a” e “d” foi dada por pesos estabelecidos de uma distribuição uniforme.

Cada característica possui média em torno de 100, grau médio de dominância (gmd) um e herdabilidade variada.

Foram considerados dois modelos. O primeiro admitindo o estabelecimento de características poligênicas, cujo valor fenotípico foi dado por:

$$Y_i = \mu + \sum \alpha_j + \sum_j \alpha_j \alpha_{j+1} + e_i.$$

Em que:

Y_i é o valor fenotípico para observação i;

μ é a média geral;

α_j é o efeito do alelo favorável no loco j e assume os valores $u + a_i, u + d_i$ e $u - a_i$ para os valores genotípicos associados às classes AA, Aa e aa, respectivamente. As classes foram identificadas pela codificação 1, 0 ou -1, respectivamente.

$\alpha_j \alpha_{j+1}$ representa a interação entre alelos favoráveis em locos diferentes.

$e \sim N(0, V_e)$ representa a estrutura de variância dos resíduos, no qual $V_e = ((1 - h^2)V_g)/h^2$, sendo V_e a variância residual, V_g a variância genotípica e h^2 a herdabilidade.

O primeiro modelo criou as características poligências x_1, x_3 e x_5, controladas por 40 genes com herdabilidade de 0,4, 0,6 e 0,8, respectivamente. Estudos anteriores

usaram menos de vinte locos de caracteres quantitativos (QTL) (GIANOLA *et al.*, 2011; LONG *et al.*, 2010, 2011), sendo o maior número de QTL um diferencial desse trabalho.

O segundo modelo admite características poligênicas com a ação adicional de alguns poucos genes de efeitos maiores. Este, gerou as características x_2, x_4 e x_6 pelo acréscimo de quatro genes de efeito maior nos poligenes criados no primeiro modelo.

Foi estabelecido que os quatro genes atuavam de forma a simular a ação de genes recessivos sem efeito cumulativo, de forma que o padrão dominante (A-B-C-D-) proporcionou um acréscimo no valor Y_i equivalente ao máximo estabelecido pela ação dos poligenes e os demais padrões um acréscimo equivalente ao mínimo estabelecido pelos poligenes. Estes valores de acréscimos são arbitrários e estabelecidos de forma coerente com a grandeza e escala da variável que está sendo analisada.

Os 40 poligenes presentes em x_2, x_4 e x_6 tinham herdabilidades 0,4, 0,6 e 0,8, respectivamente. Os quatro genes de efeito maior não foram afetados pelo efeito ambiental, ou seja, possuíam herdabilidade igual a um.

As características foram altamente correlacionadas. Todas compartilhavam os mesmos quarenta genes. As variáveis controladas por quarenta e quatro genes, compartilhavam também os genes de efeito maior.

Cada gene controlador tinha um marcador em seu interior, havendo, portanto, quarenta SNPs relacionados diretamente aos genes das características x_1, x_3 e x_5 e quarenta e quatro SNPs relacionados diretamente aos genes das características x_2, x_4 e x_6.

Os quarenta marcadores em comum estavam nos oito primeiros GL, sendo cinco marcadores em cada grupo. Os quatro marcadores de maior efeito estão nos quatro primeiros GL. A posição dos marcadores nos grupos está representada na Tabela 1.

Tabela 1. Posição correspondente dos cinco marcadores controladores das características, em cada grupo de ligação. Os marcadores de maior efeito para as características x_2, x_4 e x_6 estão destacados em negrito.

GL/POSIÇÃO DO MARCADOR(*)	1º	2º	Efeito maior	3º	4º	5º
GL1	11	56	**80**	101	146	191
GL2	212	257	**280**	302	347	392
GL3	413	458	**480**	503	548	593
GL4	614	659	**680**	704	749	794
GL5	815	860	-	905	950	995
GL6	1016	1061	-	1106	1151	1196
GL7	1217	1262	-	1307	1352	1397
GL8	1418	1463	-	1508	1553	1598

(*) Cada grupo de ligação apresenta 201 marcadores

Os dois últimos grupos de ligação não têm marcadores relacionados às características. Logo, há 402 marcadores desnecessários à predição.

4- Qualidade dos dados

Para informações genotípicas, consideramos como critério de qualidade a quantidade de distorção de segregação mendeliana, a capacidade de reconstrução de mapa genético (com recuperação do número de grupos de ligação, tamanho do grupo, distância e ordenamento de marcas) e mapa de grupo de ligação.

Para constatação da qualidade dos valores fenotípicos simulados, foi avaliado o padrão de distribuição fenotípica das seis características e recuperado o grau de herdabilidade de cada uma, dado pelo quadrado da correlação entre o valor fenotípico e genotípico simulados. Também foram analisadas as variâncias e médias das características.

RESULTADOS E DISCUSSÃO

1- Qualidade dos dados genotípicos

A segregação de 97,75% dos marcadores estava dentro da relação esperada (1:2:1), a 1% de significância pelo teste de qui-quadrado. Considerou-se testes individuais, sem nenhuma proteção do nível de significância conjunto para os testes múltiplos.

A alta frequência de marcadores segregando corretamente indica que os marcadores são adequados para a genotipagem. Os 2,25% de marcadores que apresentaram distorção de segregação foram o 798,801,803, 1175 e 1176. Apesar da distorção, eles foram mantidos nos bancos de dados para análises futuras.

WANG *et al* (2019) fizeram diferente. Eles usaram dados genômicos para prever a capacidade geral de combinação (CGC). Do total de 319.668 marcadores SNPs em linhagens de milho, somente 61.468 marcadores foram usados posteriormente para as análises, após serem aplicados filtros.

É importante analisar a ocorrência de distorções, porque elas podem impactar resultados de análise genômica e estudos de predição.

Nas análises genômicas, apesar dos testes serem robustos, a existência de ligação fatorial é medida do desvio do esperado da segregação de dois locos, supostamente independentes e que segregam segundo as proporções mendelianas. Caso estes pressupostos sejam alterados, a estimativa da distância está comprometida.

Nas análises de predição, estas distorções podem também comprometer os resultados. Assim, referenciando, por exemplo, técnicas de árvores de decisão, pode-se ter um desvio da quantidade de exemplos fenotípicos associados às classes genotípicas, que, numa população, espera-se ser de 25% de exemplos para AA, 50% de exemplos para Aa e 25% de exemplos para aa.

Apesar das distorções observadas, a qualidade dos dados no contexto de mapeamento genético mostrou ser adequada, ou seja, na análise de mapeamento foram recuperadas satisfatoriamente as informações genômicas relativas ao número e tamanho de grupo de ligação e as relativas a ordenamento e distanciamento entre pares de marcas. Até mesmo os marcadores com distorção de segregação foram posicionados e ordenados adequadamente.

Uma forma alternativa de avaliar as associações entre pares de marcadores é por meio das estatísticas r^2 e D' (SANTOS, 2020) que medem o desequilíbrio gamético entre pares de marcadores que, numa população, é estabelecido pela ligação fatorial entre eles. Na Figura 2, obtida a partir da estatística r^2, verifica-se que foram recuperados padrões de desequilíbrio que refletem a organização dos marcadores nos 10 grupos de ligação fatorial simulados.

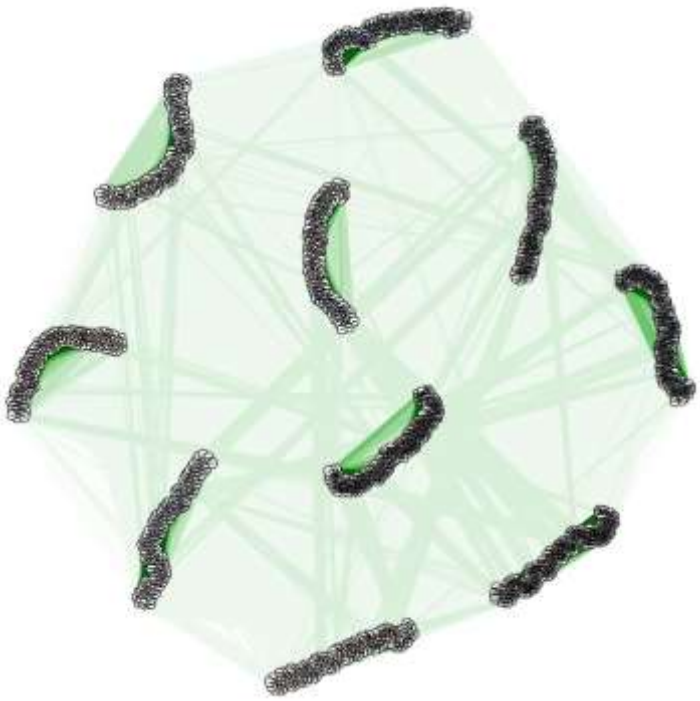


Figura 2. Mapa de desequilíbrio gamético, expressos pelas medidas r^2 refletindo 10 grupos de ligação no banco de dados.

2- Qualidade dos dados fenotípicos

Para o conjunto de informações fenotípicas simuladas, a qualidade dos dados foi constatada observando, principalmente, os aspectos relativos à distribuição e as herdabilidades.

Para as características básicas (x_1, x_3 e x_5), verificou-se que as médias foram iguais a 127,97 (Tabela 2). Estas médias são definidas como medidas de posição ou medidas de tendência central, no qual os dados tendem a se concentrar (REGAZZI, 2010).

As características com mesma média diferenciavam-se somente pelas herdabilidades, ou seja, foram influenciadas diferentemente pelo meio. Esses efeitos

ambientais eram cancelados ao se analisar a média, como se é esperado nas características quantitativas (CRUZ, 2012).

Uma das propriedades da média é ficar somada ou subtraída de uma constante, quando os valores de uma série são acrescentados ou subtraídos dessa constante. Evidenciou-se essa propriedade nas características x_2, x_4 e x_6.

Como abordado anteriormente, x_2, x_4 e x_6 possuíam quatro genes de maior efeito que juntos acrescentavam 120 (próximo ao máximo) unidades ao valor fenotípico ou acrescentavam 80 (próximo ao mínimo) unidades, aproximadamente na proporção de 81/256:175/256 dos indivíduos da população F_2 em relação as características x_1, x_3 e x_5.

As médias das três características abordadas representam o valor de 127,97 acrescentado dos valores anteriores, na proporção em que aparecem em F_2 (REGAZZI, 2010). A consequência foi a observância de valores médios para as características x_2, x_4 e x_6 mais elevados (220,21, conforme Tabela 2).

