UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

Instituto de Biociências de Botucatu - Unesp

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA APLICADA AO CONTEÚDO DE MACRONUTRIENTES EM GRÃOS DE *Coffea arabica* L.

MARIANE SILVA FELICIO

BOTUCATU – SP
2020

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de Botucatu

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA APLICADA AO CONTEÚDO DE MACRONUTRIENTES EM GRÃOS DE *Coffea arabica* L.

ALUNA: MARIANE SILVA FELICIO
ORIENTADOR: PROF. DR. DOUGLAS SILVA DOMINGUES
COORIENTADOR: PROF. DR. LUIZ FILIPE PROTASIO PEREIRA

Tese apresentada ao Instituto de Biociências, Campus de Botucatu, UNESP, para obtenção do título de Doutor no Programa de Pós-Graduação em Ciências Biológicas (Genética).

BOTUCATU – SP
2020

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Felicio, Mariane Silva.
Estudo de associação genômica ampla aplicada ao conteúdo de macronutrientes em grãos de Coffea arabica L. / Mariane Silva Felicio. - Botucatu, 2020

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Douglas Silva Domingues
Coorientador: Luiz Filipe Protasio Pereira
Capes: 20203004

1. Café. 2. Genômica. 3. Estresse vegetal. 4. Critica de imputação de dados (Estatística). 5. Marcadores genéticos. 6. Ausência de dados (Estatística).

Palavras-chave: Dados ausentes; Espécies não-modelo; Estresses bióticos e abióticos; Imputação.

“Contudo, seja qual for o grau a que chegamos, o que importa é prosseguir decididamente. ”
(Fl 3, 16)

“Não vos conformeis com este mundo, mas transformai-vos pela renovação do vosso espírito, para que possais discernir qual é a vontade de Deus, o que é bom, o que lhe agrada e o que é perfeito. ”
(Rm 12, 2)

Dedico à Deus, minha fortaleza. Ao meu pai Benedito Felicio Filho (*in memoriam*) que sempre me incentivou a buscar meus sonhos.

Agradecimentos

À Deus, por todas as graças recebidas e por nunca me abandonar.

À minha família, meu pai Benedito (*in memoriam*) que sempre me incentivou a estudar e foi um grande exemplo para mim. À minha mãe Maria Izabel, pelo suporte para realizar esse curso. À minha irmã Marina, por todo amor e paciência que tem comigo e teve durante o desenvolvimento desse trabalho. Ao meu avô Benedito, um grande homem, meu melhor amigo e conselheiro.

Ao meu orientador Dr. Douglas Silva Domingues, pela amizade, por me confiar esse trabalho, me incentivar a buscar a capacitação em diferentes cursos, por todos os conselhos e pela orientação valiosa durante o mestrado e doutorado.

Ao meu coorientador Dr. Luiz Filipe Protasio Pereira, pela amizade, pelos ensinamentos, por todas as discussões que contribuíram para o andamento do trabalho e também por me coorientar desde a graduação.

Ao Dr. Gustavo Sant'Ana por todos os conhecimentos transmitidos e auxílio para a realização das análises de dados.

Aos amigos Lucinéia Maria da Silva e Manuel Luiz Martins, pela amizade, conselhos e dedicação às coletas que foram fundamentais para a realização desse trabalho.

Ao Dr. Gabriel Rodrigues Alves Margarido, e seus alunos Amanda Avelar, Lorena Guimarães Batista, Fernando Henrique Correr e Guilherme Kenichi Hosaka, por me receberem tão bem no Laboratório de Bioinformática Aplicada a Bioenergia e por todos os conhecimentos que me passaram que foram essenciais para o desenvolvimento desse trabalho.

Ao Leandro Carrijo Cintra, por me auxiliar com o necessário para a realização das análises de Bioinformática utilizando o servidor da EMBRAPA.

Aos funcionários e estagiários do laboratório de solos e tecido vegetal do IAPAR, especialmente à Rosineia Aparecida de Souza pelos ensinamentos transmitidos e dedicação para realizar as análises de nutrientes.

Aos Drs. Eduardo Fermino Carlos, Nelson da Silva Fonseca, Paula Cristina da Silva Ângelo, Leandro Simões Azeredo Gonçalves pela amizade e conhecimentos transmitidos que auxiliaram para o desenvolvimento desse trabalho.

Ao servidor do IAPAR Ovidio Mantoani, pela amizade, por todos os conhecimentos passados e auxílios para a moagem dos grãos de café.

A equipe do programa de melhoramento do cafeeiro do IAPAR pela manutenção do germoplasma utilizado nesse trabalho. Especialmente ao Eugênio Brandt e Fernando Carducci

pelo auxílio para a identificação das plantas no campo e com as coletas de frutos. À Luciana Harumi pelo auxílio com as informações sobre a manutenção do banco de germoplasma.

Aos Drs. Celso Luis Marino, Eveline Teixeira Caixeta, Roberto Fritsche Neto e João Ricardo Bachega Feijó Rosa, pela participação da banca de tese e por todas as sugestões para a melhoria do trabalho.

Às amigas Darley de Souza, Aline Silveira, Nathália Caroline Rodrigues e Nara Moreira, pelo suporte, amizade e orações durante essa caminhada.

Às amigas Jordana Oliveira e Najila Nolie, por me acolherem em casa em tantas viagens para Botucatu. E a todos os amigos que me acolheram na cidade, especialmente a Bruna Jerônimo, Vanessa Jacob, Camila Moreira, Marco Soares, Adauto Cardoso, Arno Butzge, Viviani Sene e Isabel Silverio.

À amiga Jessica Delfini, por todas as conversas, partilhas, desabafos e scripts compartilhados.

Aos amigos do Laboratório de Biotecnologia Vegetal (LBI) do IAPAR, pela amizade e companheirismo. Especialmente Bruna Silvestre Rodrigues da Silva, Caroline Ariyoshi, Rafaelle Vecchia Ferreira e Lívia Maria Nogueira Brito pela parceria e auxílios para o desenvolvimento desse trabalho.

Às amigas do Laboratório de Cultura de Tecidos do IAPAR, Cícera Martimiano e Suely Ario Kudo pela amizade e por me auxiliarem sempre que precisei.

A todos os professores do Programa de Pós-Graduação em Ciências Biológicas (Genética), especialmente ao coordenador Dr. Ivan de Godoy Maia, pela dedicação para proporcionar uma formação de qualidade para tantos alunos e por todos os eventos científicos produzidos para nosso desenvolvimento como pesquisadores.

Aos funcionários da Seção técnica de Pós-Graduação, pelo trabalho de excelência, por terem me assessorado e tirado minhas dúvidas sempre que preciso.

À Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) e todos os seus funcionários, pela infraestrutura fornecida e a manutenção das atividades na universidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos que possibilitou minha dedicação exclusiva ao trabalho.

Ao Instituto Agronômico do Paraná (IAPAR) pela infraestrutura e pelo acesso às plantas utilizadas nesse trabalho.

Muito obrigada!

RESUMO

O café é uma das commodities agrícolas tropicais mais comercializadas no mundo. *Coffea arabica* é a principal espécie utilizada para a produção comercial de café. A espécie é originária da Etiópia. Ela é única espécie alotetraploide do gênero (2n = 4x = 44) e se reproduz predominantemente por autofecundação. As cultivares comerciais de *C. arabica* possuem baixa diversidade genética, o que indica a necessidade de introgressão de alelos de germoplasma para o melhoramento dessas cultivares. Acessos do centro de origem da espécie possuem maior diversidade que as cultivares comerciais e podem ser utilizados para a identificação de novos alelos. O conteúdo de macronutrientes em grãos do cafeeiro tem impacto direto na qualidade do produto. No entanto, a base molecular da composição mineral de grãos de cafeeiro ainda é pouco conhecida. Com isso, o objetivo desse trabalho foi identificar marcadores SNP possivelmente associados com a composição de macronutrientes em grãos de *C. arabica*. Para alcance deste objetivo, foram comparados três métodos de imputação de genótipos, bem como foi realizado o mapeamento associativo em estudo de associação genômica ampla (GWAS). Foi utilizado um painel de 110 genótipos de *C. arabica*, composto por genótipos elite do programa de melhoramento do Instituto Agronômico do Paraná (3), cultivares comerciais (11) e acessos selvagens (96). Foram realizadas análises da composição de cinco macronutrientes (N, P, K, Ca e Mg) em grãos de cafeeiro coletados de 70 e 105 genótipos de *C. arabica* nos anos de 2017 e 2018, respectivamente. Foram calculados valores de BLUP para 65 genótipos, para os quais foram realizadas coletas nos dois anos. Foi identificado que os acessos selvagens de *C. arabica* possuem maior variabilidade genética e de composição de macronutrientes nos grãos que as cultivares comerciais. Foram identificados mais marcadores SNPs quando o mapeamento dos dados de GBS foi realizado com o genoma de referência da cultivar Caturra (11.230 SNPs) do que com o diaploide Et39 (9.991 SNPs). O método Beagle apresentou o melhor desempenho para a imputação dos painéis de marcadores e a imputação contribuiu para reduzir a taxa de associações espúrias identificadas nas análises de GWAS. Foram identificados 5, 12, 13, 6, 6 e 1 marcadores em regiões codificantes associados a composição de N, P, K, Ca, Mg, Mg e K em grãos de *C. arabica*, respectivamente. As regiões genômicas identificadas foram relacionadas a diversas rotas metabólicas em plantas, incluindo o transporte de nutrientes, a germinação de sementes, o controle da época de florescimento e algumas vias de resposta a estresses bióticos e abióticos. As proteínas identificadas nesse trabalho podem ser utilizadas como alvos em futuros estudos para a caracterização de vias metabólicas que atuam no controle do acúmulo de macronutrientes em plantas da espécie *C. arabica*. Além disso, os genótipos

selvagens de *C. arabica* utilizados nesse trabalho podem ser utilizados para identificação de alelos favoráveis a serem introduzidos em cultivares comerciais.

Palavras-chave: imputação, dados ausentes, espécies não-modelo, estresses bióticos e abióticos.

ABSTRACT

Coffee is one of the most traded tropical commodities in the world. *Coffea arabica* is the main species used for commercial production. The species is originally from Ethiopia. In the *Coffea* genus, *C. arabica* is the only allotetraploid species (2n = 4x = 44) and it reproduces predominantly by self-fertilization. The commercial cultivars of *C. arabica* have a narrow genetic base that indicates the need for the introgression of new alleles from germplasm into coffee breeding programs. Wild accessions of *C. arabica*, from Ethiopia, have higher genetic diversity and can be used to identify new alleles. The macronutrient composition of the coffee grains has a direct impact on grain quality. However, the molecular basis for the mineral composition in coffee grains still poorly understood. Thus, the aim of this work was to perform mapping association analyses using the genome-wide association study (GWAS) technique to identify single nucleotide polymorphisms (SNPs) associated with macronutrient content in coffee grains from *C. arabica*. We also tested three imputation methods (haplotype missing allele imputation - Beagle, K-nearest neighbors, and Random Forest) in the genotypic data, and mapped it to two *C. arabica* reference genomes from the cultivar Caturra red and the spontaneous dihaploid Et39. We used a panel of 110 *C. arabica* genotypes, including elite landraces from the IAPAR coffee breeding program (3), commercial cultivars (11) and wild accessions (96). Analysis of the composition of five macronutrients (N, P, K, Ca and Mg) was carried out on coffee grains collected from 70 and 105 genotypes in the years 2017 and 2018, respectively. BLUP values were estimated for 65 genotypes in which the grains were collected in both years. Our results indicate that the *C. arabica* wild accessions have higher genetic variability and higher diversity for grains macronutrient content than the commercial cultivars. More SNPs markers were identified when the GBS data were mapped to the Caturra reference genome (11,230 SNPs) than to the Et39 reference genome (9,991 SNPs). Beagle presented the best performance in markers dataset imputation. The imputation reduced the false discovery rate in GWAS. We identified 5, 12, 13, 6, 6 and 1 markers in coding regions associated with the content of N, P, K, Ca, Mg, and Mg and K in coffee grains, respectively. The proteins identified participates in several metabolic pathways, including nutrient transport, seed germination, flowering time control and plant response to biotic and abiotic stresses. These proteins can be used as targets in further studies for the characterization of metabolic pathways controlling macronutrient accumulation in *C. arabica* plants. Also, the wild *C. arabica* genotypes used in this work can be used to identify favorable alleles to be introduced in

commercial cultivars, and for the selection of promising genotypes with altered levels of macronutrients in their grains than the observed among *C. arabica* commercial cultivars.

Keywords: imputation, missing genotypes, non-model species, biotic and abiotic stress.

Lista de figuras

Capítulo 1

Capítulo 2

Lista de tabelas

the GWAS models and the minimum and maximum values of LOD score, QTN effects, minor allele frequency (MAF), and correlation (r^2) to the associated trait estimated for each marker.

Sumário

1. INTRODUÇÃO

O café é uma das *commodities* agrícolas mais comercializadas no mundo. O produto possui grande importância para o desenvolvimento socioeconômico do Brasil, que é o principal país produtor e exportador de café no mundo. A principal espécie utilizada para a produção comercial de café é *Coffea arabica* L. (café arábica) (ICO 2020).

A espécie *C. arabica* pertence à família Rubiaceae, gênero *Coffea* (Davis et al. 2011). Estudos indicam que a espécie tenha uma origem relativamente recente, estimada entre 10 mil e 600 mil anos atrás (Yu et al. 2011; Scalabrin et al. 2020) e originada na região da Etiópia por meio da hibridação natural entre genótipos ancestrais das espécies *Coffea canephora* Pierre ex A. Froehner e *Coffea eugenioides* S. Moore (Lashermes et al. 1999). *Coffea arabica* é a única espécie alotetratetraploide natural (2n = 4x = 44) e com reprodução predominantemente autógama do gênero *Coffea*, enquanto o restante das espécies foi caracterizado como diploide e a maioria auto incompatível (Charrier e Berthaud 1985).

Embora o centro de origem do café arábica seja a região da Etiópia, o cultivo comercial dessa espécie teve início no Iêmen, região que foi classificada como o centro secundário de dispersão da espécie (Meyer et al. 1968). A partir de plantas cultivadas no Iêmen, surgiram as subpopulações Típica e Bourbon, que deram origem a maior parte das cultivares comerciais de *C. arabica* utilizadas no mundo (Anthony et al. 2002).

A forma de dispersão do cultivo de *C. arabica*, baseado em poucos genótipos, aliada a origem relativamente recente da espécie e o modo de reprodução predominantemente autógamo, contribuíram para a base genética estreita observada atualmente entre as cultivares comerciais (Anthony et al. 2002; Silvestrini et al. 2007; Setotaw et al. 2013). A baixa variabilidade genética entre cultivares comerciais representa um problema para a cultura, pois a maioria das cultivares são suscetíveis à estresses bióticos e abióticos (van der Vossen et al. 2015).

Para ampliar os recursos genéticos a serem explorados pelos programas de melhoramento de *C. arabica*, foram realizadas coletas de genótipos na região do centro de origem da espécie, a Etiópia (Meyer et al. 1968). Análises realizadas em grãos de café revelaram que os acessos selvagens da Etiópia abrigam ampla variabilidade para aspectos morfológicos e de compostos que afetam a qualidade da bebida de café (Gaspari-Pezzopane 2014; Sant'Ana et al. 2018; dos Santos Scholz et al. 2016).

A composição mineral do grão do cafeeiro tem impacto direto na qualidade do produto. No entanto, até o momento, a base genética da composição de macronutrientes em grãos de *C. arabica* ainda é pouco conhecida.

A análise de mapeamento associativo, ou mapeamento por desequilíbrio de ligação (DL), é uma alternativa que pode ser utilizada para a identificação de regiões genômicas associadas a características fenotípicas que possuem controle genético complexo, como o conteúdo de macronutrientes em grãos (Ziegler et al. 2017; Ziegler et al. 2018). Análises de mapeamento associativo que utilizam a informação de marcadores moleculares distribuídos ao longo de regiões do genoma recebem o nome de associação genômica ampla (*genome wide association study* - GWAS).

A técnica de GWAS geralmente é aplicada a populações com genótipos diversos (Hayward et al. 2015), como por exemplo, populações que incluem acessos de coleções de germoplasma (Rafalski 2010). Nessa técnica são identificados marcadores de polimorfismos de nucleotídeo único (*single nucleotide polymorphisms* - SNP) que possuem associação significativa com características fenotípicas complexas (Hayward et al. 2015).

O desenvolvimento das tecnologias de sequenciamento de nova geração contribuiu para o avanço de técnicas destinadas a identificação de marcadores SNPs. Em plantas, a técnica de genotipagem por sequenciamento (*genotyping-by-sequencing* - GBS) tem sido amplamente utilizada para essa finalidade (Rasheed et al. 2017; Nadeem et al. 2018). Nessa técnica, enzimas de restrição são utilizadas para reduzir a complexidade dos genomas, limitando o sequenciamento às regiões que flanqueiam o sítio de corte das enzimas. A ligação de adaptadores específicos às extremidades dos insertos de DNA permite a multiplexação de amostras para o sequenciamento em uma mesma reação, reduzindo assim o custo da genotipagem (Elshire et al. 2011). No entanto, um limitante para o uso de dados de GBS está na proporção de dados ausentes por marcador, que tende a ser alta quando o sequenciamento é realizado com baixa cobertura (Fu 2014; Torkamaneh et al. 2018).

Os dados ausentes podem reduzir a acurácia de análises de associação, como GWAS, contribuindo para a identificação de falsos positivos (Rahimi et al. 2019). A imputação de genótipos é uma alternativa que pode reduzir os prejuízos causados pelos dados ausentes, e consiste na substituição desses dados por genótipos prováveis (Marchini e Howie 2010; Torkamaneh et al. 2018). A imputação pode ser realizada por métodos desenvolvidos especificamente para substituir dados ausentes em painéis de marcadores ou por métodos estatísticos gerais (He et al. 2015; Nazzicari et al. 2016). No primeiro grupo é possível citar o método de imputação de alelos em haplótipos localizados, implementado no *software* Beagle (Browning e Browning 2016). Dentre os métodos estatísticos gerais estão incluídos o método da média ponderada entre os marcadores mais próximos (*K-nearest neighbor* - KNN) e o

método não paramétrico de regressões de florestas aleatórias (*Random forest* - RF) (Rutkoski et al. 2013; Nazzicari et al. 2016).

A acurácia da imputação de marcadores pode ser alterada de acordo com o método de imputação adotado, o genoma de referência utilizado para o mapeamento dos dados de GBS, a proporção de dados ausentes permitida por marcador, o número de marcadores identificados e as características intrínsecas às espécies e populações genotipadas, como por exemplo a relação de parentesco entre os acessos da população (Rutkoski et al. 2013; Torkamaneh e Belzile. 2015; Nazzicari et al. 2016).

Estudos recentes indicam que a imputação de marcadores pode contribuir para o aumento da acurácia na identificação de associações significativas em GWAS, como observado em populações de soja (*Glycine max* L., Torkamaneh e Belzile 2015) e trigo (*Triticum aestivum* L., Rahimi et al. 2019).

No primeiro trabalho de GWAS realizado com a espécie *C. arabica*, em uma população com 107 acessos provenientes do centro de origem da espécie, foram identificados 21 marcadores possivelmente associados à composição de lipídeos nos grãos, compostos que interferem na qualidade da bebida de café (Sant'Ana et al. 2018). Em trabalhos com soja e milho (*Zea mays* L.) foram identificados marcadores associados a regiões genômicas que controlam o acúmulo de nutrientes nos grãos (Ziegler et al. 2017; Ziegler et al. 2018).

Até o momento não foi publicado nenhum estudo em que o desempenho de métodos de imputação tenha sido analisado em painéis de marcadores da espécie *C. arabica*. Além disso, para a mesma espécie também não se sabe qual é o efeito da imputação de marcadores em análises de GWAS. Com isso, os objetivos desse trabalho foram: i) mapear dados de genotipagem de *C. arabica* em genomas de referência da própria espécie; ii) identificar o melhor método de imputação em uma população de 110 genótipos de *C. arabica*; iii) aplicar a técnica de GWAS para identificar regiões genômicas em associação com a composição de nutrientes em grãos; iv) verificar o efeito da imputação de marcadores sobre os resultados de GWAS.

2. OBJETIVOS

O principal objetivo do presente trabalho foi realizar a análise de mapeamento associativo para identificar marcadores moleculares do tipo SNPs associados a composição de cinco macronutrientes (N, P, K, Ca e Mg) em grãos de *C. arabica*, com os seguintes objetivos específicos:

- Mapear dados de sequenciamento obtidos pela técnica de GBS ao genoma de referência de *C. arabica*.
- Identificar o melhor método de imputação de genótipos para dados de GBS de *C. arabica*.
- Identificar marcadores associados com a composição de macronutrientes em grãos de *C. arabica* por meio de GWAS.
- Analisar o efeito do melhor método de imputação sobre análises de estrutura genética populacional e de GWAS.

3. CAPÍTULO 1 - Revisão de Literatura

3.1. Aspectos econômicos da cafeicultura

O café é produzido em mais de 70 países, dos quais a maior parte está em desenvolvimento, como o Brasil, Vietnã e Colômbia, que são os três principais produtores de café no mundo (FAO 2015; ICO 2020). A bebida de café é a segunda mais consumida no Brasil, precedida apenas pelo consumo de água (ABIC 2015). O Brasil também lidera a produção e exportação de café no mundo (ICO 2020).

A produção mundial de café para a safra de 2018/19 foi estimada em 169 milhões de sacas de 60 Kg de café beneficiado (ICO 2019a). Estima-se que a produção brasileira contribua com cerca de 28% desse montante, o equivalente a 48,99 milhões de sacas de café beneficiado (CONAB 2019).

Duas espécies principais dominam a produção comercial de café: *C. arabica* (café arábica) e *C. canephora* (café conilon ou robusta), que contribuem com cerca de 58% e 42% do café comercializado no mundo, respectivamente (ICO 2019b). No Brasil, o café arábica é produzido em maior escala, e representa cerca de 70% do total produzido no país (CONAB 2019). A bebida de café arábica é considerada de melhor qualidade, devido ao menor amargor quando comparado ao da bebida de café robusta, com isso, os grãos de café arábica possuem maior valor comercial (FAO 2015). Atualmente os principais estados produtores de café arábica em ordem decrescente de produção são: Minas Gerais, São Paulo, Espírito Santo, Bahia, Paraná e Rio de Janeiro (CONAB 2019).

As pequenas propriedades representam cerca de 70% da área destinada à produção de café no Brasil (Matiello et al. 2015). No país, a renovação das lavouras com variedades mais produtivas contribuiu para um aumento significativo na produção da safra de 2018, que foi estimada em 61,7 milhões de sacas de 60 Kg de café beneficiado (CONAB 2018).

3.2. Origem da espécie *Coffea arabica*

O café arábica pertence à família Rubiaceae, gênero *Coffea*, no qual foram descritas mais 123 espécies (Davis et al. 2011). Nesse gênero, a espécie *C. arabica* é a única tetraploide (2n = 4x = 44) com reprodução predominantemente autógama, enquanto o restante das espécies foi caracterizado como diploide e a maioria auto incompatível (Charrier e Berthaud 1985).

O conteúdo 2C de DNA cromossômico da espécie foi estimado em 2,62 pg por meio da análise de citrometria de fluxo. Esse valor corresponde a aproximadamente 1.58 Gb (Clarindo e Carvalho 2008). Além disso, recentes montagens de genomas de referência da espécie, a partir

das cultivares Caturra e Bourbon, possuem 1.09 e 1.54 Gb, respectivamente (RefSeq assembly accession: GCF_003713225.1, https://www.ncbi.nlm.nih.gov/, 2019; Scalabrin et al. 2020).

A espécie *C. arabica* é nativa da região sudoeste da Etiópia, sudeste do Sudão e norte do Quênia (Guerreiro-Filho et al. 2008). Estima-se que *C. arabica* seja uma espécie relativamente recente, formada entre 10 mil e 600 mil anos atrás (Yu et al. 2011; Scalabrin et al. 2020), por meio da hibridação natural entre as espécies ancestrais de *C. canephora* e *C. eugenioides* (Lashermes et al. 1999).

Embora o café arábica seja nativo da Etiópia, os primeiros relatos de cultivo da planta para a exploração comercial remetem ao Iêmen, por volta do ano de 1500 (Meyer et al. 1968). A partir de poucas plantas cultivadas no Iêmen foram formadas as variedades Típica (*C. arabica* L. var. typica) e Bourbon (*C. arabica* L. var. bourbon), os quais são grupos genotípicos que deram origem a maior parte das cultivares comerciais de *C. arabica* plantadas em todo o mundo (Anthony et al. 2002).

Visando aumentar a diversidade de material genético de *C. arabica* para o uso em programas de melhoramento e conservar os genótipos naturais do centro de origem da espécie, foram realizadas algumas expedições à Etiópia para a coleta de sementes de plantas nativas (Meyer et al. 1968; Anthony et al. 1999).

Uma das expedições para a Etiópia foi organizada pela FAO em 1964 (Meyer et al. 1968). A partir dessa expedição, sementes coletadas de *C. arabica* foram distribuídas para diversos institutos de pesquisa, incluindo o Centro Agronômico Tropical de Pesquisa e Ensino (CATIE), na Costa Rica, onde foram multiplicadas. Em 1975 alguns exemplares foram cedidos para o Instituto Agronômico de Campinas (IAC). No ano seguinte, parte desses acessos foram também plantados no Instituto Agronômico do Paraná (IAPAR) (dos Santos Scholz et al. 2016). Estudos indicam que essas plantas possuem grande diversidade de caracteres fenotípicos e genotípicos que após devidamente caracterizados poderão ser explorados para o melhoramento genético da espécie (Silvestrini et al. 2007; dos Santos et al. 2015; dos Santos Scholz et al. 2016).

3.2.1. Melhoramento de cultivares de *C. arabica* no Brasil

O início do cultivo do cafeeiro no Brasil e no mundo foi marcado pela introdução de poucos genótipos dos grupos Típica e Bourbon (Anthony 2002). No Brasil, incialmente, foi introduzida a variedade Típica, em 1727, em Belém do Pará (Guerreiro-Filho et al. 2008; Pereira et al. 2010). A partir dessa introdução, outros estados também passaram a cultivar o cafeeiro, incluindo São Paulo e Paraná (Mendes et al. 2008). Após mais de cem anos, foram

introduzidas também as cultivares Bourbon e Sumatra, em 1859 e 1896, respectivamente. Mais tarde foi verificado que a cultivar Sumatra é uma linhagem de Típica (Anthony et al. 2002; Guerreiro-Filho et al. 2008). Esses genótipos constituíram a base genética para a produção de cultivares de *C. arabica* no Brasil (Vidal et al. 2010; Setotaw et al. 2013).

O primeiro programa de melhoramento genético do cafeeiro no país foi iniciado em 1927 pelo Instituto Agronômico de Campinas (IAC). Com o decorrer dos anos, outras instituições de pesquisa começaram seus próprios programas de melhoramento, incluindo o Instituto Agronômico do Paraná (IAPAR), a Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG) e o Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (INCAPER) (Pereira et al. 2010).

Como a maioria das cultivares comerciais de *C. arabica* são propagadas via semente, é necessário que as mesmas apresentem um número elevado de locos em homozigose para a formação de lavouras uniformes (Medina Filho et al. 2008). Em função disso, as progênies dos cruzamentos são submetidas a várias gerações de autofecundações e selecionadas pelo método genealógico. Todo esse processo requer de 20 a 24 anos até a formação de uma nova cultivar comercial (Sera 2001; Moncada et al. 2016).

O uso recorrente dos mesmos genótipos parentais no início dos programas de melhoramento, aliado a forma de reprodução da espécie e ao longo período necessário para a obtenção de novas cultivares, são fatores que contribuíram para a base genética estreita que é observada atualmente entre as cultivares comerciais de *C. arabica* (Anthony et al. 2002; Silvestrini et al. 2007; Setotaw et al. 2013). Devido à essa baixa variabilidade genética, embora as cultivares comerciais sejam produtivas, a maioria é suscetível a estresses bióticos e abióticos, como por exemplo ao ataque da bactéria *Pseudomonas syringae*, causadora da mancha aureolada, e a suscetibilidade a períodos de seca (van der Vossen et al. 2015).

Com a detecção de focos de ferrugem alaranjada do cafeeiro (*Hemileia vastatrix*) na década de 1970, os programas de melhoramento começaram a investir na introdução de alelos de resistência a essa doença, principalmente pelo cruzamento com os genótipos Icatu e Híbrido de Timor. Esses dois genótipos foram provenientes de cruzamentos interespecíficos entre *C. arabica* e *C. canephora*, e carregam alelos de resistência a ferrugem, herdados de *C. canephora*. A inserção desses genótipos nos programas de melhoramento do cafeeiro levou à ampliação da diversidade genética das cultivares lançadas a partir de 1980. No entanto, em geral a diversidade encontrada entre cultivares comerciais de *C. arabica* ainda é considerada baixa (Setotaw et al. 2013).

Estudos com marcadores moleculares revelaram que genótipos de *C. arabica* coletados no centro de origem da espécie possuem maior diversidade genética do que a observada entre cultivares comerciais, por isso, podem ser utilizados para a introdução de novos alelos em programas de melhoramento (Anthony et al. 2001; Silvestrini et al. 2007). Os acessos da Etiópia também foram estudados quanto a diversidade de características de interesse agronômico, dentre eles, alguns aspectos morfológicos dos grãos (Gaspari-Pezzopane 2004; Tran et al. 2017), composição de nutrientes em folhas (dos Santos et al. 2015), e composição química de compostos que influenciam na qualidade da bebida (Tessema et al. 2011; dos Santos Scholz et al. 2016; Tran et al. 2017; Sant'Ana et al. 2018). A partir desses acessos, foi identificada uma variedade de *C. arabica* naturalmente descafeinada (Silvarolla et al. 2004).

Embora seja relatada maior diversidade genética entre os acessos selvagens de *C. arabica* quando comparada a diversidade encontrada entre cultivares comerciais (Silvestrini et al. 2007), ainda são escassas as análises a nível molecular para a identificação de regiões genômicas utilizando genótipos selvagens de *C. arabica*, os quais podem ser utilizados para a identificação de novos alelos que não estão presentes entre cultivares comerciais. No entanto, devido ao desmatamento, parte dos genótipos selvagens de *C. arabica* não são mais encontrados em seu habitat de origem (Davis et al. 2012). Além disso, de acordo com as estimativas de mudanças climáticas, é provável que esses genótipos estejam em risco de extinção futuramente (Davis et al. 2019; Moat et al. 2019). Com isso, ressalta-se a importância de preservar o material genético da espécie e a necessidade de maiores investimentos para o desenvolvimento de estudos na área.

