ISSN 2317-2029

CIRCULAR TÉCNICA

7

Brasília, DF
Abril, 2022

Tamanho amostral e detecção de genes via GWAS em características quantitativas do cafeeiro

Marcos Deon Vilela de Resende
Antonio Carlos Baião de Oliveira
Eveline Teixeira Caixeta
Emilly Ruas Alkimim
Tiago Vieira Sousa
Antonio Alves Pereira
Rodrigo Silva Alves
Camila Ferreira Azevedo

Embrapa

Tamanho amostral e detecção de genes via GWAS em características quantitativas do cafeeiro[1]

Introdução

O melhoramento genético do cafeeiro (espécie perene e de ciclo reprodutivo longo) demanda agilidade e eficácia no lançamento de novas cultivares. Para alcançar esse objetivo, a otimização do melhoramento e a maximização da eficiência seletiva são essenciais.

A seleção genômica ampla (do inglês, *genome-wide selection* – GWS) foi proposta por Meuwissen et al. (2001) como uma forma de aumentar a eficiência e acelerar o melhoramento genético. A GWS enfatiza a predição simultânea (sem o uso de testes de significância para marcas individuais) dos efeitos genéticos de milhares de marcadores genéticos de DNA dispersos em todo o genoma de um organismo, de forma a capturar os efeitos de todos os locos (tanto de pequenos quanto de grandes efeitos) e explicar toda a variação genética de uma característica quantitativa (Resende; Alves, 2020).

Além da GWS, a associação genômica ampla (do inglês, *genome-wide association* – GWAS) é relevante nesse contexto. A evolução da tecnologia genômica é previsível e a mutação causal de uma variação genética em nível de nucleotídeo poderá ser acessada em um futuro próximo. Assim, a seleção genômica poderá ser refinada e aperfeiçoada pelo uso direto dos nucleotídeos

[1] Marcos Deon Vilela de Resende, Engenheiro-agrônomo/estatístico, doutor em Genética, pesquisador da Embrapa Café; Antonio Carlos Baião de Oliveira, Engenheiro-agrônomo, doutor em Genética e Melhoramento, pesquisador da Embrapa Café; Eveline Teixeira Caixeta, Engenheira-agrônoma, doutora em Genética e Melhoramento, pesquisadora da Embrapa Café; Emilly Ruas Alkimim, Engenheira-agrônoma, doutora em Genética e Melhoramento, professora da Universidade Federal do Triângulo Mineiro; Tiago Vieira Sousa, Engenheiro-agrônomo, doutor em Genética e Melhoramento, professor do Instituto Federal Goiano; Antonio Alves Pereira, Engenheiro-agrônomo, doutor em Fitopatologia, pesquisador da Empresa de Pesquisa Agropecuária de Minas Gerais; Rodrigo Silva Alves, Engenheiro florestal, doutor em Genética e Melhoramento, pós doutorando do Instituto Nacional de Ciência e Tecnologia do Café; Camila Ferreira Azevedo, Matemática, doutora em Estatística Aplicada e Biometria, professora da Universidade Federal de Viçosa.

de característica quantitativa (do inglês, *quantitative trait nucleotides* – QTNs) em lugar dos polimorfismos de nucleotídeo único (do inglês, *single-nucleotide polymorphisms* – SNPs). Esforços na detecção de genes em características quantitativas do cafeeiro também são relevantes e contribuem para a otimização do melhoramento e maximização da eficiência seletiva. Nesse sentido, os estudos de associações genômicas são essenciais, pois visam à detecção de locos controladores de características quantitativas e/ou regiões genômicas contendo blocos poligênicos associados a fenótipos de importância econômica.

O presente documento refere-se à avaliação prática do potencial da GWAS no melhoramento dos cafeeiros das espécies *Coffea arabica* e *Coffea canephora*. O objetivo geral é orientar a aplicação eficiente da GWAS em características produtivas, agronômicas e de resistência a doenças e pragas do cafeeiro. Para isso, um novo método para o estudo de associações genômicas foi desenvolvido e avaliado via simulação, em termos de poder de detecção de genes. O objetivo específico é possibilitar a identificação de marcadores genéticos úteis ao melhoramento do cafeeiro.

Estimativa do número de locos de característica quantitativa e do tamanho amostral para a maximização da acurácia seletiva

A confiabilidade da seleção genômica é dada pela expressão: $r^2_{gg} = Nh^2 / (Nh^2 + N_{QTL})$, em que r_{gg} é a acurácia da GWS; N é o número de indivíduos da população, N_{QTL} é o número de QTL que controla a característica e h^2 é a herdabilidade individual. A estimativa do número de indivíduos que deve ser avaliado para se obter uma acurácia desejada pode ser obtida pela seguinte expressão, derivada da anterior: $N = r^2_{gg}N_{QTL}/(1 – r^2_{gg})h^2$ (Resende et al., 2014).

Na Figura 1 são apresentados gráficos mostrando a curva do número de indivíduos (N) em vários cenários (funções de h^2, N_{QTL} e r_{gg}). Com base nesses gráficos e em informações genéticas das características, os melhoristas podem dimensionar adequadamente seus estudos de herança e de maximização do ganho genético com o melhoramento feito por seleção.