As variâncias das seis características foram diferentes (Tabela 2). As variâncias são medidas de dispersão que quantificam a variabilidade dos dados, sendo dadas pela soma dos quadrados dos desvios em relação à média aritmética, dividida pelo número de graus de liberdade.

Como as características eram controlados pelos mesmos locos gênicos e com mesmos efeitos, o aumento na variância era explicado pela influência dos efeitos ambientais. Assim, observou-se o padrão de quanto menor a herdabilidade, maior a variância, para características controladas pelo mesmo número de genes.

Tabela 2. Valores da média, variância, máximo, mínimo e herdabilidade de cada característica.

Característica	Média	Máximo	Mínimo	Variância	Herdabilidade
x_1	127,97	177,41	84,42	229,83	0,3921
x_3	127,97	165,32	93,83	153,75	0,6109
x_5	127,97	162,78	91,89	121,43	0,81556
x_2	220,21	296,03	168,47	491,49	0,7159
x_4	220,21	285,32	176,96	426,78	0,8598
x_6	220,21	277,39	179,82	396,18	0,9434

As herdabilidades estimadas no conjunto de dados, com base no quadrado da correlação entre os valores fenotípicos e genotípicos estimados, demonstram a eficiência da simulação em imitar a realidade.

Os valores foram muito próximos àqueles assumidos na simulação para x_1, x_3 e x_5 que eram 0,4, 0,6 e 0,8 (Tabela 2). As herdabilidades das características x_2, x_4 e x_6, obtidas pela inclusão dos efeitos de quatro genes de efeitos maiores, foram aumentadas.

Realmente, era esperado que as características controladas por quarenta e quatro tivessem maior herdabilidade, tendo em vista o acréscimo variável de valores (120 unidades em cerca de 81/256 indivíduos da população e de 80 unidades em cerca de 175/256 indivíduos da população).

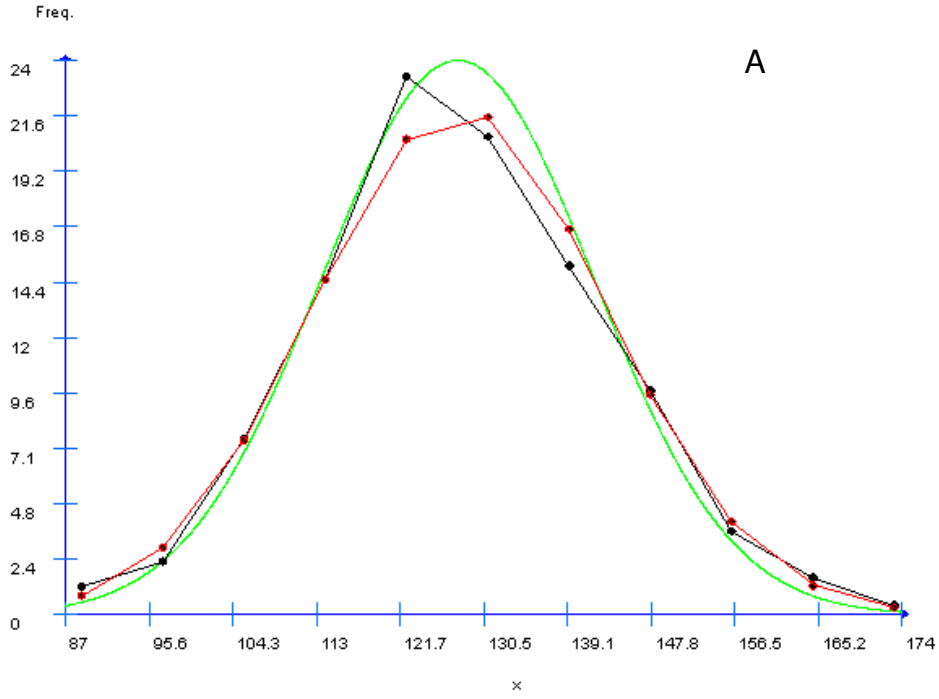
Pode-se imaginar que a presença destes locos de maior efeito possa aumentar a eficiência de predições em estudos futuros, pela ampliação da variabilidade genética e pela maior facilidade de estabelecimentos de classes de discriminação genotípica em regiões.

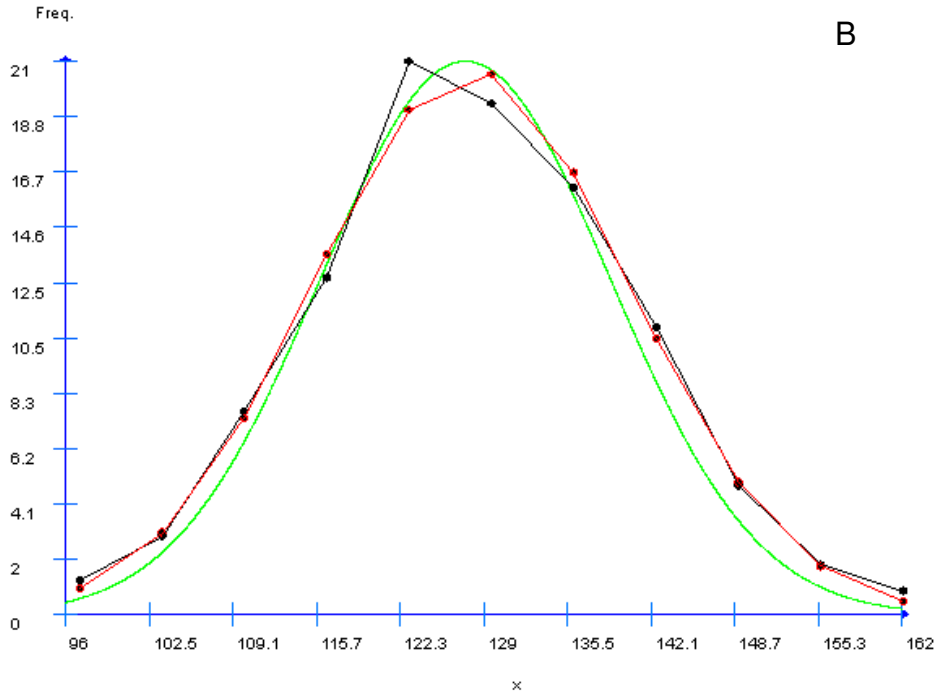
Em análises de predição, como a baseada em árvores de decisão e suas derivações, a maior discriminação pode facilitar a predição, pois estas técnicas se preocupam em dividir o espaço preditor em regiões de similaridade.

A evidência da qualidade dos dados, em termos de maior discriminação genotípica, também pode ser avaliada em relação a distribuição fenotípica. Observou-se padrão contínuo para as características x_1, x_3 e x_5 controladas por 40 genes de pequeno efeito, como mostrado na Figura 3.

Apesar dos dados terem sido simulados considerando a existência de efeitos de dominância (grau médio da dominância igual a um) e de epistasia, que contribuem para o aparecimento de padrão assimétrico de distribuição, não foi observado esse padrão.

A ação de poligenes que aumenta consideravelmente o número de classes genotípicas, os efeitos menores de locos individuais e a ação modificadora do ambiente conduzem a padrão típico de curva simétrica com distribuição típica de curva normal, amplamente relatada na literatura para variáveis quantitativas (FALCONER E MACKAY, 1996; KEARSEY E POONI, 1996).


Freq.
A
24
21.6
19.2
16.8
14.4
12
9.6
7.1
4.8
2.4
0
87
95.6
104.3
113
121.7
130.5
139.1
147.8
156.5
165.2
174
x


Freq.
B
21
18.8
16.7
14.6
12.5
10.5
8.3
6.2
4.1
2
0
96
102.5
109.1
115.7
122.3
129
135.5
142.1
148.7
155.3
162
x

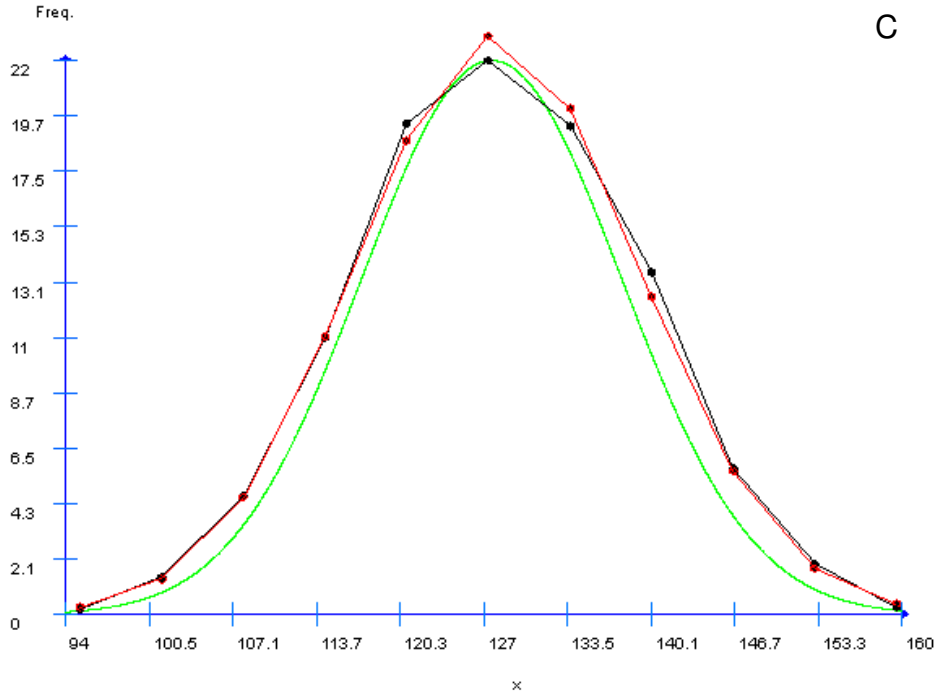
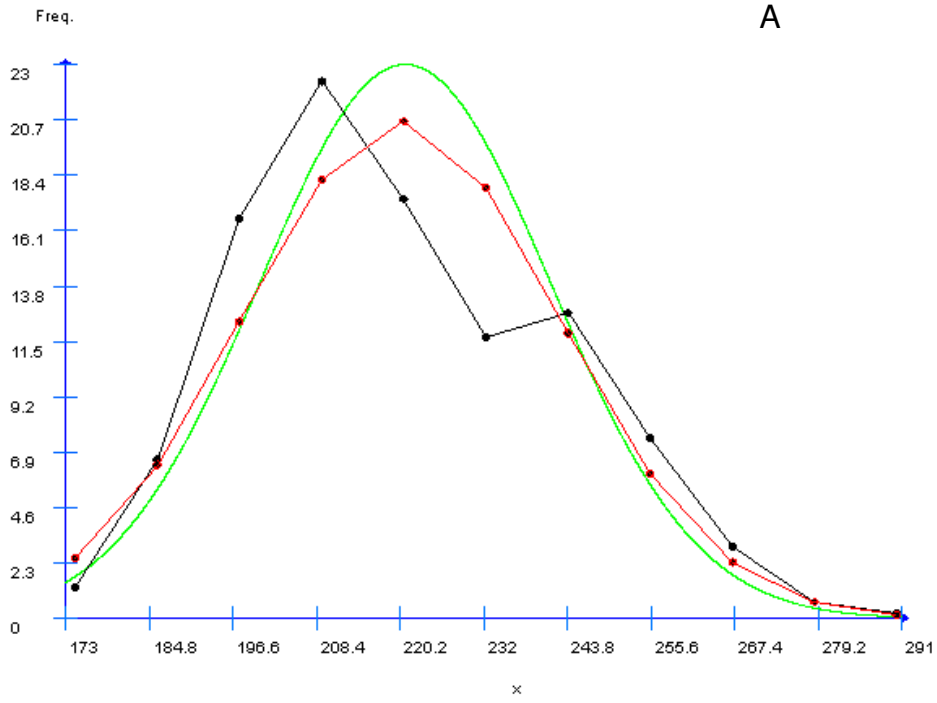