3.2.2. Composição de nutrientes em grãos de *C. arabica*

A nutrição mineral é um dos principais fatores responsáveis pela produção do cafeeiro (Laviola et al. 2007). No Brasil, em geral, é necessário que sejam aplicados fertilizantes sintéticos às lavouras de café, para que atinjam altos índices de produção com grãos de qualidade (Matiello et al. 2015). O uso de corretivos e fertilizantes representa de 25 a 30% do custo por saca de café beneficiado, de forma que compõe o segundo maior gasto para a manutenção da lavoura cafeeira (Guimarães e Reis 2010).

A análise da composição de nutrientes em grãos de *C. arabica* das cultivares Catucaí Amarelo e Catucaí Vermelho indicam a seguinte ordem decrescente de extração de macronutrientes pelos grãos: N > K > Mg > P > Ca (Garcia et al. 2009).

O nitrogênio é o nutriente exigido em maior quantidade pelo cafeeiro (Matiello et al. 1985), e também presente em maior concentração em grãos de *C. arabica*, correspondendo a

cerca de 49% do conteúdo total de macronutrientes nesse tecido (Garcia et al. 2009). Esse nutriente faz parte da estrutura de ácidos nucleicos, aminoácidos e proteínas (Kusano et al. 2011). O N também faz parte de compostos essenciais que influenciam na qualidade da bebida do café como a cafeína e a trigonelina (Clifford 1985).

Embora o K seja o nutriente presente em maior concentração em frutos inteiros de café arábica (Cantani et al. 1967; Dias Chaves e Sarruge 1984; Malavolta 1986), quando analisados apenas os grãos, esse é o segundo nutriente com maior acúmulo, e representa cerca de 40% do total de macronutrientes acumulados nos grãos (Garcia et al. 2009). O K participa no processo de transporte de carboidratos das folhas para os frutos do cafeeiro (Guimarães e Reis 2010).

O magnésio (Mg) é o terceiro nutriente com maior acúmulo nos grãos de café arábica. Sua concentração corresponde a aproximadamente 3,5% do total de macronutrientes acumulados nos grãos (Garcia et al. 2009). Esse nutriente é um dos constituintes da clorofila e está relacionado ao processo de fotossíntese, que também ocorre nos frutos quando ainda verdes. Consequentemente, a concentração de Mg também afeta a produtividade da planta e a qualidade dos grãos de *C. arabica* (Guimarães e Reis 2010).

O fósforo (P) corresponde a cerca de 3% dos macronutrientes totais acumulados nos frutos e nos grãos de café arábica (Cantani et al. 1967; Dias Chaves e Sarruge 1984; Malavolta 1986; Garcia et al. 2009) e embora seja um nutriente pouco exportado para os frutos, plantas com deficiência de P produzem uma quantidade menor de frutos e com qualidade inferior (Guimarães e Reis 2010).

O cálcio (Ca) é o macronutriente com menor acúmulo nos grãos de café arábica, e representa cerca de 2% do total de macronutrientes acumulados nesse tecido (Garcia et al. 2009). Sua função está relacionada ao processo de divisão celular e estabilização da parede celular (Laviola et al. 2007).

O acúmulo de nutrientes em grãos de *C. arabica* é influenciado por diversos fatores genéticos e ambientais. No entanto, ainda faltam trabalhos a nível molecular que revelem os genes envolvidos nas vias metabólicas relacionadas com a composição de macronutrientes em grãos de *C. arabica*.

3.3. Marcadores SNP

O polimorfismo de nucleotídeo único (*single nucleotide polymorphism* - SNP) corresponde a alteração de um nucleotídeo por substituição, deleção ou inserção, quando o mesmo loco é comparado entre dois indivíduos da mesma espécie (Nadeem et al. 2018). Alguns

autores classificam apenas as substituições como SNPs, outros também incluem pequenas inserções e deleções de bases nessa classificação (Caixeta et al. 2016).

Os SNPs são as formas de polimorfismos mais abundantes nos genomas de plantas (Huang e Han 2014), e estão presentes em regiões codantes e não codantes (Patel et al. 2015). Os SNPs em regiões codantes podem levar a alterações denominadas sinônimas ou não sinônimas. SNPs que marcam variações sinônimas não alteram o aminoácido codificado, enquanto SNPs que marcam variações não sinônimas alteram (Patel et al. 2015). Em plantas a frequência de SNPs tende a ser maior em regiões não codantes (Ching et al. 2002; Huang e Han 2014).

No genoma de *C. arabica*, o primeiro trabalho em que foram identificados marcadores SNPs em larga escala (Vidal et al. 2010), foi realizado *in silico* a partir de etiquetas de sequências expressas (*expressed sequence tags* - EST) provenientes de 33 bibliotecas de cDNA, que incluíram diversos órgãos e tecidos de *C. arabica* (raízes, folhas, flores, frutos, calos embriogênicos) em diferentes estágios de desenvolvimento, bem como alguns tecidos submetidos a estresses bióticos e abióticos (Vieira et al. 2006). Nesse trabalho, foi observada a frequência de 1 SNP a cada 260 pb, dos quais a maioria foi proveniente de polimorfismos entre os subgenomas de *C. arabica*, *C. canephora* e *C. eugenioides*. Além disso, não foi observado polimorfismo entre as cultivares Mundo Novo e Catuaí, ambas derivadas de Típica e Bourbon (Vidal et al. 2010).

Os resultados da análise *in silico* em larga escala de Vidal et al. (2010) reforçam a baixa diversidade genética encontrada entre genótipos descendentes de Típica e Bourbon. No entanto, nesse estudo a frequência de polimorfismo apresentada para a espécie está subestimada, considerando que os SNPs foram obtidos a partir apenas de regiões codificantes do genoma (Patel et al. 2015).

As tecnologias de sequenciamento de alto rendimento contribuíram para a identificação de variantes genéticos que podem ser aplicados aos programas de melhoramento de plantas cultivadas (Torkamaneh et al. 2018). Atualmente, por meio do sequenciamento de baixa a média cobertura, é possível identificar SNPs distribuídos por todo o genoma de plantas de uma mesma população, por um custo relativamente baixo por amostra (≈ 20 dólares) utilizando por exemplo, a técnica de genotipagem-por-sequenciamento (Elshire et al. 2011; Poland et al. 2012; Torkamaneh et al. 2018).

3.3.1. Genotipagem por sequenciamento

Entre as metodologias disponíveis para o sequenciamento de DNA, a técnica de genotipagem por sequenciamento (*genotyping-by-sequencing* - GBS) tem se destacado em estudos em plantas, incluindo as perenes como citrus (*Citrus* spp.) (Imai et al. 2018) e *C. arabica* (Moncada et al. 2016; Sant'Ana et al. 2018).

O princípio dessa técnica envolve a redução na complexidade do genoma, por meio da digestão com enzimas de restrição, seguida da ligação de adaptadores comuns e *barcodes*, que permitem a multiplexação de várias amostras para serem sequenciadas em uma mesma reação. Em seguida é realizada a amplificação via PCR e o sequenciamento em uma plataforma de alto rendimento (Elshire et al. 2011). Essa metodologia pode ser aplicada a praticamente qualquer espécie, incluindo aquelas com genoma grande e complexo (Elshire et al. 2011; Poland e Rife 2012).

Para aplicar a técnica, a escolha das enzimas de restrição deve ser de acordo com o tamanho e complexidade do genoma da espécie a ser sequenciada. São preferíveis enzimas que evitem o corte frequente em regiões repetitivas, como por exemplo, as enzimas sensíveis à metilação (Davey et al. 2011; Elshire et al. 2011; Poland e Rife 2012). Desse modo, regiões com baixo número de cópias também podem ser representadas no sequenciamento (Gore et al. 2007).

Após o corte com as enzimas de restrição, na mesma reação são ligados adaptadores *barcode* e comuns às extremidades dos insertos de DNA. Os adaptadores *barcode* com sentido 3' para 5', possuem na extremidade 5' uma sequência complementar à deixada pelo corte com a enzima de restrição, e uma combinação única de 4 a 8 pb (*barcode*) *upstream* para a identificação de cada amostra. Os adaptadores *barcode* sintetizados no sentido oposto possuem a sequência complementar ao barcode na extremidade 3' (Elshire et al. 2011). Os adaptadores comuns são oligonucleotídeos comuns que também são ligados por complementariedade às extremidades dos insertos de DNA (Elshire et al. 2011). Após a ligação dos adaptadores, as amostras são multiplexadas e amplificadas por PCR. Os fragmentos amplificados podem então ser sequenciados em uma plataforma de alto rendimento (Elshire et al. 2011), como por exemplo a Illumina, que gera *reads* de 50 a 300 pb (He et al. 2014).

Como o sequenciamento retorna um grande número de *reads* (na escala de milhões), ferramentas de bioinformática específicas são necessárias para a identificação dos sítios variantes nessas sequências (Glaubitz et al. 2014). Atualmente, estão disponíveis diversos *softwares* com *pipelines* voltados para a análise de dados de GBS (Glaubitz et al. 2014; Kim et

al. 2016; Torkamaneh et al. 2018). Nessas análises, a descoberta de SNPs e a genotipagem ocorrem simultaneamente (Elshire et al. 2011; Deschamps et al. 2012; Kim et al. 2016).

Em resumo, os *reads* de cada amostra são alinhados ao genoma de referência da espécie, ou de uma espécie relacionada, e os sítios polimórficos entre os diferentes indivíduos da população são identificados. O alinhamento ao genoma de referência fornece a posição dessas variantes (Elshire et al. 2011; Glaubtiz et al. 2014). Também é possível identificar variantes sem o uso de um genoma de referência, por meio de *pipelines* de descoberta de SNPs em rede (Catchen et al. 2013; Lu et al. 2013), nesse caso, a eficiência para a identificação de SNPs tende a ser menor (Kim et al. 2016).

Recentemente, o número de trabalhos que utilizaram a técnica de GBS para a identificação de marcadores SNPs em populações de *C. arabica* tem crescido. No trabalho de Moncada et al. (2016), marcadores SNPs identificados por GBS em uma população F_2 foram utilizados para a construção de um mapa genético e identificação de QTL associados a produtividade, tamanho dos grãos e altura da planta. Sant'Ana et al (2018) genotiparam uma população com 107 acessos provenientes do centro de origem da espécie e identificaram por meio de GWAS marcadores SNPs associados à composição de lipídeos nos grãos. SNPs identificados por GBS também foram utilizados em análises de diversidade de uma população com 787 genótipos de *C. arabica* (Scalabrin et al. 2020).

Embora a técnica de GBS tenha contribuído significativamente para o avanço de análises genéticas com a espécie *C. arabica*, os painéis de genótipos geralmente são caracterizados por altos índices de dados faltantes por marcador (Glaubitz et al. 2014). Como os dados faltantes podem reduzir a eficiência de análises de associação entre genótipo e fenótipo (Teo 2008), essa é uma questão chave que deve ser levada em consideração quando dados de GBS são utilizados para esse tipo de análise.

3.3.2. Dados faltantes por marcador

Em análises de dados de GBS, os dados faltantes podem ser decorrentes de características naturais da população e/ou da técnica de sequenciamento (Poland e Rife 2012). Considerando o aspecto biológico, devido às inserções e deleções (pequenas ou a nível estrutural), os sítios polimórficos de reconhecimento das enzimas de restrição podem estar presentes em alguns indivíduos da população, enquanto em outros não, de modo que nem todas as sequências serão representadas em todos os indivíduos (Poland e Rife 2012; Swarts et al. 2014).

Com relação à técnica de GBS, a proporção de dados faltantes varia principalmente em função do tamanho e complexidade do genoma da espécie, da escolha da enzima de restrição e do nível de multiplexação das amostras sequenciadas (Poland e Rife 2012; Rutkoski et al. 2013). Esses fatores são correlacionados e afetam a profundidade da cobertura do sequenciamento, que em média, tende a ser baixa em dados de GBS, e tem como resultado as altas proporções de dados faltantes (Davey et al. 2011; Torkamaneh et al. 2018).

Enzimas de restrição de corte frequente proporcionam um grande número de fragmentos, porém com baixa cobertura de *reads* por fragmento, enquanto as enzimas de corte raro, geram menor número de fragmentos com maior profundidade de cobertura de sequenciamento (Elshire et al. 2011). Quanto maior o número de genótipos por linha de sequenciamento, menor será o número de *reads* por genótipo e, consequentemente, maior a proporção de dados faltantes (Elshire et al. 2011; Poland et al. 2012; Chan et al. 2016). No entanto, reduzir o nível de multiplexação das amostras para obter maior cobertura, aumenta o custo de sequenciamento por amostra, o que pode ser um limitante em estudos com muitos indivíduos (Davey et al. 2011; Fu 2014).

Proporções elevadas de dados faltantes podem reduzir o poder de análises genéticas incluindo os estudos de associação genômica ampla (Rahimi et al. 2019). Por isso, os dados de sequenciamento utilizados para esse tipo de análise devem passar por um controle de qualidade, visando evitar associações espúrias (falsos positivos) (Teo, 2008). Em dois dos estudos em que a técnica de GBS foi utilizada para a genotipagem de populações de *C. arabica*, foram excluídos marcadores com mais de 20% de dados faltantes (Moncada et al. 2016; Sant'Ana et al. 2018). Porém, não foi avaliado se a proporção de dados faltantes que permaneceu nos painéis de marcadores influenciou no desempenho das análises de associação desenvolvidas.

3.3.3. Imputação de dados ausentes

Nesse trabalho, a palavra imputação foi traduzida do termo em inglês *impute*, o qual na área de negócios remete a cálculos realizados com dados incompletos ou que não estão totalmente corretos. Com base nesse termo, aqui a imputação de marcadores se refere a substituição de dados ausentes em painéis de marcadores por dados prováveis (Marchini e Howie 2010; Torkamaneh e Belzile 2015; Chan et al. 2016; Das et al. 2018). Dessa forma, a imputação pode ser aplicada a dados de GBS para substituir os dados ausentes sem que haja um aumento do custo com o sequenciamento.

A imputação de dados ausentes em painéis de marcadores pode ser realizada por métodos desenvolvidos especificamente para esse fim ou também por métodos estatísticos gerais de imputação, como o método da média por exemplo (Nazzicari et al. 2016).

Dentre os métodos desenvolvidos para a imputação de marcadores, pode-se citar o método de imputação de marcadores pela identificação de haplótipos localizados, utilizado pelo software Beagle (Browning e Browning 2016). Nesse método a imputação é realizada de acordo com a posição e o desequilíbrio de ligação (DL) entre marcadores presentes em cada cromossomo. O DL é definido como a associação não aleatória entre marcadores em diferentes locos de um mesmo cromossomo (Slatkin et al. 2008). Quanto maior a proximidade entre os marcadores, maior o DL entre eles. Marcadores próximos tendem a ser herdados em conjunto, formando haplótipos. No método utilizado pelo software Beagle, os haplótipos podem ser identificados a partir dos marcadores genotipados na própia população, e esses mesmos haplótipos são utilizados como referência para a imputação dos dados ausentes no painel de marcadores (Browning e Browning 2016), conforme representado na figura 1.

O método de imputação do software Beagle, foi criado inicialmente para a imputação de painéis para estudos de associação genômica ampla em humanos (Browning e Browning 2007). Posteriormente, esse método também foi testado com marcadores de algumas populações de plantas (Torkamaneh e Belzile, 2015, Chan et al. 2016, Nazzicari et al. 2016).

Estudos indicam que a imputação com o método do software Beagle apresenta alta acurácia quando utilizada em dados de genotipagem de espécies que possuem um genoma de referência bem anotado (Torkamaneh e Belzile 2015; Nazzicari et al. 2016). Como exemplo, foram observados valores de acurácia entre 85% e 94% quando o Beagle foi utilizado para a imputação de dados de GBS de uma população com 301 acessos de soja que possuíam entre 20% e 80% de dados ausentes permitidos por marcador, respectivamente (Torkamaneh e Belzile 2015).

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
Genótipo 1	0	1	1	0	1	1	1
	1	0	0	1	1	1	1
Genótipo 2	1	?	1	0	0	1	0
	1	1	1	0	?	1	0
Genótipo 3	0	1	1	0	1	1	1
	0	0	0	1	1	?	1
Genótipo 4	0	1	?	0	1	1	1
	1	1	1	0	0	1	?

1) Identificação dos haplótipos localizados

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
Genótipo 1	0	1	1	0	1	1	1
	1	0	0	1	1	1	1
Genótipo 2	1	**?**	1	0	0	1	0
	1	1	1	0	**?**	1	0
Genótipo 3	0	1	1	0	1	1	1
	0	0	0	1	1	**?**	1
Genótipo 4	0	1	**?**	0	1	1	1
	1	1	1	0	0	1	**?**

2) Imputação de dados ausentes

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
Genótipo 1	0	1	1	0	1	1	1
	1	0	0	1	1	1	1
Genótipo 2	1	**1**	1	0	0	1	0
	1	1	1	0	**0**	1	0
Genótipo 3	0	1	1	0	1	1	1
	0	0	0	1	1	**1**	1
Genótipo 4	0	1	**1**	0	1	1	1
	1	1	1	0	0	1	**0**

Fig 1. Método de imputação de alelos em haplótipos localizados. Em um painel de marcadores com dados ausentes, foram identificados os haplótipos a partir dos marcadores genotipados (1), em seguida esses haplótipos foram usados como referência para a imputação de alelos nos pontos ausentes (2). Adaptado de Marchini e Howie (2010).

Como diversas espécies de plantas cultivadas não possuem genoma de referência, outros métodos estatísticos gerais de imputação, que não dependem do mapeamento ao genoma,

também foram testados. Dois métodos têm se destacado pela acurácia da imputação de marcadores não ordenados: o K-vizinho mais próximo (*K-nearest neighbor* - KNN) (Troyanskaya et al. 2001), e a regressão de florestas aleatórias (*random forest* - RF) (Stekhoven e Bühlmann 2012).

Em um estudo em que foram comparados cinco métodos para a imputação de marcadores não ordenados em populações de trigo, milho e cevada (*Hordeum vulgare* L.), Rutkoski et al. (2013) observaram que, em geral, os melhores métodos foram KNN e RF, com os quais foram obtidas acurácias mínimas de 82% e 84%, respectivamente, para a imputação de genótipos de milho em painéis com até 70% de dados ausentes permitidos por marcador. Para as populações de trigo e cevada, a acurácia obtida com esses métodos foi ainda maior em painéis com até 70% de dados ausentes por marcador, para a primeira espécie foram observados valores entre 93% e 95%, e para cevada entre 97% e 99%.

Estudos indicam que a imputação pode aumentar o número de marcadores identificados em análises de mapeamento associativo, como GWAS. Em uma população com 139 genótipos de soja, o número de SNPs identificados em associação com a composição de óleo nos grãos aumentou de um para 11 após a imputação de marcadores (Torkamaneh e Belzile 2015). Nesse trabalho, o painel sem imputação foi composto por 7.152 SNPs identificados por GBS e com até 20% de dados ausentes por marcador, enquanto o painel imputado foi formado por aproximadamente 83 mil SNPs obtidos pela união dos dados de GBS com SNPs identificados por chips de genotipagem. Em um outro trabalho recente, dados de GBS de 298 acessos de trigo foram utilizados para a identificação de marcadores associados a características agronômicas, incluindo a produtividade (Rahimi et al. 2019). Nesse estudo a imputação também contribuiu para o aumento de associações significativas identificadas por GWAS. No painel de marcadores sem imputação, composto por 10.938 SNPs, foram identificados 313 SNPs em associação com as características analisadas, esse número aumentou para 394 SNPs quando foi utilizado um painel com dados imputados contendo 46.862 SNPs. Até o momento nenhum trabalho foi realizado para verificar o efeito da imputação de genótipos em painéis de marcadores de populações de *C. arabica*.

3.4. Estudos de associação genômica ampla

Diversas características de interesse agronômico são complexas e controladas por múltiplos genes (Khan e Korban 2012), como tamanho de grãos, composição de nutrientes nas sementes e produtividade. Estudos que correlacionam variações fenotípicas a polimorfismos genéticos são parte integrante de programas de melhoramento de plantas, que por meio da

identificação de genes e QTL visam compreender melhor a arquitetura genética envolvida no controle dessas características (Ogura e Busch 2015).

Atualmente, os dois principais métodos utilizados para a identificação de regiões gênicas que controlam características fenotípicas de interesse são o mapeamento de QTL e o mapeamento associativo (Korte e Farlow 2013; Lipka et al. 2015; Nadeem et al. 2018). No mapeamento de QTL são explorados eventos de recombinação em populações segregantes como por exemplo populações F_2, de linhagens endogâmicas recombinantes ou de retrocruzamento de linhagens puras. Já no mapeamento associativo são utilizadas populações naturais que usualmente possuem maior diversidade entre os indivíduos (Huang e Han 2014; Lipka et al. 2015).

Embora o mapeamento de QTL seja uma ferramenta poderosa para a identificação dessas regiões no genoma de plantas, a técnica apresenta algumas limitações. As populações segregantes utilizadas nesses estudos geralmente são formadas pelo cruzamento de poucos genótipos. Assim, a diversidade alélica tende a ser reduzida, quando comparada a populações naturais. Além disso, apenas eventos de recombinação recentes são analisados, aqueles que ocorreram durante a formação da população. Dessa forma, a resolução do mapeamento tende a ser limitada (Korte e Farlow 2013; Lipka et al. 2015).

Para plantas perenes, como o cafeeiro, que em comparação com plantas anuais possui um período juvenil mais longo até atingir a produção efetiva (cerca de 3 anos), um alto investimento é necessário para a formação e condução de populações segregantes. Com isso, a exploração da diversidade genética presente em populações naturais constitui uma alternativa que requer menor investimento para a identificação de regiões genômicas que controlam características fenotípicas de interesse para o melhoramento (Khan e Korban 2012).

O mapeamento associativo ou mapeamento por DL, corresponde a análise da associação estatística entre marcadores, geralmente SNPs ou haplótipos de SNPs, e características fenotípicas da população (Hayward et al. 2015), incluindo indivíduos derivados de populações selvagens, acessos de coleções de germoplasma e parentais utilizados em programas de melhoramento (Rafalski 2010). A associação pode ser realizada para uma região específica do genoma, como uma sequência gênica específica, ou com marcadores distribuídos ao longo dos cromossomos da espécie, nesse caso é também denominada de estudo de associação genômica ampla (*genome-wide association study* - GWAS) (Rafalski 2010).

Os primeiros trabalhos de GWAS foram realizados com o intuito de identificar genes relacionados a doenças complexas em humanos (Huang e Han, 2014). O desenvolvimento de técnicas de sequenciamento, contribuiu para a popularização de GWAS também em plantas

(Rafalski et al. 2010). Recentemente, o GWAS também foi aplicado com êxito para a identificação de genes e QTL em diversas culturas perenes, como citrus (Minamikawa et al. 2017; Imai et al. 2018), maçã (*Malus domestica*) (McClure et al. 2018) e café arábica (Sant'Ana et al. 2018).

Como em GWAS normalmente são utilizadas populações diversas, mais eventos de recombinação devem ser explorados pela análise de associação; aqueles que ocorreram durante a evolução dos indivíduos genotipados (Hayward et al. 2015; Lipka et al. 2015). Estudos assim requerem um painel denso de marcadores que cubram todo o genoma. Cada marcador é testado para verificar se um ou mais locos são responsáveis pela alteração da característica fenotípica em estudo, ou se estão em DL com o loci causal (Rafalski et al. 2010).

Os eventos de recombinação que ocorreram na população e são captados na genotipagem, determinam a taxa de decaimento do DL, o qual indica a resolução que pode ser obtida com o mapeamento (Huang e Han 2014). O decaimento do DL em plantas autógamas tende a ser mais lento que em alógamas, conforme observado em arroz e milho, que possuem taxas de decaimento de DL de aproximadamente 100 Kb e 2 Kb, respectivamente. Devido a ampla extensão do DL em espécies autógamas, a resolução do mapeamento pode não atingir o nível de gene específico. Apesar disso, um ponto favorável do lento decaimento do DL, é a possibilidade de realizar a genotipagem dos indivíduos da população com menor cobertura, conforme observado em arroz, em que o sequenciamento de baixa cobertura seguido de imputação de genótipos, proporcionou resultados significativos em análises de GWAS (Huang e Han 2014).

Em GWAS, o uso de populações que possuem indivíduos relacionados ou com subpopulações estruturadas pode levar à identificação de falsas associações, pois genótipos relacionados compartilham alelos, sejam esses alelos causais ou não, com isso, a estimativa do efeito de um loco pode ser afetada pelo efeito de outros locos herdados em conjunto (Vilhjálmsson e Nordborg 2013, Korte e Farlow 2015).

Uma alternativa para reduzir os falsos positivos em GWAS é o uso do modelo linear misto (MLM), no qual podem ser adicionadas as informações da estrutura genética populacional (matriz Q) e da relação de parentesco (matriz K) entre os indivíduos da população (Yu et al. 2006). Nesse método são testados os ajustes de diferentes modelos nos quais os efeitos das matrizes Q e K são adicionados separadamente e em outro modelo esses efeitos são adicionados em conjunto (Q+K). O modelo com melhor ajuste é selecionado para a associação (Yu et al. 2006).

Após o MLM, diversos outros modelos foram criados visando aumentar o poder e eficiência das análises de GWAS e reduzir o tempo de execução dessas análises, incluindo os seguintes: MLM multi-locus com SNPs de efeito aleatório (*multi-locus random effect mixed linear model* - mrMLM) (Wang et al. 2016); o FASTmrMLM, uma versão mais rápida e eficiente do mrMLM, que utiliza transformações matriciais (Tamba e Zhang 2018); e o ISIS EM-BLASSO (*Iterative Sure Independence Screening EM-Bayesian LASSO*) que utiliza uma abordagem Bayesiana para a identificação de associações (Tamba et al. 2017).

O primeiro GWAS realizado com a espécie *C. arabica*, proporcionou a identificação de 21 marcadores SNPs possivelmente associados à composição de lipídeos e diterpenos nos grãos (Sant'Ana et al. 2018). Não foi ainda relatado nenhum estudo de associação para a identificação de marcadores associados a regiões gênicas que controlam as vias metabólicas de acúmulo de nutrientes em grãos de café arábica. Porém, em análises recentes realizadas com outras espécies como, feijão (*Phaseolus vulgaris*) (Katuuramu et al. 2018), soja (Ziegler et al. 2018) e milho (Ziegler et al. 2017), foi observado que a técnica de GWAS constitui uma ferramenta poderosa para a identificação de marcadores associados à composição de nutrientes de grãos.

Considerando a escassez de dados a nível molecular que revelem a arquitetura genética envolvida nas vias metabólicas de acúmulo de macronutrientes em grãos de *C. arabica*, os resultados dos trabalhos apresentados acima, indicam que a técnica de GWAS pode ser utilizada também com genótipos de café arábica para a identificação de marcadores relacionados à essas características de controle genético complexo.

Os resultados obtidos no presente trabalho foram apresentados no capítulo 2 em forma de artigo científico formatado de acordo com as normas para submissão à revista *Molecular Breeding*.

4. CAPÍTULO 2 - Genome-wide association analysis of macronutrient content on coffee grains from wild *Coffea arabica* germplasm and commercial cultivars

Mariane Silva Felicio[1,2]; Bruna Silvestre Rodrigues da Silva[2,3]; Caroline Ariyoshi[2,3]; Gustavo César Sant'Ana[4]; Rafaelle Vecchia Ferreira[2,3]; Lívia Maria Nogueira Brito[2,3]; Lorena Guimarães Batista[5]; Amanda Avelar de Oliveira[5]; Fernando Henrique Correr[5]; Guilherme Kenichi Hosaka[5]; Gabriel Rodrigues Alves Margarido[5]; Luiz Filipe Protasio Pereira[2,6], Douglas Silva Domingues[1,7].

[1] Universidade Estadual Paulista, Programa de Pós-graduação em Ciências Biológicas (Genética), UNESP, Botucatu, SP, CEP 18618-689, Brazil.

[2] Instituto Agronômico do Paraná, Laboratório de Biotecnologia Vegetal, Londrina, PR, CEP 86047-902, Brazil.

[3] Universidade Estadual de Londrina, Programa de Pós-graduação em Genética e Biologia Molecular, Londrina, PR CEP 86057-970, Brazil.

[4] Tropical Melhoramento & Genética (TMG), Londrina, PR, CEP 86188-000, Brazil.

[5] Universidade de São Paulo (USP), Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Departamento de Genética, Piracicaba, SP, CEP 13418-900, Brazil.

[6] Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA Café), Brasília, DF CEP 70770-901, Brazil.

[7] Universidade Estadual Paulista, Instituto de Biociências, Departamento de Biodiversidade, Rio Claro, SP CEP 13506-900, Brazil.

Abstract

Coffea arabica is the main commercial species from the *Coffea* genus. The mineral composition of coffee grain affects the beverage quality. Despite its importance, the genetic basis of macronutrient accumulation in coffee grains still poorly understood. Therefore, the aim of this work was to perform GWAS analysis to identify genomic regions associated with the macronutrient content in Arabica coffee grains. We used a panel of 110 *C. arabica* genotypes, including commercial cultivars (11) and wild accessions (96). We quantified N, P, K, Ca, and Mg content in coffee grains in 2017 and 2018. SNPs markers were identified by genotyping-by-sequencing (GBS) using two reference genomes of *C. arabica* for SNP calling. We tested three imputation methods and compared the GWAS analysis before and after the imputation. We observed that the wild genotypes harbor higher variability for grains macronutrient content than the commercial cultivars. Beagle was the best imputation method. The imputation reduced the false discovery rate in GWAS. We identified 5, 12, 13, 6, 6, and 1 markers in coding regions associated with the content of N, P, K, Ca, Mg, Mg and K in coffee grains, respectively. The proteins identified participate in diverse metabolic pathways, including nutrient transport, seed germination, control of flowering time and response to biotic and abiotic stresses. We suggest that the proteins identified here could be used as targets in further studies to characterize the metabolic pathways of macronutrient accumulation in coffee grains.

Keywords: nutrient transport, genotype imputation, non-model species, biotic and abiotic stress

4.1. Introduction

Coffee is one of the most important tropical commodities traded in the world, it is grown and processed in more than 70 countries, most of which are in development, including Brazil, Vietnam, and Colombia, which are the main coffee producer countries (ICO 2020). The coffee production provides a livelihood for between 12 to 25 million farmers and their families (ICO 2019b).

The coffee commercial production can be divided into two main species, Arabica (*Coffea arabica*) and Robusta (*Coffea canephora*), which produces around 60% and 40% of the coffee traded worldwide, respectively (ICO 2020). Due to the superior characteristics of the Arabica coffee cup quality, which have lower bitterness and caffeine levels, coffee grains from this species have higher commercial value than grains from Robusta coffee (FAO 2015).