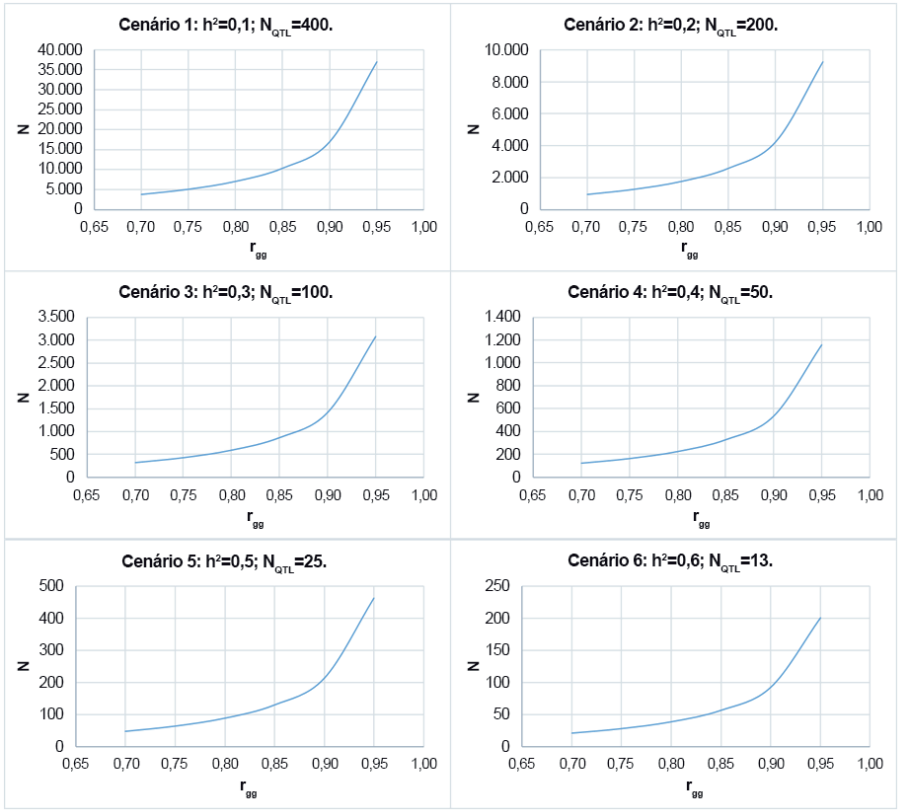

Figura 1. Tamanho amostral para a seleção genômica com acurácias variando de 0,70 a 0,95, em seis cenários.

Várias informações podem ser obtidas a partir da Figura 1. Por exemplo, considerando o cenário 3, verifica-se que, para uma característica com herdabilidade individual igual a 0,30 e controlada por 100 QTLs, pode ser obtida uma acurácia de 90% se o tamanho amostral for igual a 1.500 indivíduos genotipados e fenotipados.

A estimativa do número de QTL que controla cada característica pode ser calculada com base na expressão: $N_{QTL} = [(1 - r^2_{gg})Nh^2] / r^2_{gg}$. Tendo-se estimado a acurácia seletiva e a herdabilidade, dado o N praticado no presente estudo, foi estimado o número de QTL para várias características.

Outro exercício que pode ser feito é a determinação teórica do número de QTL, dados o N praticado e a h^2 estimada, porém variando-se r_{gg}. Para o caso de h^2 igual a 0,30 e N igual a 1.500, os números de QTL podem ser inferidos conforme Figura 2. Nessa mesma figura é apresentado o caso de h^2 igual a 0,20 e N igual a 1.500.

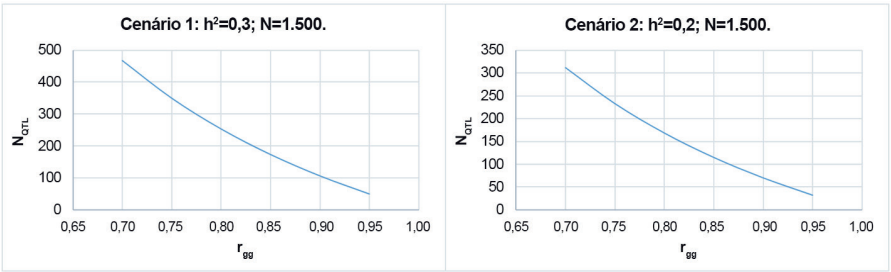

Figura 2. Número de QTL (N_{QTL}) para a seleção genômica com acurácias variando de 0,70 a 0,95, em dois cenários.

Com base na Figura 2, verifica-se que, para N = 1.500 indivíduos genotipados, os números de QTL variam de 49 a 468 (cenário 1, h^2 = 0,30) e de 32 a 312 (cenário 2, h^2 = 0,20) quando as acurácias variaram de 0,95 a 0,70, respectivamente.

Café arábica

Material genético

A partir de cruzamentos entre três genitores Catuaí e três genitores Híbrido de Timor, contrastantes em relação à característica ferrugem do cafeeiro, 13 progênies foram obtidas pelo programa de melhoramento genético de *C. arabica* da Empresa de Pesquisa Agropecuária de Minas Gerais (Epamig) em parceria com a Universidade Federal de Viçosa (UFV) e a Empresa Brasileira de Pesquisa Agropecuária (Embrapa Café). Essas progênies são gerações de retrocruzamentos resistentes (RCr), retrocruzamentos suscetíveis (RCs) e F_2. Em cada progênie foram selecionados 15 genótipos, totalizando 195 indivíduos (Sousa et al., 2019).

Fenotipagem

A avaliação de 15 características – 9 contínuas (produção de frutos, comprimento de uma folha, comprimento de um ramo plagiotrópico, número de nós vegetativos, número de frutos por ramo plagiotrópico, volume de frutos por ramo plagiotrópico, altura da planta, diâmetro da copa e diâmetro do caule) e 6 categóricas (tamanho dos frutos, incidência de ferrugem, incidência de cercospora, infestação de bicho-mineiro, vigor vegetativo e ciclo de maturação) – foi realizada nos 195 indivíduos (Sousa et al., 2019).

O tamanho dos frutos foi avaliado por notas, em que as notas 1, 2 e 3 foram atribuídas aos frutos pequenos, médios e graúdos, respectivamente. O ciclo de maturação foi avaliado por notas, em que as notas 1, 2, 3, 4 e 5 foram atribuídas para ciclo precoce, entre precoce e médio, médio, entre médio e tardio e tardio, respectivamente. A incidência de ferrugem, de cercosporiose e de bicho-mineiro também foi avaliada por notas de 1 a 5, em que se atribuíram a nota 1 aos genótipos assintomáticos e a nota 5 aos genótipos altamente suscetíveis. Da mesma forma, o vigor vegetativo foi avaliado por notas de 1 a 10, atribuindo-se a nota 1 às plantas totalmente depauperadas e a nota 10 às plantas altamente vigorosas.

Número de locos de característica quantitativa e tamanho amostral

O número estimado de QTLs que controlam as características avaliadas variou de 149 (diâmetro da copa) a 3.981 (comprimento da folha) (Tabela 1). Foi observado elevado número de QTL controlando as características agronômicas, sendo estimado, para as características produção de grãos e incidência de ferrugem – que são as características-alvo desse programa de melhoramento –, 751 e 221 QTLs, respectivamente.

