

## SELEÇÃO GENÔMICA AMPLA NO MELHORAMENTO DE *Coffea canephora*<sup>1</sup>

Fernanda de Araújo Carneiro<sup>2</sup>, Sinara Oliveira de Aquino<sup>3</sup>, Nathália Gomes Mattos<sup>4</sup>, Jéssica Coelho Valeriano<sup>5</sup>, Wendel William de Jesus Carneiro<sup>6</sup>, Gustavo Costa Rodrigues<sup>7</sup>; Milene Alves de Figueiredo Carvalho<sup>8</sup>; Adriano Delly Veiga<sup>9</sup>, Dario Grattapaglia<sup>10</sup>, Orzenil Bonfim da Silva Júnior<sup>11</sup>, Pierre Marraccini<sup>12</sup>, Indalecio Cunha Vieira Junior<sup>13</sup>, Márcio Balestre<sup>14</sup>, Alan Carvalho Andrade<sup>15</sup>

<sup>1</sup> Trabalho financiado pelo Consórcio Pesquisa Café e INCT-Café (CNPq/FAPEMIG).

<sup>2</sup> Doutora em Biotecnologia Vegetal, UFLA, Lavras-MG, Brasil, [fearca14@gmail.com](mailto:fearca14@gmail.com)

<sup>3</sup> Bolsista Pós-Doutorado PNP/CAPEL, PhD, UFLA, Lavras-MG, Brasil, [saquinobiotech@gmail.com](mailto:saquinobiotech@gmail.com)

<sup>4</sup> Doutoranda em Biotecnologia Vegetal, UFLA, Lavras-MG, Brasil, [nagm7@hotmail.com](mailto:nagm7@hotmail.com)

<sup>5</sup> Doutoranda em Biotecnologia Vegetal, UFLA, Lavras-MG, Brasil, [jessicacoelho\\_bio@hotmail.com](mailto:jessicacoelho_bio@hotmail.com)

<sup>6</sup> Bolsista Consórcio Pesquisa Café, [wendelwill11@gmail.com](mailto:wendelwill11@gmail.com)

<sup>7</sup> Pesquisador, PhD, Embrapa Informática Agropecuária, Campinas-SP, Brasil, [gustavo.rodrigues@embrapa.br](mailto:gustavo.rodrigues@embrapa.br)

<sup>8</sup> Pesquisadora, PhD, Embrapa Café/INOVACAFÉ, Lavras-MG, Brasil, [milene.carvalho@embrapa.br](mailto:milene.carvalho@embrapa.br)

<sup>9</sup> Pesquisador, PhD, Embrapa Cerrados, Planaltina-DF, Brasil, [adriano.veiga@embrapa.br](mailto:adriano.veiga@embrapa.br)

<sup>10</sup> Pesquisador, PhD, Embrapa Recursos Genéticos e Biotecnologia, Brasília-DF, Brasil, [dario.grattapaglia@embrapa.br](mailto:dario.grattapaglia@embrapa.br)

<sup>11</sup> Pesquisador, PhD, Embrapa Recursos Genéticos e Biotecnologia, Brasília-DF, Brasil, [orzenil.silva@embrapa.br](mailto:orzenil.silva@embrapa.br)

<sup>12</sup> Pesquisador, PhD-HDR, CIRAD, UMR IPME, Instituto of Agricultural Genetics (AGI) LMI RICE2, Hanoi, Vietnã, [marraccini@cirad.fr](mailto:marraccini@cirad.fr)

<sup>13</sup> Doutorando em Genética e Melhoramento de Plantas, UFLA, Lavras-MG, Brasil, [indasjunior@hotmail.com](mailto:indasjunior@hotmail.com)

<sup>14</sup> Professor, PhD, UFLA, Lavras-MG, Brasil, [marciobalestre@ufla.br](mailto:marciobalestre@ufla.br)