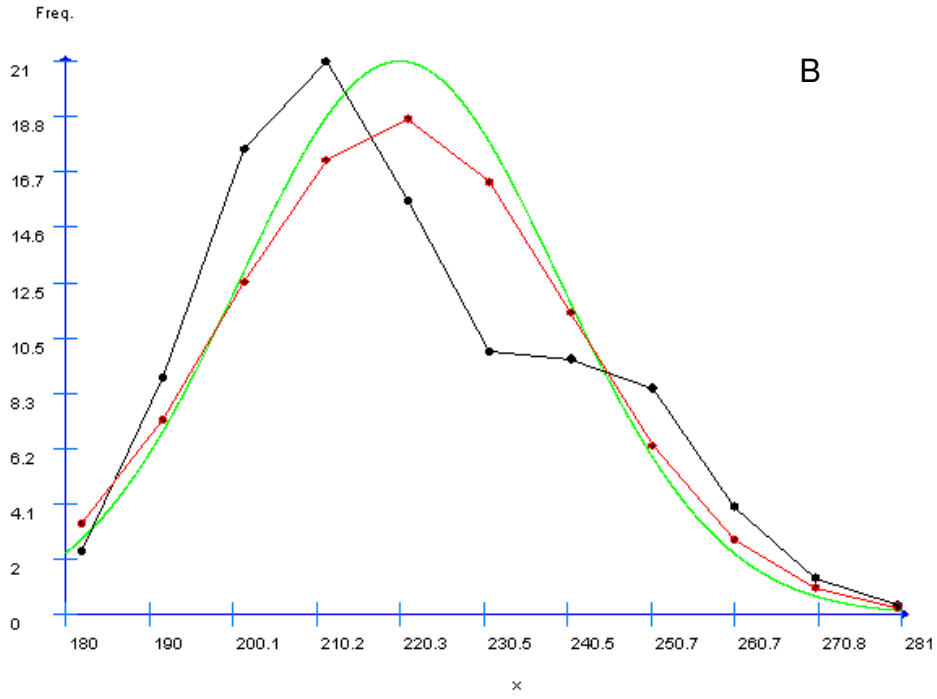
Figura 3. Distribuição dos valores mensurados em indivíduos F_2 em relação a x_1 (3A), x_3 (3B) e x_5 (3C). As linhas pretas representam as frequências observadas, as linhas vermelhas representam as frequências esperadas e as linhas verdes representam a distribuição normal.

Nas características poligênicas com ação adicional de genes de efeito maiores (x_2, x_4 e x_6), observou-se também padrão contínuo, mas com tendência de perda de simetria e formação de duas regiões modais, como pode ser observado na Figura 4.

Esse padrão com particularidade de grupos discretos aumentou junto com a herdabilidade. Por exemplo, para x_6, característica com maior herdabilidade, há duas modas nítidas na frequência observada.

As modas podem ser definidas como medidas de posição, que representam o valor de maior ocorrência dentro do conjunto de dados (REGAZZI, 2010). Essas modas devem-se aos genes de maior efeito, pois elas não ocorrem na característica x_5, com mesma herdabilidade para os quarenta poligenes.


A
Freq.
23
20.7
18.4
16.1
13.8
11.5
9.2
6.9
4.6
2.3
0
173
184.8
196.6
208.4
220.2
232
243.8
255.6
267.4
279.2
291
x


B
Freq.
21
18.8
16.7
14.6
12.5
10.5
8.3
6.2
4.1
2
0
180
190
200.1
210.2
220.3
230.5
240.5
250.7
260.7
270.8
281
x

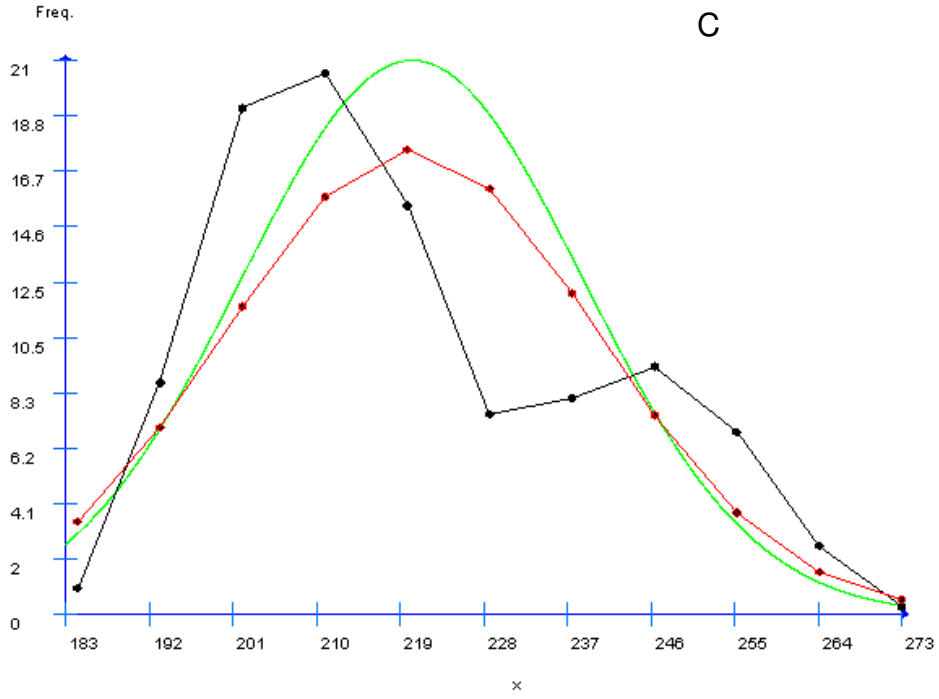


Figura 4. A figura 4A é o gráfico para a variável x_2, a 4B para variável x_4 e a 4C para a variável x_6. As linhas pretas representam as frequências observadas, as linhas vermelhas representam as frequências esperadas e as B s verdes representam a distribuição normal.

As características quantitativas são definidas como aquelas controladas por vários genes e bastante influenciada pelo meio, sendo a distribuição dos valores genotípicos das diferentes classes geralmente contínua (CRUZ, 2012).

Há características em que os efeitos dos genes são diferentes, havendo genes com efeitos maiores. Esses genes podem afetar a distribuição das classes fenotípicas, tornando-as próximas de padrão discreto, ou seja, caracteres quantitativos que exibem gene maior, cujo efeito pode ser avaliado em classes discretas, e, mesmo sob efeito do ambiente, podem ser estudados qualitativamente.

Técnicas de predição poderiam se comportar de forma diferente em situações ou em grupos de características que apresentassem maior ou menor quantidade de genes de efeitos maiores.

CONCLUSÃO

Os dados simulados podem ser eficientemente usados nas análises preditivas, pois preservam as características genéticas da população (padrão de segregação), do genoma e dos marcadores genotipados. Também possuem qualidade em relação aos dados fenotípicos simulados, tendo em vista os valores simulados e observados da herdabilidade e o padrão de distribuição de dados com e sem ação de genes modificadores.

REFERÊNCIAS

Bhering, LL. Mapeamento genético em famílias simuladas de irmãos completes. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, 166p, 2008.

Chagas, FEO. Análises genômicas e biométricas para escolha de genitores e predição de híbridos não realizados, 2018.

Cruz, CD. Princípios de Genética Quantitativa. 2ª edição, capítulo 1. 2012.

Falconer, DS. Mackay, TF. Introduction to quantitative genetics, [S.I.: s.n.], 1996.

Gianola, D. Okut, H. Weigel, KA. Rosa, GJM. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics. 2011

Hirschhorn, JN. Lohmueller, K. Byrne, E. Hirschhorn, K. A comprehensive review of genetic association studies. Genet Med 4:45–61. 2002.

Kearsey, MG. Pooni, HS. The genetical analysis of quantitative traits. London: Chapman & Hall; 1996.

Lee, Sh. Van Der Werf, JHJ. Hayes, BJ. Goddard, Me. Visscher, PM. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet 4(10):e1000231. 2008.

Long, N. Gianola, D. Rosa, GJ. Weigel, KA. Marker-assisted prediction of non-additive genetic values. Genetica. 2011.

Long, N. Gianola, D. Rosa, GJ. Weigel, KA. Kranis, A. Gonzalez-Recio, O. Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics research. 2010

Mendel, G. Experiments in plant hybridization. Cambridge, MA: Harvard University Press, 1965.

Meuwissen, THE. Hayes, BJ. Goddard, ME. Prediction of total genetic value using genome-wide dense marker maps. 2001.

Oliveira, GF. Nascimento, ACC. Nascimento, Moysés. Sant'anna, I de C. Romero, JV. Azevedo, CF. Bhering, LL. Moura, ETC. Quantile regression in genomic selection for oligogenic traits in autogamous plants: A simulation study. 2021.

Regazzi, AJ. Curso de Iniciação à estatística. Roteiro de aulas da disciplina EST 105. 2010.

Santos, IG de. Oliveira, M da S. Silva, MJ da. Cruz, CD. Genética de Populações com o aplicativo GPOP. Capítulo 7: Ligação e desequilíbrio de ligação. 2020.