Tabela 1. Estimativas do número de indivíduos (N) a serem avaliados para se obter determinada (0,5; 0,6; 0,7; 0,8; 0,9) acurácia seletiva para as 15 características morfoagronômicas avaliadas na população de melhoramento de *Coffea arabica*.

Característica	h^2_a	r_{gg_gws}	N_{QTL}	N(0,5)	N(0,6)	N(0,7)	N(0,8)	N(0,9)
PF	0,26	0,25	751	964	1.626	2.778	5.140	12.326
IF	0,31	0,46	221	237	400	684	1.265	3.033
IC	0,44	0,47	304	231	390	666	1.233	2.957
BM	0,3	0,33	476	536	904	1.544	2.858	6.852
VV	0,34	0,36	440	437	738	1.260	2.332	5.592
CM	0,31	0,21	1.313	1.434	2.421	4.134	7.650	18.345
TF	0,36	0,39	394	370	624	1.066	1.973	4.730
CF	0,29	0,12	3.981	4.530	7.644	13.057	24.160	57.935
CRP	0,41	0,5	244	198	335	572	1.058	2.538
NNV	0,46	0,56	199	143	242	413	765	1.834
NFRP	0,34	0,33	1.157	1.134	1.913	3.267	6.046	14.498
VFRP	0,25	0,21	1.081	1.418	2.393	4.087	7.562	18.133
AP	0,46	0,56	202	146	246	420	777	1.864
DCo	0,45	0,61	149	112	189	322	596	1.429
DCa	0,16	0,14	1.658	3.363	5.674	9.692	17.934	43.006

h^a_2: herdabilidade genômica; r_{gg_gws}: acurácia seletiva da GWS; N_{QTL}: estimativa do número de QTL que controla a característica; PF: produção de frutos; IF: incidência de ferrugem; IC: incidência de cercospora; BM: infestação de bicho-mineiro; VV: vigor vegetativo; CM: ciclo de maturação; TF: tamanho dos frutos; CF: comprimento de uma folha; CRP: comprimento de um ramo plagiotrópico; NNV: número de nós vegetativos; NFRP: número de frutos por ramo plagiotrópico; VFRP: volume de frutos por ramo plagiotrópico; AP: altura de planta; DCo: diâmetro da copa; DCa: diâmetro do caule(2017).

Os resultados evidenciaram uma relação inversamente proporcional entre a acurácia seletiva da GWS (r_{gg_gws}) e número de QTL. Esse fato pode ser explicado pelo aumento da complexidade preditiva em função do maior do número de genes que controlam a característica. Quando muitos genes estão envolvidos no controle de uma característica, em geral, tais genes são de efeito menor, e, consequentemente, a estimação precisa de seus efeitos é dificultada (Goddard, 2009). Isso evidencia a importância da utilização de

alta densidade de marcadores SNP nas análises preditivas. Em estudos com espécies florestais (Grattapaglia; Resende, 2011; Iwata et al., 2011) e em milho (*Zea mays*) (Riedelsheimer et al., 2012), não foi possível verificar relação entre número de QTL e acurácia seletiva.

Observa-se a necessidade de avaliar um maior número de indivíduos quando se almejam maiores estimativas de acurácia seletiva (r_{ggd}) (Tabela 1). Para se obter estimativa de acurácia seletiva de 0,7, considerada de alta magnitude (Resende; Duarte, 2007), os resultados demonstraram a necessidade de se avaliar de 322 (diâmetro da copa) a 13.057 (comprimento de uma folha) indivíduos. Para a maioria das características, há a necessidade de se avaliar mais de 1.000 indivíduos para se obter r*ggd* igual a 0,7.

Em *C. arabica*, análises da capacidade preditiva, utilizando-se diferentes densidades de marcadores, evidenciaram, de maneira geral, o aumento da acurácia seletiva (r_{gg}) ao se elevar o número de SNP nas análises (Sousa et al., 2019). Assim, como observado em outros estudos, o aumento do número de marcadores não aumenta linearmente a acurácia da GWS (Fernando et al., 2007; Cavalcanti et al., 2012). Esses resultados demonstraram a necessidade de avaliar um maior número de genótipos, de forma a aumentar as acurácias seletivas das características de interesse

Café canéfora

Material genético

A população de *C. canephora* foi composta por 51 clones de café conilon, 32 clones de café robusta e 82 híbridos (conilon versus robusta) (Alkimim et al., 2020). Os clones de café conilon foram obtidos do Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (Incaper) e os clones de café robusta, do Centro Agronómico Tropical de Investigación y Enseñanza (Catie). Por sua vez, os híbridos foram obtidos por meio de cruzamentos controlados entre cinco clones de café conilon (genitores masculinos) e cinco clones de café robusta (genitores femininos), em delineamento fatorial interpopulacional. Essa população compõe o programa de melhoramento da Epamig em parceria com a UFV e a Embrapa Café.

Fenotipagem e genotipagem

A avaliação de cinco características – duas contínuas (altura da planta e diâmetro da copa) e três categóricas (vigor vegetativo, incidência de ferrugem e incidência de cercospora) – foi realizada durante 3 anos consecutivos (2014 a 2016), na época de maturidade fisiológica dos frutos (Alkimim et al., 2020).

O vigor vegetativo foi avaliado com notas de 1 a 10, em que a nota 1 foi atribuída às plantas totalmente depauperadas e a nota 10 foi atribuída às plantas altamente vigorosas. A incidência de ferrugem, causada pelo fungo *Hemileia vastatrix* Berk. & Br., foi avaliada com notas de 1 a 5, com a atribuição da nota 1 para plantas sem nenhum sintoma do patógeno e da nota 5 para plantas altamente suscetíveis. A incidência de cercospora, causada pelo fungo *Cercospora coffeicola* Berk. & Cooke, foi avaliada com notas de 1 a 5, em que a nota 1 foi atribuída às plantas sem nenhum sintoma do patógeno e a nota 5 foi atribuída às plantas altamente suscetíveis.

Foram genotipados 165 genótipos (51 cafés conilon + 32 cafés robusta + 82 híbridos interpopulacionais) utilizando-se a plataforma de sequenciamento da empresa RAPiD Genomics. Foram identificados 18.111 marcadores SNP, restando, após análises de qualidade, 14.429 marcadores.