The species *C. arabica* is native to the southwestern Ethiopia, southeast Sudan, and northern Kenya (Guerreiro-Filho et al., 2008). It is estimated that the species is a relatively recent, formed by the natural hybridization between the ancestral species *C. canephora* and *C. eugenioides* (Lashermes et al. 1999), which probably occurred between 10 thousand and 600 thousand years ago (Yu et al. 2011, Scalabrin et al. 2020).

Coffea arabica is a tetraploid species (2n = 4x = 44) that reproduces predominantly by self-fertilization (Medina Filho et al. 2008). The analysis of flow cytometry from leaf cells of the Catuaí Vermelho cultivar indicated that the species have a mean 2C nuclear DNA content of 2,62 pg, which corresponds to approximately 1,58 Gb (Clarindo and Carvalho 2008). In addition, in the recent sequencing of the species reference genome from Catrurra red cultivar were identified 1.09 Gb (RefSeq assembly accession: GCF_003713225.1, https://www.ncbi.nlm.nih.gov/, 2019).

Most of the Arabica cultivars grown worldwide were originated from two subpopulations, Typica and Bourbon (Anthony et al. 2002), which in addition to the species

recent origin, and its reproductive system, led to a narrow genetic diversity among commercial cultivars (Anthony et al. 2001, 2002; Silvestrini et al. 2007; Setotaw et al. 2013). Therefore, most of the Arabica commercial cultivars are susceptible to biotic and abiotic stresses, which already are problems for coffee production, considering the recent climate changes (i.e. long periods of drought), and the spread of severe diseases as the coffee leaf rust (*Hemileia vastatrix*) (van der Vossen et al. 2015).

Studies using molecular markers revealed that wild *C. arabica* genotypes collected from its center of origin have a wide genetic diversity that can be used to introduce new favorable alleles in coffee breeding programs (Anthony et al. 2001; Silvestrini et al. 2007; Sant'Ana et al. 2018). *C. arabica* wild genotypes were also studied regarding the diversity of important agronomic traits, including the level of chemical compounds that influence the cup quality (Tesseema et al. 2011; dos Santos Scholz et al. 2016; Tran et al. 2017; Sant'Ana et al. 2018).

Due to deforestation, part of the wild *C. arabica* genotypes was lost from its center of origin, the Ethiopian forests (Gole 2003; Davis et al. 2012). Also, according to climate change predictions, the remaining genotypes may be further in extinction risk (Davis et al. 2019, Moat et al. 2019). Therefore, preserving the Arabica coffee wild germplasm and investing in research should be a priority for coffee breeders and governance at coffee-producing countries.

The composition of different macronutrients in coffee grains are complex traits that directly impact grain quality. However, until now there is little information about the genomic regions related to the composition of macronutrients in coffee grains.

Genome-wide association study (GWAS) is an association mapping method that has contributed to the identification of genomic regions that regulate complex quantitative traits. GWAS statistical models are used to identify significant associations between markers, usually SNPs, with phenotypic traits from diverse populations with high genetic variability (Hayward et al. 2015).

The development of high-throughput sequencing technologies has contributed to the advancement of association analysis (Kumar et al. 2012; Rasheed et al. 2017). In that context, genotyping-by-sequencing (GBS) has been used as a technique to identify SNP markers in plant populations. In this technique restriction enzymes are used to cut the genomic DNA, reducing its complexity; than barcoded adaptors are used to identify each sample, thus, different genotypes can be polled in the same amplification reaction and sequenced in a multiplex (Elshire et al. 2011). However, when the GBS coverage of sequencing is low, the marker panel tends to present missing data (Fu 2014), which can limit the accuracy of associations found by GWAS (Rahimi et al. 2019).

Imputation of missing data is an alternative to use low-coverage GBS data without increasing analysis cost, the technique consists of the replacement of a missing data by a putative data (Marchini and Howie 2010; Torkamaneh and Belzile 2015; Chan et al. 2016; Das et al. 2018). Imputation of missing markers can be performed using general statistical methods or methods specially developed for genotype imputation, as the imputation based on localized haplotypes identification. To the best of our knowledge, until now, no study tested methods for marker imputation in panels of markers from *C. arabica* genotypes.

In a recent study, 107 *C. arabica* wild genotypes were used for GWAS analysis, in which 21 SNPs were found in association with the content of lipids and diterpenes in coffee grains (Sant'Ana et al. 2018). GWAS analysis also revealed several genomic regions in association with the macronutrient content in grains from another species, as common bean (*Phaseolus vulgaris*, Katuuramu et al. 2018), soybean (*Glycine max*, Ziegler et al. 2018), and maize (*Zea mays*, Ziegler et al. 2017).

The aim of this study was to identify genomic regions in association with the content of five macronutrients (N, P, K, Ca, and Mg) in coffee grains using GWAS analysis. We used a

panel of 110 *C. arabica* genotypes, which most of it was wild accessions (96). We also tested three imputation methods to replace missing data.

To the best of our knowledge, this is one of the first studies that used complete genomes of *C. arabica* as reference for SNP calling, which contributed to improved identification of genetic groups and significant SNPs associated with the phenotypic traits.

4.2. Material and Methods

4.2.1. Plant material

This study was performed at the experimental station of Instituto Agronômico do Paraná (IAPAR), in Londrina, PR, Brazil (23°22' S; 51°10' W, 585 m a.s.l.). We used a panel of 110 *C. arabica* genotypes from the germplasm bank of IAPAR, including three elite landrace, 11 commercial cultivars, and 96 accessions derived from the FAO collection mission at Ethiopia 1964-1965 (Meyer et al. 1968). The panel of genotypes was described in detail in the supplementary table S1. Similar genotype panels from the same population were also used in a previous study from our group (Sant'Ana et al. 2018) and for phenotypic analysis of the concentration of chemical compounds in coffee grains, as the caffeine content (dos Santos Scholz et al. 2016).

The genotypes collected during the FAO mission to Ethiopia were maintained in the germplasm bank of The Tropical Agricultural Research and Higher Education Center (CATIE, Costa Rica). In 1976, seeds from open-pollinated fields in CATIE were introduced in Brazil, at the Instituto Agronomico de Campinas (IAC). From these plants, 132 accessions were introduced at IAPAR and are maintained until nowadays (dos Santos Scholz et al. 2016; Sant'Ana et al. 2018). Each genotype was represented by individual plants cultivated in the same field in a completely randomized experimental design. The plants were maintained by the coffee breeding program at IAPAR with routine cultural practices used for coffee plants cultivation.

Coffee fruits were collected in the mature cherry stage (S7) which was described by Salmona et al. (2008) as fruits completely ripped with a pericarp coloration from red to purple. We collected fruits in two years, 2017 and 2018. In 2017 the fruits were collected from three elite landraces (BA10, SEL 106 and M7846), seven commercial cultivars (IPR100, IPR101, IPR102, IPR103, IPR105, IPR99 and, Mundo Novo) and 46 accessions from the FAO

collection. In 2018 we collected fruits from tow elite landraces (BA10 and SEL106), eleven cultivars (Catuaí Vermelho, IAPAR 59, IPR99, IPR100, IPR101, IPR102, IPR103, IPR104, IPR105, IPR107, and Mundo Novo) and 92 accessions from the FAO collection.

The climate conditions registered during the sampling periods in 2017 and 2018 were: cumulative precipitation of 678.7 mm and 291.8 mm, mean temperature of 19.6 °C and 21.3 °C, respectively (data provided by the IAPAR meteorological station). For 65 genotypes, we collected coffee fruits in both years. The phenotypic data from the samples collected in both years were used to calculate the Best Linear Unbiased Predictions (BLUPs), similar to Katuuramu et al. (2018) rationale.

4.2.2. Elemental composition analysis

After harvest, the coffee fruits were dried in a forced-air circulation oven, at 70°C, until they reach a constant weight, then the fruits were processed in a peeling machine DRC2 (Pinhalense, Espírito Santo do Pinhal, Brazil), to remove the pericarp. After that, the grains were dipped in liquid nitrogen and grounded in a disc mill (PERTEN 3600, Kungens Kurva, Sweden). The grounded grains were sieved in a 0.5 mm mesh and divided into three technical replicates for the estimation of macronutrient content.

The elemental composition analysis was performed by the Soil and Vegetal Tissues Laboratory at IAPAR, following the procedures described by Miyazawa et al. (1999). The Nitrogen (N) composition was determined using the Kjeldahl method: samples were sulfuric digested in a heated digester block at 350°C for 1h. The ammonium (NH_4^+) concentration was measured by spectrophotometry. Calcium (Ca), Magnesium (Mg), Phosphorus (P) and Potassium (K) were extracted in a hydrochloric acid solution, and the concentration measured by inductively coupled plasma - optical emission spectroscopy (ICP-OES, Optima 83000, Perkin Elmer, Waltham) (Miyazawa et al. 1999).

The normality and homogeneity of variances from the phenotypic data were tested with Shapiro-Wilk and Bartlett test ($\alpha = 0.05$), respectively, using the software R (R Core Team 2017). The normal quantile-quantile plot of residuals was used to identify outliers using the software R (R Core Team 2017). For selected samples that presented outliers, we removed the outlier, remaining only the duplicates for further analyses. In the phenotypic data from 2017, we removed the Ca and Mg data content from the genotypes E450_235 and E511_157, respectively. In the phenotypic data from 2018, the Ca and K content from the genotypes E270_044 and E118_213 were removed, in that order. These genotypes were also removed to estimate BLUPs for these traits. To estimate BLUPs of Mg content, we also excluded the data from the genotype E159_180. In these cases, the GWAS was performed with missing phenotypic points.

To visualize the phenotypic data distribution, we plotted the Z-score (number of standard deviations from the mean) of the traits, dividing it into commercial and non-commercial genotypes, using the R package ggplot2 (Wickham 2016). We also performed a correlation analysis between nutrients in each year and for each nutrient between the years. The trait mean values from replicates were used to calculate the Spearman correlation using the function *rcorr* from the R package Hmisc (Harrel Jr et al. 2019).

The variance components were estimated by Restricted Maximum Likelihood (REML) using the software Selegen REML-BLUP (Resende 2016). We used a model for experiments in completely randomized design (model 83): $y = Xu + Zg + e$, where y is the response trait value; u is the mean, considered as fixed; g is a vector of the genotypic effect, and e is the experimental error, both (e and g) considered random. The capital letters represent the incidence of matrices for the referred effects. The Best Linear Unbiased Predictor (BLUPs) were also calculated by REML using the same software, and the model 121: $y = Xf + Za + e$, where y is the response trait value; f is the vector of environment effect plus the mean, assumed as

fixed; *a* is the vector of the genotypic effect, and *e* is the experimental error, both considered as random. The capital letters represent the incidence of matrices for the referred effects. For GWAS, we used the trait mean values from each year separately (2017 and 2018) and the BLUPs (2017/2018) from the 65 genotypes sampled in the two years.

4.2.3. Genotyping-by-sequencing

The genotyping-by-sequencing library used in this work was sequenced in a previous study from our group (Sant'Ana et al. 2018) using the methodology described by Elshire et al. (2011). Coffee leaves were collected from 159 *C. arabica* genotypes and the DNA extracted by a modified CTAB protocol (Healey et al. 2014). The DNA samples were digested with the *PstI* restriction enzyme, and two 96-well plates were sequenced as single-end reads. The genomic library was prepared by the Genomic Diversity Facility at Cornell University using an Illumina HiSeq 2000 equipment.

We used the pipeline Tassel 5 GBS v2 (Glaubitz et al. 2014) to perform the SNP calling with the software Tassel v 5.2.37 (Bradbury et al. 2007). The raw FASTQ samples were trimmed to select barcoded reads with the following parameters: kmer length = 90, minimum kmer length = 20, and quality score > 20. Bowtie2 v. 2.3.4.2 (Langmead and Salzberg 2012; Langmead et al. 2019) was used to align the sequences into two reference genomes of *C. arabica*: the reference genome from the commercial cultivar Caturra red (Ref Seq assembly: GCF_003713225, https://www.ncbi.nlm.nih.gov/), and the reference genome from Et39, a spontaneous *C. arabica* genotype that only have one set of chromosomes from each subgenome (dihaploid) (Berthaud 1976). This genotype was derived from a plant collected in Ethiopia in 1966 by Guillaumet and Hallé (Guillaumet and Hallé 1978). The sequences from Et39 were provided by the Arabica Coffee Genome Consortium (ACGC, de Kohchko 2018).

After the alignment, the parameters used for SNP calling were: minor allele frequency (MAF) > 0.01 and locus coverage > 0.1. From the alignment to the reference genomes we obtained two Variant Call Format (VCF) files.

Although *C. arabica* is a tetraploid species the markers were obtained considering diploids. Thus, in the VCF file the homozygous genotypes AA and BB were represented by 0/0 and 1/1, where 0 and 1 are the reference (A) and alternative alleles (B), respectively. The heterozygous genotypes were represented by 0/1 (AB) or 1/0 (BA).

From the raw VCF files with 159 *C. arabica* accessions, we selected only the 110 accessions that were also phenotyped in this work, and then removed from the markers panel the monomorphic sites, INDELs, sites with more than one reference allele, and sites aligned to unplaced scaffolds (chromosome zero).

To perform the imputation and population structure analysis we used the panel with 110 genotypes, including all accessions that were phenotyped in this work. For GWAS analysis, the panel was divided according to the number of samples phenotyped in each year: 70 genotypes in 2017, 105 in 2018, and 65 used for BLUP which were collected in both years. An exception to the BLUP panel, all of the datasets were analyzed before and after imputation. Thus, we used a total of 14 marker datasets that are described in table 1. In each dataset, we applied the following quality control filters: MAF > 0.05, markers call rate > 0.80 and heterozygosity < 0.90 as described by Sant'Ana et al. (2018). On imputed datasets, we performed another quality control to remove imputed alleles with MAF < 0.05 and heterozygosity > 0.9 (Table 1).

To analyze the distribution of SNPs that were identified using each *C. arabica* reference genome, in the datasets D1 and D2, we measured the density of SNPs in windows of 300 Kb and plotted the number of SNPs per windows across the chromosomes. We used the software VCFtools 0.1.16 (Danecek et al. 2011). The density of SNPs in the panel distributed across the two *C. arabica* reference genomes was plotted using the package beeswarm (Eklund 2016) in

software R (R Core Team 2017). We used the R script developed by Hu et al. (2019) to plot the figure showing the distribution of SNPs across the *C. arabica* reference genomes; this script is available at GitHub (https://github.com/zhenbinHU/Sorghum_SNP_dataset).

Table 1 Datasets used in the present study. The GBS data was aligned to two different reference genomes (Caturra and Et39). The complete genotype datasets were used to test the accuracy of imputation methods and to calculate the population linkage disequilibrium. For GWAS analysis the datasets were divided including only the accessions phenotyped at each year.

Dataset	Reference genome	Genotypes	Year of phenotyping	Imputed	Number of SNPs	Analysis[a]
D1	Caturra	110	-	No	11230	Imputation, PS, LD
D2	Et39	110	-	No	9991	Imputation, PS, LD
D3	Caturra	110	-	Yes	10715	PS
D4	Et39	110	-	Yes	9581	PS
D5	Caturra	70	2017	No	11229	PS, GWAS
D6	Caturra	70	2017	Yes	10911	PS, GWAS
D7	Caturra	105	2018	No	11142	PS, GWAS
D8	Caturra	105	2018	Yes	10168	PS, GWAS
D9	Caturra	65	2017/2018	Yes	9984	PS, GWAS
D10	Et39	70	2017	No	10016	PS, GWAS
D11	Et39	70	2017	Yes	9766	PS, GWAS
D12	Et39	105	2018	No	9944	PS, GWAS
D13	Et39	105	2018	Yes	9097	PS, GWAS
D14	Et39	65	2017/2018	Yes	8988	PS, GWAS

[a] **Imputation:** datasets used to test the accuracy of imputation methods. **LD:** datasets used to calculate the population linkage disequilibrium decay. **PS:** datasets submitted to population structure analysis. **GWAS:** datasets used for GWAS.

4.2.4. Imputation methods

The datasets including all genotypes (D1 and D2) were used to test the imputation methods. We tested three methods for missing data imputation: allele imputation in localized haplotypes from Beagle software v. 4.1, the K-nearest neighbor (KNN), and Random Forest (RF). The first is a genotype-specific method, while KNN and RF are general imputation methods. Therefore, to perform the imputation with KNN and RF, the genotypes AA, AB, BA, and BB were replaced by the numbers 1, 2, 3, and 4, respectively. The imputation procedures adopted were based on the methodology described by Nazzicari et al. (2016).

The method of allele imputation in localized haplotypes was performed using the software Beagle v. 4.1 (Browning and Borwning 2016). In this method to estimate the localized haplotypes, the initial GBS target data is considered as phased. Thus the imputation is performed as the imputation of missing alleles, rather than genotypes considering that each sample will be composed of haplotype pairs. Then, the model construct clusters of aggregated markers for each sample, following the order of genotyped markers in the chromosome. Those aggregates can be formed by up to 5000 markers. After forming the localized aggregates, a Hidden Markov Model (HMM) forward-backward algorithm is used to estimate the haplotypes probabilities for each sample. As for *C. arabica* there is no haplotype reference panel, the aggregated genotyped markers from the own samples in the population were used as a reference for haplotype inference in this work. The algorithm was performed with 10 burn-in iterations followed by 5 phasing iterations. In each iteration, the haplotypes were estimated for a single sample at a time, while the haplotypes from the other samples were used as a reference panel. Therefore the estimated haplotypes were updated at each iteration. The localized haplotype pairs for each sample were estimated according to its similarity to the localized haplotypes at the reference panel. After the 15 iterations, the missing data were replaced according to the HMM estimated probabilities in the last iteration (Browning and Browning 2016).

In the KNN imputation method (Troyanskaya et al. 2001), the imputation of a missing point in a marker is based on the weighted average of the K-nearest markers that are genotyped at that same point and presents similar genotype profiles to the marker to be imputed, where K is the number of nearest markers chosen for imputation. The distance between pair-wise markers was calculated based on a simple matching coefficient (Schwender 2007). We used the function *knncaimpute* from the R package Scrime (Schwender and Fritsch 2013) to perform the KNN imputation with the parameters nn=5 (K=5), and wheights= T. The value present in each nearest marker was weighted by the reciprocal of its distance to the marker that presented a missing point. The weighted mean among the five nearest markers was used to impute the missing point (Schwender and Fritsch 2013).

The RF imputation method (Breiman 2001) is a non-parametric method developed to impute categorical, continuous, or mixed type data. We used the modified RF imputation method described by Stekhoven and Bühlmann 2012, which performs the fitting of the regression trees in the observed genotypes of the data. In this method, to perform the imputation of a missing point (y) in an accession sample (Y) the dataset is divided into four parts: 1) the observed genotypes in the accession Y (y_{obs}); 2) the missing genotypes in the accession Y (y_{mis}); 3) the observed genotypes in an accession (X) other than the accession Y (x_{obs}); 4) the observed genotypes in an accession (X) other than accession Y that is in the same position of the missing point y (x_{mis}). To initiate the imputation, first, the missing points are imputed by the mean value. Then, the observed data (y_{obs} and x_{obs}) other than the missing point (y) in X and Y are used to fit regression models. The number of samples used to construct each regression tree is a random sample of $\sqrt{x_{obs}}$. After that, the observed genotype in X at the same location as the missing point y (x_{mis}) is used to predict y with the model that was fitted in the genotyped data. We selected a number of 100 regression trees to predict a missing point y. Thus, the missing point was replaced by the mean of the 100 values predicted from the random forest regressions. After

the imputation, the new matrix replaces the previous. This process is repeated until it reaches a stipulated number of iterations or until a stop criterion, which occurs when the difference between the new imputed matrix increases from the previous matrix. When the stop criterion is reached, the matrix before the last is used for imputation, because the last matrix tends to have the lowest level of accuracy. We used the function *missForest* from the R package missForest (Stekhoven and Bühlmann 2012) to perform this imputation method. The maximum interaction was set to 10 and the number of trees in each regression forest was set as 100. As this method requires more computer power, the imputation procedure was parallelized into 10 CPUs. The parameters used in the function was ntree=100, maxiter=10, and parallelize = "variables". The data was considered continuous, therefore after the imputation, the imputed values in the matrix were rounded up to the nearest integer (1, 2, 3, or 4).

To measure the imputation accuracy of each method, the markers datasets were masked randomly by the insertion of three proportions of missing data (0.01, 0.05, and 0.15) with three replications for each proportion. Then the imputed datasets were compared to the initial, and the masked ones. The imputation accuracy was measured by the ratio between points correctly imputed and the total number of points imputed. We measured the total imputation accuracy and for each genotype AA, BB, and AB or BA. The masking of the datasets and the measure of imputation accuracy were performed in the software Tassel v 5.2.37 (Bradbury et al. 2007).

All programs were run on a Linux server from the Bioinformatic Laboratory of EMBRAPA (https://www.agropediabrasilis.cnptia.embrapa.br/web/lmb/home), with four Intel Xeon E5-2683 processor chips, each with 16 CPU cores, and 512 GB RAM. As the RF imputation was run in parallel in 10 cores, the time of imputation from RF was multiplied by 10.

4.2.5. Population structure and linkage disequilibrium

The population structure from the markers datasets described in table 1 was inferred using the program sNMF, available as the function *snmf* in the R package LEA (Frichot and François 2015). In this function, for each subpopulation (K), the ancestry coefficients and genotype frequencies are estimated by a sparse non-negative matrix factorization (sNMF) which is calculated using a modified least-squares algorithm as described by Frichot et al. (2014). The subpopulations were estimated from K=1 to 10, with 100 runs by K ($\alpha = 100$). The number of subpopulations was chosen based on the cross-entropy criterion: the smallest values of cross-entropy indicate the best K. We also analyzed the barplot of the Q matrix and considered the origin of the genotypes to choose the best K. After selecting the K, the best run was also selected based on the smallest cross-entropy value. From the sNMF analysis performed in the datasets with 110 genotypes (D1 to D4, Table 1), a threshold value of 0.6 for the ancestry coefficient was stipulated to divide the genotypes into groups, as described by Sant'Ana et al (2018). The genotypes that presented values below that threshold were added to a mixed group.

To measure the linkage disequilibrium (LD) decay of the population, we used the two datasets with 110 genotypes after the quality control, without imputation (D1 and D2, Table 1). Thus, the measure of LD was restricted to original markers. The LD decay measure was based on the pairwise distance and r^2 between SNPs at the same chromosome, with MAF > 0.1 and a maximum distance of 300 Kb. The values of total LD decay were calculated for r^2=0.2. The measures of LD and r^2 were performed with the software PopLDdecay (Zhang et al. 2019, https://github.com/BGI-shenzhen/PopLDdecay).

4.2.6. Genome-wide association studies

To identify SNPs associated with the content of the macronutrients N, P, K, Ca, and Mg in coffee grains, we used five multi-locus GWAS methods: multi-locus random-SNP-effect Mixed Linear Model (mrMLM, Wang et al. 2016); FASTmrMLM (Tamba and Zhang 2018); Fast multi-locus random-SNP-effect EMMA (FASTmrEMMA, Wen et al. 2018); polygene-background-control-based least angle regression plus empirical Bayes (pLARmEB, Zhang et al. 2017); Iterative Sure Independence Screening EM-Bayesian LASSO (ISIS EM-BLASSO, Tamba et al. 2017).

The procedures used by the five multi-locus models for GWAS can be divided into two stages. Considering that most of the markers probably are not associated with the quantitative trait, in the first step the potentially associated markers are selected. These markers are then added to a single multi-locus model which uses the expectation-maximization (EM) empirical Bayesian method to estimate the effect of each marker (Xu 2010). The markers that show nonzero effects are tested by the likelihood ratio test and the ones that present a LOD score value higher than a threshold are selected as candidate QTN.

The difference between the five models is basically in the first step. In mrMLM, first, a single marker random effect mixed linear model (RMLM) is used to identify the potentially associated markers. The markers that present a p-value < 0.01 are selected. Also, to reduce the effect of collinearity, the consecutive markers with p-value up to 0.01, which are in a range of ±20 Kb from an already selected marker are eliminated (Wang et al. 2016). The FASTmrMLM is similar to mrMLM, but it uses matrix transformation and identities to select the markers in the first step, therefore it performs the selection faster than mrMLM (Tamba and Zhang 2018). In ISIS EM-BLASSO, the first stage has two screening phases: one using the iterative sure independence screening (ISIS) method and the other using the smoothly clipped absolute deviation (SCAD) method (Tamba et al. 2017). The algorithms used in the first step of

FASTmrEMMA and pLARmEB whitens the covariance matrix of the kinship matrix (K) and environmental noise (Wen et al. 2018; Zhang et al. 2017). Then, in FASTmrEMMA the markers that present a p-value ≤ 0.005 are selected for the second step (Wen et al. 2018). In pLARmEB, the least angle regression (LAR) algorithm is used to select the potentially associated markers. In this method, while the markers in one chromosome are added to a multi-locus model, the markers from the other chromosomes are used to calculate the K matrix to control the polygenic background (Zhang et al. 2017). In the second step of the five methods, the effects of markers are estimated using an EM empirical Bayesian method (Xu 2010).

These methods were implemented in the R package mrMLM (Zhang et al. 2018a), which was used here for GWAS analysis. For all methods, the critical LOD score to identify significant QTNs was 3. In FASTmrEMMA, the QTN variances were estimated by REML. For pLARmEB, the number of potentially associated variables selected for each chromosome was 50.

The datasets used for GWAS are described in Table 1. For each dataset, the Q matrix from sNMF and the kinship matrix (K) were added as fixed effects to almost all models to correct the bias from population structure and kinship relationship among genotypes, respectively. The K matrix was calculated by the mrMLM package (Zhang et al. 2018a). In almost all analyses, we used the Q matrix with K=2, in exception to the GWAS analysis with dataset D10, in which the lowest cross-entropy was for K=1, thus for this dataset, the Q matrix was not included in the GWAS models. SNPs that were detected by at least two methods or in two years were considered as significant.

4.2.7. Functional annotation of candidate genes

From the significant SNPs, we selected the ones that were placed inside of coding sequences (CDS). The protein coded from this CDS were annotated according to the conserved

domains (InterPro) and the gene ontology (GO) term for molecular function (Ashburner et al. 2000; The Gene Ontology Consortium 2019). The functional annotation was performed using InterProScan v5.25-64.0 (Jones et al. 2014). For some proteins that there was no annotated GO term for the domains identified, we used studies that characterized and described the same domains as references for functional annotation. The significant SNPs that were colocalized to CDS of annotated proteins were named by the chromosome number in the subgenomes followed by its position, i.e. S1c_1000 refers to the base 1000 in chromosome 1 from subgenome C^a. Proteins from the Caturra reference genome were identified with the NCBI protein ID "XP_", while the proteins from the Et39 reference genome have the prefix "Cara".

4.3. Results

4.3.1. Macronutrient concentration in coffee grains

The content of N, P, K, Ca, and Mg in coffee grains, were analyzed in samples from *C. arabica* genotypes collected in the years of 2017 and 2018, which are described in the supplementary table S1. We observed significant ($p < 0.05$) genotypic and environmental variations for all the elements (Table 2). The Z-score distribution also revealed that the commercial cultivars present different contents of macronutrients in comparison to the non-commercial genotypes. Commercial cultivars accumulated less N than the other genotypes, in both years (Fig. 1). The differences in macronutrient content among the years can be explained by the natural coffee biennially trend added to the differences in environmental conditions in each year. The decreasing order of macronutrient content in coffee grains for both years was N > K > Mg > P > Ca (Table 2).

In both years, the Mg content had the highest percentage of broad-sense heritability: 94% in 2017 and 97% in 2018 (Table 2). The lowest values of heritability observed in 2017 and 2018 were found for N (72%) and Ca (93%), respectively. The correlation analysis indicated that in 2017 the P content had a moderate positive correlation to Ca (0.43) and K contents (0.47) (Table S2). In 2018, we observed a higher number of significant correlations: P showed a moderate positive correlation to K (0.39), N (0.41), Mg (0.31), and a weak correlation to Ca (0.27). Mg and Ca concentrations had also a moderate positive correlation (0.45), and the content of Mg and K had a weak positive correlation (0.28) (Table S2). All of the correlation coefficients for each nutrient between years were positive. Ca presented the lowest correlation (0.27) and N the highest (0.61) between the years (Table S2).

Table 2 Summary statistics of the coffee grain macronutrient content (mg.Kg-1), estimated heritability (h^2), likelihood ratio test (LRT) of the genotypic effect, and F-test of the environmental effect. The coffee grains were collected in 2017 and 2018 from 70 and 105 *C. arabica* genotypes, respectively. The plants were cultivated at IAPAR, in Londrina, PR, Brazil.

Trait	Year	Mean	SD[a]	Minimum	Maximum	h^2	Genotype[b]	Environment[c]
N	2017	20.75	1.84	16.75	24.66	0.72	251.03***	4.303*
	2018	20.82	1.97	14.60	26.31	0.94		
P	2017	1.51	0.12	1.28	1.98	0.87	223.03***	434.19***
	2018	1.34	0.12	1.04	1.69	0.94		
K	2017	17.47	1.11	15.30	20.35	0.89	207.97***	498.79***
	2018	15.87	1.15	13.54	19.43	0.96		
Ca	2017	1.47	0.16	1.24	1.98	0.92	156.92***	976.57***
	2018	1.16	0.18	0.85	1.73	0.93		
Mg	2017	1.85	0.16	1.41	2.25	0.94	280.75 ***	50.375***
	2018	1.77	0.19	1.24	2.19	0.97		

[a] SD = standard deviation.