Silva Júnior, AC da. Silva, MJ da. Cruz, CD. Souza, PRE de. Silva, HAD da. Leite, FCB. Maia, M de MD. Garcia, ACL. Montes, MA. Genética Geral para Universitários. 2015.

Silva Júnior, AC da. Sant'anna, I de C. Silva, GN. Nascimento, M. Azevedo, CF. Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. 2021.

Whang, X. Zhang, Z. Xu, Y. Li, P. Zhang, X. Xu, Chenwu. Using genomic data to improve the estimation of general combining ability based on sparse partial diallel cross designs in maize. 2019.

CAPÍTULO 2

RR-BLUP E APRENDIZADO DE MÁQUINA PARA ESTUDOS DE PREDIÇÃO DE CARACTERÍSTICAS POLIGÊNICAS COM OU SEM O EFEITO ADICIONAL DE GENES DE MAIOR EFEITO

RESUMO

DE MORAES, Francyse Edite de Oliveira Chagas, D.Sc., Universidade Federal de Viçosa, julho de 2022. **RR-BLUP e aprendizado de máquinas para estudos de predição de características poligênicas com ou sem o efeito adicional de genes de maior efeito.** Orientador: Cosme Damião Cruz.

O presente estudo avaliou o impacto do uso de diferentes conjuntos de marcadores sobre a eficiência da predição utilizando as técnicas RR-BLUP, árvore de decisão, *bagging, boosting* e *random forest*. As técnicas foram analisadas em relação a seis características. As características foram controladas pelos mesmos quarenta genes com diferentes herdabilidades (0,4, 0,6 e 0,8) acrescidos, ou não, por quatro genes com efeitos maiores de herdabilidade igual a um. O grau médio de dominância adotado foi igual a um para todas as características. Dentro de cada gene havia um marcador. Os genes controladores de efeito menor estavam distribuídos equitativamente nos oito primeiros grupos de ligação (GL) e os quatro de efeito maior estavam nos quatro primeiros GL. Como foram simulados dez grupos de ligação com 201 marcas codominantes em cada, havia 1608 marcas diretamente ou indiretamente relacionadas aos genes e 402 marcas desnecessárias a predição. A formação dos conjuntos de marcadores levou essas informações como critério. No grupo um, estavam todos os marcadores. No grupo dois, os 1608 marcadores direta ou indiretamente relacionados aos genes. No grupo três, os quarenta e quatro marcadores dentro dos genes e os 402 marcadores não relacionados. No grupo quatro, os 402 marcadores desnecessários a predição. No grupo cinco, os quarenta e quatro marcadores diretamente relacionados aos genes controladores. A eficiência da predição foi avaliada pela capacidade preditiva (r^2) e a raiz do erro quadrado médio (REQM) das diferentes técnicas nos cinco diferentes conjuntos de marcadores. Ao se analisar o r^2 e REQM das técnicas, observou-se que a maioria delas promoveu resultados péssimos na situação quatro. A técnica árvore de decisão chegou a não obter os valores em algumas repetições. Como nessa situação não havia marcadores relacionados as características, era esperado que em nenhuma técnica fosse possível obter resultados. A explicação veio pelo RR-BLUP. Ele forneceu o efeito dos marcadores sobre as características. Foram encontrados efeitos falsos positivos relacionados às 402 marcas desnecessárias a predição. Continuando-se as análises, foi observado que as técnicas *bagging* e *boosting* obtiveram os maiores valores de r^2

entre todas as técnicas (0,880 e 0,815, respectivamente) e os menores valores de REQM (5,852 e 5,853). A maioria dos valores estimados foi obtida do quinto conjunto de dados e, ou não diferiu significativamente dos outros conjuntos, ou foi diferente apenas do conjunto quatro (sem marcadores relacionados). Resultado diferente foi observado para a *random forest*. Ela foi a mais suscetível, tanto aos difentes subconjuntos de marcadores quanto às diferentes características. Para o quinto conjunto de marcadores, obteve r^2 para as características x_3, x_4, x_5 e x_6, respectivamente iguais a 0,371; 0,720; 0,514 e 0,788. Para REQM, obteve, naquele mesmo conjunto, em x_3 e x_5, respectivamente, 10,289 e 8,371. Esses valores foram os melhores e diferentes significativamente dos obtidos para as mesmas características nos outros quatro conjuntos. Os resultados obtidos mostram que o uso de diferentes técnicas explora melhor o conjunto de dados. Também mostra que o descarte de marcadores desnecessários não prejudica o processo preditivo, algumas vezes até o melhora, sendo recomendável. Trabalhos futuros deveriam se concentrar na identificação dos marcadores diretamente envolvidos com as características.

Palavras-chave: Capacidade preditiva. Raiz do erro quadrado médio. Conjunto de marcadores. Herdabilidade. Grau médio de dominância.

ABSTRACT

DE MORAES, Francyse Edite de Oliveira Chagas, D.Sc., Universidade Federal de Viçosa, July 2022. **RR-BLUP and machine learning for predictive studies of polygenic traits with or without the additional effect of higher effect genes.** Advisor: Cosme Damião Cruz.

The present study evaluated the impact of the use of different sets of markers on the prediction efficiency using the RR-BLUP, decision tree, bagging, boosting and random forest techniques. The techniques were analyzed in relation to six characteristics. The traits were controlled by the same forty genes with different heritability (0.4, 0.6 and 0.8) plus, or not, by four genes with greater heritability effects equal to one. The average degree of dominance adopted was equal to one for all characteristics. Within each gene was a marker. The minor-effect controller genes were evenly distributed in the first eight linkage groups (GL) and the four major-effect genes were in the first four GL. As ten linkage groups were simulated with 201 codominant markers in each, there were 1608 markers directly or indirectly related to genes and 402 markers unnecessary for prediction. The formation of the marker sets took this information as a criterion. In group one, there were all the markers. In group two, the 1608 markers directly or indirectly related to the genes. In group three, the forty-four markers within genes and the 402 unrelated markers. In group four, the 402 markers unnecessary the prediction. In group five, the forty-four markers were directly related to the controlling genes. The prediction efficiency was evaluated by the predictive capacity (r^2) and the root mean square error (REQM) of the different techniques in the five different sets of markers. When analyzing the r^2 and REQM of the techniques, it was observed that most of them promoted poor results in situation four. The decision tree technique did not obtain the values in some repetitions. As in this situation there were no markers related to the characteristics, it was expected that in no technique it would be possible to obtain results. The explanation came from RR-BLUP. It provided the effect of markers on traits. False positive effects were found related to the 402 unnecessary marks for prediction. Continuing the analysis, it was observed that the bagging and boosting techniques obtained the highest values of r^2 among all the techniques (0.880 and 0.815, respectively) and the lowest values of REQM (5.852 and 5.853). Most of the estimated values were obtained from the fifth dataset and either did not differ significantly from the other sets or differed only from set four (no related markers).

Different result was observed for random forest. She was the most susceptible, both to different subsets of markers and to different characteristics. For the fifth set of markers, r^2 was obtained for the characteristics x_3, x_4, x_5 and x_6, respectively equal to 0.371; 0.720; 0.514 and 0.788. For REQM, he obtained, in that same set, at x_3 and x_5, respectively, 10.289 and 8.371. These values were the best and significantly different from those obtained for the same characteristics in the other four sets. The results obtained show that the use of different techniques better explores the dataset. It also shows that discarding unnecessary markers does not harm the predictive process, sometimes even improves it, which is recommended. Future work should focus on identifying the markers directly involved with the traits.

Keywords: Predictive ability. Root mean square error. Set of markers. Heritability. Mean degree of dominance.

INTRODUÇÃO

Antes dos avanços nos estudos sobre o material genético, a manipulação dos organismos consistia exclusivamente na modificação dos fenótipos e por meio de experimentos, ocorria a seleção dos que carregavam a melhor informação genética. Atualmente, pode-se contar com as técnicas preditivas, utilizando como critério de seleção as informações fenotípicas e moleculares.

O uso das informações moleculares, junto com o uso de duplos haploides, é considerado um dos maiores impactos no melhoramento de plantas que ocorreu nos últimos tempos (CHAKRADHAR *et al.*,2017).

Há várias metodologias que se utilizam das ferramentas mencionadas, como o mapeamento associativo e a seleção genômica ampla (GWS).

A GWS realiza uma seleção assistida por marcadores sem a identificação de marcas com efeitos significantes (MASSMAN, JUNG & BERNARDO, 2012). Ela pode ser aplicada em técnicas baseadas em BLUP e no aprendizado de máquinas, como por exemplo, o RR-BLUP e as árvores de decisão (BEUCHER *et al.*, 2019; PARMLEY *et al.*, 2019) e seus refinamentos, como *bagging*, *random forest* e *boosting* (DEGENHARDT *et al.*, 2019; SILVA JUNIOR *et al.*, 2021).

A seleção genômica enfrenta o problema de estabelecer modelos preditivos que considerem abordagem biométrica apropriada e que levem em consideração a complexidade genética das características, minimizando os efeitos perturbadores da seleção, em especial, o ambiente. É conhecido por outros trabalhos que a sua eficácia é diretamente influenciada pela herdabilidade e grau médio de dominância, sendo interessante avaliar esses fatores aplicados em características poligênicas com ou sem o efeito adicional de genes de maior efeito (CROSSA *et al.*, 2017).

Questões relativas à possibilidade da redução de dimensionalidade podem ser tratadas para nortear estudos futuros com bancos de dados de menor dimensão, evitando-se problemas de dimensionalidade e multicolinearidade que comumente ocorrem em situações em que o número de informações moleculares superam, em muito, o número de indivíduos genotipados e fenotipados.

Diante do exposto, analisar o comportamento da GWS frente à diferentes técnicas de predição e conjuntos de marcadores, é interessante para os programas de melhoramento. Essa análise demonstraria que os resultados obtidos com

informações reduzidas podem melhorar a predição ou preservar as mesmas conclusões biológicas quando se utiliza maiores conjuntos de dados (SILVA, 2021).

Assim, neste estudo será realizado a seleção genômica ampla a partir do emprego de diferentes técnicas preditivas para determinar o valor genético genômico (VGG) dos indivíduos, tais como a metodologia RR-BLUP e procedimento de aprendizado de máquina como árvore de decisão, *bagging*, *boosting* e *random forest,* considerando-se diferentes características e conjuntos de marcadores.