Número de locos de característica quantitativa e tamanho amostral

O número estimado de QTLs que controlam as características avaliadas variou de 35 (incidência de cercospora) a 87 (vigor vegetativo) (Tabela 2). Por meio desses resultados, é possível verificar a natureza quantitativa das características avaliadas. Além disso, observa-se que os menores valores de acurácia da GWS, 67% (vigor vegetativo) e 68% (altura da planta), foram obtidos para as características que apresentaram maior número de QTL.

Tabela 2. Estimativas do número de indivíduos (N) a serem avaliados para se obter determinada (0,5; 0,6; 0,7; 0,8; 0,9) acurácia seletiva para as cinco características morfoagronômicas avaliadas na população de melhoramento de *Coffea canephora*.

r_{ggd}	VV	IF	IC	AP	DC
0,5	67	33	27	64	31
0,6	114	56	46	108	52
0,7	194	96	78	184	89
0,8	360	178	145	341	164
0,9	863	426	347	817	394
N_{QTL}	87	37	35	69	49
h^2_a	0,43	0,37	0,43	0,36	0,53
r_{gg_gws}	0,67	0,79	0,82	0,68	0,8

r_{ggd}: acurácia desejada da GWS; N_{QTL}: número de QTLs; h^2_a: herdabilidade genômica; r_{gg_gws}: acurácia seletiva da GWS; VV: vigor vegetativo; IF: incidência de ferrugem; IC: incidência de cercospora; AP: altura da planta; DC: diâmetro da copa.

Em trabalho realizado com dendê (*Elaeis guineensis* Jacq.), foi também verificado que a acurácia da GWS é inversamente proporcional ao número de QTL que controla a característica (Wong; Bernardo, 2008). Isso é esperado porque, quando a característica é governada por maior número de QTL, mais complexa ela é. Além disso, em características poligênicas, em geral, verifica-se que cada gene contribui com pequeno efeito para a manifestação do fenótipo. No entanto, mesmo em cenário com baixa herdabilidade e maior número de QTL, os resultados foram promissores.

Para se obter acurácia desejada (r_{ggd}) de 70%, valor considerado de alta magnitude (Resende; Duarte, 2007), seria necessário avaliar 194 indivíduos para a característica vigor vegetativo, 96 para incidência de ferrugem, 78 para incidência de cercospora, 184 para altura da planta e 89 para diâmetro da copa. Dessa forma, observa-se que, para a maioria das características (incidência de ferrugem, incidência de cercospora e diâmetro da copa), foram avaliados mais indivíduos do que o necessário para alcançar acurácia de 70%. Também foi possível verificar para todas as características que quanto maior a acurácia desejada, maior número de indivíduos deve ser avaliado.

Detecção de genes no controle de uma característica

O modelo tradicional para a GWAS é: , em que $y = Xb + Ts + Jm_i + Zg + e$, em que y, b, s, m_i e g são vetores de dados, de efeitos fixos de natureza ambiental, de covariável de efeitos fixos referente à estrutura de população, de efeito fixo do marcador i e de efeitos aditivos poligênicos genômicos (aleatórios), respectivamente, com matrizes de incidência X, T, J e Z.

Sob esse modelo, a marca *i* deve ser excluída da composição da matriz G (a qual é computada à parte e não via modelagem) usada na estimação de *g*. Assim, se há 50 mil marcadores, deverão ser realizadas 50 mil análises de G-BLUP e a matriz G deverá ser construída 50 mil vezes. O cômputo de G sem a inclusão do marcador candidato visa evitar o ajuste duplo ou contaminação proximal.

Um modelo alternativo é o modelo da GWS, dado por $y = Xb + Ts + Wm_i + e$, em que m é o vetor de efeitos aleatórios de marcadores. Sob esse modelo, propõe-se a seguir um teste estatístico para os efeitos aleatórios de cada marca. Essa abordagem demanda a realização de apenas uma análise G-BLUP ou RR-BLUP e possui propriedade estatística ótima, pois não supõe herdabilidade da média de marcador igual a 1, suposição essa implícita quando os efeitos m_i são tratados como fixos.

O teste estatístico baseado em F_i (descrito no tópico seguinte) só é valido sob um modelo com efeitos aleatórios de marcas e predição simultânea de todos os efeitos m_i. Esses efeitos podem ser preditos pelo RR-BLUP. Sendo tratados como de efeitos aleatórios, há a consideração do desequilíbrio de ligação e contorna-se a multicolinearidade entre os marcadores, garantindo-se a estimabilidade; isso é feito regressando cada marcador, um na direção do outro e todos na direção de zero, conduzindo a uma estimação precisa, acurada e não viesada, ou seja, com mínimo erro quadrático médio.

Proposição de um teste *F* para a detecção simultânea de genes no controle de uma característica

O seguinte teste F foi proposto por Resende et al. (2014) e Resende (2015) para a detecção simultânea de genes no controle de um característica

$$F_i = 1 + \frac{N\,2p_iq_i\tilde{m}_i^2}{\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2} = 1 + \frac{Nh_{mi}^2}{1-h^2} \approx \chi^2$$, em que $h_{mi}^2 = \frac{2p_iq_i\tilde{m}_i^2}{\sigma_y^2}$ e $h^2 = \frac{\sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2}{\sigma_y^2}$,

sendo p_i e q_i as frequências alélicas do marcador *i*, σ_y^2 a variância fenotípica, $\tilde{m}_i^2$ o quadrado do efeito predito do marcador *i*, como efeito aleatório pelo método RR-BLUP (Resende et al., 2014; Resende, 2015).

A derivação da Estatística de Teste advém da relação entre o teste *F* e o quadrado da acurácia, dado por $F = \frac{1}{(1-r_{gg}^2)}$, conforme demonstrado por Resende e Duarte (2007), usando esperança de quadrados médios e dados fenotípicos. De forma análoga, para marcadores, tem-se $F(\tilde{m}_i) = \frac{1}{(1-r_{mi\tilde{m}i}^2)}$ (Resende et al., 2014; Resende, 2015).