[b] Depicts relevant effect (LRT). *** $p < 0.0001$

[c] Depicts relevant effect (F value). * $p < 0.05$; *** $p < 0.0001$

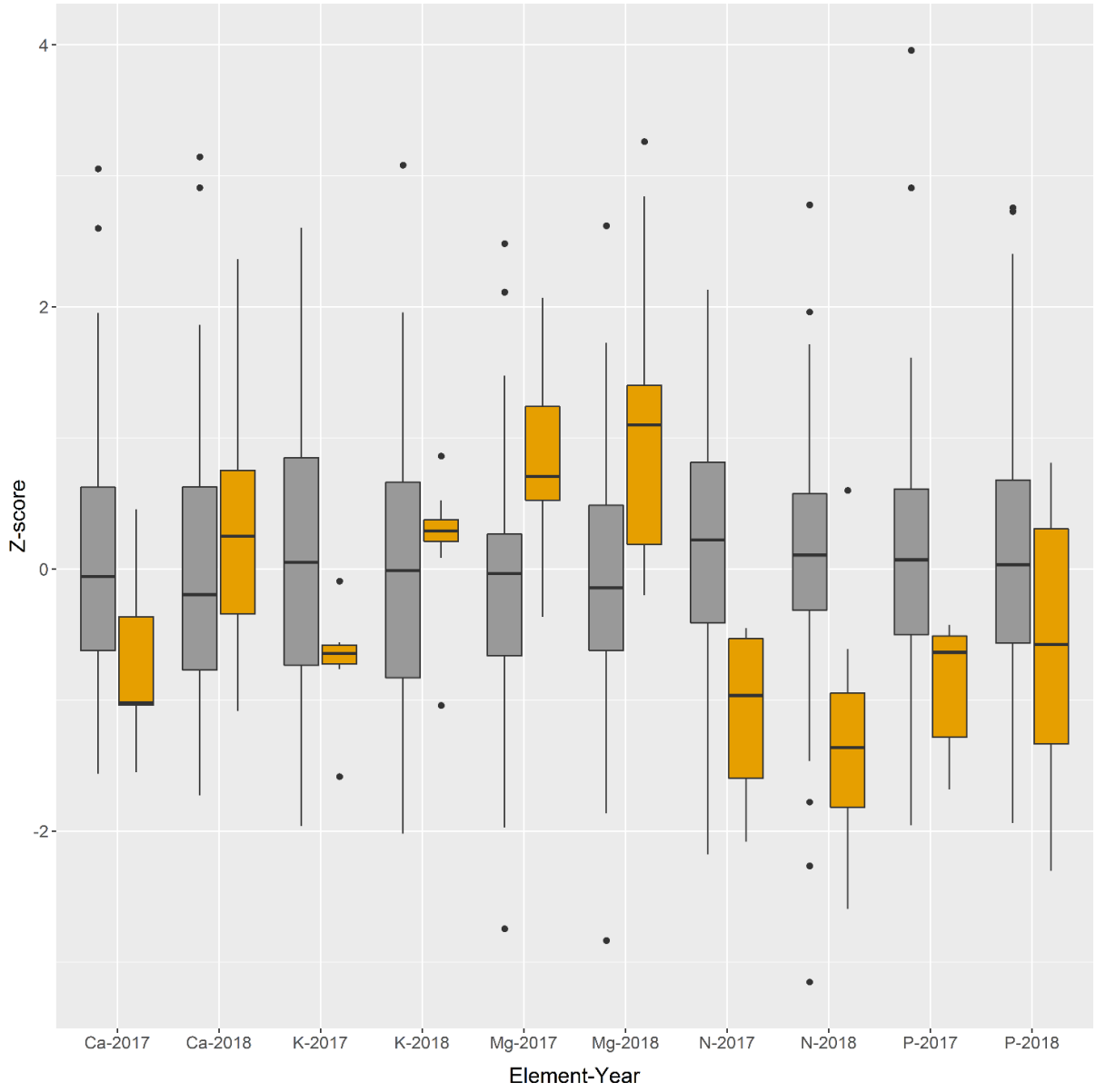

Fig 1 Z-score distribution of macronutrient content in coffee grains collected from 70 and 105 *C. arabica* genotypes in 2017 and 2018, respectively. Boxplots represent commercial cultivars (yellow) and non-commercial genotypes (grey). Phenotypic data from 2017 was collected from 7 commercial cultivars and 63 non-commercial genotypes. Samples from 2018 were collected from 11 commercial cultivars and 94 non-commercial genotypes. The y-axis represents the z-score distribution, and x-axis the macronutrient content by year of collection. This figure was generated using the R package ggplot2.

4.3.2. SNPs mapped to the *C. arabica* reference genomes

The GBS library previously sequenced by Sant'Ana et al. (2018) from the 159 *C. arabica* genotypes, generated approximately 540 million single-end reads, from which 1,083,002 tags were retrieved with high-quality reads. The percentage of tags aligned to Caturra and Et39 reference genomes were 24,45% and 24,07%, respectively. In this percentage were included the alignments to unique and to multiple positions. The raw number of SNPs was higher in the alignment to the Caturra reference genome (158,554 SNPs) than for Et39 (157,560 SNPs).

After the selection of the 110 samples that were phenotyped in this work, we applied the quality control and the number of SNPs was also higher in the dataset aligned to the Caturra reference genome (D1), with 11,230 SNPs. While for the alignment with the Et39 reference genome (D2), 9,991 SNPs were maintained (Table 3). The mean depth per site (averaged across all individuals) were: 15.60 x in D1 and 15.51 x in D2.

For both alignments, the total number of SNPs identified in each subgenome was similar and the markers were distributed across the two *C. arabica* reference genomes (Table 3, Fig. 2). Among the 22 chromosomes, the chromosome 2c and 2e had a higher number of mapped SNPs (> 900), while the chromosomes 9c and 9e from both alignments and 10e from Caturra alignment had the lowest number of mapped SNPs (> 270, Table 3, Fig. 2).

Although a similar proportion of tags were aligned to both reference genomes, we observed that the density of SNPs along the chromosomes was different when comparing the datasets, suggesting that different regions of the *C. arabica* genome were represented in each reference genome, i.e. at the initial part of the chromosomes 2c and 2e there was a higher density of SNPs in the dataset aligned to the Caturra reference genome (D1) when compared to the dataset aligned to the Et39 reference genome (D2, Fig. 2).

The missing rate and the proportion of alternative allele (BB) were the same in both panels: 10% and 8%, respectively (Table 3). While the proportion of heterozygous (AB) and homozygous for the reference allele (AA) varied between the panels. In the dataset D1, aligned to the Caturra reference genome, we observed a lower proportion of genotypes AA (67%) than the observed in the dataset D2 (72%), aligned to the Et39 reference genome. For the heterozygous genotype, the inverse occurred, a higher proportion was observed in D1 (25%) than in D2 (20%) (Table 3).

Table 3 Number of SNPs identified in a population of 110 *C. arabica* genotypes. The SNPs were identified by GBS, and the data was aligned to two *C. arabica* reference genomes (Caturra and Et39). The markers were separated according to the chromosomes in the *C. arabica* subgenomes: *C. canephora* (C^a) and *C. eugenioides* (E^a). From the total number of SNPs per reference genome the minor allele frequency (MAF), the proportion of missing data, homozygous (pAA, pBB), and heterozygous (pAB) genotypes were estimated.

Chromosome	**Reference genome**			
	Caturra (D1)		**Et39 (D2)**	
	C^a	**E^a**	**C^a**	**E^a**
1	567	858	417	476
2	1213	1105	1022	973
3	559	403	331	360
4	438	522	443	498
5	485	359	378	353
6	510	495	566	440
7	399	465	444	398
8	589	510	470	519
9	252	238	207	271
10	244	313	349	380
11	335	371	352	344
Total	5591	5639	4979	5012
Total n° of SNPs per reference genome	11230		9991	
Missing rate	0.10		0.10	
MAF	0.15		0.12	
p(AA)	0.67		0.72	
p(AB)	0.25		0.20	
p(BB)	0.08		0.08	

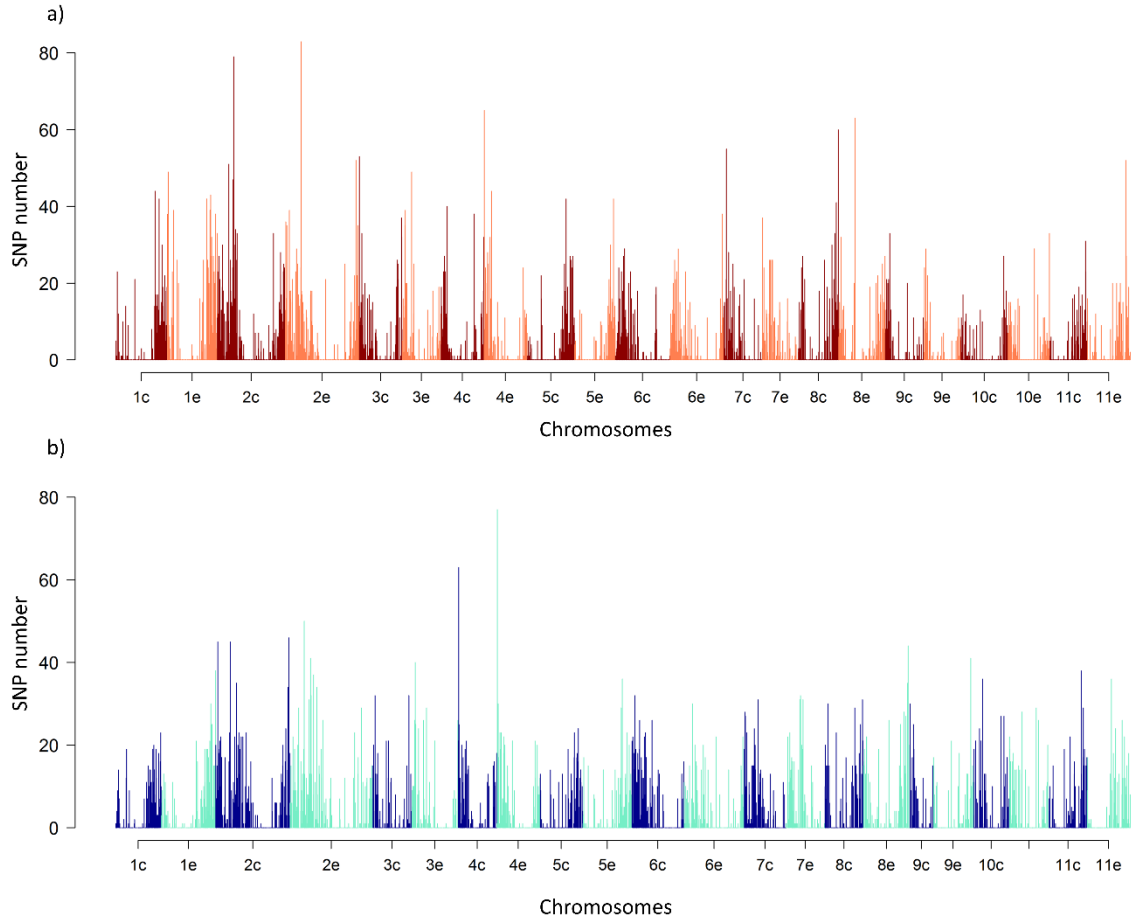

Fig 2 Distribution of SNPs across the chromosomes of *C. arabica* in 300 Kb windows. The SNPs were identified in 110 *C. arabica* genotypes by GBS. (a) Alignment to the Caturra reference genome. (b) Alignment to the Et39 reference genome. The letters c and e indicate the subgenomes *C. canephora* and *C. eugenioides*, respectively. The density of SNPs in the chromosomes from the *C. canephora* and *C. eugenioides* subgenomes were represented by darker and lighter colors, respectively.

4.3.3. Imputation accuracy

The datasets D1 and D2 were also used to measure the imputation accuracy of three imputation methods: Beagle, KNN, and RF. The three methods presented similar levels of total imputation accuracy for both panels, from 79% to 82% (Table S3). In general, the imputation accuracy was higher for the dataset of SNPs aligned to the Caturra reference genome than for

the dataset aligned to Et39 (Fig. 3, Table S3). For the panel aligned to the Caturra reference genome (D1), the mean imputation accuracy from Beagle and KNN was 82% and from RF was 80%. While for the panel aligned to the Et39 reference genome (D2), Beagle was the best imputation method with a mean accuracy of 82%, followed by KNN (80%) and RF (79%) (Table S3).

When considering the imputation accuracy for the separated genotypes (AA, AB, BB), we observed differences among the methods (Fig. 3). In the imputation of genotypes AA, Beagle had the highest accuracy levels (98%). While for the heterozygous genotype (AB), RF surpassed the other methods with a mean imputation accuracy of 53% for D1 and 31% for D2 (Fig. 3, Table S3). For the genotypes homozygous for the alternative allele (BB), Beagle and KNN presented the highest imputation accuracy percentage (38%) for D1, and KNN was the best imputation method for D2 (29% of accuracy).

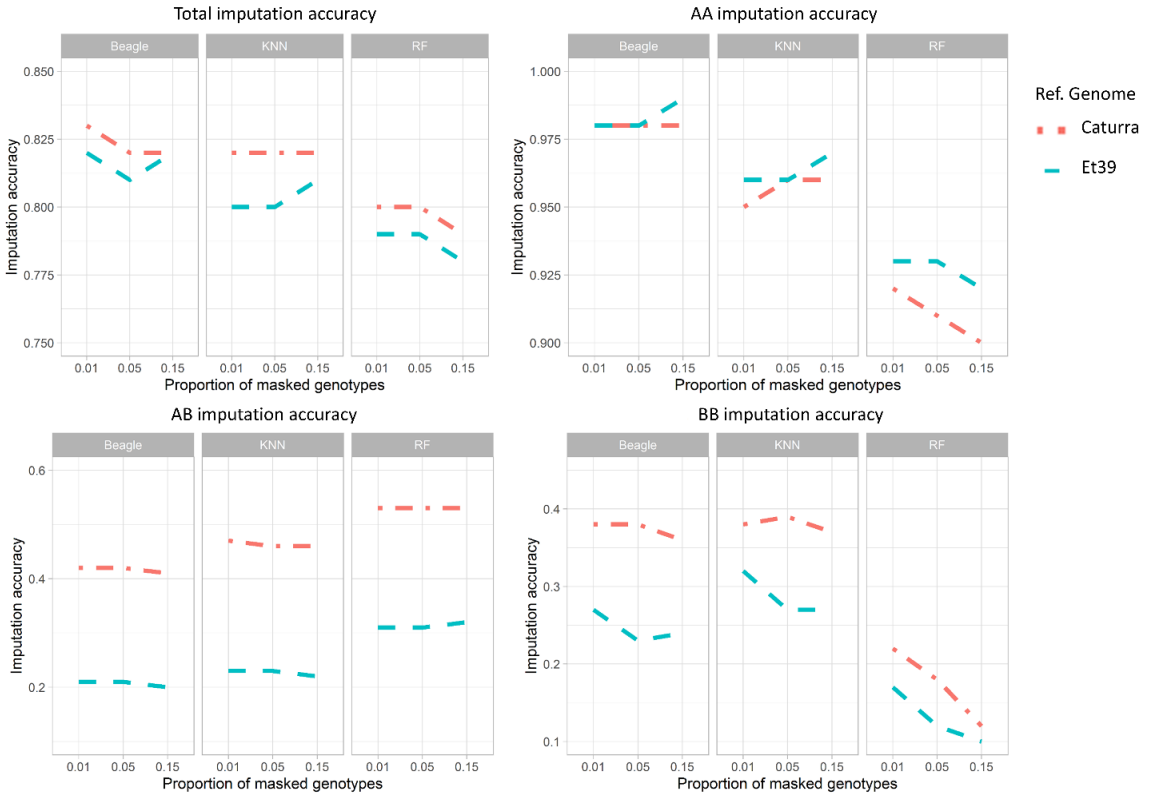

Fig 3 Comparison of three imputation methods (Beagle, KNN, and RF) using two SNPs datasets identified in a population with 110 C. arabica genotypes. The SNPs were called from alignments to the Caturra (pink) and Et39 (blue) reference genomes. Imputation accuracies are the mean values from three replicates from each level of masked genotypes (0.01; 0.05; 0.15) inserted in the panels. The y-axis were plotted according to the range of imputation accuracy in each category analyzed (total, AA, AB, and BB). This figure was generated using the R package ggplot2.

Beagle was the fastest imputation method in all the panels tested, with an average of 2.76 min and 2.51 min to impute D1 and D2, respectively (Table S3). KNN was in the second position, it took an average of 8.36 min to impute the dataset D1, and 5.48 min to impute the dataset D2. RF was the method that required more computational time to impute both datasets:

5.30 h to complete D2 imputation, and 6.75 h to impute D1. As the Beagle imputation required less computational time and had the highest percentage of total imputation accuracy for D2, and also one of the highest for D1 (82%, equal to KNN), this method was selected to impute the datasets used for GWAS, which are described in table 1.

4.3.4. Population structure and linkage disequilibrium

Due to the number of datasets analyzed in this work, to evaluate the efficiency of markers imputation for group identification, we plotted the Q matrix from the datasets with 110 genotypes before (D1, D2) and after imputation (D3, D4) (Fig. 4). Although the lowest cross-entropy value observed was for K=2 (Fig. S1), based on the origin of the genotypes and the analysis of the Q matrix, we also presented the results using K=3 (Fig. 4, Table S4).

When the population was divided into two groups, the alignment to the Et39 reference genome (D2) resulted in more genotypes allocated to groups Q1 (blue) and Q2 (green) and less in the mixed group than the observed from the alignment to the Caturra reference genome (D1, Fig 4, Table S4, and S5). In the dataset D2, 58 individuals were allocated to Q1 (blue), 42 to Q2 (green) and 10 to the mixed group (Fig 4), while in dataset D1, 66 genotypes were allocated into Q1, 26 in Q2 and 18 in the mixed group (Table S5). In both cases, the 11 commercial cultivars were placed in Q1, while most of the genotypes of Q2 were wild accessions from the FAO collection, in exception for BA10_057 which was included in Q2 in both datasets (D1 and D2) and SEL_106 that was included in Q2 in the dataset D2 (Table S4). Both of these genotypes are elite landraces of IAPAR coffee breeding program.

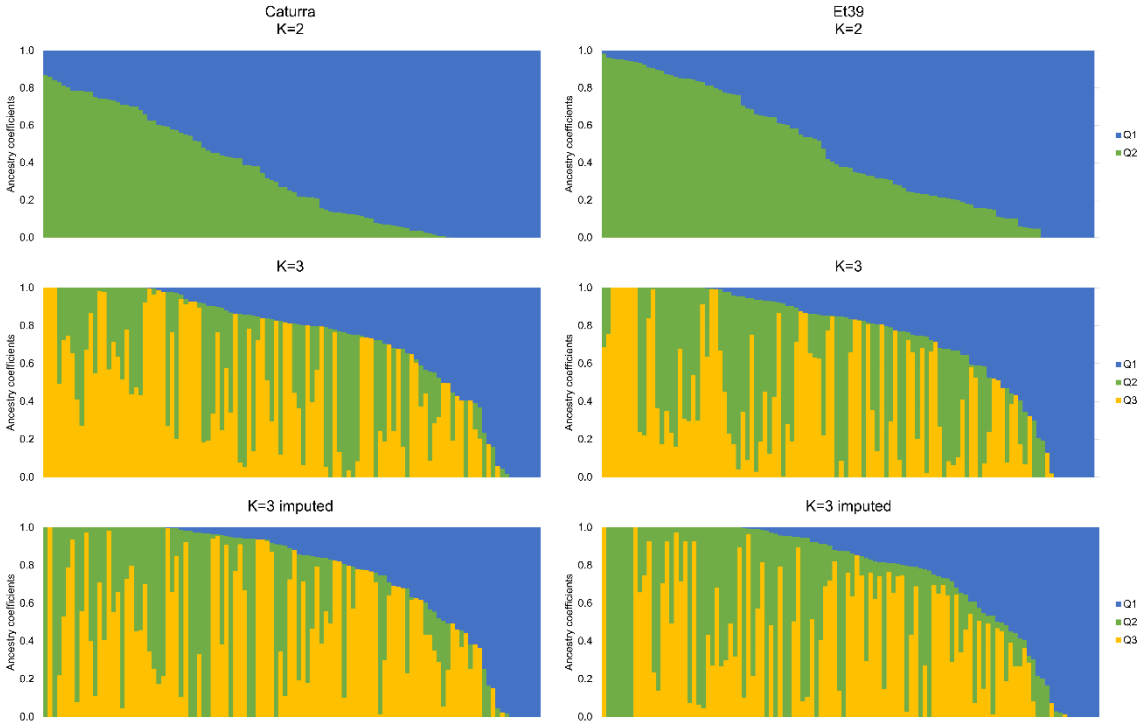

Fig 4 Population structure of 110 C. arabica genotypes. Group assignment (Q) from sNMF algorithms (K=2 and 3) using datasets aligned to the Caturra (left) and Et39 (right) reference genomes before and after imputation (K=3). The y-axis represents the values of the ancestry coefficients and the x-axis are the genotypes.

When the population was subdivided into three groups, for the datasets without imputation, the alignment to Et39 (D2) also resulted in fewer genotypes into the mixed group (M=23), while for the dataset aligned to Caturra (D1), 33 genotypes were included in that group (Table S5). The imputation reduced the number of genotypes placed into the mixed group for the dataset aligned to the Caturra reference genome (D3, M=27), while the opposite occurred to the dataset aligned to Et39 (D4) after the imputation more genotypes were placed into the mixed group (M=25, Table S5). We observed that 81 genotypes were placed in the same groups when comparing the datasets D1, D2, D3, and D4 using K=3. Among those genotypes, 17.28% were allocated in Q1 (blue), 23.46% in Q2 (green), 45.68% in Q3 (yellow), and 13.58% in the

mixed group. An exception for the cultivar Mundo Novo, all the other 10 commercial cultivars were placed into Q1 in all datasets, reinforcing the high similarity among these genotypes (Table S4).

We observed that the alignment to different reference genomes resulted in significant differences in the population LD decay. For the dataset aligned to the Caturra reference genome (D1), the mean LD decay among the 22 chromosomes for r^2=0.2 was 41 Kb, while for the dataset aligned to the Et39 reference genome it was 21 Kb.

4.3.5. GWAS Results

We used 10 datasets to identify SNPs associated with the content of macronutrients in coffee grains (Table 1). The number of SNPs identified with a LOD score above 3 was higher for the GWAS analysis with datasets aligned to the Et39 reference genome than to the Caturra reference genome with 300 and 247 associations, respectively.

Four datasets were compared before (D5, D7, D10, and D12) and after (D6, D8, D11, and D13) imputation. Considering the raw set of associations found in those datasets with LOD score higher than 3, more associations were found in the non-imputed (264) than in the imputed datasets (201). Also, in the GWAS results from the non-imputed datasets, the markers genotype information was missing for 79 associations (NN, Tables S6, and S7), which can be spurious associations. Thus, this indicates that the imputation helped to control the false discovery rate (FDR) of GWAS analysis as the missing data were replaced by putative genotypes and fewer associations were found in the imputed datasets.

The raw panel of SNPs associated with each trait was analyzed together to select the SNPs that were found at least by two methods or in two years (2017, 2018, and 2017/2018). From the datasets aligned to the Caturra reference genome, 75 markers were considered as significant (Table S5), of which 24 were placed in CDS. Similar results were observed for the

datasets aligned to the Et39 reference genome: 76 significant markers (Table S5) and 32 inside CDS.

We retrieved the protein sequences from markers placed inside of CDS. From the alignment to the Caturra and Et39 genomes, 21 and 22 proteins were annotated, respectively (Table 4). Interestingly, different genomic regions from each reference genome were associated with the coffee grains macronutrient content.

Among the annotated proteins, 24 were placed in the C^a subgenome and 19 in the E^a subgenome. The content of the macronutrients N, P, K, Ca, and Mg, in the coffee grains, were associated with 5, 12, 13, 6, and 6 annotated proteins, respectively. One protein (Cara010g024130) was associated with the contents of Mg and K, which were marked by two close SNPs 5e_36481790 and 5e_36481803. The association with both nutrients were found in datasets from 2018. For K, one association was also identified in the dataset of 2017/2018 (BLUP). Mg and K did not have a significant correlation in 2017, while in 2018 the nutrients presented a weak positive correlation of 0.28 (Table S2). This protein was annotated as a DNA Topoisomerase I (Table 4).

Table 4 Functional annotation of candidate genes colocalized with SNPs associated with the coffee grain macronutrient content**.** The SNPs were identified by GWAS analysis using datasets aligned to two *C. arabica* reference genomes (Caturra and Et39). The phenotypic traits used for GWAS were the content of N, P, K, Ca, and Mg. The candidate genes were described by the InterPro entry (IPR), protein or family domain, and the function according to the gene ontology (GO) terms. References in the literature were used to annotate the proteins in which the GO terms were missing.

Ref. genome	**Trait**	**Chr**	**Position**	**Protein ID**	**IPR**	**Protein/family domain**	**Function**	**GO/ Reference**
Et39	N	7c	7474388	Cara013g014850	IPR003035	RWP-RK domain	DNA binding	Chardin et al. (2014)
Et39	N	7c	12758930	Cara013g020080	IPR002401	Cytochrome P450, E-class, group I	Heme binding |Iron ion binding |Oxidoreductase activity.	GO:0020037 GO:0005506 GO:0016705
Caturra	N	7e	2651208	XP_027079538.1	IPR008630	Glycosyltransferase 34	Transferase activity, transferring glycosyl groups	GO:0016757
Et39	N	8c	30292619	Cara015g024690	IPR007493	Protein of unknown function DUF538	-	-
Caturra	N	9e	2042013	XP_027088358.1	IPR000504	RNA recognition motif domain	Nucleic acid binding	GO:0003676
Et39	P	1c	1045971	Cara001g005730	IPR006045	Cupin 1	Nutrient reservoir acitivity	GO:0045735
					IPR001929	Germin	Manganese ion binding	GO:0030145
Caturra	P	4c	3072247	XP_027122792.1	IPR000387	Tyrosine specific protein phosphatases domain	Phosphatase activity	GO:0016791
					IPR000340	Dual specificity phosphatase, catalytic domain	Protein tyrosine/serine/threonine phosphatase activity	GO:0008138
					IPR015275	Actin-fragmin kinase, catalytic domain	Phosphatase activity	Cheek et al. (2002)
					IPR020422	Dual specificity protein phosphatase domain	Protein tyrosine/serine/threonine phosphatase activity	GO:0008138
Et39	P	5c	29867171	Cara009g020230	IPR005333	Transcription factor, TCP	DNA-binding transcription factor activity	GO:0003700
Et39	P	6c	7442606	Cara011g013800	IPR001164	Arf GTPase activating protein	GTPase activator activity	GO:0005096
Caturra	P	6e	2219258	XP_027070752.1	PF00773[a]	RNB domain	Exonuclease	Schneider et al. (2007)

Ref. genome	Trait	Chr	Position	Protein ID	IPR	Protein/family domain	Function	GO/ Reference
Et39	P	7c	1131684	Cara013g006540	IPR019442	Domain of unknown function DUF2428, death-receptor-like	-	-
Caturra	P	7c	2399011	XP_027075580.1	IPR000719	Protein kinase domain	Protein kinase activity | ATP binding	GO:0004672 GO:0005524
					IPR001611	Leucine-rich repeat	Protein binding	GO:0005515
					IPR013210	Leucine-rich repeat-containing N-terminal, plant-type		
Caturra	P	7e	1384928	XP_027077691.1	IPR000719	Protein kinase domain	Protein kinase activity | ATP binding	GO:0004672 GO:0005524
Et39	P	7e	6318625	Cara014g013290	IPR011084	DNA repair metallo-beta-lactamase	Nuclease	Dominski (2007)
					IPR012308	DNA ligase, ATP-dependent, N-terminal	DNA binding| DNA ligase (ATP) activity.	GO:0003677 GO:0003910
					IPR012309	DNA ligase, ATP-dependent, C-terminal	DNA ligase (ATP) activity	GO:0003910
					IPR012310	DNA ligase, ATP-dependent, central	DNA binding | DNA ligase (ATP) activity.	GO:0003677 GO:0003910
Et39	P	10e	3044516	Cara020g008840	IPR000504	RNA recognition motif domain	Nucleic acid binding	GO:0003676
Caturra	P	10e	3347245	XP_027092098.1	IPR001932	PPM-type phosphatase domain	Catalytic activity	GO:0003824
Caturra	P	11e	24925979	XP_027099926.1	IPR001841	Zinc finger, RING-type	DNA, RNA, protein or lipid binding| Ubiquitination	Wu et al. (2014)
					PF13920[a]	Zinc finger, C3HC4 type (RING finger)	Protein-protein interaction | Ubiquitination	Wu et al. (2014)
Et39	K	1c	1009617	Cara001g028630	IPR013126	Heat shock protein 70 family	Assist in the protein folding process.	Usman et al. (2017)
Caturra	K	1e	34837440	XP_027081647.1	IPR006868	Domain of unknown function DUF630	-	-
					IPR006867	Domain of unknown function DUF632	-	-
Et39	K	2c	3036194	Cara003g008820	IPR006153	Cation/H+ exchanger	Solute:proton antiporter activity	GO:0015299
Et39	K	2e	15411452	Cara004g022170	IPR019787	Zinc finger, PHD-finger	Chromatin mediated gene regulation.	Sanchez and Zhou (2011)
					IPR004343	Plus-3 domain	DNA binding	GO:0003677
					IPR003121	SWIB/MDM2 domain	Protein binding	GO:0005515
					IPR000571	Zinc finger, CCCH-type	Metal ion binding	GO:0046872
					IPR003169	GYF domain	Protein binding	GO:0005515

Ref. genome	Trait	Chr	Position	Protein ID	IPR	Protein/family domain	Function	GO/ Reference
Et39	K	3c	2252158	Cara005g007870	IPR001611	Leucine-rich repeat	Protein binding	GO:0005515
Caturra	K	3c	11159390	XP_027117001.1	IPR023210	NADP-dependent oxidoreductase domain	Oxidoreductase activity	Sengupta et al. (2015)
					IPR020471	Aldo/keto reductase	Oxidoreductase activity	GO:0016491
Et39	K	5c	14986549	Cara009g010230	IPR017986	WD40-repeat-containing domain	Protein binding	GO:0005515
Caturra	K	5c	41982008	XP_027063151.1	IPR003245	Phytocyanin domain	Electron transfer activity	GO:0009055
Caturra	K	6e	36057694	XP_027072480.1	PR01217[b]	Proline rich extensin signature	Cell wall formation	Kishor et al. (2015)
Caturra	K	7e	45587	XP_027077681.1	IPR000719	Protein kinase domain	Protein kinase activity | ATP binding	GO:0004672 GO:0005524
					IPR001611	Leucine-rich repeat	Protein binding	GO:0005515
Et39	K	10c	4582941	Cara019g010540	IPR032552	Acin1, RNSP1-SAP18 binding (RSB) motif	Transcriptional regulation.	Murachelli et al. (2012)
Et39	K	11c	2494757	Cara021g006470	IPR000073	Alpha/beta hydrolase fold-1	Catalytic activity.	Mindrebo et al. (2016)
Caturra	K	11e	7373080	XP_027098638.1	IPR000639	Epoxide hydrolase-like	Catalytic activity	GO:0003824
					IPR000073	Alpha/beta hydrolase fold-1	Catalytic activity.	Mindrebo et al. (2016)
Caturra	Ca	1c	44829225	XP_027109136.1	IPR000225	Armadillo	Protein binding	GO:0005515
					IPR003613	U box domain	Ubiquitin-protein transferase activity	GO:0004842
Caturra	Ca	2c	51564194	XP_027107360.1	IPR007656	GTD-binding domain	Protein binding	Zhang et al. (2018b)
Et39	Ca	5c	24419353	Cara009g014640	IPR001214	SET domain	Protein binding	GO:0005515
Et39	Ca	6e	7581121	Cara012g014340	IPR000209	Peptidase S8/S53 domain	Serine-type endopeptidase activity | Proteolysis	GO:0004252 GO:0006508
Et39	Ca	8c	9060394	Cara015g010440	IPR005135	Endonuclease/exonuclease/phosphatase	Nuclease	Kaye et al. (2011)
Caturra	Ca	11e	39282028	XP_027098311.1	IPR001594	Palmitoyltransferase, DHHC domain	Palmitoyltransferase activity	GO:0016409
Et39	Mg	2c	13968839	Cara003g020670	IPR000719	Protein kinase domain	Protein kinase activity | ATP binding	GO:0004672 GO:0005524
					IPR001480	Bulb-type lectin domain superfamily	Carbohydrate binding	Zhao et al. (2017)
					IPR003609	PAN/Apple domain	Protein binding.	Cheng et al. (2003)
Caturra	Mg	3e	2970187	XP_027121290.1	IPR001180	Citron homology (CNH) domain	Protein interaction.	Hu et al. (2006)

Ref. genome	Trait	Chr	Position	Protein ID	IPR	Protein/family domain	Function	GO/ Reference
					IPR000547	Clathrin, heavy chain/VPS, 7-fold repeat	Intracellular protein transport	GO:0006886
					IPR019452	Vacuolar sorting protein 39/Transforming growth factor beta receptor-associated domain 1	Not characterized	-
					IPR019453	Vacuolar sorting protein 39/Transforming growth factor beta receptor-associated domain 2	Not characterized	-
Caturra	Mg	4c	39262132	XP_027123688.1	PF00773[a]	RNB domain	Exonuclease	Schneider et al. (2007)
					IPR022143	Protein of unknown function DUF3675	-	-
					IPR011016	Zinc finger, RING-CH-type	Zinc ion binding	GO:0008270
Caturra	Mg	4e	493778	XP_027126092.1	IPR018108	Mitochondrial substrate/solute carrier	Mitochondrial carrier protein	Lunetti et al. (2013)
Caturra	Mg	5e	36075183	XP_027127522.1	IPR003245	Phytocyanin domain	Electron transfer activity	GO:0009055
					PR01217[b]	Proline rich extensin signature	Cell wall formation	Kishor et al. (2015)
Et39	Mg	7c	777009	Cara013g006090	IPR000719	Protein kinase domain	Protein kinase activity | ATP binding	GO:0004672 GO:0005524
Et39	Mg[c]	5e	36481790	Cara010g024130	IPR008336	DNA topoisomerase I, DNA binding, eukaryotic-type	DNA binding |DNA topoisomerase type I (single strand cut, ATP-independent) activity.	GO:0003677 GO:0003917
	K[c]	5e	36481803		IPR013500	DNA topoisomerase I, catalytic core, eukaryotic-type	DNA binding |DNA topoisomerase type I (single strand cut, ATP-independent) activity.	GO:0003677 GO:0003917
					IPR025834	Topoisomerase I C-terminal domain	DNA breaking-rejoining enzyme	Kupriyanova et al. (2017)
					IPR001631	DNA topoisomerase I	DNA binding |DNA topoisomerase type I (single strand cut, ATP-independent) activity.	GO:0003677 GO:0003917

[a] Protein domain identified in Pfam database.