MATERIAIS E MÉTODOS

1- Conjunto de dados originais

Neste estudo será utilizado o mesmo conjunto de dados descritos no Capítulo 1. Trata-se do estudo de uma população F_2, obtida por simulação, representada por 1000 indivíduos, em que 2010 marcadores estão distribuídos equitativamente em 10 grupos de ligação.

Seis características foram simuladas, manifestando efeito de dominância completa e herdabilidade variada. Dentre as seis características, três (x_1, x_3 e x_5) são controladas por 40 genes de pequeno efeito e com herdabilidades iguais a 0,4, 0,6 e 0,8, respectivamente. As demais características (x_2, x_4 e x_6), além dos 40 genes de pequeno efeito, na mesma ordem de herdabilidade apresentada por x_1, x_3 e x_5, possuem quatro outros genes de efeitos maiores para o controle. Os quatros genes possuem herdabilidade igual a um e todos os genes possuem grau médio de dominância um.

2- Partição do conjunto de dados

Para estabelecimento de modelos preditivos, por diferentes abordagens biométricas, o conjunto de dados foi particionado, estabelecendo conjuntos para fins de treinamento, ou de ajuste propriamente dito, e de validação permitindo o emprego da técnica de validação cruzada considerando cinco partições ($k = 5$).

Assim, a população foi subdividida em cinco subconjuntos mutuamente exclusivos, sendo que a cada rodada, quatro desses subconjuntos constituíram a população de treinamento (80% dos indivíduos) e o subconjunto restante constituirá a população de validação (20% dos indivíduos). A forma de divisão do arquivo está representada na Tabela 1 a seguir:

Tabela 1. Distribuição dos indivíduos F_2 nos cinco arquivos de validação e treinamento.

	Validação	**Treinamento**
Partição	Indivíduos envolvidos	Indivíduos envolvidos
1	1 até 200	201 até 1000
2	201 até 400	1 até 200 e 401 até 1000
3	401 até 600	1 até 400 e 601 até 1000
4	601 até 800	1 até 600 e 801 até 1000
5	801 até 1000	1 até 800

3- Metodologias para predição das características

a- Random Regression Best Linear Unbiased Predictor (RR-BLUP)

O RR-BLUP foi usado para predizer o valor genético estimado dos 1000 indivíduos. O modelo ajustado abaixo foi usado para estimar o efeito dos marcadores:

$$y = Xb + Zm + \varepsilon$$

Em que:

y é o vetor de observações fenotípicas;

b é o vetor de efeitos fixos (média geral) com matriz de incidência X, composta pelos valores 1, 0 e -1 de acordo com o número de alelos marcadores dos genótipos MM, Mm e mm, respectivamente;

m é o vetor de efeitos aleatórios de marcadores;

Z é a matriz de incidência para o vetor dos efeitos aleatórios de marcadores. Contém os valores 0 e 1 para ausência ou presença do marcador;

e ε é o vetor de erros aleatórios.

As equações de modelos mistos para o modelo completo foram dadas por (RESENDE *et al.* ,2014):

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I\dfrac{\sigma_e^2}{\sigma_{gi}^2} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

O valor genético estimado (GEBV) do indivíduo i foi definido como:

$$\widehat{GEBV_i} = \widehat{y_i} = \sum_{i=1}^{N} Z_i \widehat{m}_i$$

b- Aprendizado de máquinas

Para estabelecer os modelos de predição das características, também foram usadas as abordagens baseadas nas árvores de decisão e seus refinamentos (*random forest*, *bagging,* and *boosting)*.

Árvore de Decisão (AD)

Baseia-se na partição do espaço do preditor de acordo com regras simples, identificando regiões com as respostas mais homogêneas aos preditores e ajustando a resposta média para observações nessa região.

O espaço de decisão é estimado pela divisão recursiva dos dados em cada nó com base em um teste estatístico que aumenta a homogeneidade dos dados de treinamento nos nós descendentes resultantes (BRODLEY e FRIED, 1997). A cada nó é atribuída uma descrição de classe e cada ramo refere-se a uma regra de decisão, ou seja, uma condição relacionada aos recursos do conjunto de dados de entrada e que descreve o caso em que cada ramo é escolhido (HASTIE *et al.*, 2009).

A estrutura da árvore de regressão foi criada a partir da busca pela árvore que levaria à partição dos dados até que fosse obtida a formação dos grupos homogêneos.

Para realizar a divisão binária recursiva, primeiro selecionamos o preditor X_j e o ponto de corte s de modo que dividindo o espaço preditor nas regiões $\{x|x_j < s\}$ e $\{x|x_j \geq s\}$ leva à maior redução possível na soma de quadrados do resíduo (RSS). Ou seja, consideramos todos os preditores $x_1, \ldots, x_p$ e todos os valores possíveis do ponto de corte s para cada um dos preditores e, em seguida, escolhe o preditor e o ponto de corte de modo que a árvore resultante tenha o menor RSS.O ponto de corte escolhido resultou em 6 nós após a poda. A equação que reflete a divisão binária é (JAMES *et al.*, 2021):

$$R_1(j,s) = \{X|X_j < s\} \, e \, R_2(j,s) = \{X|X_j \geq s\},$$

e então procuramos o valor de J e S que minimizam a equação:

$$\sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$$

em que:

$\hat{y}_{R_1}$ é a média da variável resposta das observações de treinamento pertencente a região $R_1(j, s)$,

$\hat{y}_{R_2}$ é a média da variável resposta das observações de treinamento pertencente a região $R_2(j, s)$

e y_i é o valor verdadeiro da característica de cada indivíduo.

Bagging (BA)

O refinamento *bagging* cria vários conjuntos de dados semelhantes por reamostragem (*bootstrapping*), para obter uma média das várias árvores de regressão que são realizadas sem poda para cada conjunto de dados (BREIMAN, 1996; PRASAD *et al.*, 2006).

Assim, são obtidos um número B de modelos: $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$. Esses modelos gerados são usados para obter um modelo médio, dado por: $\hat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x)$.

O número de árvores amostradas para o *bagging* foi fixado em 500 árvores.

Random Forest (RF)

As amostras de *bootstrap* são utilizadas para construir várias árvores e cada uma é estabelecida com um subconjunto aleatório de preditores. O número de preditores (m) usados para determinar a melhor divisão em cada nó é escolhido por $m = \frac{v}{3}$, sendo v o número total de preditores. Como foi feita a escolha de marcadores, o 1/3 do número de preditores foi diferente em cada situação. O número de árvores para a RF foi fixado em 500. Para essa metodologia, as árvores crescem até o tamanho máximo sem poda, e a agregação é feita pela média das árvores (COSTA *et al.*, 2021).

Boosting (BO)

O refinamento *boosting* cria árvores sequencialmente utilizando informações das árvores anteriores (JAMES *et al.*, 2021). Essa é uma abordagem treinada repetidamente na mesma amostra para que a cada iteração, uma medida de erro de

previsão seja calculada para cada SNP, e, na próxima iteração, SNPs que proporcionaram maiores erros, recebam maior peso no treinamento do modelo.

A técnica utilizada visa a otimização numérica para minimizar a função de perda adicionando, a cada passo, uma nova árvore que melhor reduz a função de perda (GHAFOURI-KESBI *et al.*, 2017).

A previsão é realizada ponderando os resultados do conjunto de todas as árvores de regressão.

O número de árvores amostradas foi de 500, com uma taxa de aprendizado de 0.01 e profundidade igual a 2. O seguinte modelo foi utilizado para ajustar o BO (HASTIE *et al.,* 2009):

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m)$$

Onde:

β_m, m = 1, 2, ..., M são os coeficientes de expansão da base e $b(x; \gamma_m)$ são funções simples do argumento multivariado x, com um conjunto de parâmetros $\gamma = \gamma_1, \gamma_2, \ldots, \gamma_m$.

As análises foram feitas utilizando o software GENES em integração com o R (CRUZ, 2016).

4- Avaliação de cenários que incluem, ou não, marcadores relacionados as características

Dentro de cada gene controlador da característica havia um marcador. Os genes controladores de efeito menor estavam distribuídos equitativamente nos oito primeiros grupos de ligação (GL) e os quatro de efeito maior estavam nos quatro primeiros GL.

Como foram simulados dez grupos de ligação com 201 marcas codominantes, em cada uma havia 1608 marcas direta ou indiretamente relacionadas aos genes e 402 marcas desnecessárias a predição. A formação dos cenários foi realizada com base nesse critério:

Cenário 1: 2010 marcadores dos dez grupos de ligação foram usados para testar a eficiência das técnicas;

Cenário 2: 1608 marcadores direta ou indiretamente relacionados aos genes;

Cenário 3: quarenta e quatro marcadores localizados no interior dos genes e os 402 marcadores dos dois últimos GL;

Cenário 4: 402 marcadores desnecessários à predição que estão nos GL 9 e 10;

Cenário 5: quarenta e quatro marcadores diretamente relacionados aos genes controladores.

A análise dos parâmetros possibilitou identificar a importância de cada preditor (marcador molecular) no modelo estimado. A comparação da importância prévia e a fornecida pela técnica de predição apontaram procedimentos mais eficazes para seleção prévia de marcadores em procedimentos de predição.

5- Comparação da eficiência da predição

Para estabelecer a eficiência do modelo, foi utilizado os parâmetros de confiabilidade ou acurácia seletiva (r^2) e raiz do erro quadrático (REQM).

A confiabilidade é uma medida do quadrado da correlação entre os valores estimados ($\hat{y}$) e os valores verdadeiros (y), ou seja, mede o quanto a estimativa obtida é relacionada com o valor real do parâmetro que, em genética quantitativa, expressam a herdabilidade da característica (CRUZ *et al.*, 2012).

$$r^2 = \left(cor(\hat{y}, y)\right)^2$$

A raiz do erro quadrático médio (REQM) foi adotada para expressar a acurácia preditiva dos modelos, uma vez que apresenta as estimativas do erro na mesma escala da variável de interesse.

$$REQM = \sqrt{\frac{\sum(\hat{y} - y)^2}{n}}$$

Onde n representa o número de observações.

SILVA (2021) analisou a influência de diferentes conjuntos de marcadores e características sobre a eficiência da predição com os mesmos parâmetros. Por outro lado, SILVA JUNIOR (2021) e CHAGAS (2018) utilizaram, respectivamente, o coeficiente de determinação e a correlação de Pearson para medir a eficiência dos seus modelos preditivos.

6- Teste estatístico

Cada uma das cinco partições (k = 5) foram consideradas como repetições. Procedeu-se à análise de variância dos dados, obtendo-se os quadrados médios dos fatores pelo teste F.