<sup>15</sup> Pesquisador, PhD, Embrapa Café/INOVACAFÉ, Lavras-MG, Brasil, [alan.andrade@embrapa.br](mailto:alan.andrade@embrapa.br)

**RESUMO:** O café é globalmente uma das *commodities* mais importante e sua produção é baseada em duas principais espécies, *Coffea arabica* e *Coffea canephora*, sendo o Brasil o maior produtor mundial. Acredita-se que toda essa produção possa ser comprometida, devido aos efeitos das mudanças climáticas, afetando a viabilidade das flores, o desenvolvimento dos frutos, a produção e a qualidade da bebida. Uma alternativa para auxiliar na obtenção de cafeeiros que sejam mais adaptados às condições climáticas futuras e, ao mesmo tempo, sem redução da produção e a qualidade da bebida pode ser a seleção genômica ampla (SGA). A implementação desse programa requer muitos marcadores genéticos, que agora são mais facilmente descobertos com a publicação do genoma de referência de *C. canephora*. A maioria das características que são economicamente importantes são quantitativas e complexas, requerindo uma cobertura genômica completa de marcadores, e, desta forma, a SGA, através do cálculo dos valores genéticos genômicos (VGG), se apresenta como uma ferramenta importante para o melhoramento de plantas. Na SGA, esses marcadores são usados para construir um modelo preditivo usando indivíduos com informações genotípicas e fenotípicas conhecidas. Com este modelo, os VGGs para as características desejadas podem ser calculados e usados para classificar os indivíduos com fenótipo desconhecido para posterior seleção. A SGA é uma nova abordagem no melhoramento genético de plantas que permite a seleção precoce de materiais de elite, maximizando ganhos genéticos ao longo de gerações, principalmente no caso de plantas perenes como o café. Para este propósito, neste trabalho, um grande número de indivíduos e marcadores foram selecionados, para se avaliar a aplicabilidade da SGA em termos de resposta de seleção das características alvo e precisão da predição em uma população de *C. canephora*. Aproximadamente 1.300 indivíduos de *C. canephora* de uma população localizada na Embrapa Cerrados (Planaltina, DF, Brasil) foram selecionados e avaliados quanto as características como produção, precocidade dos frutos, tamanho de grão, tamanho da cereja e peso de 100 grãos. O DNA dos indivíduos foi extraído e os dados genotípicos foram obtidos a partir do chip desenvolvido para *C. canephora* com aproximadamente 26K SNPs. A população amostrada foi utilizada para estimação e validação no modelo preditivo. Equações de predição de VGGs foram construídas para cada característica de interesse, previsto pelo G-BLUP. Embora os estudos de SGA ainda não tenham sido capazes de explicar a variação fenotípica total, eles identificaram regiões genômicas essenciais, e para espécies perenes como café, o uso de SGA pode aumentar muito a eficiência de seleção, reduzindo o tempo de introdução de características de interesse.

**PALAVRAS-CHAVE:** *C. canephora*, Seleção Genômica Ampla (SGA), SNP, DNA array.

## GENOME WIDE SELECTION IN *Coffea canephora* BREEDING

**ABSTRACT:** Coffee is one of the most heavily globally traded commodities and its production is based on *Coffea arabica* and *Coffea canephora* and Brazil being the world's largest coffee producer. It is believed that all this production has already been affected due to climatic changes, affecting the flower viability, fruit development, yield and beverage quality. An alternative to assist in obtaining coffee plants that would support future climatic conditions and, at the same time, would not reduce production and quality of the beverage would be genomic selection (GS). The implementation of these programs requires a lot of genetic markers, which are more readily discovered now after the reference genome of *C. canephora* became available. Most of the traits which are economically important are quantitative and complex in nature and require whole genome coverage of markers thus, GS, through the calculation of genomic estimated breeding