[b] Protein domain identified in PRINTS database.

[c] Traits associated with SNPs in close positions inside of the same CDS.

Among the proteins annotated the main functions were: binding to DNA, proteins, or other elements (i.e. zinc and iron); protein kinase activity; transport of metabolites; nucleases; cell wall formation; and catalytic activity (Table 4).

As many proteins were identified, for discussion, we selected a set of proteins involved in diverse important pathways including nutrient transport, embryogenesis, seed development, flowering time, and plant response to biotic and abiotic stress. The summary of the GWAS results from the markers colocalized to the candidate genes coding for these proteins is described in table 5. The manhattan plots, QQ-plots, and LOD score plots obtained from the GWAS analysis in which these markers were identified can be seen in the supplementary figures 2 to 12.

Table 5 Summary of the GWAS results from markers found in association with the coffee grain macronutrient content. The SNPs were identified by GBS using two *C. arabica* reference genomes (Caturra and Et39). The table presents the datasets in which the SNPs were identified, the GWAS models and the minimum and maximum values of LOD score, QTN effects, minor allele frequency (MAF), and correlation (r^2) to the associated trait estimated for each marker.

Ref. Genome	Trait	Chr	Position	Protein ID	Candidate gene	Datasets[a]	Models[b]	LOD score	QTN effect	r^2 (%)	MAF[c]
Et39	N	7c	7474388	Cara013g014850	RWP-RK	D12, D13	1, 2, 3, 4, 5	4.04 -8.56	0.00 - 1.48	0-28.60	0.22 - 0.23
Et39	P	1c	1045971	Cara001g005730	Cupin, Germin-like	D10, D11	2, 4	3.05 - 5	-0.05 - 0.04	0.07 - 9.89	0.14 - 0.22
Caturra	P	11e	24925979	XP_027099926.1	RING-type zinc finger	D5	1, 4	3.54 - 4.39	0.05 - 0.07	9.72 - 12.04	0.21
Et39	K	2e	15411452	Cara004g022170	PHD zinc finger	D10	1, 2, 4, 5	3.17 - 4.34	0.54 - 0.86	9.86 - 24.25	0.11 - 0.12
Et39	K	5e	36481803	Cara010g024130	Topoisomerase I	D13, D14	2, 4, 5	3.08 - 4.52	-0.57 - -0.44	9.48 - 16.11	0.14 - 0.16
Et39	Mg	5e	36481790	Cara010g024130	Topoisomerase I	D12	1, 2, 3, 5	3.32 - 6.44	0 - 0.13	0 - 18.68	0.12
Caturra	Mg	4c	39262132	XP_027123688.1	RING-type zinc finger	D9	2, 5	4.10 - 9.4	-0.05 - -0.06	8.57 - 8.59	0.15
Caturra	Ca	11e	39282028	XP_027098311.1	Palmitoyl transferase DHHC	D6	1, 2	3.04 - 4.11	0 - 0.11	0 - 4.25	0.10 - 0.12

[a] Number of the dataset in which was detected the association. The datasets are described in table 1.

[b] Models used for GWAS. 1 - mrMLM, 2 - FASTmrMLM, 3 - FASTmrEMMA, 4 - pLARmEB, 5 - ISIS EM-BLASSO.

[c] MAF: minor allele frequency.

4.4. Discussion

The coffee grain is an important tropical commodity traded worldwide. The macronutrient content in coffee grains directly interferes in the grain quality, however, the genomic regions associated with those complex traits are still unknown. Therefore, here we used a panel of 110 *C. arabica* genotypes, most of it wild accessions (96), to identify genomic regions associated with the content of five macronutrients in coffee grains: N, P, K, Ca, and Mg. As *C. arabica* is a non-model plant, complete annotated reference genomes from the species were released only recently. Thus, using GBS data we performed the SNP calling from the same panel using two *C. arabica* reference genomes and tested three imputation methods to replace missing data. The GWAS results indicated that imputation favored the control of FDR and that using two reference genomes is a good approach to identify different genomic regions that regulate important traits, some of which we discussed here.

Our results showed that the wild *C. arabica* genotypes harbor a higher range of variability considering the genotypic and phenotypic data. The analysis of the macronutrient content distribution shows the effect of selection and domestication of the commercial cultivars, i.e. the N content was reduced and the Mg content increased in comparison to the wild genotypes (Fig 1). In the wild genotypes, we found a variability that could not be found in the commercial cultivars.

Analysis of other important traits for coffee production, as the level of biochemical compounds that affect the coffee cup quality (Tesseema et al. 2011; dos Santos Scholz et al. 2016; Tran et al. 2017; Sant'Ana et al. 2018) and characteristics of grain morphology (Gaspari-Pezzopane 2004; Tran et al. 2017) also revealed high phenotypic variability among wild *C. arabica* genotypes.

An initial panel of 159 *C. arabica* genotypes (Sant'Ana et al. 2018), was used to identify SNPs via GBS with two *C. arabica* reference genomes for SNP calling. The percentage of total

reads aligned to multiple and unique positions in both of the reference genomes was similar (~ 24%). When the same GBS library was mapped to the *C. canephora* reference genome, 22% of the reads aligned to unique positions (Sant'Ana et al. 2018). Considering that *C. arabica* is a tetraploid species and *C. canephora* one of its diploid relatives, we expected a higher percentage of alignment to the own species reference genome. However, the similarity in the percentage of reads aligned to *C. arabica* and *C. canephora* reference genomes, may be due to the relatively recent origin of *C. arabica* (< 0.665 million years ago) (Yu et al. 2011) and the low sequence divergence between the genomes of *C. arabica* and *C. canephora* and the *C. arabica* subgenomes, C^a and E^a that have a high synteny (Cenci et al. 2012). Thus, probably the same genomic regions were covered in both genomes, resulting in a similar percentage of reads alignment.

After the selection of the 110 genotypes that were also phenotyped in this work and the application of the quality control, several high-quality SNPs were retrieved from the alignment to both reference genomes: 9,900 and 11,230 in the datasets D1 and D2 aligned to Caturra and Et39, respectively. The use of the reference genome from the own species revealed a higher number of polymorphisms than the identified by Sant'Ana et al. (2018) in 107 genotypes of the same population, in which 2,587 high-quality SNPs were identified from the alignment to the *C. canephora* reference genome.

In the population structure analysis, we observed that when the genotypes were divided into three groups (K=3), most of the commercial cultivars were allocated into a small group (Q1, blue) in all datasets (Fig. 4, Table S4), confirming the narrow genetic base found among commercial cultivars, and the higher genetic variability among wild accessions. Therefore, our results confirm previous works that revealed a higher genetic variability among *C. arabica* wild genotypes using different types of markers, including microsatellites (Anthony et al. 2002; Silvestrini et al. 2007; da Silva et al. 2019), AFLP (Anthony et al. 2002), RAPD (Anthony et

al. 2001), and more recently SNPs (Sant'Ana et al. 2018; Merot-L'anthoene et al. 2019). These results reinforce the need for maintaining and preserving the wild *C. arabica* genotypes considering that part of the diversity harbor by those genotypes can not be found among the commercial cultivars.

The measure of LD among pairs of markers revealed a significant difference in LD decay between the datasets aligned to each reference genome: the dataset aligned to the Et39 reference genome presented a faster decay (21 Kb) than the dataset aligned to Caturra (41 Kb). Also, the distribution of SNPs along the genome was slightly different when comparing the two reference genomes (Fig. 2). Therefore, probably, the alignment to the Et39 reference genome revealed more sites of recombination than the alignment to Caturra, leading to a faster decay of LD. We suggest that this is due to the different origins of the genotypes used to construct the reference genomes.

Caturra is a commercial cultivar from the Bourbon group. On the other hand, the genotype Et39 was obtained from an accession collected in the species center of origin (Guillaumet and Hallé 1978), thus it did not pass by the same domestication bottleneck occurred in the plants from the Bourbon group. Therefore, probably, there were some sequences in the Et39 reference genome, which were not present in the Caturra reference genome. These sequences may have a higher frequency of recombination. Thus, the maintenance of these regions in the genome of commercial cultivars would lead to higher segregation rates during the selection of commercial cultivars, which is not the objective for most of the coffee breeding programs that seek to produce uniform commercial cultivars.

The previous alignment to the *C. canephora* genome showed a lower LD decay for a similar panel composed of 107 *C. arabica* genotypes (~ 280 Kb, Sant'Ana et al. 2018). The faster LD decay observed in our datasets can be explained also by the higher number of polymorphisms found, which revealed more sites of recombination, and probably leading to an

estimation of faster LD decay. The measure of LD in a diverse panel of genotypes from *Medicago truncatula*, which is also an autogamous species, also revealed a non-extensive measure of LD (from 1 to 10 Kb) among SNP markers (Branca et al. 2011).

Low-coverage GBS data generally presents a high rate of missing data that can interfere in the accuracy of associations found by GWAS (Rahimi et al. 2019), thus we tested three imputation methods to replace the missing points in the SNPs datasets: Beagle, KNN, and RF. The three methods presented similar results for the total imputation accuracies of the datasets (~80%, Fig. 2, Table S3).

Although KNN and RF are general methods, the LD between markers and the population structure also interfere in the imputation accuracy (Nazzicari et al. 2016). In KNN the distance between markers was used to weigh the average between the nearest markers, and in RF the multiple regressions training was performed with SNPs from the same population that were genotyped in the point to be imputed. Thus, it explains the similarity in the performance of the three methods, as the Beagle imputation is based on the identification of localized haplotypes, and in the other two methods, the LD and population were also used indirectly (Rutkoski et al. 2013; Nazzicari et al. 2016).

Despite the similarity between the methods, Beagle was selected as the best imputation method due to its higher accuracy level (82%) and the minor computational time required for imputation (Table S3).

In general, the imputation accuracies observed in our dataset were lower than the registered when Beagle was used for the imputation of GBS data from species with a well-annotated reference genome. As an example, when Beagle was used to impute a panel of markers identified by GBS in a population of 301 soybean accessions, the accuracy of imputation ranged from 85.6% to 94% in panels that presented from 20% to 80% of missing data, respectively (Torkamaneh and Belzile 2015).

However, low imputation accuracy levels were observed when Beagle was used to impute GBS data from non-model species. For cassava (*Manihot esculenta*) the imputation accuracy with Beagle was 76.48% in a GBS dataset aligned to the own species reference genome (Chan et al. 2016). The imputation of GBS data from alfalfa (*Medicago sativa*) aligned to the reference genome of the related species *M. truncatula* resulted in an accuracy of 73.14% (Nazzicari et al. 2016). As *C. arabica* is a non-model species, with reference genomes recently sequenced and that until now do not have a reference haplotype panel for imputation, Beagle presented a relatively high accuracy rate in the imputation of the datasets used in our work.

Analyzing the raw number of associations with the LOD score higher than 3, more associations were found in the datasets aligned to the Et39 reference genome (300) than in the datasets aligned to the Caturra reference genome (247) (Table S5 and S6). The higher number of associations found using the datasets aligned to the Et39 reference genome can be related to the regions of the genome that were covered by the GBS. Considering that probably in Et39 the regions identified have a higher proportion of recombination than the identified using the Caturra reference genome, it contributed to the identification of more markers associated with the traits.

The imputed datasets led to the identification of fewer associations in GWAS analysis (201) compared to the associations retrieved from non-imputed datasets (264) (Table S5 and S6). Opposite results were obtained when imputed SNPs datasets from soybean (Torkamaneh and Belzile 2015), and bread wheat (Rahimi et al. 2019) were used for GWAS, providing a higher number of significant associations than from non-imputed datasets. However, the number of genotypes and SNPs used in those GWAS analyses was considerably higher than the used here: more than 298 genotypes and 46,000 SNPs (Rahimi et al. 2019), which explains the better performance of GWAS for those imputed datasets. We hypothesize that the imputation of missing genotypes in the *C. arabica* datasets helped to control the false-positive

associations identified in non-imputed datasets, explaining the reduced number of associations found in the imputed panels.

From the GWAS analysis, we identified 151 significant SNPs associated with the coffee grains macronutrient content (Table S6 and S7), of which 43 were annotated (Table 4). One interesting result was found from the association with the N content in coffee grains, the SNP S7c_7474388 placed in the CDS of the protein Cara013g014850, which has a domain of the transcription factor RWP-RK protein domain (RKDs) (Table 5, Fig. S2, and S3). In some plants, the RKDs regulates the embryogenesis and the female gametogenesis, as reported for *A. thaliana* (Waki et al. 2011; Tedeschi et al. 2017), and wheat (Kumar et al. 2018). In wheat, the RKDs also have transcriptional activity during seed germination (Kumar et al. 2018). In the genome of the algae *Chlamydomonas reinhardtii*, the RKDs regulates the gametogenesis in response to N starvation (Lin and Goodenough 2007). Therefore, as the protein Cara013g014850 was identified in association with the N content in coffee grains, it is likely that this protein is also involved in the gametogenesis regulation in response to N availability.

Another important result was the identification of a protein (Cara001g005730) with the Cupin 1 and Germin domains. This protein was found in association with the P content in coffee grains marked by the SNP S1c_1045971 (Table 5, Fig. S4 and S5). The Cupin is a superfamily that includes the Germin-like proteins (GLPs), enzymes that have diverse functions, including oxalate oxidase and superoxide dismutase activities (Dunwell et al. 2004). GLPs are ubiquitous proteins expressed in various plant organs and stages of development and are also regulated in response to biotic and abiotic stresses (Barman and Banerjee 2015). A study with *A. thaliana* showed that two GLPs genes (*At5g39160* and *At5g39130*) were upregulated in leaves of P-starved plants cultivated with and without sucrose, suggesting that these genes may be directly involved in P remobilization process (Müller et al 2007), which agrees to the association found

in our study. Therefore, the protein Cara001g005730 probably acts as a P transporter in the cells of coffee grains.

The transcriptional level of a GLP gene was measured in germinating seeds of two *C. arabica* cultivars (Catuaí and Icatu) under Al^{3+} stress conditions. The study revealed that in Icatu seedlings the GLP is upregulated in response to the entry of Al^{3+} into the cytoplasm, acting as an antioxidant enzyme (Bazzo et al. 2013). The content of GLP was also altered in leaves of *C. arabica* plants infected with important fungal diseases: coffee leaf rust (*Hemileia vastatrix*) and coffee berry disease (*Colletotrichum kahawae*); suggesting that the GLP proteins may be involved in the pathway of fungal resistance response in coffee plants (Diniz et al. 2015; Guerra-Guimarães et al. 2015), as also observed for several other plants (reviewed by Ilyas et al. 2016). In other species, some *GLP* genes presented specific transcriptional profile in seed tissues: in *A. thaliana* the genes *AtGLP3-5* and *AtGLP5-10* were highly expressed during seed development; in rice, two genes (*OsGLP3-3* and *OsGLP8-2*) had their expression restricted to the seeds (Li et al. 2016), thus GLPs participate in the process of seed maturation. As this protein have diverse functions in plant metabolism, and for *C. arabica* the studies related to that family domain at the molecular level still scarce, we suggest that GLPs should be a target of more research to reveal metabolic pathways related to the P transport, seed germination, and the response to stress in non-model plants.

Among the annotated proteins, one protein (Cara010g024130) was found in association with the Mg and K content, marked with SNPs separated by only seven bases (S5e_36481790 and S5e_36481803) (Table 5, Fig. S6-S8). The protein Cara010g024130 was annotated as a Topoisomerase I (Table 5). This enzyme is responsible for modulating the DNA topology, which is essential for replication and transcription. Topoisomerase I uses Mg^{2+} ions as cofactors to cut one strand of the DNA double helix, removing supercoils and entanglements, which preserve the chromosome structure (Vos et al. 2011; Gong et al. 2017). Recent studies revealed

that the gene TOPOISOMERASE 1α of *A. thaliana* regulates the transcription of genes that control the flowering time (Gong et al. 2017; Zhong et al. 2019). As the flowering time is essential for plant reproduction and seed formation, the protein Cara010g024130 from *C. arabica* may also develop a similar function as TOPOISOMERASE 1α, regulating the coffee flowering time and probably also the grain development.

We also found three proteins annotated with Zinc finger domains which were distributed in distinct genomic regions and associated with the content of three macronutrients in coffee grains: P, K, and Mg. The CDS for the protein Cara004g022170 was colocalized with the SNP S2e_15411452 associated with the K content. This protein was annotated with a PHD zinc finger domain (Table 5, Fig S9). The other two proteins (XP_027123688.1 and XP_027099926.1) have a RING-type zinc finger domain and were colocalized to the SNPs S4c_39262132 and S11e_24925979 associated to the Mg and P content in coffee grains, respectively (Table 5, Fig. S10, and S11).

Zinc fingers are a large family of transcriptional factors (Wu et al. 2014). In *A. thaliana* a study revealed that a PHD zinc finger domain from the transcription factor *HSI2* probably represses the expression of a specific subset of seed maturation genes, including genes coding for the Cupin family protein (Veerappan et al. 2012). In another study, a PHD finger protein homolog (PFP) also repressed the expression of a gene that controls the flowering time (*FLOWERING LOCUS C*) in *A. thaliana* (Yokoyama et al. 2019).

The other type of zinc finger identified in our work, the RING domain, seems to be involved in the control of organ size. In *A. thaliana* was observed that the RING finger protein BIG BROTHER (BB) is a repressor of plant growth (Disch et al. 2006), and another RING finger protein, the FLYING SAUCER1 regulates the deposition of pectin in the cell wall of seeds (Voiniciuc et al. 2013). For *Nicotiana benthamiana*, the silencing a C3HC4-type RING finger protein reduced the fruit development (Wu et al. 2014).

We also identified the SNP S11e_39282028 in association with the Ca content in coffee grains, this SNP was placed in the CDS of a Palmitoyl transferase DHHC domain (XP_027098311.1) (Table 5 and Fig. S12). The Ca function is related to the cell division and the stabilization of the cell wall (Laviola et al. 2007). A study revealed that in *Aspergillus nidulans* a Palmitoyl transferase DHHC motif participates in the regulation of the Ca^{2+} homeostasis (Zhang et al. 2016). Thus, it indicates that in the coffee grains the DHHC domain may also participate in a similar process, regulating Ca homeostasis.

We suggest that the proteins listed in Table 4 can be targets for further analysis to reveal the metabolic pathways related to macronutrient content accumulation in coffee grains.

4.5. Conclusions

Our work showed that the macronutrient content in coffee grains of commercial cultivars has less variability than observed among wild genotypes. The genetic variability of the wild germplasm was also higher than the observed in commercial cultivars. Thus, we reinforce the need for conserving coffee wild germplasm, which can be used for the identification and introduction of favorable alleles to improve the coffee grain macronutrient content.

Here, we used two reference genomes for SNP calling and identified several significant associations with genomic regions distributed along the *C. arabica* genome. For each reference genome, different annotated genes were found in association with the macronutrient content. Thus, we demonstrated that the approach of using more than one reference genome for SNP calling can favor the identification of more genomic regions that are related to the phenotypic traits. We identified some of the proteins that regulate the macronutrient content in coffee grains, which are also involved in diverse pathways of plant metabolism. We also demonstrated that the imputation of missing data helped to control the FDR in GWAS. We suggest that the genomic regions identified should be targets for further studies to increase the knowledge about the molecular basis that controls the macronutrient content in coffee grains.

Acknowledgments: The authors would like to thank Lucineia Maria da Silva and Manuel Luiz Martins who helped with the coffee grains sampling. Rosineia Aparecida de Souza for sharing her knowledge in macronutrient quantification analysis. We also thank Leandro Carrijo Cintra for helping with the bioinformatic tools.

Funding: This study was financed in part from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Consórcio Pesquisa Café. Mariane

Silva Felicio was a recipient of a scholarship from CAPES. Douglas Silva Domingues and Luiz Filipe Protasio Pereira are CNPq research fellows.

Conflicts of interest: The authors declare that they have no conflict of interest.

5. CONCLUSÕES GERAIS

Os resultados obtidos nesse trabalho indicam que os genótipos selvagens de *C. arabica* abrigam uma ampla variabilidade para o conteúdo de macronutrientes em grãos. Além disso, as análises genotípicas também confirmaram maior variabilidade genética entre genótipos selvagens quando comparados às cultivares comerciais. Com isso, esse trabalho pode contribuir para a identificação de genótipos com diferentes níveis de concentração de macronutrientes nos grãos a serem introduzidos aos programas de melhoramento do café, visando maior qualidade do grão e da bebida. Portanto, a preservação de genótipos selvagens de *C. arabica* é essencial para o melhoramento da espécie e sua sobrevivência em frente a diferentes fatores bióticos e abióticos.

O alinhamento dos dados de GBS de *C. arabica* à dois genomas de referência completos da própria espécie favoreceu a identificação de um grande número de polimorfismos na população (9000 - 11230 SNPs). O desempenho dos métodos de imputação de marcadores Beagle, KNN e RF foi similar, com aproximadamente 80% de acurácia para os dois alinhamentos. Como *C. arabica* é uma espécie não modelo e não possui painéis de referência para a imputação de marcadores, a acurácia obtida pode ser considerada elevada. Além disso, a imputação de marcadores com o software Beagle favoreceu o controle da identificação de associações espúrias, as quais foram identificadas nos painéis não imputados.

Nesse trabalho também foram identificadas diversas regiões genômicas associadas à composição de cinco macronutrientes (N, P, K, Ca e Mg) em grãos de *C. arabica*. Portanto, a estratégia de utilizar dois genomas para o alinhamento dos dados de GBS se mostrou eficaz, considerando que diferentes regiões foram identificadas através das análises de GWAS a partir de cada genoma de referência.

As proteínas codificadas pelos genes candidatos anotados nesse trabalho desempenham diferentes funções no metabolismo da planta, por isso, sugerimos que esses genes sejam

utilizados como alvos de estudos futuros para a melhor caracterização da base genética da composição de macronutrientes em grãos de *C. arabica*. O presente trabalho contribuiu para a identificação de diversos marcadores associados com a composição de macronutrientes nos grãos que após validados poderão ser utilizados para a seleção assistida por marcadores de genótipos de *C. arabica*.

6. REFERÊNCIAS

Anthony F, Astorga C, Berthaud J (1999) Los recursos genéticos: las bases de una solución genética a los problemas de la caficultura latinoamericana. In: Desafíos de la caficultura centroamericana. IICA-PROMECAFE-CIRAD, San José, pp 369-406.

Anthony F, Bertrand B, Quiros O, et al (2001) Genetic diversity of wild coffee (Coffea arabica L.) using molecular markers. Euphytica 118:53–65. https://doi.org/10.1023/A:1004013815166

Anthony F, Combes MC, Astorga C, et al (2002) The origin of cultivated Coffea arabica L. varieties revealed by AFLP and SSR markers. Theor Appl Genet 104:894–900. https://doi.org/10.1007/s00122-001-0798-8

Ashburner M, Ball CA, Blake JA, et al (2000) Gene Ontology : tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29. https://doi.org/10.1038/75556

Associação Brasileira da Indústria de Café (ABIC) (2015) Café é a segunda bebida mais consumida no Brasil. http://www.consorciopesquisacafe.com.br/index.php/imprensa/noticias/580-cafe-e-a-segunda-bebida-mais-consumida-no-brasil. Accessed 24 January 2018

Barman AR, Banerjee J (2015) Versatility of germin-like proteins in their sequences, expressions, and functions. Funct Integr Genomics 15:533–548. https://doi.org/10.1007/s10142-015-0454-z

Bazzo BR, Eiras A de L, DeLaat DM, et al (2013) Gene Expression Analysis Suggests Temporal Differential Response to Aluminum in Coffea arabica Cultivars. Trop Plant Biol 6:191–198. https://doi.org/10.1007/s12042-013-9120-6

Berthaud J (1976) Etude cytogénétique d'un haploïde de Coffea arabica L. Café, Cacao, Thé

(Francia) v. 20 (2):91-96.

Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Branca A, Paape TD, Zhou P, et al (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. Proc Natl Acad Sci U S A 108:864–870. https://doi.org/10.1073/pnas.1104032108

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Browning BL, Browning SR (2016) Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98:116–126. https://doi.org/10.1016/j.ajhg.2015.11.020

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097. https://doi.org/10.1086/521987

Caixeta ET, Oliveira AD, Brito GD, Sakiyama N S (2016). Tipos de marcadores moleculares. In: Marcadores moleculares, Editora UFV, Viçosa, pp 9-93.

Catani RA, Pellegrino D, Alcarde, JC, Graner CAF (1967). Variação na concentração e na quantidade de macro e micronutrientes no fruto do cafeeiro, durante o seu desenvolvimento. Anais da escola superior de agricultura Luiz de Queiroz 24:249-263. https://doi.org/10.1590/S0071-12761967000100024

Catchen J, Hohenlohe PA, Bassham S, et al (2013) Stacks: An analysis tool set for population genomics. Mol Ecol 22:3124–3140. https://doi.org/10.1111/mec.12354

Cenci A, Combes MC, Lashermes P (2012) Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. Plant Mol Biol 78:135–145. https://doi.org/10.1007/s11103-011-9852-3

Chan AW, Hamblin MT, Jannink JL (2016) Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. PLoS One 11:1–17. https://doi.org/10.1371/journal.pone.0160733

Chardin C, Girin T, Roudier F, et al (2014) The plant RWP-RK transcription factors: Key regulators of nitrogen responses and of gametophyte development. J Exp Bot 65:5577–5587. https://doi.org/10.1093/jxb/eru261

Charrier A, Berthaud J (1985) Botanical classification of coffee. In: Coffee, Springer, Boston, pp 13-47.

Cheek S, Zhang H, Grishin N V. (2002) Sequence and structure classification of kinases. J Mol Biol 320:855–881. https://doi.org/10.1016/S0022-2836(02)00538-7

Cheng Q, Sun MF, Kravtsov DV, Aktimur LA, Gailani D (2003) Factor XI apple domains and protein dimerization. Journal of Thrombosis and Haemostasis, 1(11): 2340-2347. https://doi.org/10.1046/j.1538-7836.2003.00418.x

Ching ADA, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S et al. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC genetics, 3(1), 19.

Clarindo WR, Carvalho CR (2008). Comparison of the Coffea canephora and C. arabica karyotype based on chromosomal DNA content. Plant Cell Rep, 28:73-81. https://doi.org/10.1007/s00299-008-0621-y

Clifford MN (1985) Chemical and physical aspects of green coffee and coffee products. In: Coffee Botany, Biochemistry and Production of Beans and Beverage. Springer, Boston, pp 305-374.

Companhia Nacional de Abastecimento (CONAB) (2018) Acompanhamento da Safra Brasileira: Café, Safra 2018 - Quarto Levantamento Dezembro 2018. Brasília, pp 1-84.