A confiabilidade $r_{mi\tilde{m}i}^2$ é dada por:

$$r_{mi\tilde{m}}^2 = \frac{2p_iq_i\tilde{m}_i^2}{2p_iq_i\tilde{m}_i^2 + (\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2)/N} = \frac{1}{1 + \frac{(\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2)}{2p_iq_i\tilde{m}_i^2}\frac{1}{N}} = \frac{1}{1 + \frac{(\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2)}{N\,2p_iq_i\tilde{m}_i^2}}.$$

Assim, $F(\tilde{m}_i) = \frac{1}{(1-r_{mi\tilde{m}i}^2)} = \frac{(\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2) + N\,2p_iq_i\tilde{m}_i^2}{\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2} = 1 + \frac{N\,2p_iq_i\tilde{m}_i^2}{\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2}.$

$F(\tilde{m}_i) = 1 + \frac{N\,2p_iq_i\tilde{m}_i^2}{\sigma_y^2 - \sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2} = 1 + \frac{Nh_{mi}^2}{1-h^2} \approx \chi^2$ tem distribuição assintótica qui-quadrado e, nesse caso, propicia um LRT para o efeito da marca, em que

$h_{mi}^2 = \frac{2p_iq_i\tilde{m}_i^2}{\sigma_y^2}$ e $h^2 = \frac{\sum_{i=1}^{n} 2p_iq_i\tilde{m}_i^2}{\sigma_y^2}$.

A estatística $F(\tilde{m}_i)$ está associada a (1, N – 1) graus de liberdade. Na prática da GWAS, com m_i de efeito fixo, o nível de significância a ser adotado deve ser bem menor, da ordem de 10^{-5} (F = 18.80). Na ausência de qualquer efeito real do SNP, a estatística *F* tem valor esperado igual a 1.

A estatística F_{mi} está associada a (1, N – 1) graus de liberdade. A estatística *F* tem distribuição assintótica (com N grande) qui-quadrado e, nesse caso, propicia um LRT (essa estatística tem distribuição qui-quadrado com 1 grau de liberdade) para o efeito da marca *i*. Decorrente dessas propriedades, há as equivalências entre distribuição qui-quadrado com 1 grau de liberdade e distribuição F com (1, N – 1) graus de liberdade. Por exemplo, para significância 5%, os valores tabelados para ambas as distribuições são 3.84. Nesse caso, os testes são equivalentes.

A derivação via PEV é feita a seguir.

Com $m \sim N(0, I\sigma_m^2)$, os efeitos em *m* são não correlacionados e a PEV é dada por: $PEV_{\tilde{m}_i} = (1 - r^2_{m\tilde{m}})2p_iq_i\tilde{m}_i^2$. Sendo $F(\tilde{m}_i) = (2p_iq_i\tilde{m}_i^2)/PEV_{\tilde{m}_i} = \dfrac{1}{(1 - r^2_{mi\tilde{m}i})}$, em que $2p_iq_i\tilde{m}_i^2$ é a variância genética do loco *i*. Comparando

$F(\hat{m}_i + \bar{e}) = 1 + \dfrac{N\,"\,h^2_{m}\,"}{1-"\,h^2_{m}\,"}$ para o modelo fixo (Resende, 2015) e

$F(\tilde{m}_i) = 1 + \dfrac{Nh^2_{mi}}{1-h^2}$ para o modelo aleatório, verifica-se que $F(\tilde{m}_i) > F(\hat{m}_i + \bar{e})$ e, portanto, o modelo aleatório tem maior probabilidade de detectar significância do efeito do marcador.

A qualidade desse estimador (F_i) pode ser avaliada por meio de simulação, por exemplo, considerando o cenário herdabilidade individual igual a 0,30 e 100 QTL controlando a característica. Nesse caso, considerando a simulação com 10 réplicas, 1.000 indivíduos genotipados para 20.000 marcadores, sendo 100 correspondentes a QTL, obtiveram-se os resultados apresentados nas Tabelas 3 e 4.

Tabela 3. Poder, significância (p-valor) e número de QTL detectado considerando o cenário herdabilidade individual igual a 0,30 e 100 QTL controlando a característica.

p-valor corte	N_m*	N_{QTL}*	N_{QTL}	N_{QTL}/N_m*	Poder
1%	9,9	0,4	9,7	0,98	0,1
2%	26,3	1,4	22,4	0,85	0,22
3%	46,4	2,2	36,9	0,8	0,37
4%	64,6	3,3	45,5	0,7	0,46
5%	87,5	4,5	53,9	0,62	0,54
10%	216,6	11,7	79,1	0,37	0,79

N_m*: número de marcadores declarados como associados a efeitos genéticos significativos na característica; N_{QTL}*: número de marcadores significativos, os quais correspondem aos próprios QTL; N_{QTL}: número total de QTL com efeitos significativos considerando os próprios QTL somados aos QTL carregados pelos demais marcadores significativos, dado um desequilíbrio de ligação (r^2) = 0,90. Poder: $N_{QTL}/100$.

Tabela 4. Poder, significância (p-valor) e número de QTL detectado considerando o cenário herdabilidade individual igual a 0,30 e 100 QTL controlando a característica.

p-valor corte	N_m*	N_{QTL}*	N_{QTL}	N_{QTL}/N_m*	Poder
1%	10,3	0,4	17,8	1,73	0,18
2%	27,3	1,4	37,5	1,37	0,38
3%	47,9	2,2	57,9	1,21	0,58
4%	67,8	3,3	67,6	1	0,68
5%	92,3	4,5	74,6	0,81	0,75
10%	230,9	11,7	87,7	0,38	0,88

N_m*: número de marcadores declarados como associados a efeitos genéticos significativos na característica; N_{QTL}*: número de marcadores significativos, os quais correspondem aos próprios QTL; N_{QTL}: número total de QTL com efeitos significativos considerando os próprios QTL somados aos QTL carregados pelos demais marcadores significativos, dado um desequilíbrio de ligação (r^2) = 0,81. Poder: $N_{QTL}/100$.

Dentre os parâmetros das Tabelas 3 e 4, os mais importantes são o poder e o nível de significância (p-valor). Esses dois parâmetros medem a confiabilidade (1 – p-valor) estatística da análise e a capacidade (poder) de se de-

tectar significância. De grande importância prática para o melhoramento é a identificação de marcadores que são genes ou QTL somados àqueles genes (alelos) que são carregados ou arrastados pelos marcadores, quando o desequilíbrio de ligação (r^2) é alto. Embora um alto poder tenha sido obtido apenas com p-valor igual a 10%, na simulação com $r^2 = 0,90$, a combinação p-valor = 10% e poder = 79% (Tabela 3) são adequados estatisticamente, podendo ser adotado o teste F_i nessas condições. Nesse caso, interpreta-se, com boa precisão, que o método foi adequado, permitindo identificar muitos genes e/ou blocos gênicos controlando a característica.