values (GEBVs), are becoming important and effective tools for plant breeding. In the case of GS, these markers are used to build a predictive model using individuals with known genotypic and phenotypic information. With this model, GEBVs for the desired trait can be calculated and used to rank the individuals with unknown phenotype for subsequent selection. GS is a new approach in coffee breeding that enable early selection of elite materials, maximizing genetic gains over generations. For this purpose, in this work, a large number of individuals and markers were selected to assess GS applicability in terms of selection response of the targeted traits and prediction accuracy in a population of *C. canephora*. Approximately 1,300 individuals of *C. canephora* from a breeding population located at Embrapa Cerrados (Planaltina, DF, Brazil) were selected and evaluated for characteristics such as productivity, fruit precocity, grain size, cherry fruit size and weight of 100 grains. DNA of each of the 1,300 individuals selected of breeding population was extracted and the genotypic data were obtained from 26K Axiom SNP array developed for *C. canephora*. The sampled population performed two functions at the same time, estimation and validation of the GS model. Prediction equations of GEBVs were constructed for each character of interest, predicted by G-BLUP. Although GS studies have not yet been able to explain the total phenotypic variation, they have identified essential genomic regions, and for perennial species such as coffee, the use of GS can greatly increase selection efficiency, thus reducing the time to introduction of characteristics of interest.

**KEY WORDS:** *C. canephora*, Genomic Selection, SNP, DNA array.

## INTRODUÇÃO

O grande atrativo da genética molecular em benefício do melhoramento de plantas é a utilização direta das informações do DNA na seleção, permitindo assim, alta eficiência seletiva, maior rapidez na obtenção de ganhos genéticos e baixo custo, isto em comparação com a seleção fenotípica tradicional (RESENDE et al., 2008). Nesse contexto, a seleção genômica ampla (SGA) usa uma alta cobertura de marcadores no genoma e captura seus efeitos simultaneamente.

Hoje, com os avanços nas plataformas de genotipagem de alto rendimento, a SGA foi implementada com sucesso, primeiramente no melhoramento de bovinos, e posteriormente adaptada para outros programas de melhoramento, tanto animal (HAYES et al., 2009; HAYES; GODDARD, 2010) como vegetal (BERNARDO; YU, 2007, HEFFNER, SORRELS, JANNINK, 2009; JANNINK; LORENZ; IWATA, 2010). Esse sucesso foi atribuído principalmente às técnicas atuais de genotipagem de DNA, que reduziram os custos e permitiram que um grande número de genótipos pudessem ser genotipados a um preço acessível. A SGA assume que, com uma alta densidade de marcadores, os genes de interesse provavelmente apresentem desequilíbrio de ligação (DL) com parte importante desses marcadores (GODDARD; HAYES, 2007) e, como somente os marcadores em DL são úteis para explicar a variação genética, é esperado que grande parte ou quase a totalidade desta variação seja explicada pelos mesmos.

*C. canephora* é um bom ponto de partida para desenvolver estudos de SGA dentro do gênero *Coffea*, por razões econômicas e genéticas, incluindo a ploidia e a maior variabilidade genética, quando comparado ao *C. arabica* (TRAN et al., 2016). A motivação econômica baseia-se no fato de que a produção de *C. canephora* representa 40% da produção mundial de café, com o Brasil ocupando a primeira posição na produção e exportação desses grãos (ICO, 2018).

O melhoramento genético do cafeeiro via SGA representa uma mudança de paradigma, pois permite a seleção precoce de plantas elite ainda no estágio de mudas, principalmente para características de expressão tardia tais como, produção, qualidade da bebida e tolerância a estresses bióticos e abióticos. Avanços pioneiros na área de genômica do cafeeiro foram alcançados nos últimos anos, como por exemplo, o genoma de referência de *C. canephora* (DENOEUDE et al., 2014), gerando as ferramentas básicas para o estabelecimento de um programa de melhoramento com base na SGA. Diante do exposto, este estudo objetivou: (i) estimar os parâmetros genéticos e (ii) avaliar a aplicabilidade e acurácia da SGA na predição dos valores genéticos genômicos (VGGs) para *C. canephora*.