As médias do quadrado da correlação de validação (r^2) e da raiz do erro quadrático médio de validação (REQM) resultantes de cada estudo foram submetidas ao teste de Tukey, a 5% de probabilidade.

A análise de variância serve para verificar se há alguma diferença significativa entre as médias dos níveis de um fator no nível de significância estipulado. Sendo não significativa, não há diferença entre todos os possíveis contrastes entre médias. Quando significativa, há pelo menos um contraste entre médias estatisticamente diferente de zero.

A diferença mínima significativa (dms) do teste Tukey é dada por Δ e pode ser calculada da maneira a seguir (RESENDE, 2007):

$$\Delta = q\sqrt{\frac{QMRes}{k}}$$

Em que:

q = $q_{\alpha}(l,n_2)$ é o valor tabelado da amplitude total estudentizada, que é obtido em função do nível α de significância do teste, número de níveis do fator em estudo (l) e número de graus de liberdade do resíduo (n_2) da análise de variância;

QMRes = quadrado médio do resíduo obtido na análise de variância;

K = número de repetições dos tratamentos.

RESULTADOS E DISCUSSÃO

1- Predição por diferentes abordagens

O r^2 de todas as metodologias é mostrado na Tabela 2. Baseado nessa tabela, pode-se definir e comparar as variáveis e técnicas que forneceram as mais eficientes predições.

Tabela 2. Capacidade preditiva (r^2) de todas as metodologias em relação as características analisadas.

	AM				BLUP
V	AD	RF	BA	BO	RR-BLUP
x_1	0,201 B	0,199 B	0,589 A	0,508 AB	0,267 B
x_2	0,548 B	0,570 B	0,771 A	0,641 A	0,224 C
x_3	0,247 C	0,320 BC	0,642 A	0,592 A	0,389 ABC
x_4	0,656 B	0,699 AB	0,839 A	0,722 AB	0,277 C
x_5	0,267 C	0,449 BC	0,717 A	0,689 A	0,503 A
x_6	0,703 B	0,754 AB	0,867 A	0,771 AB	0,301 C

Médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. AM: Aprendizado de máquinas; BLUP: *Best Linear Unbiased Prediction*; AD: Árvore de decisão; RF: *Random forest*; BA: *Bagging*; BO: *Boosting*; RR-BLUP: *Ridge Regression-Best Linear Unbiased Prediction*.

As técnicas árvore de decisão e RR-BLUP não se destacaram em relação as suas capacidades preditivas ou suas percepções a todas as diferenças entre as características. Os maiores valores encontrados para as técnicas foram, respectivamente, 0,703 e 0,503.

O RR-BLUP desconsiderou os efeitos não aditivos presentes nas características, sendo abordado na literatura que ignorar esses efeitos traz consequências como estimação superestimada da variância aditiva (PALUCCI *et al.*, 2007), falha ao detectar QTLs, resultados replicados de estudos de associação

(HIRSCHHORN *et al.*, 2002), perda da acurácia preditiva dos fenótipos (LEE *et al.*, 2008), dentre outros problemas.

A árvore de decisão se baseia em somente um resultado e geralmente, uma única árvore não possui boa capacidade preditiva quando comparada com outras abordagens (SOUSA *et al.*, 2020). Essa metodologia sofre alta variação em termos de predição em comparação a seus refinamentos (JAMES *et al.*, 2013).

Os refinamentos da árvore de decisão mostrados a seguir levaram aos melhores resultados. O uso dos refinamentos aumentou a precisão preditiva da árvore, pois combinaram múltiplas árvores para reduzir a variabilidade (BREIMAN, 1996; SOUSA *et al.*, 2020).

As técnicas *bagging* e *boosting* obtiveram os maiores valores de r^2 (0,867 e 0,771, respectivamente) ou estavam entre as três técnicas com os maiores valores. A capacidade preditiva entre as variáveis não mudou bruscamente, apesar de todas as diferenças simuladas entre elas.

SILVA JUNIOR *et al*. (2021) demonstraram que o bagging e *boosting* forneceram as melhores estimativas de R^2. Esses autores encontraram estimativas superiores a 80% usando a cultura de arroz para a variável rendimento de grãos. Além disso, esses autores argumentaram que a *random forest* não obteve os melhores resultados no experimento.

Assim como no trabalho anteriormente mencionado, a *random forest* não obteve os maiores resultados, chegando a se comparar a árvore de decisão. Entretanto, ela surpreendeu em relação a sua resposta as diferenças entre as variáveis, obtendo melhor acurácia em relação as características com maior número de genes e herdabilidade.

Em ordem crescente de herdabilidade, obteve-se nas características controladas por quarenta genes r^2 de 0,199, 0,320 e 0, 449, respectivamente. Nas características controladas por quarenta e quatro genes, novamente em ordem crescente de herdabilidade, obteve r^2 de 0,570, 0,699 e 0,754.

A técnica captou a presença dos quatro locos de maior efeito, aumentando a eficiência da predição. Eles foram usados para facilitar o estabelecimento de classes, uma vez que a técnica abordada se preocupa em dividir o espaço preditor em regiões de similaridade (MITCHELL, 1997).

Em relação a herdabilidades, pode-se notar que nas características com mesmo número de genes, a menor herdabilidade sempre esteve associada a menor

capacidade preditiva. No trabalho de SANT'ANNA (2021), as consequências negativas para a predição do maior efeito ambiental também foram observadas. Nas características controladas por menores herdabilidade, as técnicas preditivas também obtiveram os menores resultados preditivos.

A raiz do erro quadrado médio de todas as metodologias é mostrada na Tabela 3. Baseado nessa tabela, pode-se novamente comparar as características e técnicas que forneceram as mais eficientes predições.

Tabela 3. Raiz do erro quadrado médio (REQM) de todas as metodologias em relação as características analisadas.

	AM				**BLUP**
V	AD	RF	BA	BO	RR-BLUP
x_1	14,054 A	13,736 A	9,566 B	10,446 AB	12,986 AB
x_2	14,952 B	14,653 B	9,978 C	12,842 BC	19,687 A
x_3	11,116 A	10,545 A	7,374 B	7,695 B	9,674 AB
x_4	12,177 B	11,496 B	7,804 C	10,523 BC	17,638 A
x_5	9,669 A	8,729 A	6,074 B	5,853 B	7,774 AB
x_6	10,848 B	9,981 B	6,757 C	9,200 BC	16,762 A

As médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. AM: Aprendizado de máquinas; BLUP: *Best Linear Unbiased Prediction*; AD: Árvore de decisão; RF: *Random forest*; BA: *Bagging*; BO: *Boosting*; RR-BLUP: *Ridge Regression-Best Linear Unbiased Prediction*.

As técnicas se comportaram de maneira equivalente para essa medida e para a capacidade preditiva. No *bagging* e *boosting* foram obtidos os menores valores (5,853 e 6,074, respectivamente). Como demonstrado na Tabela 2, a *random forest* novamente teve qualidade intermediária, mas diferenciou bem as características.

A REQM diferenciou-se da capacidade preditiva por ter sido influenciada pela dimensão das características. Foi observado o maior erro para as características controladas por quarenta e quatro genes.

Os resultados encontrados foram adequados, tendo em vista o acréscimo variável de valores (120 unidades em cerca de 81/256 indivíduos da população e de

80 unidades em cerca de 175/256 indivíduos da população) nas características estudadas, influenciando diretamente a medida do erro.

2- Predição utilizando diferentes conjuntos de marcadores

Os diferentes conjuntos de marcadores foram, junto com a fenotipagem, utilizados para predição dos valores genéticos dos indivíduos.

A Tabela 4 mostra a capacidade preditiva das técnicas da árvore de decisão e RR-BLUP. Como mostrado nessa tabela, as técnicas novamente não se destacaram.

O cenário quatro (marcadores desnecessários a predição) de ambas as técnicas foi significativamente diferente da maioria dos demais. Nele, se observou os piores valores preditivos. No RR-BLUP, ele forneceu valor 0,007 e na árvore de decisão, esse cenário chegou a não fornecer resultados para algumas ou todas as validações cruzadas.

Os valores de r^2 foram iguais estatisticamente entre a maioria dos demais cenários de ambas as técnicas. O maior valor obtido para o RR-BLUP foi em x_5 (0,522) e para AD foi em x_6 (0,703).

GRANATO *et al*. (2013) obtiveram resultados diferentes. Esses autores analisaram a capacidade preditiva da GWS pelas técnicas RR-BLUP e BLASSO em diferentes situações de redução do número de marcadores. Quando retiram das análises marcadores desnecessários a predição, a capacidade preditiva foi aumentada.

CAVALCANTI (2012) também percebeu melhora da estimativa ao utilizar diferentes subconjuntos marcadores. Havia em seu trabalho inicialmente 238 marcadores. Ao utilizar diferentes conjuntos, a acurácia seletiva foi aumentando. Ao chegar a conjuntos menores que 70 marcadores, foi observado redução da medida.

Tabela 4. Capacidade preditiva (r^2) de todos os cenários em relação as características analisadas nas metodologias RR-BLUP e árvore de decisão.

	RR-BLUP					AD				
V	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
x_1	0,267 A	0,263 A	0,295 A	0,015 B	0,268 A	0,201 A	0,189 A	0,134 AB	0,024 B	0,121 AB
x_2	0,224 A	0,210 A	0,256 A	0,014 B	0,229 A	0,548 A	0,548 A	0,545 A	0,014 B	0,545 A
x_3	0,389 A	0,390 A	0,396 A	0,009 B	0,393 A	0,247 A	0,242 A	0,179 A	0,009*	0,174 A
x_4	0,277 A	0,271 A	0,293 A	0,010 B	0,262 A	0,656 A	0,656 A	0,652 A	0,006**	0,652 A
x_5	0,503 A	0,500 A	0,522 A	0,011 B	0,494 A	0,267 A	0,274 A	0,227 A	0,012***	0,223 A
x_6	0,301 A	0,292 A	0,324 A	0,007 B	0,289 A	0,703 A	0,703 A	0,703 A	-****	0,703 A

Médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. RR-BLUP: *Ridge Regression-Best Linear Unbiased Prediction*. AD: Árvore de decisão. C_1: Cenário 1 com os 2010 marcadores genotipados; C_2: Cenário 2 com os 1608 marcadores direta ou indiretamente relacionados às características; C3: Cenário 3 com os 44 marcadores diretamente relacionados às características e os 402 marcadores desnecessários a predição; C_4: Cenário 4 com os 402 marcadores desnecessários a predição; C_5: Cenário 5 com os 44 marcadores diretamente relacionados às características. *média de quatro valores, pois uma repetição não forneceu resultados; **média de um valor, pois quatro repetições não forneceram resultados; ***média de três valores, pois duas validações cruzadas não geraram resultados; ****nenhum valor foi obtido nas cinco validações cruzadas.