Entretanto, um desequilíbrio de ligação (r^2) = 0,81 está associado a uma alta acurácia de 90% e, portanto, a interpretação baseada na Tabela 4 e na Figura 3 parece mais adequada, ou seja, menos rígida sem perder precisão. Com base nessa tabela, o poder obtido com p-valor igual a 10% é de 88% (Figura 3), permitindo recuperar quase todos os genes e/ou blocos gênicos associados à característica. Verifica-se também que com p-valor igual a 5% já se obtém poder de 75%, mostrando uma contribuição balanceada de significância e poder. Assim, as inferências com base na Tabela 4 são preferíveis.

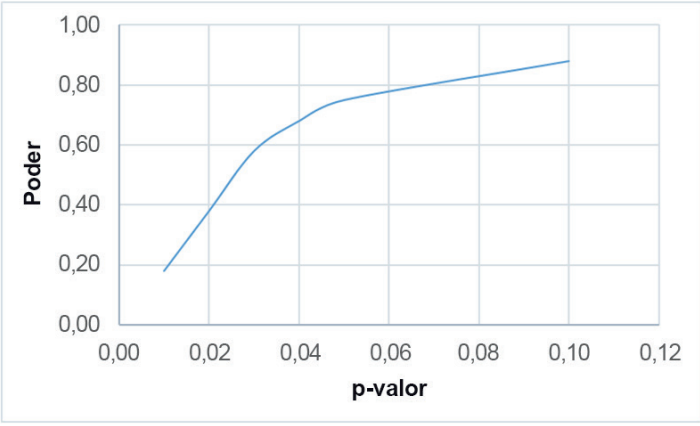

Figura 3. Comportamento do poder estatístico de detecção de genes em função do nível de significância (p-valor usado como ponto de corte para significância).

Outra informação importante refere-se ao número de marcadores significativos (N_m*) que serão selecionados visando carregar os genes identificados. Para o caso da Tabela 4, foram necessários 92 marcadores para capturar 75 genes, gerando uma relação N_{QTL}/N_m* igual a 0,81. O valor ideal dessa relação é em torno de 1, significando que, em média, cada marcador carrega um QTL. Na Tabela 4, os valores dessa relação variaram de 0,81 a 1,37 para

p-valores variando de 5% a 2%.

Verifica-se que 81% dos marcadores declarados como significativos capturaram QTL e apenas 19,2% [(92,3-74,6)/92,3] foram inúteis. Os resultados discutidos até aqui são úteis para GWS, via uso da LD-MAS (seleção assistida por marcadores via desequilíbrio de ligação), que se refere à seleção genômica por meio do uso apenas dos genes/blocos gênicos identificados com significância estatística. Este tipo de seleção pode ser comparado com a GWS tradicional.

Outro ponto de vista dos resultados das Tabelas 3 e 4 é a própria identificação de genes diretamente, sem o uso do desequilíbrio de ligação (LD). Nesse caso, em ambas as tabelas verifica-se que níveis de significância iguais ou superiores a 2% devem ser adotados para que se detecte pelo menos um gene afetando a característica. Isso pode ser visto na coluna N_{QTL}* para ambos os níveis de LD (0,90 e 0,81). Nesse caso, o interesse é a descoberta de genes para uso direto (via retrocruzamentos, por exemplo) ou para clonagem em vetores para uso em transgenia.

Tamanho amostral adequado para detecção de QTL de efeito fixo

Os livros de estatística fornecem a expressão para cálculo do tamanho amostral (N) adequado (Snedecor; Cochran, 1967; Steel; Torrie, 1980): $N = \frac{(z_\alpha + z_\beta)^2 \sigma_D^2}{\delta^2}$, em que z_∂ e z_β são valores da função distribuição acumulada da distribuição normal padrão, associados às probabilidades de erro tipo I (α) e erro tipo II (β), em teste de hipótese unilateral ou $N = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma_D^2}{\delta^2}$ para testes bilaterais, em que σ_D^2 é a variância da diferença entre duas médias de tratamentos; δ é o tamanho da diferença verdadeira entre duas médias, que se deseja discriminar como significativa. A quantidade $(1-\beta)$ é a probabilidade de que o experimento exiba uma diferença estatisticamente significativa entre médias de tratamentos. Probabilidades de 0,80 e 0,90 são comuns e adequadas na prática. A variância σ_D^2 é função da variância residual e δ^2 pode ser tomada como o quadrado do contraste entre um efeito e o centro de massa zero.

No contexto da GWAS tradicional com a análise de uma marca (de efeito fixo) por vez tem-se que $N = \frac{(z_{(1-\alpha/2)} + z_{(1-\beta)})^2 \sigma_D^2}{\delta^2} = \frac{(z_{(1-\alpha/2)} + z_{(1-\beta)})^2 \sigma^2 / Var(W_i)}{m_i^2} = \frac{(z_{(1-\alpha/2)} + z_{(1-\beta)})^2 \sigma^2}{2 p_i q_i m_i^2}$

em que $\sigma^2 = (1 - "h_{mi}^2")\sigma_y^2$ e, portanto, $N = (Z_{(1-\alpha/2)} + Z_{(1-\beta)}) \frac{1 - "h_{mi}^2"}{"h_{mi}^2"}$.

Assumindo $\sigma_y^2 - 2 p_i q_i m_i^2 \approx \sigma_y^2$, tem-se que $1 - "h_{mi}^2" = 1$ e $N \approx \frac{(Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{"h_{mi}^2"}$.

Com r^2 menor que 1 (baixa densidade de marcas), tem-se: $N = (Z_{(1-\alpha/2)} + Z_{(1-\beta)}) \frac{1 - r^2 "h_{mi}^2"}{r^2 "h_{mi}^2"}$ e $N \approx \frac{(Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{r^2 "h_{mi}^2"}$.