## MATERIAL E MÉTODOS

Um total de 1.319 indivíduos de *C. canephora* selecionados a partir de uma população de melhoramento localizada no campo experimental da Embrapa Cerrados (Planaltina, DF) foram fenotipados para diversas características de interesse agrônomo (CARNEIRO et al., 2013) e genotipados através da plataforma Coffee Axiom chip – 26K (ANDRADE et al., 2017).

Ao longo do genoma de *C. canephora* foram identificados 25.456 SNPs e após um processamento mínimo, utilizando a função *A.mat* do pacote rrBLUP (ENDELMAN, 2011), no software R (R Core Team, 2017), obteve-se um conjunto de 16.688 marcadores com *call rate* acima de 90% e MAF (*Minimum Allele Frequency*) superior a 1%. As análises de SGA foram realizadas utilizando o método G-BLUP, usando o pacote rrBLUP, e o comando *mixed.solve*.

Uma vez estimado o efeito de cada marca para os diferentes caracteres, foi possível estimar o valor genético genômico (VGG) dos indivíduos. A validação cruzada é um método usado para avaliar a capacidade de generalização de um modelo preditivo para um conjunto de dados. Ao aplicar esse método, o conjunto de dados é dividido em subconjuntos. Nesse trabalho, utilizou-se a validação *10-fold*, realizando a predição dos VGGs para um dos grupos a partir do modelo de predição gerado com o uso dos outros 9 grupos, portanto para a população de 1.319 indivíduos, 1.189 foram utilizados para treinamento ou estimação e 130 indivíduos na validação.

## RESULTADOS E DISCUSSÃO

Na análise dos componentes de variância e de herdabilidade (Tabela 1) das marcas para os 1.319 indivíduos, nota-se que as maiores herdabilidades foram encontradas para as características Moca Miúdo (56%), seguidas por Chato Graúdo (55%) e Moca Médio (50%). As maiores variâncias genética e ambiental foram obtidas para a característica Chato Graúdo e os caracteres de Eixo Maior e Menor, apresentaram os menores valores.

Tabela 1. Resumo dos resultados obtidos para os componentes de variância e herdabilidade na análise dos caracteres avaliados para os 1.319 indivíduos de *C. canephora*.

Caracteres	Componentes da variância		Herdabilidade
	Genético	Ambiental	
Prod12	1.47	11.73	0.11
Prod13	5.17	16.77	0.24
Prod14	8.84	32.59	0.21
Mean1_2_3	3.49	7.76	0.31
P100	2.52	2.78	0.48
Boia	11.44	130.45	0.08
Área	0.04	0.07	0.35
Perímetro	0.12	0.21	0.37
Eixo Maior	0.00286	0.00437	0.40
Eixo Menor	0.00078	0.00224	0.26
Chato Graúdo	210.71	168.97	0.55
Chato Médio	45.56	152.08	0.23
Chato Miúdo	30.33	34.96	0.46
Moca Graúdo	57.30	63.76	0.47
Moca Médio	20.09	19.95	0.50
Moca Miúdo	29.76	23.17	0.56

Como apresentado na Tabela 2, as acurácias do modelo proposto foram de 61% para a característica Eixo Maior, seguido pela característica de Chato Graúdo, explicando 60% da variação fenotípica. As acurácias obtidas para a característica de Produção variaram de 0,18, para Produção de 2012, a 0,4, para a média dos três anos avaliados.

Tabela 2. Acurácia preditiva do modelo, ganho com a seleção (GS) e intensidade de seleção (IS) para cada característica analisada nos 1.319 indivíduos de *C. canephora*.