A Tabela 5 mostra a raiz do erro quadrado médio das técnicas árvore de decisão e RR-BLUP. As técnicas novamente se comportaram de maneira equivalente para essa medida e para a capacidade preditiva, exceto pela sensibilidade a dimensão das características.

Foram observados os maiores erros para as características controladas por quarenta e quatro genes e a ausência de algumas repetições na árvore de decisão para o quarto cenário. O menor valor no RR-BLUP e AD foi em x_5 (7,625 e 9,592, respectivamente).

Tabela 5. Raiz do erro quadrado médio (REQM) de todos os cenários em relação as características analisadas nas metodologias RR-BLUP e árvore de decisão.

	RR-BLUP					AD				
V	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**	**C_1**	**C_2**	**C_3**	**C_4**	**C_5**
x_1	**12,986B**	**13,011B**	**12,778B**	**14,999A**	**12,940B**	**14,054A**	**14,171A**	**14,484 A**	**5,145 A**	**14,473 A**
x_2	**19,687B**	**19,823B**	**19,313B**	**22,042A**	**19,514B**	**14,952B**	**14,952B**	**15,016 B**	**2,131A**	**15,016 B**
x_3	**9,674 B**	**9,660 B**	**9,645 B**	**12,333A**	**9,800 B**	**11,116A**	**11,136A**	**11,451 A**	**0,951***	**11,455 A**
x_4	**17,638B**	**17,696B**	**17,484B**	**20,575A**	**17,725B**	**12,177A**	**12,177A**	**12,243 A**	**1,004****	**12,243 A**
x_5	**7,774 B**	**7,795 B**	**7,625 B**	**10,970A**	**7,863 B**	**9,669 A**	**9,592 A**	**9,836 A**	**1,082*****	**9,855 A**
x_6	**16,762B**	**16,841B**	**16,509B**	**19,872A**	**16,814B**	**10,848A**	**10,848A**	**10,847 A**	**-******	**10,847 A**

As médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. RR-BLUP: *Ridge Regression-Best Linear Unbiased Prediction*. AD: Árvore de decisão. C_1: cenário 1 com os 2010 marcadores genotipados; C_2: cenário 2 com os 1608 marcadores direta ou indiretamente relacionados às características; C3: cenário 3 com os 44 marcadores diretamente relacionados às características e os 402 marcadores desnecessários a predição; C_4: cenário 4 com os 402 marcadores desnecessários a predição; C_5: cenário 5 com os 44 marcadores diretamente relacionados às características; *média de quatro valores, pois uma repetição não forneceu resultados; **média de um valor, pois quatro repetições não forneceram resultados; ***média de três valores, pois duas validações cruzadas não geraram resultados; ****nenhum valor foi obtido nas cinco validações cruzadas.

A técnica RR-BLUP, além de ser usada nos processos preditivos, foi útil em mostrar o efeito dos 2010 marcadores sobre as características. É mostrado na Figura 1 esse efeito nas características x_1 e x_6. O padrão observado se repetiu entre as características controladas pelo mesmo número de genes.

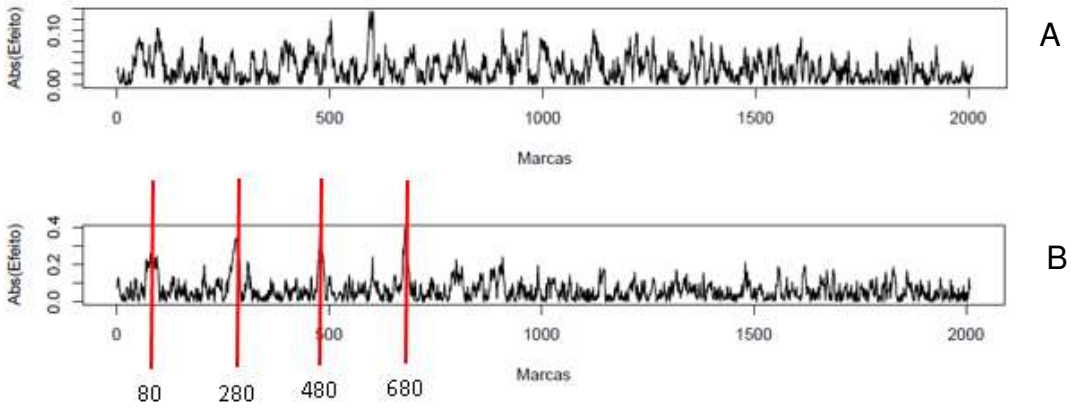


Figura 1. Na Figura 1A está o efeito dos 2010 marcadores para a característica x_1 e na Figura 1B para a característica x_6.

É possível notar que na característica x_1 há vários picos, associados aos quarenta marcadores internos aos QTL, ou a proximidade com esses marcadores. Na característica x_6, além dos picos anteriores, há quatro picos expressivos relacionados diretamente as marcas 80, 280, 480 e 680 e outros picos associados com a proximidade a essas marcas.

Há também picos entre a marca 1500 e 2000, no qual estão majoritariamente os 402 marcadores desnecessários aos processos preditivos. Esses picos representam, portanto, falsos positivos.

Na Tabela 6 é mostrado que o *bagging* e o *boosting* também obtiveram os maiores valores de capacidade preditiva entre as diferentes situações de escolha de marcadores (0,880 e 0,815 para x_6, respectivamente).

Para o *bagging*, a maioria dos cenários não se diferenciou nem do cenário quatro. Nessa técnica, se obteve os máximos valores de r^2 do cenário mencionado, sugerindo que a capacidade preditiva máxima obtida pelos efeitos falsos positivos seja em torno de 0,49.

Algumas variáveis do boosting não tiveram capacidade preditiva diferente entre as situações. As que se diferenciaram, tinham somente o cenário quatro como desigual.

Tabela 6. Capacidade preditiva (r^2) de todos os cenários em relação às características analisadas nas metodologias *bagging* e *boosting*.

	BA					BO				
V	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
x_1	0,589 A	0,588 A	0,594 A	0,488 A	0,600 A	0,508 A	0,496 A	0,486 A	0,282 A	0,453 A
x_2	0,771 A	0,771 A	0,770 A	0,490 A	0,778 A	0,641 A	0,633 A	0,636 A	0,285 B	0,640 A
x_3	0,643 A	0,647 A	0,644 A	0,489 A	0,659 A	0,592 A	0,587 A	0,569 A	0,289 A	0,561 A
x_4	0,839 A	0,842 A	0,837 A	0,487 B	0,847 A	0,722 A	0,727 A	0,725 A	0,283 B	0,741 A
x_5	0,717 A	0,717 A	0,717 A	0,487 A	0,738 A	0,689 A	0,692 A	0,698 A	0,288 B	0,704 A
x_6	0,867 A	0,870 A	0,868 A	0,487 B	0,880 A	0,771 A	0,775 A	0,788 A	0,286 B	0,815 A

Médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. BA: *bagging*; BO: *boosting*. C_1: cenário 1 com os 2010 marcadores genotipados; C_2: cenário 2 com os 1608 marcadores direta ou indiretamente relacionados às características; C3: cenário 3 com os 44 marcadores diretamente relacionados às características e os 402 marcadores desnecessários a predição; C_4: cenário 4 com os 402 marcadores desnecessários a predição; C_5: cenário 5 com os 44 marcadores diretamente relacionados às características.

A raiz do erro quadrado médio das técnicas novamente concordou com os resultados obtidos pela capacidade preditiva e repetiu o padrão de ser influenciado pela dimensão da característica, como mostrado na Tabela 7. Os menos valores foram obtidos em x_5, com REQM 5,852 para o *bagging* e 5,853 para o *boosting*.

Tabela 7. Raiz do erro quadrado médio (REQM) de todos os cenários em relação as características analisadas nas metodologias *bagging* e *boosting*.

	BA					BO				
V	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
x_1	9,566 A	9,557 A	9,607 A	10,875A	9,497 A	10,446A	10,667A	10,843 A	13,710 A	11,302 A
x_2	9,978 A	9,969 A	10,024A	16,036A	9,868 A	12,842B	13,068B	13,230 B	19,815 A	13,323 B
x_3	7,374 A	7,328 A	7,464 A	8,914 A	7,280 A	7,695 B	7,769 B	8,109 AB	10,997 A	8,247 AB
x_4	7,804 B	7,729 B	7,841 B	15,099A	7,650 B	10,523B	10,517B	10,797 B	18,884 A	10,578 B
x_5	6,074 A	6,025 A	6,098 A	7,922 A	5,852 A	5,853 B	5,861 B	5,919 B	9,746 A	5,938 B
x_6	6,757 B	6,683 B	6,730 B	14,438A	6,459 B	9,200 B	9,219 B	9,092 B	17,923 A	8,591 B

As médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. BA: *bagging*; BO: *boosting*. C_1: cenário 1 com os 2010 marcadores genotipados; C_2: cenário 2 com os 1608 marcadores direta ou indiretamente relacionados às características; C3: cenário 3 com os 44 marcadores diretamente relacionados às características e os 402 marcadores desnecessários a predição; C_4: cenário 4 com os 402 marcadores desnecessários a predição; C_5: cenário 5 com os 44 marcadores diretamente relacionados às características.

Apesar de não fornecer os maiores resultados, a técnica *random forest* novamente se destacou em relação a sua percepção as diferenças entre os cenários. Na Tabela 8 a seguir é mostrada a capacidade preditiva da técnica.

Tabela 8. Capacidade preditiva (r^2) de todos os cenários em relação as características analisadas na metodologia *random forest*.

	RF				
V	C_1	C_2	C_3	C_4	C_5
x_1	0,199 A	0,194 A	0,184 A	0,007 B	0,242 A
x_2	0,570 A	0,574 AB	0,566 B	0,005 C	0,607 A
x_3	0,322 B	0,330 B	0,306 B	0,004 C	0,371 A
x_4	0,700 BC	0,704 B	0,686 C	0,003 D	0,720 A
x_5	0,449 B	0,452 B	0,446 B	0,003 C	0,514 A
x_6	0,754 B	0,760 B	0,749 B	0,003 C	0,788 A

As médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. RF: *Random forest*. C_1: cenário 1 com os 2010 marcadores genotipados; C_2: cenário 2 com os 1608 marcadores direta ou indiretamente relacionados às características; C3: cenário 3 com os 44 marcadores diretamente relacionados às características e os 402 marcadores desnecessários a predição; C_4: cenário 4 com os 402 marcadores desnecessários a predição; C_5: cenário 5 com os 44 marcadores diretamente relacionados às características.