Verifica-se na Tabela 5 e Figura 4 que os tamanhos amostrais (< 1.000) usados corriqueiramente no melhoramento vegetal e animal somente vão detectar QTLs quando o QTL explicar 5% ou mais da variação fenotípica, fato que é pouco provável sob herança poligênica (h^2 total < 0,50). O poder 0,90 é mais adequado, pois leva a 81% ($0,90^2$) de probabilidade de que dois estudos independentes detectem o mesmo QTL.

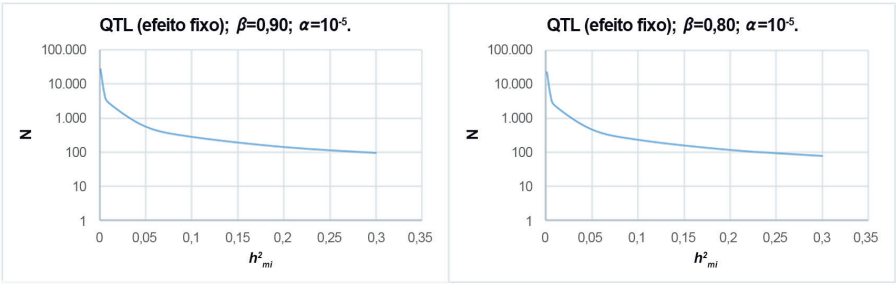

Figura 4. Tamanho amostral (N) necessário para a detecção de efeitos genéticos de marcadores (assumidos como de efeitos fixos) com diferentes herdabilidades (h^2_{mi}): valores de N como função de h^2_{mi}. Os valores de N plotados foram obtidos via transformação logarítmica para melhorar a visualização..

Tabela 5. Tamanho amostral (N) e poder para detecção de nível significância 10^{-5} de acordo com a magnitude ("h^2_{mi}") do QTL (considerado de efeito fixo): $N \approx \frac{(Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2}{"h^2_{mi}"}$.

Z para β =0,90	Z para $\alpha=10^{-5}$	$Z_{(1-\alpha/2)}+Z_{(1-\beta)}{}^2$	h^2_{mi}	N	Z para β =0,80	Z para $\alpha=10^{-5}$	$Z_{(1-\alpha/2)}+Z_{(1-\beta)}{}^2$	h^2_{mi}	N
1,28	3,99	27,78	0,001	27.773	0,84	3,99	23,33	0,001	23.329
1,28	3,99	27,78	0,005	5.555	0,84	3,99	23,33	0,005	4.666
1,28	3,99	27,78	0,01	2.777	0,84	3,99	23,33	0,01	2.333
1,28	3,99	27,78	0,05	555	0,84	3,99	23,33	0,05	467
1,28	3,99	27,78	0,1	278	0,84	3,99	23,33	0,1	233
1,28	3,99	27,78	0,2	139	0,84	3,99	23,33	0,2	117
1,28	3,99	27,78	0,3	93	0,84	3,99	23,33	0,3	78

Tamanho amostral adequado para detecção de QTL de efeito aleatório

No caso de marcas de efeitos aleatórios, $N = (Z_{(1-\alpha/2)} + Z_{(1-\beta)})\frac{1-"h^2_{mi}"}{"h^2_{mi}"}$ muda para $N = \frac{(Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2(1-h^2)}{h^2_{mi}}$.

Verifica-se na Tabela 6 e na Figura 5 que considerar marcadores como de efeitos aleatórios demanda tamanho amostral muito menor (que a análise de marca única como efeito fixo) para atingir o mesmo poder.

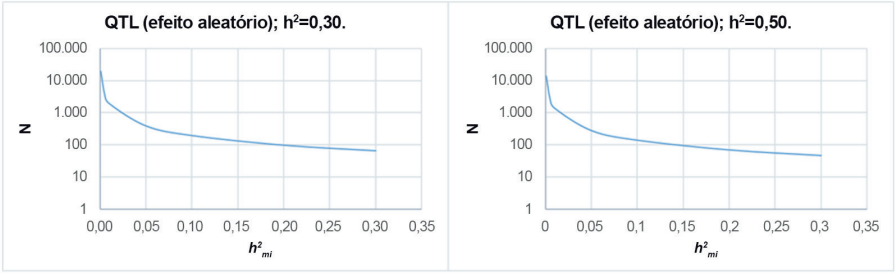

Figura 5. Tamanho amostral (N) necessário para a detecção de efeitos genéticos de marcadores (assumidos como de efeitos aleatórios) com diferentes herdabilidades (h^2_{mi}): valores de N como função de h^2_{mi}. Os valores de N plotados foram obtidos via transformação logarítmica para melhorar a visualização.

Tabela 6. Tamanho amostral (N) e poder para detecção de nível significância 10^{-5} de acordo com a magnitude ("h^2_{mi}") do QTL (considerado de efeito fixo): $N = \frac{(Z_{(1-\alpha/2)} + Z_{(1-\beta)})^2 (1 - h^2)}{h^2_{mi}}$.

h^2=0,30						h^2=0,50				
Z para β =0,90	Z para $\alpha=10^{-5}$	$Z_{(1-\alpha/2)}+Z_{(1-\beta)}{}^2$	h^2_{mi}	N		Z para β =0,90	Z para $\alpha=10^{-5}$	$Z_{(1-\alpha/2)}+Z_{(1-\beta)}{}^2$	h^2_{mi}	N
1,28	3,99	27,78	0,001	19.441		1,28	3,99	27,78	0,001	13.886
1,28	3,99	27,78	0,005	3.888		1,28	3,99	27,78	0,005	2.777
1,28	3,99	27,78	0,01	1.944		1,28	3,99	27,78	0,01	1.389
1,28	3,99	27,78	0,05	389		1,28	3,99	27,78	0,05	278
1,28	3,99	27,78	0,1	194		1,28	3,99	27,78	0,1	139
1,28	3,99	27,78	0,2	97		1,28	3,99	27,78	0,2	69
1,28	3,99	27,78	0,3	65		1,28	3,99	23,33	0,3	46

Considerações finais

Para a maioria das características de café arábica, faz-se necessária a avaliação de mais de 1.000 indivíduos para obter acurácia igual a 0,70. Por sua vez, para café canéfora, como o número estimado de QTL variou de 35 a 87, cerca de 200 indivíduos é um número adequado para obter acurácia igual a 0,70.