Caracteres	Acurácia	GS	IS
Prod12	0.18	0.17	0.02
Prod13	0.39	0.46	0.10
Prod14	0.29	0.48	0.05
Mean1_2_3	0.40	0.30	0.09
P100	0.56	0.17	0.13
Boia	0.25	0.59	0.03
Área	0.50	0.11	0.15
Perímetro	0.53	0.06	0.16
Eixo Maior	0.61	0.09	0.17
Eixo Menor	0.30	0.05	0.04
Chato Graúdo	0.60	1.03	0.16
Chato Médio	0.26	0.31	0.02
Chato Miúdo	0.50	1.94	0.10
Moca Graúdo	0.45	0.38	0.08
Moca Médio	0.55	0.69	0.12
Moca Miúdo	0.57	1.58	0.12

Para todas as características avaliadas na população foi obtido o gráfico de lagarta contendo o VGG e o erro de predição do modelo, identificando aqueles indivíduos que se encontram acima da média. Considerando a característica de tamanho e formato de frutos, as medidas de área, perímetro e eixo maior apresentaram um elevado número de indivíduos acima da média, comparado ao resultado obtido para eixo menor (Figura 1).

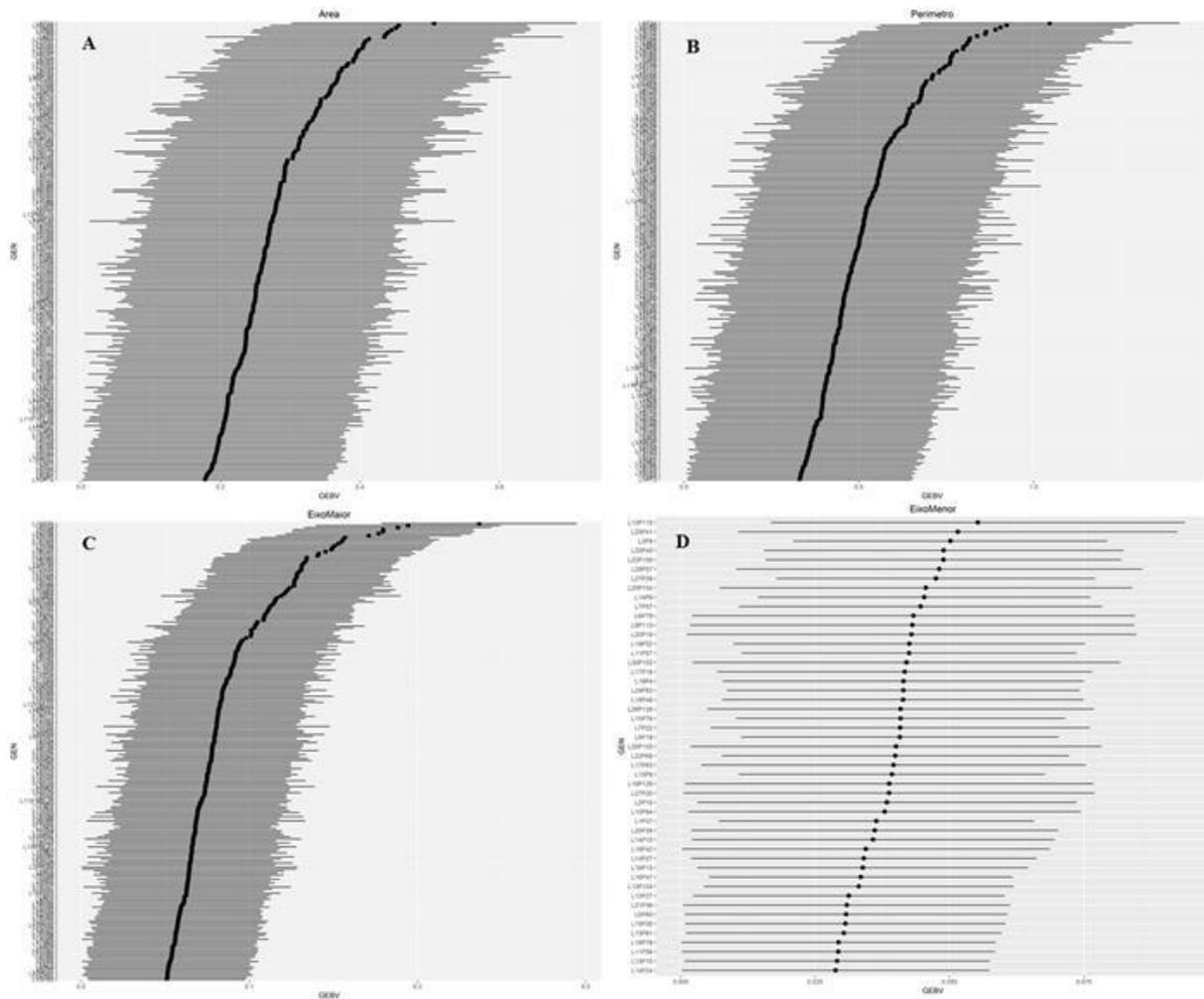


Figura 1. Gráfico com o VGG e o erro de predição do modelo para as características (A) Área, (B) Perímetro, (C) Maior Eixo e (D) Menor Eixo na análise com os 1.319 indivíduos. Somente os indivíduos que apresentaram médias acima de zero estão presentes nos gráficos.

Para peso de 100 grãos, as plantas identificadas acima da média apresentaram VGGs variando de 1 a 5. A produção avaliada em 2012 apresentou 26 plantas acima da média com o VGG variando de 1,5 a 2,5. Já para a produção (2013), o número de plantas que ficaram acima da média foi bem mais elevado, assim como o VGG, variando de 2,5 até 7,5. A produção de 2014 apresentou 56 indivíduos acima da média, com o VGG variando de 3-7 (Figura 2).