Nesta tabela, o cenário quatro se diferenciou dos demais para todas as características e teve o diferencial de obter, estatisticamente, os melhores valores no cenário cinco, para a maioria das características. Em x_3, x_4, x_5 e x_6 obteve r^2, respectivamente iguais a 0,371, 0,720, 0,514 e 0,788.

Com a redução do número de marcadores, foi diminuída a multicolinearidade. Reduziu-se a alta correlação entre os polimorfismos, o que proporcionou uma melhora do ajuste do modelo, afetando as estimativas dos parâmetros.

Resultado semelhante foi obtido por SANT'ANNA *et al.* (2021). Esses autores demonstraram que a redução da dimensionalidade também melhora a predição dos valores genéticos das técnicas RR-BLUP e inteligência computacional (IC). Eles usaram características sob efeito epistático e a redução fez a capacidade preditiva variar de 0,07 para 0,5 para a IC e de 0,06 para 0,47 para o RR-BLUP.

SILVA (2021) obteve resultados diferentes. A redução de dimensionalidade levou a resultados similares de quando se usou todos os conjuntos de marcadores.

O REQM para o cenário cinco concordou com o r^2 somente nas variáveis x_3 e x_5 (10,280 e 8,371, respectivamente). No entanto, deve-se ressaltar que essa estimativa estava sendo influenciada pela dimensão das características.

Tabela 9. Raiz do erro quadrado médio (REQM) de todos os cenários em relação as características analisadas na metodologia *random forest*.

	RF				
V	C_1	C_2	C_3	C_4	C_5
x_1	13,736 B	13,736 B	13,929 B	15,418 A	13,463 B
x_2	14,653 B	14,576 B	14,886 B	22,746 A	14,214 B
x_3	10,545 C	10,474 C	10,817 B	12,653 A	10,280 D
x_4	11,496 BC	11,411 BC	11,919 B	21,353 A	11,307 C
x_5	8,729 BC	8,656 C	8,957 B	11,210 A	8,371 D
x_6	9,981 C	9,855 CD	10,367 B	20,474 A	9,576 D

As médias seguidas pelas mesmas letras maiúsculas na horizontal não diferem estatisticamente entre si pelo teste de Tukey a 5% de probabilidade. V: variáveis; x_1: 40 genes, h^2: 40%; x_2: 44 genes, h^2: 40%; x_3: 40 genes, h^2:60%; x_4: 44 genes, h^2:60%; x_5: 44 genes, h^2:80%; x_6: 40 genes, h^2:80%. RF: *Random forest*. C_1: cenário 1 com os 2010 marcadores genotipados; C_2: cenário 2 com os 1608 marcadores direta ou indiretamente relacionados às características; C3: cenário 3 com os 44 marcadores diretamente relacionados às características e os 402 marcadores desnecessários a predição; C_4: cenário 4 com os 402 marcadores desnecessários a predição; C_5: cenário 5 com os 44 marcadores diretamente relacionados às características.

Nesse estudo, foi comparada a importância preditiva em características oligogências envolvendo metodologias BLUP e de aprendizado de máquinas. Adicionalmente, foi incluída a comparação das metodologias em relação a diferentes conjuntos de marcadores, permitindo encontrar se a redução de dimensionalidade afeta a capacidade preditiva das técnicas.

O *bagging*, apesar de ser uma técnica que obteve excelentes resultados preditivos em vários trabalhos, foi enganada pelos efeitos falsos positivos do cenário quatro. A técnica *random forest,* no entanto, explorou o conjunto de dados de forma

melhor, obtendo ótimos resultados no cenário cinco e péssimos resultados no cenário quatro.

Os resultados mostram que as técnicas reconhecidamente classificadas como ótimas devem ser analisadas antes de serem aplicadas aos diferentes conjuntos de dados, pois podem obter resultados indesejados.

Em outros trabalhos também se obteve diferença entre predições utilizando diferentes conjuntos de marcadores, mas não foi realizado um teste estatístico mostrando a real eficiência de se utilizar um menor conjunto de dados.

O uso de grandes conjuntos de marcadores envolve problemas de dimensionalidade e multicolinearidade, além de aumentar o tempo computacional das análises. Nesse contexto, a identificação de subconjuntos polimórficos na predição se faz necessário, pois é possível reduzir esforço físico, custo, mão de obra e tempo na experimentação (PALIWAL E KUMAR, 2011; FERREIRA *et al.*, 2015).

Os resultados sugerem que, para determinar a eficácia da redução de dimensionalidade, deveriam ser testadas várias técnicas e analisadas suas capacidades preditivas em cada subconjunto de SNPs.

CONCLUSÃO

As técnicas *bagging* e *boosting* obtiveram os melhores valores numéricos para r^2 e REQM em todas as características e situações com número diferenciado de marcadores usados na predição.

A técnica *random forest* se mostrou com qualidade intermediária em relação as outras técnicas, mas extremamente perceptível as variações entre as características e entre os cenários.

A exclusão de marcadores desnecessários a predição leva aos mesmos resultados ou até a resultados melhores de quando todos os marcadores são usados, podendo ocorrer a retirada desses marcadores para aumentar a rapidez das análises computacionais.

Nesse trabalho, a redução da dimensionalidade foi feita tendo-se o conhecimento da importância dos marcadores. Trabalhos futuros devem focar na identificação da melhor forma de encontrar os marcadores envolvidos diretamente na análise.

REFERÊNCIAS

Beucher, A. Møller, AB. Greve, MH. Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark, Geoderma. 2019.

Breiman, L. Bagging Predictors. Machine Learning 24: 123–140.1996

Brodley, CE. Friedl, MA. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment 61: 399–409. 1997.

Cavalcanti, JJV. Resende, MDV de. Santos, FHC dos. Pinheiro, CR. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. 2012.

Chakradhar, T. Hindu, V. Reddy, PS. Genomic-based-breeding tools for tropical maize improvement. 2017.

Costa, WG. Barbosa, IP. Souza, JE. Cruz, CD. Nascimento, M. Oliveira, ACB. Machine learning and statistics to qualify environments through multi-traits in Coffea arabica. PLOS ONE 16: 1–21. 2021.

Crossa, J. Perez-Rodríguez, P. Cuevas, J. Montesinos-Lopez, O. Jarquín, D. de Los Campos, G. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22, 961–975.2017.

Cruz, CD. Genes Sofware – extended and integrated with the R, Matlab and Selegen. Acta Scientiarum. 38(4), 547-552.2016.

Cruz, CD. Princípios de Genética Quantitativa. 2ª edição, capítulo 1. 2012.

Chagas, FE de O. Análises genômicas e biométricas para escolha de genitores e predição de híbridos não realizados.2018.

Degenhardt, F.Seifert, S.Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 2019.

Ferreira, MG. Azevedo, AM. Siman, LI. Silva, GH. Carneiro, CS. Alves, FM. Delazari, FT.Silva, DJH. Nick, C. Automation in accession classification of Braziliam Capsicum germplams through artificial neural networks. Scientia Agricola. 2015.

Ghafouri-Kesbi, F. Rahimi-Mianji, G. Honarvar, M. Nejati-Javaremi,A. Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation. Animal Production Science 57: 229. 2017.

Granato, ISC. Marinho, CD. Filho, JE de A. Resende, MDV de. Silva, FF. Ferreira, KCZ. Rosse, LN. Sansaloni, CP. Petroli, CD. Grattapaglia, D. Seleção de Marcadores Para os

Métodos RR-BLUP e BLASSO na Seleção Genômica Ampla. 2013.

Hastie, T. Tibshirani, R. Friedman, J. The elements of statistical learning: Data mining, inference, and prediction, 2. ed. Springer, New York, NY, USA, 764p.2009.

Hirschhorn, JN. Lohmueller, K. Byrne, E. Hirschhorn, K. A comprehensive review of genetic association studies. Genet Med 4:45–61. 2002.

James, G. Witten, D.Hastie, T. Tibshirani, R. An Introduction to Statistical Learning. Springer, New York. 2013.

James, G.Witten, D. Hastie, T. Tibshirani, R. An Introduction to Statistical Learning. p. 612. In Springer Texts in Statistics.2021.

Lee, Sh. Van Der Werf, JHJ. Hayes, BJ. Goddard, Me. Visscher, PM. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet 4(10):e1000231. 2008.

Massman, JM. Jung, H-JG. Bernardo, R. Genome wide Selection versus Marker-assisted Recurrent Selection to Improve Grain Yield and Stover-quality Traits for Cellulosic Ethanol in Maize. 2012.

Mitchell TM. Machine Learning. WCB – McGraw-Hill, Boston, MA. 1997.

Paliwal, M. Kumar, UA. Assessing the contribution of variables in feed forward neural network. Applied Soft Computing. 11:3690-3696.2011.

Palucci, V. Schaeffer, LR. Miglior, F. Osborne, V. Non-additive genetic effects for fertility traits in Canadian Holstein cattle. Genet Sel Evol 39(2):181–193. 2007.

Parmley, KA. Higgins, RH. Ganapathysubramanian, B.Machine Learning Approach for Prescriptive Plant Breeding. Sci Rep 9, 17132. 2019.

Prasad, AM. Iverson, LR. Liaw, A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 9: 181–199. 2006.

Resende, M.D.V. Matemática e estatística na análise de experimentos e no melhoramento genético Colombo: Embrapa Florestas. 561 p.2007.

Sant'Anna, I de C. Silva, GN. Nascimento, M. Cruz, CD. Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks. 2021.

Silva, MJ da. Efficiency of genomic prediction according to decreased the SNP´s markers and different degrees of dominance, heritability, and epistatic interactions. 2021.

Silva Júnior, AC da. Silva, MJ da. Cruz, CD. Sant'anna, I de C. Silva, GN. Nascimento, M. Azevedo, CF. Prediction of the importance of auxiliary traits using computational intelligence and machine learning: A simulation study. 2021.

Sousa, IC. Nascimento, M. Silva, GN. Nascimento, ACC. Cruz, CD. Fonseca, F. Almeida, DP. Pestana, KN. Azevedo, CF.Zambolim, L. Caixeta, ET. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola 78: 1-8. 2020.