Na simulação referente ao tamanho amostral, verificou-se que, para uma característica com herdabilidade individual igual a 0,30 e controlada por 100 QTL, pode ser obtida uma acurácia igual 90% se o tamanho amostral for igual a 1.500 indivíduos genotipados e fenotipados. Na simulação referente ao número de QTL, verificou-se que, para N = 1.500 indivíduos genotipados, os números de QTL variam de 49 a 468 (cenário 1, h^2 = 0,30) e de 32 a 312 (cenário 2, h^2 = 0,20), quando as acurácias variaram de 0,95 a 0,70, respectivamente.

Os resultados referentes ao novo teste *F* proposto para marcas de efeitos aleatórios revelaram que, dentre os parâmetros avaliados, os mais importantes são o poder e o nível de significância (p-valor). Esses dois parâmetros medem a confiabilidade (1 – p-valor) estatística da análise e a capacidade (poder) de se detectar significância.

Os resultados obtidos neste trabalho permitirão a orientação dos programas de melhoramento do cafeeiro. Os estudos de detecção de genes em características quantitativas do cafeeiro guiarão a seleção de genótipos superiores portadores de genes que governam as principais características agronômicas, utilizando-se as novas metodologias otimizadas.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig), ao Instituto Nacional de Ciência e Tecnologia do Café (INCT Café) e à Embrapa Café pelo apoio financeiro.

Referências

ALKIMIM, E. R.; CAIXETA, E. T.; SOUSA, T. V.; RESENDE, M. D. V. de; SILVA, F. L.; SAKIYAMA, N. S.; ZAMBOLIM, L. Selective efficiency of genome-wide selection in *Coffea canephora* breeding. **Tree Genetics & Genomes**, v. 16, article number 41, 2020. DOI: 10.1007/s11295-020-01433-3.

CAVALCANTI, J. J. V.; RESENDE, M. D. V. de; SANTOS, F. H. C. dos; PINHEIRO, C. R. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, v. 34, n. 3, p. 840-846, set. 2012.

FERNANDO, R. L.; HABIER, D.; STRICKER, C.; DEKKERS, J. C. M.; TOTIR, L. R. Genomic selection. **Acta Agriculturae Scandinavica, Section A - Animal Science**, v. 57, n. 4, p. 192-195, 2007. DOI: 10.1080/09064700801959395.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, v. 136, p. 245-257, 2009. DOI: 10.1007/s10709-008-9308-0.

GRATTAPAGLIA, D.; RESENDE, M. D. V. de. Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, v. 7, n. 2, p. 241-255, Apr. 2011. DOI: 10.1007/s11295-010-0328-4.

IWATA, H.; HAYASHI, T.; TSUMURA, Y. Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. **Tree Genetics & Genomes**, v. 7, p. 747-758, 2011. DOI: 10.1007/s11295-011-0371-9.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, Apr. 2001. DOI: 10.1093/genetics/157.4.1819.

RESENDE, M. D. V. de. **Genética quantitativa e de populações**. Viçosa, MG: UFV, 2015. 463 p.

RESENDE, M. D. V. de; ALVES, R. S. Linear, generalized, hierarchical, bayesian and random regression mixed models in genetics/genomics in plant breeding. **Functional Plant Breeding Journal**, v. 2, n. 2, article 1, July/Dec. 2020. DOI: 10.35418/2526-4117/v2n2a1.

RESENDE, M. D. V. de; DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182-194, set. 2007.

RESENDE, M. D. V. de; SILVA, F. F. e; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional**: modelos mistos, multivariados, categorias e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência. Viçosa, MG: UFV, 2014. 882 p.

RIEDELSHEIMER, C.; CZEDIK-EYSENBERG, A.; GRIEDER, C.; LISEC, J.; TECHNOW, F.; SULPICE, R.; ALTMANN, T.; STITT, M.; WILLMITZER, L.; MELCHINGER, A. E. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. **Nature Genetics**, v. 44, n. 2, p. 217-220, Feb. 2012. DOI: 10.1038/ng.1033.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical methods**. 6. ed. Iowa: Iowa State University Press, 1967. 593 p.

SOUSA, T. V.; CAIXETA, E. T.; ALKIMIM, E. R.; OLIVEIRA, A. C. B. de; PEREIRA, A. A.; SAKIYAMA, N. S.; ZAMBOLIM, L.; RESENDE, M. D. V. de. Early selection enabled by the implementation of genomic selection in coffea arabica breeding. **Frontiers in Plant Science**, v. 9, article 1934, 2019. DOI: 10.3389/fpls.2018.01934.

STEEL, R.G.D.; TORRIE, J.H. **Principles and procedures of statistics**: a biometrical approach. 2nd. New York: McGraw-Hill Book, 1980. 633 p.

WONG, C. K.; BERNARDO, R. Genome wide selection in oil palm: increasing selection gain per unit time and cost with small populations. **Theoretical and Applied Genetics**, v. 116, n. 6, p. 815–824, Apr. 2008. DOI: 10.1007/s00122-008-0715-5.

Exemplares desta edição podem ser adquiridos na:

Embrapa Café
Parque Estação Biológica (PqEB)
Av. W3 Norte (final), Ed. Sede
CEP: 70770-901, Brasília, DF
Fone: +55 (61) 3448-4378 / 4010
Fax: +55 (61) 3448-1797
www.embrapa.br
www.embrapa.br/fale-conosco/sac

1ª edição
Publicação digital (2022)

Comitê Local de Publicações da Embrapa Café

Presidente
Lucas Tadeu Ferreira

Vice-Presidente
Jamilsen de Freitas Santos

Secretária-Executiva
Adriana Maria Silva Macedo

Membros

Anísio José Diniz, Carlos Henrique Siqueira de Carvalho, Helena Maria Ramos Alves, Lucilene Maria de Andrade, Mauricio Sergio Zacarias, Milene Alves de Figueiredo Carvalho, Omar Cruz Rocha, Rogério Novais Teixeira, Roseane Pereira Villela

Revisão de texto
Everaldo Correia da Silva Filho

Normalização bibliográfica
Maria de Fátima da Cunha (CRB-1/2616)

Tratamento das ilustrações
Thiago Farah Cavaton

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Thiago Farah Cavaton

Foto da capa
Dos autores

Embrapa
MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO
PÁTRIA AMADA BRASIL GOVERNO FEDERAL

CGPE 017454