Companhia Nacional de Abastecimento (CONAB) (2019) Acompanhamento da Safra

Brasileira: Café, Safra 2019 - Terceiro Levantamento Setembro 2019. Brasília, pp 1-48.

da Silva BSR, Sant'Ana GC, Chaves CL, et al (2019) Population structure and genetic relationships between Ethiopian and Brazilian Coffea arabica genotypes revealed by SSR markers. Genetica 147:205–216. https://doi.org/10.1007/s10709-019-00064-4

Danecek P et al. (2011). The variant call format and VCFtools. Bioinformatics, 27(15):2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Das S, Abecasis GR, Browning BL (2018) Genotype Imputation from Large Reference Panels. Annu Rev Genomics Hum Genet 19:73–96. https://doi.org/10.1146/annurev-genom-083117-021602

Davey JW, Hohenlohe PA, Etter PD, et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510. https://doi.org/10.1038/nrg3012

Davis AP, Chadburn H, Moat J, O'Sullivan R, Hargreaves S, Lughadha EN (2019) High extinction risk for wild coffee species and implications for coffee sector sustainability. Science advances, 5(1), eaav3473. https://doi.org/10.1126/sciadv.aav3473

Davis AP, Gole TW, Baena S, Moat J (2012). The impact of climate change on indigenous arabica coffee (Coffea arabica): predicting future trends and identifying priorities. PloS one, 7(11). https://doi.org/10.1371/journal.pone.0047981

Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. Bot J Linn Soc 167:357–377. https://doi.org/10.1111/j.1095-8339.2011.01177.x

de Kochko A. and ACGC (2018) Deciphering the Allotetraploid Genome of Coffea arabica L. In Plant and Animal Genome Conference XXVI (Januray 13- 17, 2018). San Diego, CA, USA.

Deschamps S, Llaca V, May GD (2012) Genotyping-by-sequencing in plants. Biology (Basel) 1:460–483. https://doi.org/10.3390/biology1030460

Dias Chaves J, Sarruge J (1984) Alteraçoes nas concentraçoes de macronutrientes nos frutos e folhas do cafeeito durante um ciclo productivo. Pesqui agropecu bras 19:427–432

Diniz I et al. (2015) Unveilling the involvement of oxidases in the resistance of Coffea sp. to Colletotrichum kahawae. In: Proceedings of 25th International Conference on Coffee Science (ASIC). Armenia, Colombia.

Disch S, Anastasiou E, Sharma VK, et al (2006) The E3 Ubiquitin Ligase BIG BROTHER Controls Arabidopsis Organ Size in a Dosage-Dependent Manner. Curr Biol 272–279. https://doi.org/10.1016/j.cub.2005.12.026

Dominski Z (2007) Nucleases of the metallo-β-lactamase family and their role in DNA and RNA metabolism. Crit Rev Biochem Mol Biol 42:67–93. https://doi.org/10.1080/10409230701279118

dos Santos TB, Meda AR, Sitta RB, et al (2015) Nutritional characterization of Arabica coffee accession from Ethiopia. Coffee Sci 10:10–19. https://doi.org/10.25186/cs.v10i1.716

dos Santos Scholz MB, Kitzberger CSG, Pagiatto NF, et al (2016) Chemical composition in wild ethiopian Arabica coffee accessions. Euphytica 209:429–438. https://doi.org/10.1007/s10681-016-1653-y

Dunwell JM, Purvis A, Khuri S (2004) Cupins: The most functionally diverse protein superfamily? Phytochemistry 65:7–17. https://doi.org/10.1016/j.phytochem.2003.08.016

Eklund A (2016). beeswarm: The Bee Swarm Plot, an Alternative to Stripchart. R package version 0.2.3. https://CRAN.R-project.org/package=beeswarm

Elshire RJ, Glaubitz JC, Sun Q, et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:1–10. https://doi.org/10.1371/journal.pone.0019379

Food and Agriculture Organization of the United Nations (FAO) (2015) FAO Coffee Pocketbook 2015, Rome.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014). Fast and efficient estimation of individual ancestry coefficients. Genetics, 196(4), 973-983. https://doi.org/10.1534/genetics.113.160572

Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. Methods Ecol Evol 6:925–929. https://doi.org/10.1111/2041-210X.12382

Fu YB (2014) Genetic diversity analysis of highly incomplete snp genotype data with imputations: An empirical assessment. G3 Genes, Genomes, Genet 4:891–900. https://doi.org/10.1534/g3.114.010942

Garcia ALA, de CARVALHO CHS, Garcia AWR (2009) Extração de nutrientes em cafeeiros da espécie Coffea arabica. In Embrapa Café-Artigo em anais de congresso (ALICE). In: Congresso Brasileiro de Pesquisa Cafeeiras, 34., 2008, Caxambú. Anais. Brasília, DF: Embrapa Café, 2009.

Gaspari-Pezzopane CD, Medina Filho HP, Bordignon R (2004). Variabilidade genética do rendimento intrínseco de grãos em germoplasma de Coffea. Bragantia, 63(1), 39-54. https://doi.org/10.1590/S0006-87052004000100005

Glaubitz JC, Casstevens TM, Lu F, et al (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9(2): e90346. https://doi.org/10.1371/journal.pone.0090346

Gole TW (2003) Vegetation of the Yayu Forest in SW Ethiopia: impacts of human use and implications for in situ conservation of wild Coffea arabica L. populations. Ph.D. Thesis. University of Bonn, Germany.

Gong X, Shen L, Peng YZ, et al (2017) DNA Topoisomerase I a Affects the Floral Transition. Plant physiol, 173:642–654. https://doi.org/10.1104/pp.16.01603

Gore M, Bradbury P, Hogers R, et al (2007) Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci 47:135–148. https://doi.org/10.2135/cropsci2007.02.0085tpg

Guerra-Guimarães L, Tenente R, Pinheiro C, et al (2015) Proteomic analysis of apoplastic fluid of Coffea arabica leaves highlights novel biomarkers for resistance against Hemileia vastatrix. Front Plant Sci 6:1–16. https://doi.org/10.3389/fpls.2015.00478

Guerreiro-Filho O et al. (2008) Origem e classificação botânica do cafeeiro. In: Cultivares de café: origem, características e recomendações, Embrapa, Brasília, pp 27-33.

Guillaumet JL, Hallé F (1978) Echantillonnage du matériel Coffea arabica récolté en Ethiopie. Bulletin IFCC, 14:13-18.

Guimarães PT, Reis THP (2010) Nutrição e adubação do cafeeiro. In: Café arábica do plantio à colheita. Epamig, Lavras, pp 343-414.

Harrel Jr F E et al. (2019) Hmisc: Harrel Miscellaneous. R package version 4.2-0. https://CRAN.R-project.org/package=Hmisc

Hayward A C, Tollenaere R, Dalton-Morgan J, Batley J (2015) Molecular marker application in plants. In: Plant Genotyping Methods and Protocols, Humana Press, New York, pp 13-27.

He J, Zhao X, Laroche A, et al (2014) Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:1–8. https://doi.org/10.3389/fpls.2014.00484

He S, Zhao Y, Mette MF, et al (2015) Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC Genomics 16:1–12. https://doi.org/10.1186/s12864-015-1366-y

Healey A, Furtado A, Cooper T, Henry RJ (2014) Protocol : a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant

Methods, 1–8. https://doi.org/10.1186/1746-4811-10-21

Hu X, Song F, Zheng Z (2006) Molecular cloning and expression analysis of rice OsTVLP1, encoding a protein with similarity to TGF-β receptor interacting proteins and vacuolar assembly Vam6p/Vps39p proteins. DNA Seq 17:152–158. https://doi.org/10.1080/10425170600700212

Hu Z, Olatoye MO, Marla S, Morris GP (2019) An integrated genotyping-by-sequencing polymorphism map for over 10,000 sorghum genotypes. The Plant Genome, 12(1). https://doi.org/10.3835/plantgenome2018.06.0044

Huang X, Han B (2014) Natural Variations and Genome-Wide Association Studies in Crop Plants. Annu Rev Plant Biol 65:531–551. https://doi.org/10.1146/annurev-arplant-050213-035715

International Coffee Organization (ICO) (2019a) Coffee market report October 2019. http://www.ico.org/documents/cy2019-20/cmr-1019-e.pdf. Accessed 27 November 2019

International Coffee Organization (ICO) (2019b) Coffee Development Report 2019. Growing for prosperity. United Nations, New York Geneva 1–84

International Coffee Organization (ICO) (2020) Trade Statistics. Coffee production by exporting countries. http://www.ico.org/trade_statistics.asp?section=Statistics. Accessed 01 February 2020

Ilyas M, Rasheed A, Mahmood T (2016) Functional characterization of germin and germin-like protein genes in various plant species using transgenic approaches. Biotechnol Lett 38:1405–1421. https://doi.org/10.1007/s10529-016-2129-9

Imai A, Nonaka K, Kuniga T, et al (2018) Genome-wide association mapping of fruit-quality traits using genotyping-by-sequencing approach in citrus landraces, modern cultivars, and breeding lines in Japan. Tree Genet Genomes 14(2):24. https://doi.org/10.1007/s11295-018-1238-0

Jones P, Binns D, Chang H, et al (2014) InterProScan 5 : genome-scale protein function classification. Bioinformatics, 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Katuuramu DN, Hart JP, Porch TG, et al (2018) Genome-wide association analysis of nutritional composition-related traits and iron bioavailability in cooked dry beans (Phaseolus vulgaris L.). Mol Breed 38(4):44. https://doi.org/10.1007/s11032-018-0798-x

Kaye Y, Golani Y, Singer Y, et al (2011) Inositol polyphosphate 5-phosphatase7 regulates the production of reactive oxygen species and salt tolerance in arabidopsis. Plant Physiol 157:229–241. https://doi.org/10.1104/pp.111.176883

Khan MA, Korban SS (2012) Association mapping in forest trees and fruit crops. J Exp Bot 63:4045–4060. https://doi.org/10.1093/jxb/ers105

Kim C, Guo H, Kong W, et al (2016) Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Sci 242:14–22. https://doi.org/10.1016/j.plantsci.2015.04.016

Kishor PBK, Hima Kumari P, Sunita MSL, Sreenivasulu N (2015) Role of proline in cell wall synthesis and plant development and its implications in plant ontogeny. Front Plant Sci 6:1–17. https://doi.org/10.3389/fpls.2015.00544

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: A review. Plant Methods 9:1–9. https://doi.org/10.1186/1746-4811-9-29

Kumar A, Batra R, Gahlaut V, et al (2018) Genome-wide identification and characterization of gene family for RWP-RK transcription factors in wheat (Triticum aestivum L .). PloS One 13(12) https://doi.org/10.1371/journal.pone.0208409

Kumar S, Banks TW, Cloutier S (2012) SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics 2012. https://doi.org/10.1155/2012/831460

Kupriyanova E V., Albert E V., Bliznina AI, et al (2017) Arabidopsis DNA topoisomerase I

alpha is required for adaptive response to light and flower development. Biol Open 6:832–843. https://doi.org/10.1242/bio.024422

Kusano M, Fukushima A, Redestig H, Saito K (2011) Metabolomic approaches toward understanding nitrogen metabolism in plants. J Exp Bot 62:1439–1453. https://doi.org/10.1093/jxb/erq417

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

Langmead B, Wilks C, Antonescu V, Charles R (2019) Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 35:421–432. https://doi.org/10.1093/bioinformatics/bty648

Lashermes P, Combes M, Robert J, et al (1999) Molecular characterisation and origin of the Coffea arabica L . genome. Mol Gen Genet MGG, 259–266. https://doi.org/10.1007/s004380050965

Laviola BG, Martinez HEP, De Souza RB, Víctor Hugo Alvarez V (2007) Dinâmica de cálcio e magnésio em folhas e frutos de coffea arabica. Rev Bras Cienc do Solo 31:319–329. https://doi.org/10.1590/s0100-06832007000200014

Li L, Xu X, Chen, C., & Shen, Z. (2016) Genome-wide characterization and expression analysis of the germin-like protein family in rice and Arabidopsis. International journal of molecular sciences, 17(10): 1622. https://doi.org/10.3390/ijms17101622

Lin H, Goodenough UW (2007) Gametogenesis in the Chlamydomonas reinhardtii minus Mating Type Is Controlled by Two Genes, MID and MTD1. Genetics, 1:913–925. https://doi.org/10.1534/genetics.106.066167

Lipka AE, Kandianis CB, Hudson ME, et al (2015) From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol 24:110–118. https://doi.org/10.1016/j.pbi.2015.02.010

Lu F, Lipka AE, Glaubitz J, et al (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLoS Genet 9(1). https://doi.org/10.1371/journal.pgen.1003215

Lunetti P, Cappello AR, Marsano RM, et al (2013) Mitochondrial glutamate carriers from Drosophila melanogaster: Biochemical, evolutionary and modeling studies. Biochim Biophys Acta - Bioenerg 1827:1245–1255. https://doi.org/10.1016/j.bbabio.2013.07.002

Malavolta E (1986) Nutrição mineral e adubação do cafeeiro – passado, presente e perspectiva. In: Nutrição e adubação do cafeeiro. Instituto de Potassa & Fosfato (EUA), Piracicaba, pp 138-178.

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511. https://doi.org/10.1038/nrg2796

Matiello JB et al. (2015) Cultura do café no Brasil: manual de recomendações, edição 2015. MAPA, Rio de Janeiro.

McClure KA, Gardner KM, Douglas GM, et al (2018) A Genome-Wide Association Study of Apple Quality and Scab Resistance. Plant Genome 11:1–14. https://doi.org/10.3835/plantgenome2017.08.0075

Medina Filho H, Bordingnon R, Carvalho CHS (2008) Desenvolvimento de novas cultivares de café arábica. In: Cultivares de café: origem, características e recomendações, Embrapa, Brasília, pp 79-101.

Mendes ANG et al. (2008) História das primeiras cultivares de café plantadas no Brasil. In: Cultivares de café: origem, características e recomendações, Embrapa, Brasília, pp 69-78.

Merot-L'anthoene V, Tournebize R, Darracq O, et al (2019) Development and evaluation of a genome-wide Coffee 8.5K SNP array and its application for high-density genetic mapping and for investigating the origin of Coffea arabica L. Plant Biotechnol J 17:1418–1430. https://doi.org/10.1111/pbi.13066

Meyer F et al. (1968) FAO Coffee mission to Ethiopia, 1964:1965. Food and agriculture organization of the United Nations.

Minamikawa MF, Nonaka K, Kaminuma E, et al (2017) Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. Sci Rep 7:1–2. https://doi.org/10.1038/s41598-017-05100-x

Mindrebo JT, Nartey CM, Seto Y, et al (2016) Unveiling the functional diversity of the alpha/beta hydrolase superfamily in the plant kingdom. Curr Opin Struct Biol 41:233–246. https://doi.org/10.1016/j.sbi.2016.08.005

Miyazawa M et al. (1999) Análise química de tecido vegetal. In: Manual de análises químicas de solos, plantas e fertilizantes, Embrapa, Brasília, pp 171-223.

Moat J, Gole TW, Davis AP (2019) Least concern to endangered: Applying climate change projections profoundly influences the extinction risk assessment for wild Arabica coffee. Glob Chang Biol 25:390–403. https://doi.org/10.1111/gcb.14341

Moncada MDP, Tovar E, Montoya JC, et al (2016) A genetic linkage map of coffee (Coffea arabica L.) and QTL for yield, plant height, and bean size. Tree Genet Genomes 12:1–17. https://doi.org/10.1007/s11295-015-0927-1

Müller R, Morant M, Jarmer H, Nilsson L, Nielsen T H (2007) Genome-wide analysis of the Arabidopsis leaf transcriptome reveals interaction of phosphate and sugar metabolism. Plant Physiology, 143(1):156-171. https://doi.org/10.1104/pp.106.090167

Murachelli AG, Ebert J, Basquin C, et al (2012) The structure of the ASAP core complex reveals the existence of a Pinin-containing PSAP complex. Nat Struct Mol Biol 19:378–386. https://doi.org/10.1038/nsmb.2242

Nadeem MA, Nawaz MA, Shahid MQ, et al (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32:261–285.

https://doi.org/10.1080/13102818.2017.1400401

Nazzicari N, Biscarini F, Cozzi P, et al (2016) Marker imputation efficiency for genotyping-by-sequencing data in rice (Oryza sativa) and alfalfa (Medicago sativa). Mol Breed 36:1–16. https://doi.org/10.1007/s11032-016-0490-y

Ogura T, Busch W (2015) From phenotypes to causal sequences: Using genome wide association studies to dissect the sequence basis for variation of plant development. Curr Opin Plant Biol 23:98–108. https://doi.org/10.1016/j.pbi.2014.11.008

Patel DA, Zander M, Dalton-Morgan J, Batley J (2015). Advances in plant genotyping: where the future will take us. In: Plant genotyping. Humana Press, New York, pp 1-11.

Pereira AA, Carvalho GR, Moura WM, Botelho CE, Rezende JC, Oliveira ACB, Silva FL (2010). Cultivares: origem e suas características. Café arábica do plantio à colheita. Epamig, Lavras, pp167-221.

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7(2). https://doi.org/10.1371/journal.pone.0032253

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102. https://doi.org/10.3835/plantgenome2012.05.0005

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180. https://doi.org/10.1016/j.pbi.2009.12.004

Rahimi Y, Bihamta MR, Taleei A, et al (2019) Genome-wide association study of agronomic traits in bread wheat reveals novel putative alleles for future breeding programs. BMC Plant Biol 19:1–19. https://doi.org/10.1186/s12870-019-2165-4

Rasheed A, Hao Y, Xia X, et al (2017) Crop Breeding Chips and Genotyping Platforms:

Progress, Challenges, and Perspectives. Mol Plant 10:1047–1064. https://doi.org/10.1016/j.molp.2017.06.008

Resende MDV (2016) Software Selegen-REML/BLUP: a useful tool for plant breeding. Crop Breed Appl Biotechnol 16:330–339. https://doi.org/10.1590/1984

Rutkoski JE, Poland J, Jannink JL, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3 Genes, Genomes, Genet 3:427–439. https://doi.org/10.1534/g3.112.005363

Salmona J, Dussert S, Descroix F, De Kochko A, Bertrand B, Joët T (2008). Deciphering transcriptional networks that govern Coffea arabica seed development using combined cDNA array and real-time RT-PCR approaches. Plant Mol Biol, 66:105-124. https://doi.org/10.1007/s11103-007-9256-6

Sanchez R, Zhou MM (2011) The PHD finger: A versatile epigenome reader. Trends Biochem Sci 36:364–372. https://doi.org/10.1016/j.tibs.2011.03.005

Sant'Ana GC, Pereira LFP, Pot D, et al (2018) Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. Sci Rep 8:1–12. https://doi.org/10.1038/s41598-017-18800-1

Scalabrin S, Toniutti L, Di Gaspero G et al (2020). A single polyploidization event at the origin of the tetraploid genome of Coffea arabica is responsible for the extremely low genetic variation in wild and cultivated germplasm. Sci Rep, 10(1):1-13. https://doi.org/10.1038/s41598-020-61216-7

Schneider C, Anderson JT, Tollervey D (2007) The Exosome Subunit Rrp44 Plays a Direct Role in RNA Substrate Recognition. Mol Cell 27:324–331. https://doi.org/10.1016/j.molcel.2007.06.006

Schwender H (2007) Statistical Analysis of Genotype and Gene Expression Data. Dissertation, University of Dortmund.

Schwender H, Fritsch A (2013) scrime: Analysis of High-Dimensional Categorical Data such as SNP Data. R package version 1.3.3. https://CRAN.R-project.org/package=scrime

Sengupta D, Naik D, Reddy AR (2015) Plant aldo-keto reductases (AKRs) as multi-tasking soldiers involved in diverse plant metabolic processes and stress defense: A structure-function update. J Plant Physiol 179:40–55. https://doi.org/10.1016/j.jplph.2015.03.004

Sera T (2001) Coffee Genetic Breeding at IAPAR. Crop Breed Appl Biotechnol 1:179–199. https://doi.org/10.13082/1984-7033.v01n02a08

Setotaw TA, Caixeta ET, Pereira AA, et al (2013) Coefficient of parentage in Coffea arabica L. cultivars grown in Brazil. Crop Sci 53:1237–1247. https://doi.org/10.2135/cropsci2012.09.0541

Silvarolla MB, Mazzafera P, Fazuoli LC (2004) A naturally decaffeinated arabica coffee. Nature 429(6994):826. https://doi.org/10.1038/429826a

Silvestrini M, Junqueira MG, Favarin AC, et al (2007) Genetic diversity and structure of Ethiopian, Yemen and Brazilian Coffea arabica L. accessions using microsatellites markers. Genet Resour Crop Evol 54:1367–1379. https://doi.org/10.1007/s10722-006-9122-4

Slatkin M (2008) Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–485. https://doi.org/10.1038/nrg2361

Stekhoven DJ, Bühlmann P (2012) Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics 28:112–118. https://doi.org/10.1093/bioinformatics/btr597

Swarts K, Li H, Alberto Romero Navarro J, et al (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7:1–12. https://doi.org/10.3835/plantgenome2014.05.0023

Tamba CL, Ni YL, Zhang YM (2017) Iterative sure independence screening EM-Bayesian

LASSO algorithm for multi-locus genome-wide association studies. PLoS Comput Biol 13:1–20. https://doi.org/10.1371/journal.pcbi.1005357

Tamba CL, Zhang Y (2018) A fast mrMLM algorithm for multi-locus genome-wide association studies. bioRxiv, 341784 https://doi.org/10.1101/341784

Tedeschi F, Rizzo P, Rutten T, et al (2017) RWP-RK domain-containing transcription factors control cell differentiation during female gametophyte development in Arabidopsis. New Phytol 213:1909–1924. https://doi.org/10.1111/nph.14293

Teo YY (2008) Common statistical issues in genome-wide association studies: A review on power, data quality control, genotype calling and population structure. Curr Opin Lipidol 19:133–143. https://doi.org/10.1097/MOL.0b013e3282f5dd77

Tesseema A, Alamerew S, Kufa T, Garedew W (2011) Variability and association of quality and biochemical attributes in some promising Coffea arabica germplam collections in Southwestern Ethiopia. Int J Plant Breed and Genetics, 4:302-316. https://doi.org/10.3923/ijpbg.2011.302.316

The Gene Ontology Consortium (2019) The Gene Ontology Resource : 20 years and still GOing strong. Nucleic Acids Res, 47:330–338. https://doi.org/10.1093/nar/gky1055

Torkamaneh D, Belzile F (2015) Scanning and filling: Ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. PLoS One 10:1–16. https://doi.org/10.1371/journal.pone.0131533

Torkamaneh D, Boyle B, Belzile F (2018) Efficient genome-wide genotyping strategies and data integration in crop plants. Theor Appl Genet 131:499–511. https://doi.org/10.1007/s00122-018-3056-z

Tran HTM, Vargas CAC, Slade Lee L, et al (2017) Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (Coffea arabica L.). Tree Genet Genomes 13(3):54. https://doi.org/10.1007/s11295-017-1138-8

Troyanskaya O, Cantor M, Sherlock G, et al (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Usman MG, Rafii MY, Martini MY, et al (2017) Molecular analysis of Hsp70 mechanisms in plants and their function in response to stress. Biotechnol Genet Eng Rev 33:26–39. https://doi.org/10.1080/02648725.2017.1340546

van der Vossen H, Bertrand B, Charrier A (2015) Next generation variety development for sustainable production of arabica coffee (Coffea arabica L.): a review. Euphytica 204:243–256. https://doi.org/10.1007/s10681-015-1398-z

Veerappan V, Wang J, Kang M, et al (2012) A novel HSI2 mutation in Arabidopsis affects the PHD-like domain and leads to derepression of seed-specific gene expression. Planta 236:1–17. https://doi.org/10.1007/s00425-012-1630-1

Vidal RO, Mondego JMC, Pot D, et al (2010) A High-Throughput Data Mining of Single Nucleotide Polymorphisms in Coffea Species Expressed Sequence Tags Suggests Differential Homeologous Gene Expression in the Allotetraploid Coffea arabica. Plant Physiol 154:1053–1066. https://doi.org/10.1104/pp.110.162438

Vieira LGE et al. (2006). Brazilian coffee genome project: an EST-based genomic resource. Braz. J. of Plant Physiol., 18(1): 95-108. https://doi.org/10.1590/S1677-04202006000100008

Vilhjálmsson BJ, Nordborg M (2013) The nature of confounding in genome-wide association studies. Nat Rev Genet 14:1–2. https://doi.org/10.1038/nrg3382

Voiniciuc C, Dean GH, Jonathan S. Griffiths, et al (2013) FLYING SAUCER1 Is a Transmembrane RING E3 Ubiquitin Ligase That Regulates the Degree of Pectin Methylesteri fi cation in Arabidopsis Seed Mucilage. The Plant Cell, 25:944–959. https://doi.org/10.1105/tpc.112.107888

Vos SM, Tretter EM, Schmidt BH, Berger JM (2011) All tangled up: how cells direct, manage

and exploit topoisomerase function. Nature reviews Molecular cell biology, 12(12), 827-841. https://doi.org/ 10.1038/nrm3228

Waki T, Hiki T, Watanabe R (2011) Report The Arabidopsis RWP-RK Protein RKD4 Triggers Gene Expression and Pattern Formation in Early Embryogenesis. Curr Biol, 4:1277–1281. https://doi.org/10.1016/j.cub.2011.07.001

Wang SB, Feng JY, Ren WL, et al (2016) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep 6:1–10. https://doi.org/10.1038/srep19444

Wen YJ, Zhang H, Ni YL, et al (2018) Methodological implementation of mixed linear models in multi-locus genome-wide association studies. Brief Bioinform 19:700–712. https://doi.org/10.1093/bib/bbw145

Wickham H (2016) Elegant Graphics for Data Analysis, Springer-Verlag, New York.

Wu W, Cheng Z, Liu M, et al (2014) C3HC4-type RING finger protein NbZFP1 is involved in growth and fruit development in Nicotiana benthamiana. PLoS One 9(6). https://doi.org/10.1371/journal.pone.0099352

Xu, S (2010). An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity, 105(5):483-494. https://doi.org/ 10.1038/hdy.2009.180

Yokoyama Y, Kobayashi S, Kidou S ichiro (2019) PHD type zinc finger protein PFP represses flowering by modulating FLC expression in Arabidopsis thaliana. Plant Growth Regul 88:49–59. https://doi.org/10.1007/s10725-019-00487-1

Yu J, Pressoir G, Briggs WH, et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702

Yu Q, Guyot R, Kochko A De, et al (2011) Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species

(Coffea). 305–317. https://doi.org/10.1111/j.1365-313X.2011.04590.x

Zhang C, Dong SS, Xu JY, et al (2019) PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 35:1786–1788. https://doi.org/10.1093/bioinformatics/bty875

Zhang J, Feng JY, Ni YL, et al (2017) PLARmEB: Integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. Heredity 118:517–524. https://doi.org/10.1038/hdy.2017.8

Zhang N, Yao LL, Li X dong (2018b) Regulation of class V myosin. Cell Mol Life Sci 75:261–273. https://doi.org/10.1007/s00018-017-2599-5

Zhang Y, Zheng Q, Sun C, et al (2016) Palmitoylation of the Cysteine Residue in the DHHC Motif of a Palmitoyl Transferase Mediates Ca2+ Homeostasis in Aspergillus. PLoS Genet 12:1–30. https://doi.org/10.1371/journal.pgen.1005977

Zhang Y, Li P, Ren W, Ni Y, Zhang Y (2018a) mrMLM: Multi-Locus Random-SNP-Effect Mixed Linear Model Tools for Genome-Wide Association Study. R package version 3.1. https://CRAN.R-project.org/package=mrMLM

Zhao Y, Jian Y, Liu Z, et al (2017) Network analysis reveals the recognition mechanism for dimer formation of bulb-type lectins. Sci Rep 7:1–9. https://doi.org/10.1038/s41598-017-03003-5

Zhong P, Li J, Luo L, et al (2019) TOP1 α regulates FLOWERING LOCUS C expression by coupling histone modification and transcription machinery. Development 146(4):dev167841. https://doi.org/10.1242/dev.167841

Ziegler G, Kear PJ, Wu D, et al (2017) Elemental accumulation in kernels of the maize nested association mapping panel reveals signals of gene by environment interactions. bioRxivd. https://doi.org/10.1101/164962

Ziegler G, Nelson R, Granada S, et al (2018) Genomewide association study of ionomic traits

on diverse soybean populations from germplasm collections. Plant Direct 2:e00033. https://doi.org/10.1002/pld3.33

7. MATERIAL SUPLEMENTAR

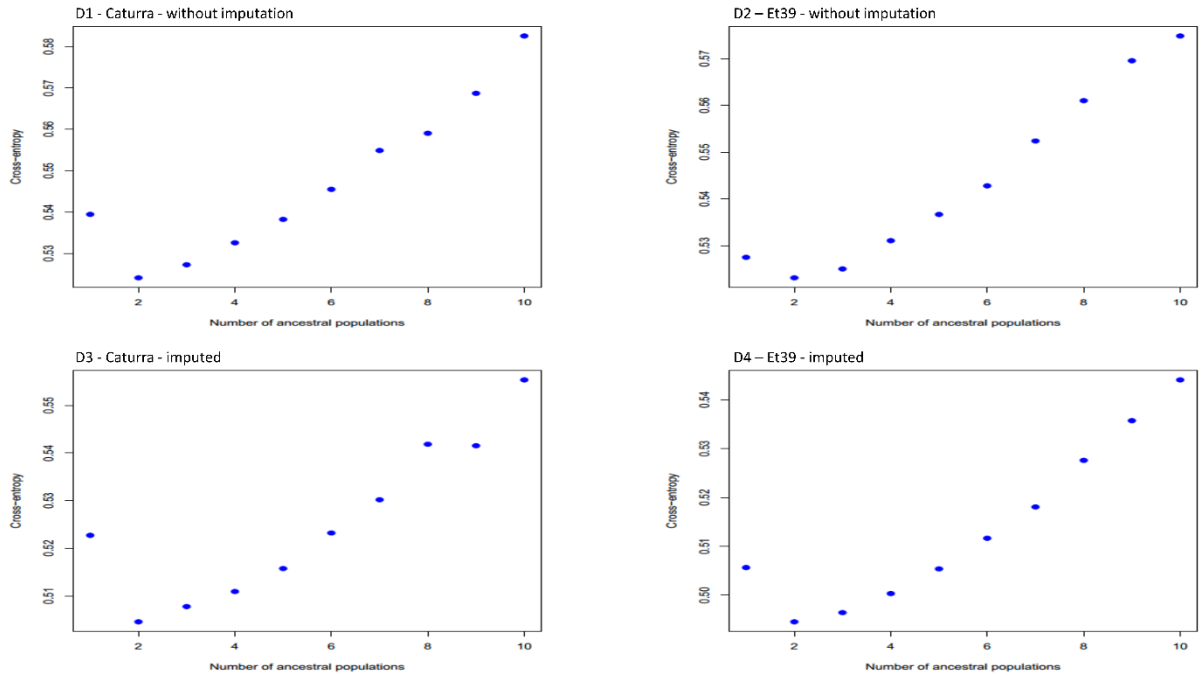

Fig S1 Values of the cross-entropy criterion for sNMF runs using markers dataset from 110 *C. arabica* genotypes. The datasets were aligned to the Caturra (left) and Et39 (right) reference genome before (up) and after (down) imputation.