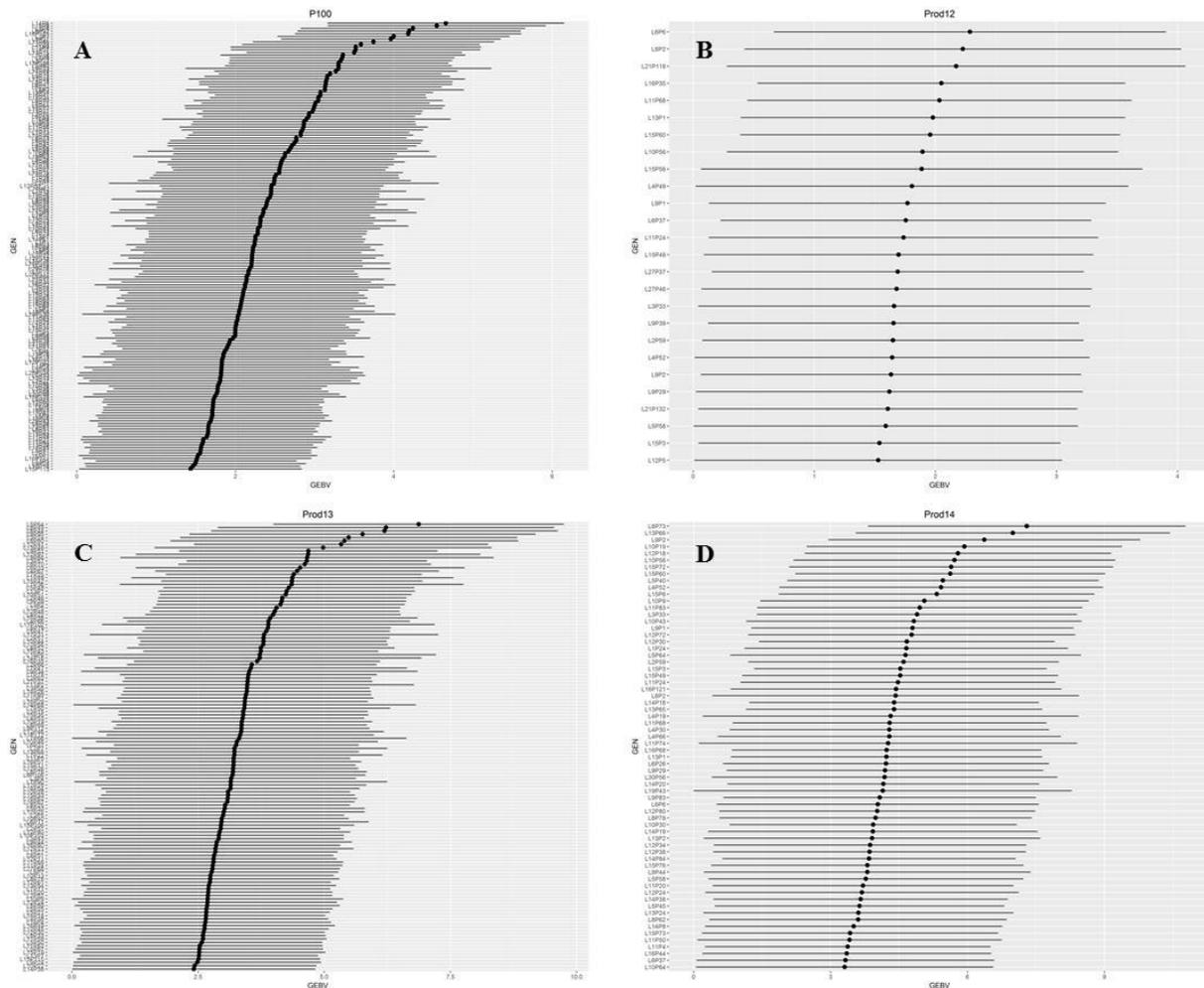


Figura 2. Gráfico com o VGG e o erro de predição do modelo para as características (A) Peso de 100 grãos, (B) Produção (2012), (C) Produção (2013) e (D) Produção (2014) na análise com os 1.319 indivíduos. Somente os indivíduos que apresentaram médias acima de zero estão presentes nos gráficos.

Dentre as diversas metodologias envolvendo marcadores moleculares no melhoramento de plantas perenes, a SGA provavelmente tem gerado a maior expectativa, em termos de possibilidade de seleção para caracteres quantitativos (RESENDE et al., 2008), uma vez que utiliza uma grande quantidade de marcas, o que aumenta a probabilidade de capturar uma maior proporção da variação genética (MEUWISSEN; GODDARD; HAYES., 2001; WIENTJES; VEERKAMP; CALUS, 2013), facilitando a seleção de genótipos superiores.

A partir das informações das equações preditivas da SGA, estimou-se a herdabilidade genômica, apresentando magnitudes baixas ou moderadas. De certo modo, a herdabilidade permite prever o sucesso a ser alcançado com a seleção, uma vez que características com baixa herdabilidade devem apresentar menor capacidade preditiva (LEGARRA et al., 2008). Em estudo realizado por Grattapaglia e de Resende (2011), algumas simulações foram realizadas e verificou-se que o aumento na herdabilidade levava a um aumento na acurácia da SG. Dentre os parâmetros genéticos avaliados, o componente que tem maior interesse por parte dos melhoristas é a herdabilidade, por refletir todos os outros componentes de variância e indicar a proporção da variação fenotípica que pode ser herdada.

Os resultados mostraram uma alta correlação entre as acurácias da validação e as herdabilidades, de mais de 90%. Uma elevada correlação (82%) entre acurácia e herdabilidade também foi reportada por Sousa et al. (2019). Considerando especificamente o resultado obtido para a média de produção, verifica-se que os resultados de herdabilidade e acurácia do modelo estão um pouco mais elevados comparado ao demonstrado por Sousa et al. (2019). Herdabilidade e acurácia foram de 0,31 e 0,4 para os 1.319 indivíduos.