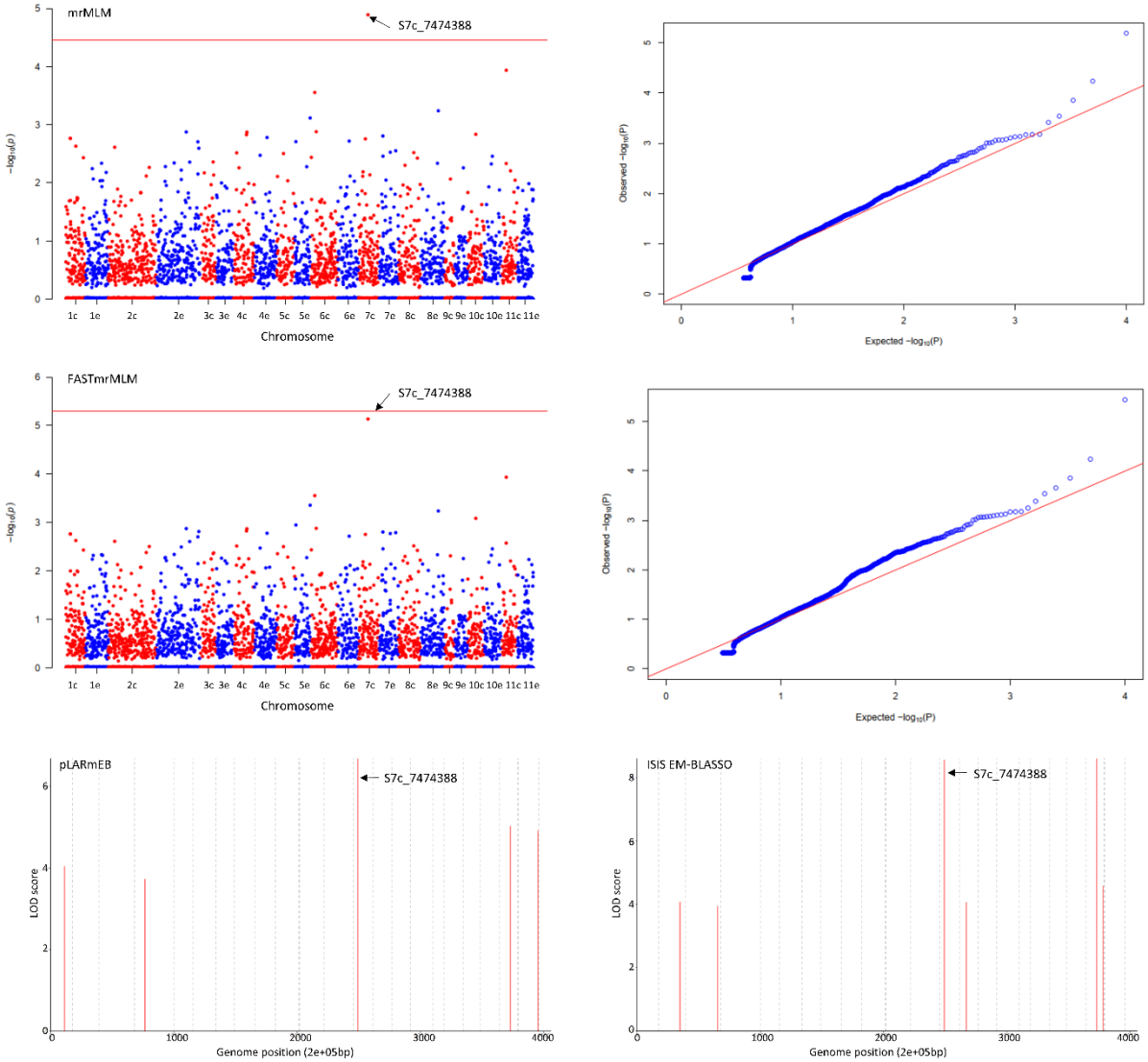

Fig S2 Manhattan plots, QQ-plots, and LOD plots score indicating the marker S7c_7474388 that identified in association with the coffee grain nitrogen content using the models mrMLM, FASTmrMLM, pLARmEB, and ISIS EM-BLASSO. The GWAS was performed with a panel of 105 *C. arabica* genotypes that were phenotyped in 2018. The marker was identified in the dataset D12 that was aligned to the Et39 reference genome.

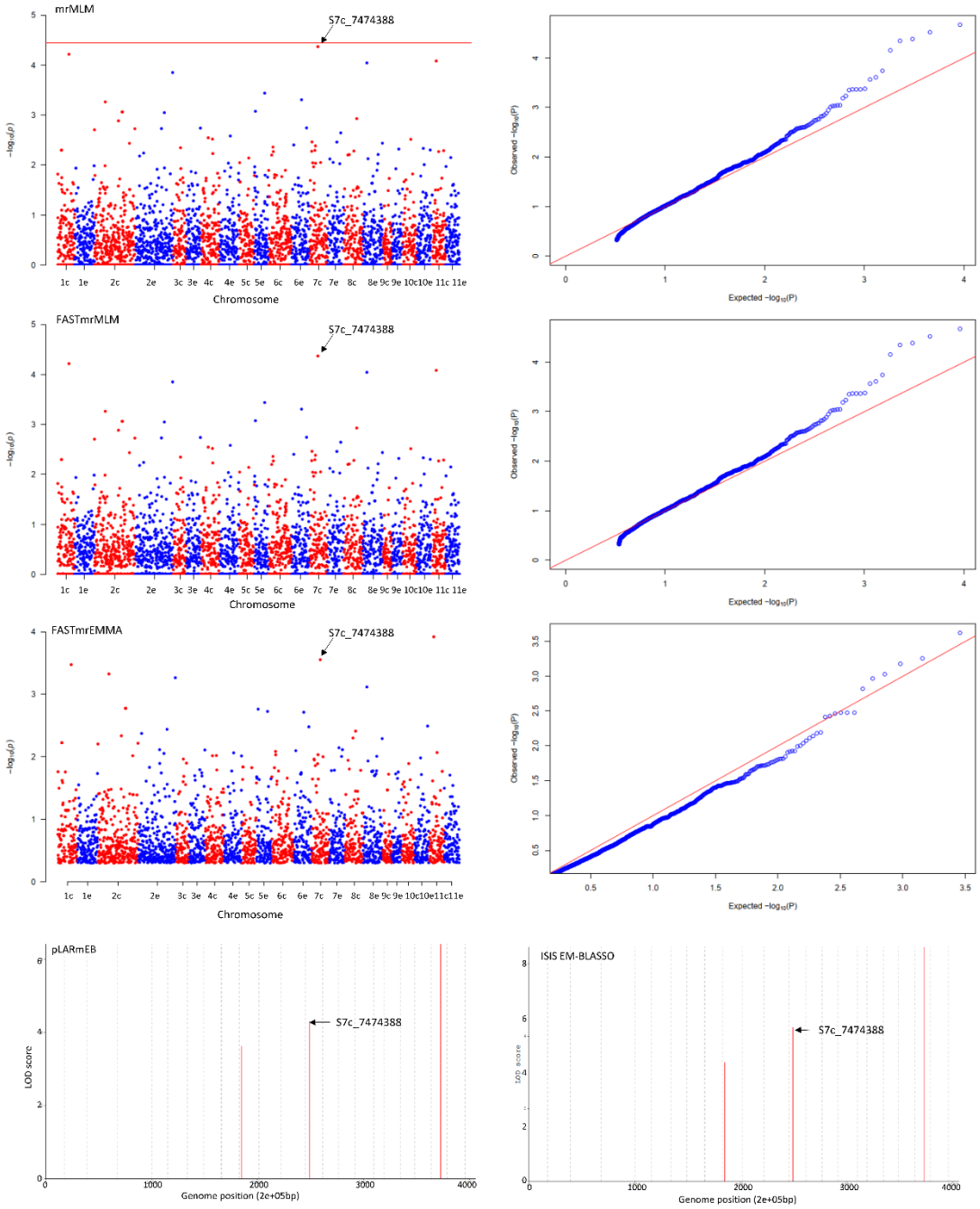

Fig S3 Manhattan plots, QQ-plots, and LOD score plots indicating the marker S7c_7474388 that was identified in association with the coffee grain nitrogen content using the models mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO. The GWAS was performed with a panel of 105 *C. arabica* genotypes that were phenotyped in 2018. The marker was identified in the dataset D13 that was aligned to the Et39 reference genome and the imputation of missing data was performed.

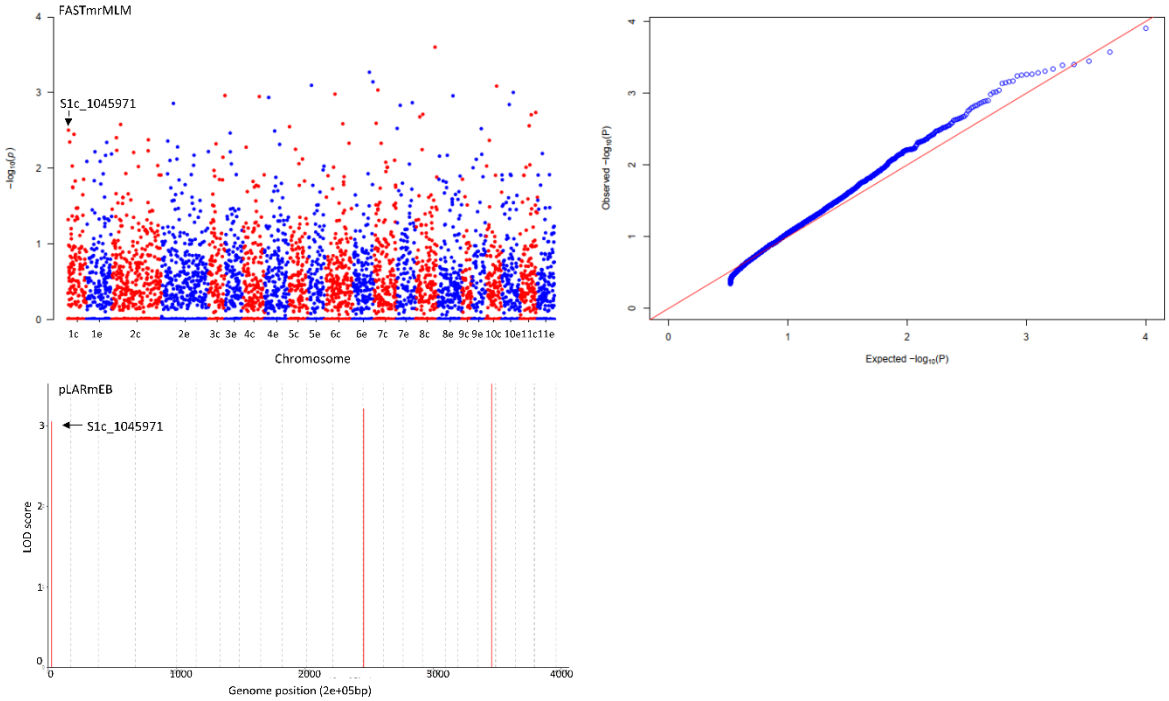

Fig S4 Manhattan plot, QQ-plot, and LOD score plot indicating the marker S1c_ 1045971 that was identified in association with the coffee grain phosphorus content using the models FASTmrMLM, and pLARmEB. The GWAS was performed with a panel of 70 *C. arabica* genotypes that were phenotyped in 2017. The marker was identified in the dataset D10 that was aligned to the Et39 reference genome.

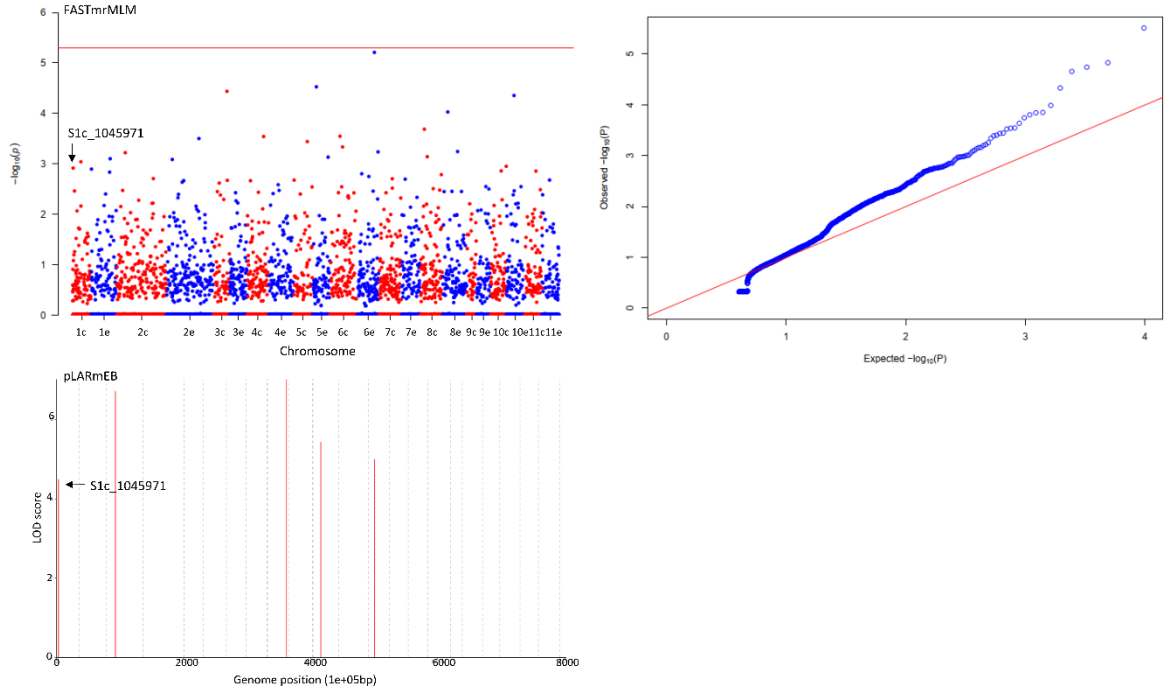

Fig S5 Manhattan plot, QQ-plot, and LOD score plot indicating the marker S1c_ 1045971 that was identified in association with the coffee grain phosphorus content using the models FASTmrMLM, and pLARmEB. The GWAS was performed with a panel of 70 *C. arabica* genotypes that were phenotyped in 2017. The marker was identified in the dataset D11 that was aligned to the Et39 reference genome and the imputation of missing data was performed.

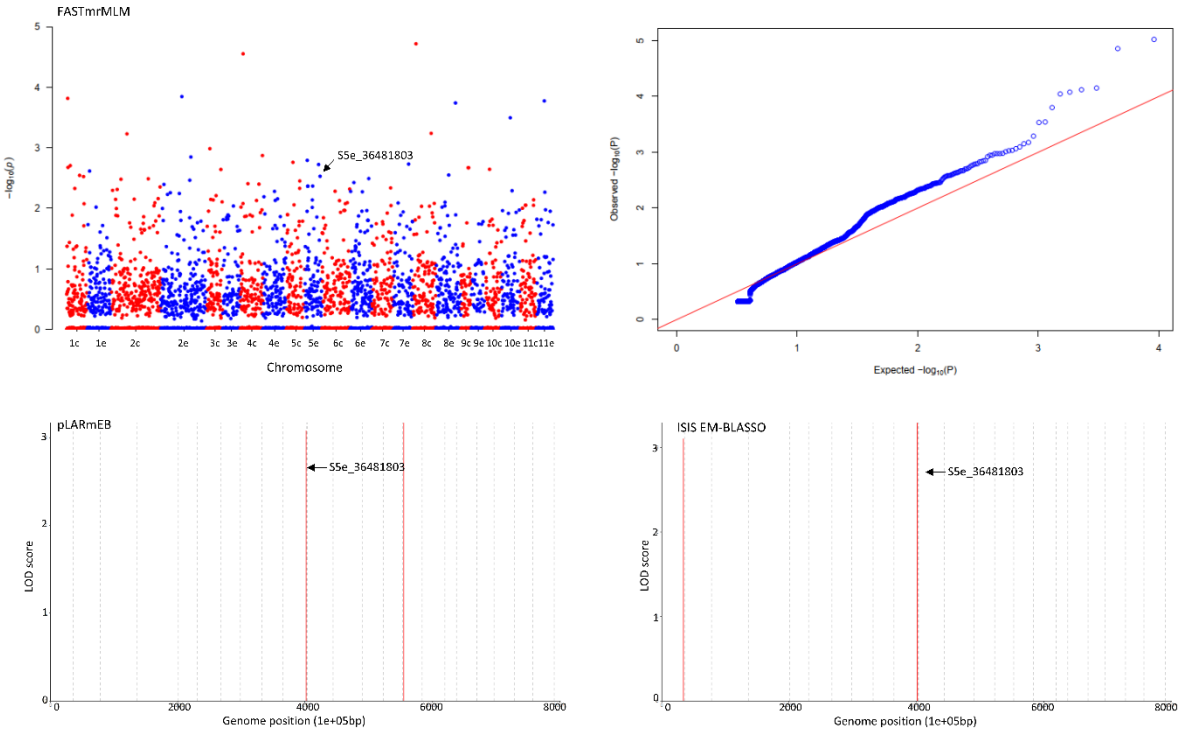

Fig S6 Manhattan plot, QQ-plot, and LOD score plots indicating the marker S5e_ 36481803 that was identified in association with the coffee grain potassium content using the models FASTmrMLM, pLARmEB, and ISIS EM-BLASSO. The GWAS was performed with a panel of 105 *C. arabica* genotypes that were phenotyped in 2018. The marker was identified in the dataset D13 that was aligned to the Et39 reference genome and the imputation of missing data was performed.

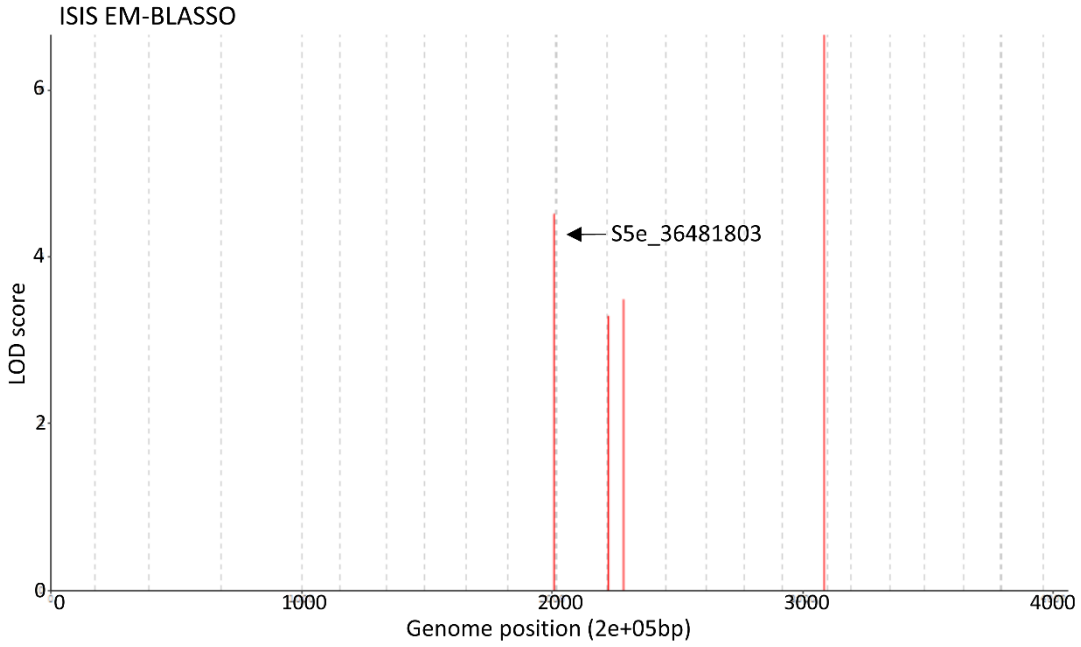

Fig S7 LOD score plot indicating the marker S5e_ 36481803 that was identified in association with the coffee grain potassium content using the model ISIS EM-BLASSO. The GWAS was performed with a panel of 65 *C. arabica* genotypes that were phenotyped in 2017 and 2018. The values of BLUP were used as phenotypic data for GWAS. The marker was identified in the dataset D14 that was aligned to the Et39 reference genome and the imputation of missing data was performed.

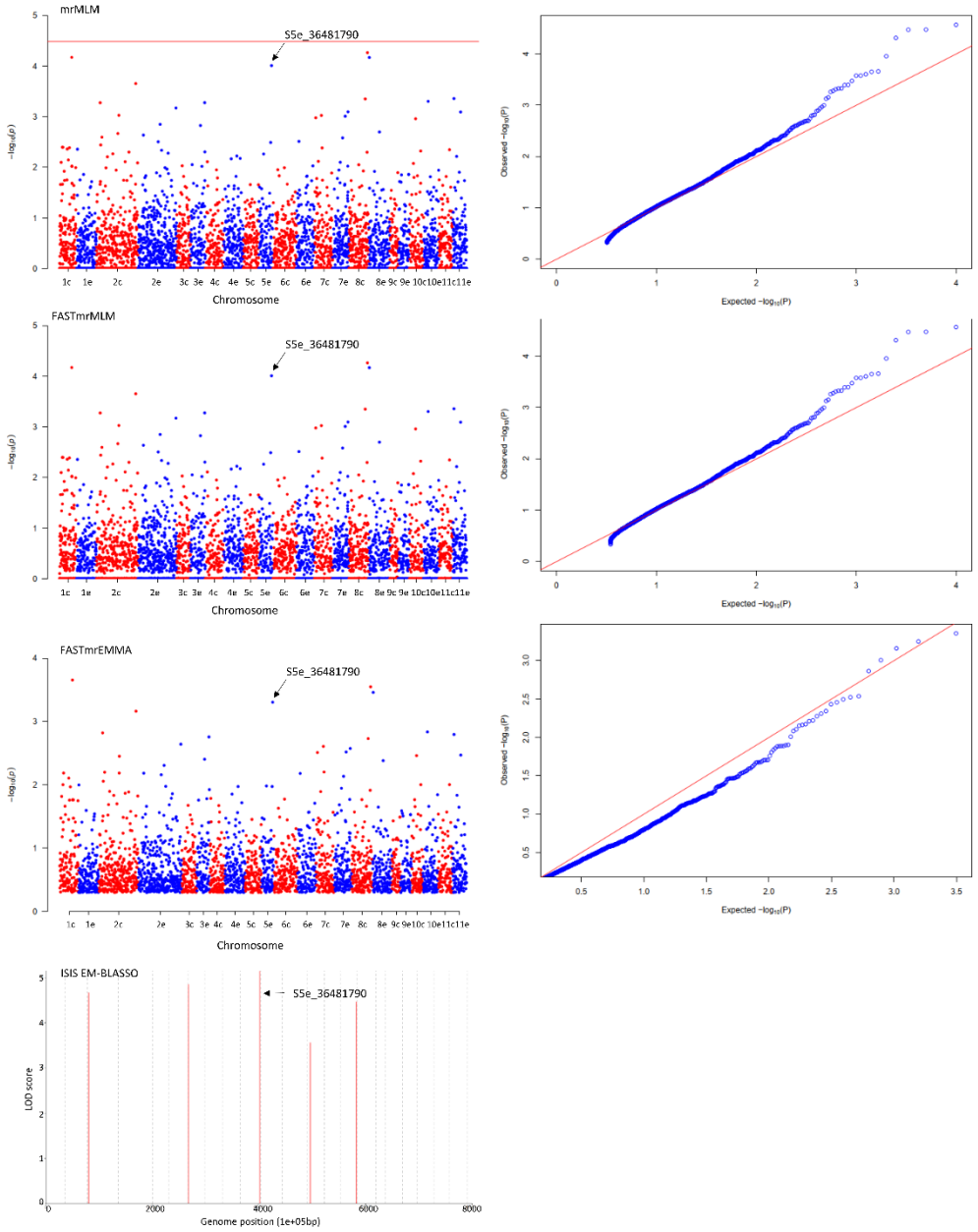

Fig S8 Manhattan plots, QQ-plots, and, LOD score plot indicating the marker S5e_ 36481803 that was identified in association with the coffee grain magnesium content using the models mrMLM, FASTmrMLM, FASTmrEMMA, and ISIS EM-BLASSO. The GWAS was performed with a panel of 105 *C. arabica* genotypes that were phenotyped in 2018. The marker was identified in the dataset D12 that was aligned to the Et39 reference genome.

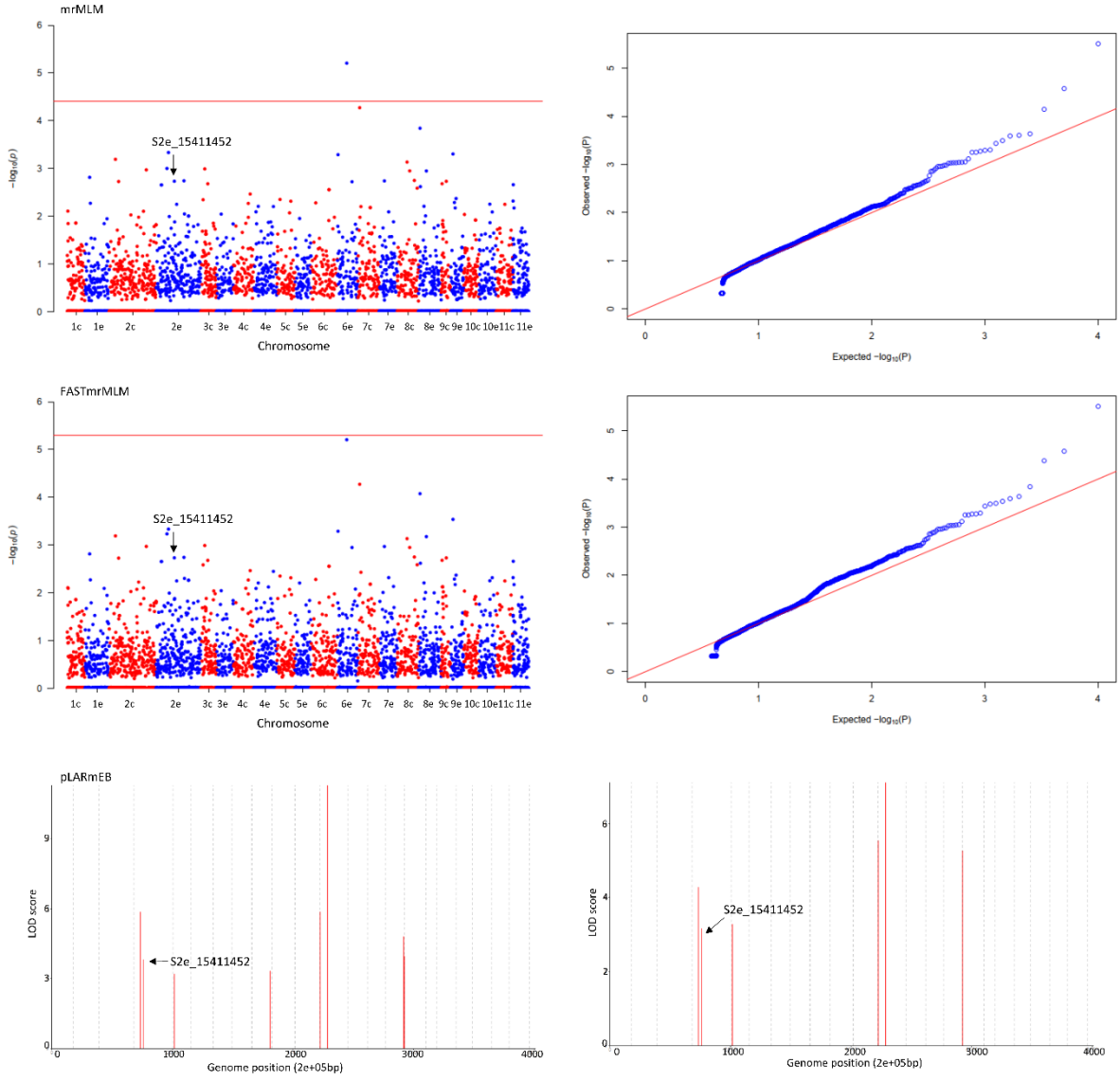

Fig S9 Manhattan plots, QQ-plots, and LOD score plots indicating the marker S2e_15411452 that was identified in association with the coffee grain potassium content using the models mrMLM, FASTmrMLM, pLARmEB, and ISIS EM-BLASSO. The GWAS was performed with a panel of 70 *C. arabica* genotypes that were phenotyped in 2017. The marker was identified in the dataset D10 that was aligned to the Et39 reference genome.

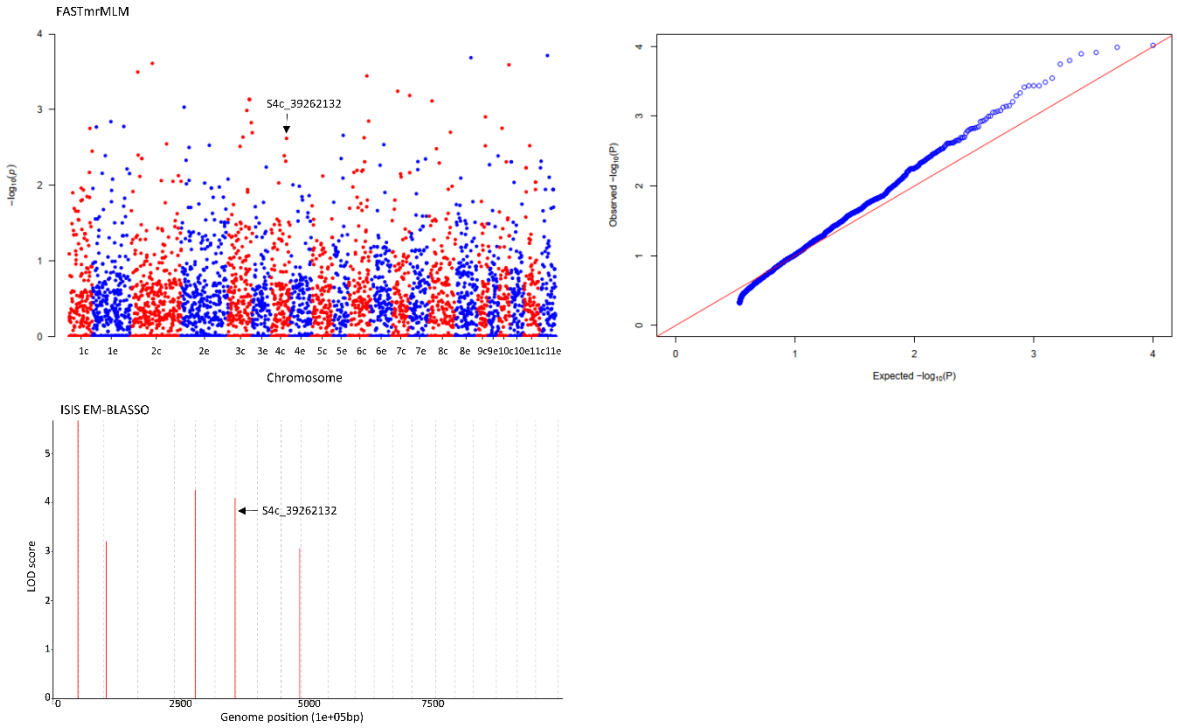

Fig S10 Manhattan plot, QQ-plot, and LOD score plot indicating the marker S4c_ 39262132 that was identified in association with the coffee grain magnesium content using the models FASTmrMLM, and ISIS EM-BLASSO. The GWAS was performed with a panel of 65 *C. arabica* genotypes that were phenotyped in 2017 and 2018. The values of BLUP were used as phenotypic data for GWAS. The marker was identified in the dataset D9 that was aligned to the Caturra reference genome and the imputation of missing data was performed.