O número de marcadores utilizados nesse estudo mostra-se suficiente, uma vez que não necessariamente o aumento progressivo de marcadores resulte em aumento da acurácia, ao contrário, Sousa et al. (2019) demonstraram que o máximo de acurácia obtida foi utilizando 16.000 marcadores, aumentando esse número para 20.000 a acurácia reduzia. O sucesso da SGA pode ser influenciado por inúmeros fatores, que consequentemente interferem na acurácia do modelo, como tamanho da população de treinamento, tamanho real da população, densidade de marcadores, herdabilidade e número de QTLs controlando as características.

## CONCLUSÕES

1 - O uso da SGA no gênero *Coffea* é uma abordagem promissora, considerando-se que a espécie é perene, de ciclo longo e, em comparação com a avaliação fenotípica tradicional, espera-se que a incorporação da SGA no melhoramento, possa acelerar o processo de seleção, com aumento do ganho genético por unidade de tempo.

## REFERÊNCIAS BIBLIOGRÁFICAS

- ANDRADE, A. C. et al. 'Towards GWAS and Genomic Prediction in Coffee: Development and Validation of a 26K SNP Chip for *Coffea canephora*', In XX INTERNATIONAL PLANT AND ANIMAL GENOME CONFERENCE, p. W173. Anais eletrônicos. Disponível em: <<https://pag.confex.com/pag/xxv/webprogram/Paper23677.html>>
- BERNARDO, R.; YU, J. Prospects for genome wide selection for quantitative traits in maize. *Crop Science*, v. 47, p. 1082-1090, 2007.
- CARNEIRO, F. A. et al. Phenotyping a *Coffea canephora* population, cultivated at high altitude, aiming at a GWS program for coffee. In: Workshop on Biotic and Abiotic Stress Tolerance in Plants: the Challenge for the 21st Century, 2013, Ilhéus. Book of abstracts, 2013.
- DENOEUDE, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, v. 345, n. 6201, p. 1181-1184, 2014.
- ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, v. 4, p. 250-255, 2011.
- GODDARD, M. E.; HAYES, B. J. Genomic selection. *Journal of Animal Breeding and Genetics*, v. 124, p. 323-330, 2007.
- GRATTAPAGLIA, D.; de RESENDE, M. D. V. Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, v. 7, n. 2, p. 241-255, 2011.
- HAYES, B. J. et al. Invited review: genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science*, v. 92, p. 433-443, 2009.
- HAYES, B. J.; GODDARD, M. E. Genome-wide association and genomic selection in animal breeding. *Genome*, v. 53, p. 876-883. 2010.
- HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J. L. Genomic selection for crop improvement. *Crop Science*, v. 49, p. 1-12, 2009.
- ICO – International Coffee Organization. Disponível em: <<http://www.ico.org/>>. Acesso em: 18 jan. 2019.
- JANNINK, J. L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics*, v. 9, p. 166-177, 2010.
- LEGARRA, A. et al. Performance of genomic selection in mice. *Genetics*, v. 180, p. 611-618, 2008.
- MEUWISSEN, T. H. E.; GODDARD, M. E.; HAYES, B. J. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819-1829, 2001.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2017 .URL <http://www.R-project.org/>.
- RESENDE, M. D. V. de et al. Seleção Genômica Ampla (GWS) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira*, n. 56, p. 63-77, 2008.
- SOUSA, T. V. et al. Early Selection Enabled by the Implementation of Genomic Selection in *Coffea arabica* Breeding. *Frontiers in Plant Science*, v. 9, p. 1-12, 2019.
- TRAN, H. T. M. et al. Advances in genomics for the improvement of quality in coffee. *Journal of the Science of Food and Agriculture*, v. 96, n. 10, p. 3300-3312, 2016.
- WIENJES, Y. C. J.; VEERKAMP, R. F.; CALUS, M. P. L. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, v. 193, n. 2, p. 621-631, 2013.