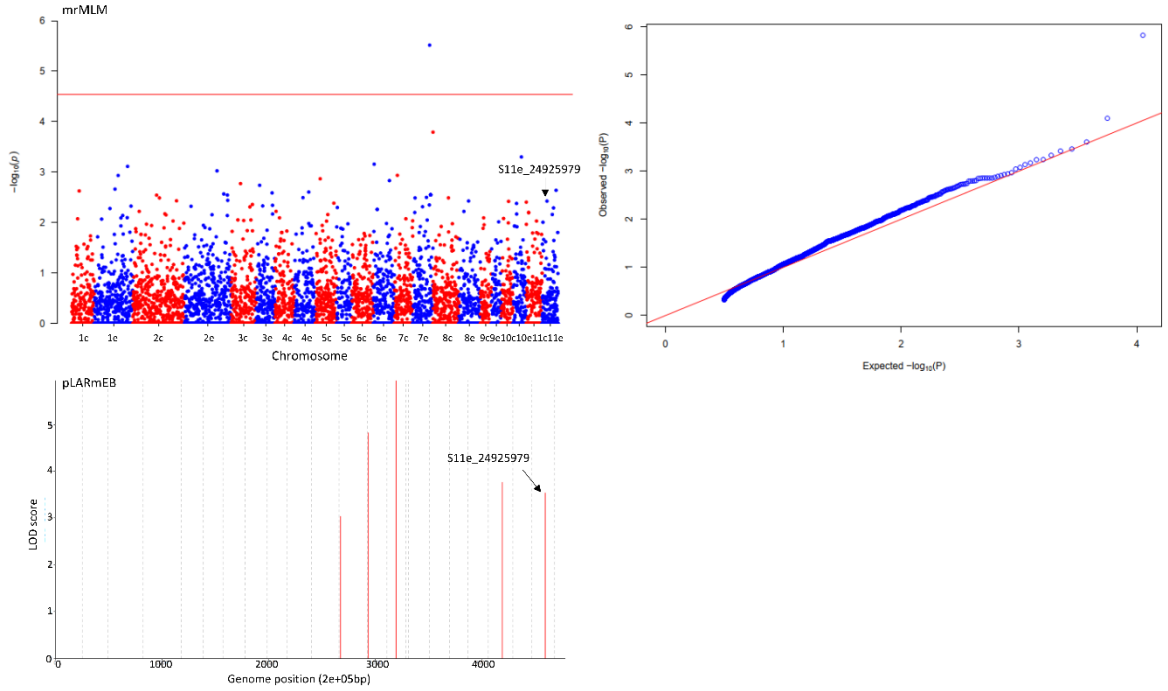

Fig S11 Manhattan plot, QQ-plot, and LOD score plot indicating the marker S11e_24925979 that was identified in association with the coffee grain phosphorus content using the models mrMLM and pLARmEB. The GWAS was performed with a panel of 70 *C. arabica* genotypes that were phenotyped in 2017. The marker was identified in the dataset D5 that was aligned to the Caturra reference genome.

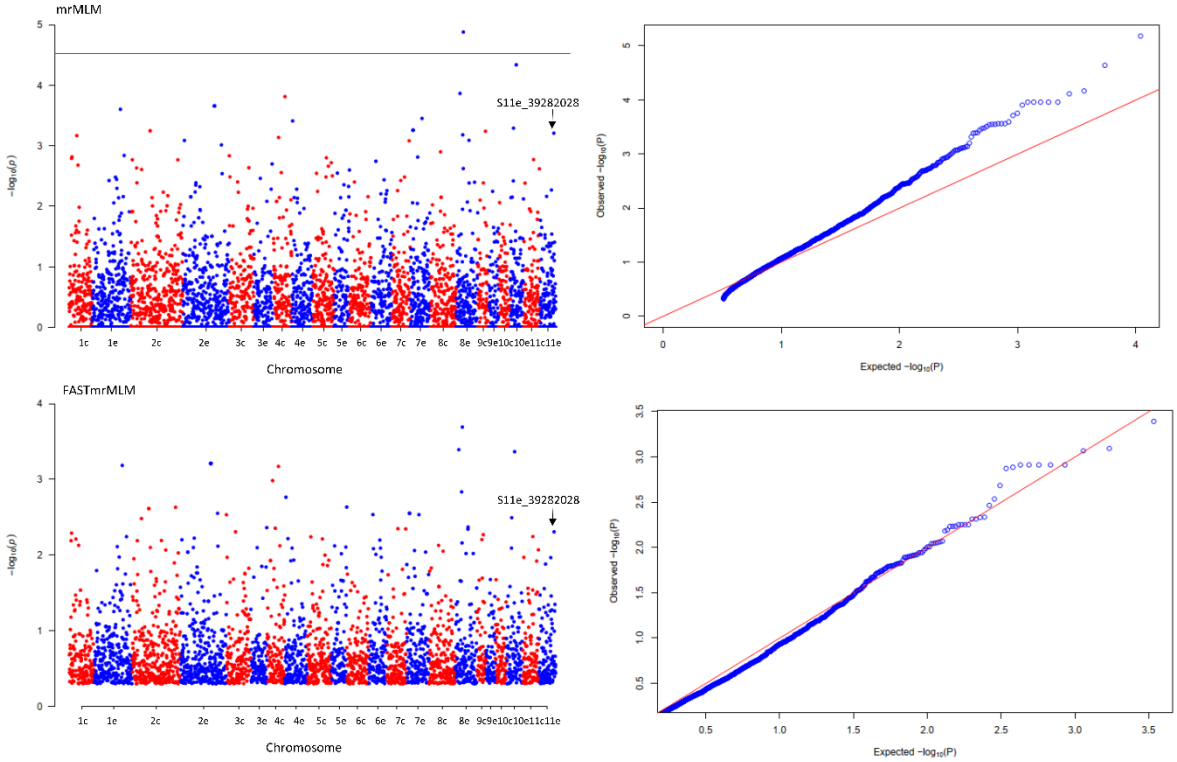

Fig S12 Manhattan plots and QQ-plots indicating the marker S11e_39282028 that was identified in association with the coffee grain calcium content using the models mrMLM, and FASTmrMLM. The GWAS was performed with a panel of 70 *C. arabica* genotypes that were phenotyped in 2017. The marker was identified in the dataset D6 that was aligned to the Caturra reference genome and the imputation of missing data was performed.

Table S1 Description of the *C. arabica* genotypes used in this work indicating the years in which the genotypes were sampled. The plants from FAO collection (Meyer et al. 1968) were first described by Sant'Ana et al. (2018).

Genotype	Sample Origin	Harvesting year
BA10_057	Elite landrace from IAPAR coffee breeding program	2017/2018
CatuaiV26	Commercial cultivar	2018
E007_087	FAO collection	2017/2018
E012_136	FAO collection	2017/2018
E016_298	FAO collection	2017
E018_494	FAO collection	2018
E021_011	FAO collection	2017/2018
E022_163	FAO collection	2018
E025_308	FAO collection	2017/2018
E037_676	FAO collection	2017/2018
E038_043	FAO collection	2017/2018
E041_079	FAO collection	2018
E044_122	FAO collection	2017/2018
E046_021	FAO collection	2017/2018
E047_267	FAO collection	2017/2018
E055_005	FAO collection	2018
E061_126	FAO collection	2018
E068_014	FAO collection	2018
E071_258	FAO collection	2018
E080_584	FAO collection	2018
E081_041	FAO collection	2017/2018
E087_194	FAO collection	2017/2018
E089_391	FAO collection	2018
E114_447	FAO collection	2018
E116_061	FAO collection	2017/2018
E118_213	FAO collection	2018
E123a_231	FAO collection	2018
E123b_121	FAO collection	2017/2018
E130_169	FAO collection	2017/2018
E131_018	FAO collection	2017/2018
E146_012	FAO collection	2017/2018
E148_254	FAO collection	2017
E159_180	FAO collection	2017/2018
E183_138	FAO collection	2018
E189_119	FAO collection	2017/2018
E190_013	FAO collection	2017/2018
E196_117	FAO collection	2018
E201_134	FAO collection	2017/2018
E209_031	FAO collection	2017/2018
E213_211	FAO collection	2017/2018
E220_127	FAO collection	2017/2018
E221_214	FAO collection	2017/2018
E233_015	FAO collection	2017/2018
E237_071	FAO collection	2017/2018
E238_022	FAO collection	2017/2018
E254_284	FAO collection	2017/2018
E267_090	FAO collection	2018
E268_067	FAO collection	2018
E270_044	FAO collection	2017/2018

Genotype	Sample Origin	Harvesting year
E272_143	FAO collection	2018
E279_618	FAO collection	2018
E283_096	FAO collection	2018
E287_029	FAO collection	2017/2018
E298_382	FAO collection	2018
E301_111	FAO collection	2017/2018
E308_049	FAO collection	2018
E315_081	FAO collection	2017/2018
E320_145	FAO collection	2017/2018
E324_093	FAO collection	2017/2018
E325_522	FAO collection	2018
E326_124	FAO collection	2017
E327_032	FAO collection	2018
E331_280	FAO collection	2017/2018
E332_023	FAO collection	2017/2018
E333_104	FAO collection	2017/2018
E338_218	FAO collection	2017
E344_008	FAO collection	2017/2018
E364_059	FAO collection	2018
E368_600	FAO collection	2017/2018
E370_196	FAO collection	2018
E383_142	FAO collection	2017/2018
E386_131	FAO collection	2017/2018
E401_643	FAO collection	2018
E408_001	FAO collection	2017/2018
E409_114	FAO collection	2018
E428_109	FAO collection	2017/2018
E439_094	FAO collection	2017/2018
E450_235	FAO collection	2017/2018
E454_107	FAO collection	2017/2018
E456_062	FAO collection	2018
E457_477	FAO collection	2018
E458_097	FAO collection	2017/2018
E467_045	FAO collection	2018
E478_408	FAO collection	2018
E481_238	FAO collection	2018
E486_189	FAO collection	2017/2018
E490_516	FAO collection	2018
E494_173	FAO collection	2017/2018
E505_140	FAO collection	2017/2018
E511_157	FAO collection	2017/2018
E514_129	FAO collection	2018
E516_069	FAO collection	2018
E534_036	FAO collection	2017/2018
E546_118	FAO collection	2017/2018
E552_323	FAO collection	2017/2018
E565_010	FAO collection	2017/2018
E571_072	FAO collection	2018
E621_139	FAO collection	2017/2018
IAPAR59	Commercial cultivar	2018
IPR100	Commercial cultivar	2017/2018
IPR101	Commercial cultivar	2017/2018
IPR102	Commercial cultivar	2017/2018

Genotype	Sample Origin	Harvesting year
IPR103	Commercial cultivar	2017/2018
IPR104	Commercial cultivar	2018
IPR105	Commercial cultivar	2017/2018
IPR107	Commercial cultivar	2018
IPR99	Commercial cultivar	2017/2018
Mundo Novo	Commercial cultivar	2017/2018
M7846_67	Elite landrace from IAPAR coffee breeding program	2017
SEL106	Elite landrace from IAPAR coffee breeding program	2017/2018

Table S2 Spearman correlation analysis among the nutrient content from coffee grains collected in 2017 (upper right) and 2018 (lower left) and for each nutrient among the years (central diagonal). The data of 2017 and 2018 were collected from 70 and 105 *C. arabica* genotypes, respectively. The coffee plants were cultivated at the experimental station of IAPAR in Londrina, Paraná, Brazil.

	N	**P**	**K**	**Ca**	**Mg**
N	**0.61 *****	0.19 ns	0.07 ns	0.12 ns	-0.11 ns
P	0.41 ***	**0.39 ****	0.47 ***	0.43 ***	0.12 ns
K	0.25 ns	0.39 ***	**0.37 ****	0.07 ns	0.11 ns
Ca	-0.07 ns	0.27 **	-0.03 ns	**0.27 ***	0.13 ns
Mg	0.00 ns	0.31 **	0.28**	0.45 ***	**0.60 *****

ns, not significant
*significance at 0.05 probability level ** significance at 0.01 probability level; *** significance of 0.001 probability level.

Table S3 Imputation accuracies of three imputation methods (Beagle, KNN, and RF) in two markers dataset identified in a population of 110 *C. arabica* genotypes. The SNPs were called from alignments to the Caturra (D1) and Et39 (D2) reference genomes.

Genotype	**Caturra (D1)**			**Et39 (D2)**		
	Beagle	**KNN**	**RF**	**Beagle**	**KNN**	**RF**
AA	0.98 ± 0.00	0.96 ± 0.00	0.91 ± 0.01	0.98 ± 0.00	0.96 ± 0.00	0.92 ± 0.01
AB	0.42 ± 0.01	0.46 ± 0.00	0.53 ± 0.00	0.20 ± 0.01	0.23 ± 0.00	0.31 ± 0.00
BB	0.38 ± 0.01	0.38 ± 0.01	0.17 ± 0.05	0.25 ± 0.02	0.29 ±0. 03	0.13 ± 0.04
Total	**0.82 ± 0.00**	**0.82 ± 0.00**	**0.80 ± 0.01**	**0.82 ± 0.00**	**0.80 ± 0.00**	**0.79 ± 0.01**
Time (min)	2.76 ± 0.07	8.36 ± 2.37	405.39 ± 35.24	2.51 ± 0.12	5.43 ± 1.38	318.13 ± 36.71

Table S4 Population structure analysis from SNPs datasets identified in 110 genotypes of *C. arabica*. The markers were identified from the alignment to Caturra (D1) and Et39 (D2) reference genomes. Subpopulations are indicated by the letter Q, and the letter M indicated the mixed group. After the imputation, the datasets D1 and D2 imputation were named as D3 and D4, respectively.

Genotype	Sample Origin	Subpopulations K=2		Subpopulation (K=3)			
		Caturra (D1)	Et39 (D2)	Caturra (D1)	Caturra imputed (D3)	Et39 (D2)	Et39 imputed (D4)
BA10_057	Elite landrace from IAPAR coffee breeding program	Q2	Q2	Q2	Q2	Q2	Q2
CatuaiV26	FAO collection	Q1	Q1	Q1	Q1	Q1	Q1
E007_087	FAO collection	Q1	Q1	Q1	Q1	Q1	Q1
E012_136	FAO collection	M	M	M	M	M	M
E016_298	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E018_494	FAO collection	Q1	Q1	M	M	M	M
E021_011	FAO collection	Q1	Q1	Q1	Q1	Q1	Q1
E022_163	FAO collection	Q1	Q1	M	M	Q1	Q1
E025_308	FAO collection	Q2	Q2	M	M	Q2	Q2
E037_676	FAO collection	Q1	Q1	M	Q1	M	M
E038_043	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E041_079	FAO collection	Q1	Q1	M	M	M	M
E044_122	FAO collection	Q1	Q1	M	M	M	M
E046_021	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E047_267	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E055_005	FAO collection	M	Q2	M	M	Q2	M
E061_126	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E068_014	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E071_258	FAO collection	M	Q2	Q3	Q3	M	M

Genotype	Sample Origin	Subpopulations K=2		Subpopulation (K=3)			
		Caturra (D1)	Et39 (D2)	Caturra (D1)	Caturra imputed (D3)	Et39 (D2)	Et39 imputed (D4)
E080_584	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E081_041	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E087_194	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E089_391	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E114_447	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E116_061	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E118_213	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E123a_231	FAO collection	M	Q2	M	M	Q2	M
E123b_121	FAO collection	Q1	Q1	Q3	Q3	M	M
E130_169	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E131_018	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E146_012	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E148_254	FAO collection	Q1	Q1	Q3	Q3	M	M
E159_180	FAO collection	M	Q2	M	M	Q2	M
E183_138	FAO collection	M	Q2	M	M	Q2	M
E189_119	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E190_013	FAO collection	Q2	Q2	M	Q2	Q2	Q2
E196_117	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E201_134	FAO collection	M	M	Q3	Q3	Q3	Q3
E209_031	FAO collection	Q2	Q2	M	M	Q2	Q2
E213_211	FAO collection	Q1	M	M	M	M	M
E220_127	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E221_214	FAO collection	M	Q2	M	M	Q2	M
E233_015	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E237_071	FAO collection	M	M	M	M	M	M
E238_022	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E254_284	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3

Genotype	Sample Origin	Subpopulations K=2		Subpopulation (K=3)			
		Caturra (D1)	Et39 (D2)	Caturra (D1)	Caturra imputed (D3)	Et39 (D2)	Et39 imputed (D4)
E267_090	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E268_067	FAO collection	M	Q2	Q3	Q3	M	M
E270_044	FAO collection	M	Q2	M	M	Q2	M
E272_143	FAO collection	Q1	Q1	Q3	Q3	M	M
E279_618	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E283_096	FAO collection	M	Q2	M	M	Q2	Q2
E287_029	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E298_382	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E301_111	FAO collection	Q1	Q1	M	M	M	M
E308_049	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E315_081	FAO collection	Q1	M	Q3	Q3	Q3	Q3
E320_145	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E324_093	FAO collection	M	Q2	M	Q2	M	Q2
E325_522	FAO collection	M	Q2	Q3	Q3	Q3	Q3
E326_124	FAO collection	Q1	Q2	M	Q3	M	M
E327_032	FAO collection	Q1	Q1	Q1	M	M	M
E331_280	FAO collection	Q1	M	Q3	Q3	Q3	Q3
E332_023	FAO collection	Q1	Q1	M	M	M	M
E333_104	FAO collection	Q1	Q1	Q1	Q1	Q1	Q1
E338_218	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E344_008	FAO collection	Q1	M	M	M	M	M
E364_059	FAO collection	Q2	Q2	M	Q2	Q2	Q2
E368_600	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E370_196	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E383_142	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E386_131	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E401_643	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3

Genotype	Sample Origin	Subpopulations K=2		Subpopulation (K=3)			
		Caturra (D1)	Et39 (D2)	Caturra (D1)	Caturra imputed (D3)	Et39 (D2)	Et39 imputed (D4)
E408_001	FAO collection	Q1	Q1	M	M	M	M
E409_114	FAO collection	Q1	M	M	M	M	M
E428_109	FAO collection	M	Q2	M	M	Q2	Q2
E439_094	FAO collection	Q2	Q2	M	M	M	M
E450_235	FAO collection	Q1	M	Q3	Q3	Q3	Q3
E454_107	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E456_062	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E457_477	FAO collection	Q2	Q2	M	Q2	Q2	Q2
E458_097	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E467_045	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E478_408	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E481_238	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E486_189	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E490_516	FAO collection	M	Q2	M	M	Q2	M
E494_173	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E505_140	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E511_157	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E514_129	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E516_069	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E534_036	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E546_118	FAO collection	M	Q2	M	Q2	M	Q2
E552_323	FAO collection	Q1	M	Q3	Q3	Q3	Q3
E565_010	FAO collection	Q2	Q2	Q2	Q2	Q2	Q2
E571_072	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
E621_139	FAO collection	Q1	Q1	Q3	Q3	Q3	Q3
IAPAR59	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR100	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1

Genotype	Sample Origin	Subpopulations K=2		Subpopulation (K=3)			
		Caturra (D1)	Et39 (D2)	Caturra (D1)	Caturra imputed (D3)	Et39 (D2)	Et39 imputed (D4)
IPR101	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR102	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR103	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR104	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR105	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR107	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
IPR99	Commercial cultivar	Q1	Q1	Q1	Q1	Q1	Q1
Mundo Novo	Commercial cultivar	Q1	Q1	M	Q1	Q1	Q1
M7846_67	Elite landrace from IAPAR coffee breeding program	Q1	Q1	Q1	Q1	Q1	Q1
SEL106	Elite landrace from IAPAR coffee breeding program	M	Q2	M	Q2	M	Q2

Table S5 Number of genotypes per group identified in the population structure analysis using GBS data of 110 *C. arabica* genotypes. The markers were identified from the alignment to Caturra (D1) and Et39 (D2) reference genomes. Subpopulations are indicated by the letter Q, and the letter M indicated the mixed group. After the imputation, the datasets D1 and D2 imputation were named as D3 and D4, respectively.

Groups	Caturra			Et39		
	K=2 (D1)	K=3 (D1)	K=3 imputed (D3)	K=2 (D2)	K=3 (D2)	K=3 imputed (D4)
Q1	66	15	16	58	16	16
Q2	26	20	26	42	34	30
Q3	-	42	43	-	37	37
M	18	33	25	10	23	27

Table S6 List of SNPs identified in association with the macronutrient content in coffee grains. The SNPs were identified by GBS data aligned to the Caturra reference genome. GWAS analyses were performed using phenotypic and genotypic data from 110 *C. arabica* genotypes. The SNPs listed were identified by more than two GWAS methods or in two environments (2017, 2018, 2017/2018).

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
1c	39640652	Ca	D5	No	2017	mrMLM	-0.19	6.41	41.52	0.12	TT
			D5	No	2017	FASTmrMLM	-0.16	6.33	39.60	0.12	NN
			D5	No	2017	pLARmEB	-0.13	5.37	34.07	0.12	NN
			D5	No	2017	ISIS EM-BLASSO	-0.12	4.14	26.16	0.12	NN
1c	42157466	Mg	D8	Yes	2018	pLARmEB	0.09	5.31	10.74	0.10	AA
			D8	Yes	2018	ISIS EM-BLASSO	0.10	4.12	10.82	0.10	AA
1c	44829225	Ca	D7	No	2018	mrMLM	-0.21	4.90	44.70	0.07	TT
			D7	No	2018	pLARmEB	-0.18	5.24	33.05	0.07	CT
			D8	Yes	2018	mrMLM	-0.25	6.41	40.79	0.05	TT
			D8	Yes	2018	pLARmEB	-0.19	3.57	11.77	0.06	TC
1c	48553001	Mg	D8	Yes	2018	pLARmEB	-0.08	3.28	5.99	0.08	GG
			D9	Yes	2017/2018	ISIS EM-BLASSO	-0.13	5.68	24.04	0.08	GG
1c	48553036	P	D7	No	2018	mrMLM	0.10	3.31	24.18	0.12	AA
			D7	No	2018	ISIS EM-BLASSO	0.08	3.71	17.99	0.12	NN
1c	50131361	P	D7	No	2018	FASTmrMLM	0.06	3.07	12.37	0.13	NN
			D7	No	2018	pLARmEB	0.05	3.20	8.88	0.13	NN
1c	50488201	Mg	D9	Yes	2017/2018	mrMLM	-0.10	5.26	27.78	0.20	A
			D9	Yes	2017/2018	FASTmrMLM	-0.06	4.54	13.09	0.20	AA
1e	33636748	Mg	D5	No	2017	mrMLM	-0.07	3.66	15.73	0.25	AA
			D5	No	2017	FASTmrMLM	-0.07	5.44	18.88	0.26	AA
			D5	No	2017	FASTmrEMMA	-0.15	3.85	13.59	0.26	AA
			D5	No	2017	pLARmEB	-0.05	3.59	9.72	0.26	AA

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D6	Yes	2017	mrMLM	-0.10	5.85	26.28	0.25	AA
			D6	Yes	2017	FASTmrMLM	-0.08	5.03	24.13	0.26	AA
			D6	Yes	2017	FASTmrEMMA	-0.16	5.87	14.93	0.26	AA
			D6	Yes	2017	pLARmEB	-0.07	3.09	18.31	0.26	AA
1e	34837440	K	D7	No	2018	mrMLM	-0.97	4.06	19.71	0.43	AA
			D7	No	2018	FASTmrMLM	-0.73	4.26	12.46	0.43	GA
			D8	Yes	2018	ISIS EM-BLASSO	-0.76	4.40	16.18	0.43	AG
1e	40472978	Ca	D5	No	2017	mrMLM	0.11	3.40	9.97	0.49	GG
			D5	No	2017	pLARmEB	0.04	3.82	1.80	0.49	NN
1e	47282019	K	D5	No	2017	mrMLM	0.82	5.10	16.15	0.41	AA
			D5	No	2017	FASTmrMLM	0.74	5.37	16.15	0.41	AC
			D5	No	2017	FASTmrEMMA	1.77	3.57	12.75	0.41	AC
			D5	No	2017	pLARmEB	0.74	5.61	17.09	0.41	AC
			D5	No	2017	ISIS EM-BLASSO	0.71	4.67	13.92	0.41	AC
			D6	Yes	2017	FASTmrMLM	0.66	3.08	19.12	0.39	AC
			D6	Yes	2017	FASTmrEMMA	2.05	3.67	19.29	0.39	AC
			D6	Yes	2017	pLARmEB	0.79	3.03	27.43	0.39	AC
			D6	Yes	2017	ISIS EM-BLASSO	0.72	3.83	19.38	0.39	AC
2c	3780046	K	D6	Yes	2017	FASTmrMLM	-0.98	4.65	23.02	0.06	GG
			D6	Yes	2017	ISIS EM-BLASSO	-1.17	6.70	28.21	0.06	GG
2c	34895792	N	D5	No	2017	mrMLM	-1.70	3.97	23.62	0.08	AA
			D5	No	2017	pLARmEB	-1.52	5.08	13.90	0.09	AA
2c	51564194	Ca	D9	Yes	2017/2018	mrMLM	-0.17	6.20	46.01	0.06	T
			D9	Yes	2017/2018	FASTmrMLM	0.00	3.82	0.00	0.05	TT
2c	54046869	Ca	D5	No	2017	ISIS EM-BLASSO	0.09	4.02	16.26	0.22	NN
			D6	Yes	2017	mrMLM	0.13	4.70	26.72	0.15	TT
			D6	Yes	2017	FASTmrMLM	0.10	4.78	23.74	0.14	TT
2c	54046885	P	D7	No	2018	mrMLM	0.09	3.03	22.50	0.22	TT

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D7	No	2018	ISIS EM-BLASSO	0.06	3.44	13.85	0.23	NN
			D8	Yes	2018	FASTmrMLM	0.00	4.13	0.00	0.15	TT
2c	64477322	N	D5	No	2017	mrMLM	-2.14	6.51	45.09	0.10	AA
			D5	No	2017	pLARmEB	-1.67	4.77	20.44	0.10	AA
			D5	No	2017	ISIS EM-BLASSO	-1.46	6.09	28.08	0.10	AA
			D6	Yes	2017	mrMLM	-2.52	7.25	52.77	0.10	AA
			D6	Yes	2017	pLARmEB	-1.28	4.60	20.53	0.09	AA
2e	456798	Ca	D9	Yes	2017/2018	FASTmrMLM	0.07	3.83	2.58	0.09	GG
			D9	Yes	2017/2018	pLARmEB	0.07	3.31	0.67	0.09	GG
			D9	Yes	2017/2018	ISIS EM-BLASSO	0.10	3.42	3.74	0.09	GG
2e	1525286	N	D5	No	2017	mrMLM	-0.72	3.17	7.51	0.45	CC
			D5	No	2017	pLARmEB	-0.66	5.69	4.66	0.46	TC
2e	1566430	Mg	D9	Yes	2017/2018	mrMLM	0.06	4.24	9.78	0.46	T
			D9	Yes	2017/2018	pLARmEB	0.06	4.79	16.14	0.44	TT
2e	1714405	N	D6	Yes	2017	pLARmEB	-1.17	5.28	12.74	0.07	TT
			D6	Yes	2017	ISIS EM-BLASSO	-1.05	4.22	10.87	0.07	TT
2e	1714463	Mg	D8	Yes	2018	mrMLM	-0.17	3.04	28.66	0.07	AA
			D8	Yes	2018	FASTmrMLM	-0.11	3.57	13.33	0.07	AA
			D8	Yes	2018	pLARmEB	-0.13	4.00	15.82	0.07	AA
			D8	Yes	2018	ISIS EM-BLASSO	-0.13	4.80	15.66	0.07	AA
2e	16364163	N	D6	Yes	2017	pLARmEB	1.14	4.80	13.14	0.08	TT
			D6	Yes	2017	ISIS EM-BLASSO	1.08	4.39	12.56	0.08	TT
2e	17227892	Mg	D9	Yes	2017/2018	mrMLM	0.08	4.12	6.38	0.36	C
			D9	Yes	2017/2018	FASTmrMLM	0.03	3.62	0.95	0.37	TC
2e	24413597	N	D6	Yes	2017	mrMLM	-1.87	3.02	10.30	0.50	GG
			D6	Yes	2017	FASTmrMLM	-1.98	8.28	16.21	0.49	GG
			D6	Yes	2017	ISIS EM-BLASSO	-1.35	4.62	8.60	0.49	GG
2e	71172143	Mg	D9	Yes	2017/2018	mrMLM	0.10	4.57	26.95	0.38	T

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D9	Yes	2017/2018	FASTmrMLM	0.06	5.86	13.93	0.37	TT
3c	11159390	K	D7	No	2018	pLARmEB	-0.43	3.13	5.95	0.22	NN
			D8	Yes	2018	pLARmEB	-0.50	3.05	5.17	0.17	AA
3c	40962154	Mg	D9	Yes	2017/2018	FASTmrMLM	0.04	4.41	2.61	0.45	CT
			D9	Yes	2017/2018	ISIS EM-BLASSO	0.11	4.26	13.50	0.45	CT
3e	2399876	N	D5	No	2017	FASTmrMLM	-0.71	3.31	5.68	0.49	GG
			D6	Yes	2017	FASTmrMLM	-0.70	4.65	5.25	0.49	GG
3e	2702885	P	D5	No	2017	ISIS EM-BLASSO	0.07	3.04	19.18	0.13	TT
			D6	Yes	2017	mrMLM	0.08	3.48	13.01	0.09	TT
			D6	Yes	2017	pLARmEB	0.02	3.33	0.43	0.09	TT
3e	2898294	K	D5	No	2017	mrMLM	0.82	7.12	13.29	0.49	AA
			D5	No	2017	FASTmrMLM	0.70	6.83	12.20	0.49	NN
			D5	No	2017	pLARmEB	0.40	3.55	4.24	0.49	NN
			D5	No	2017	ISIS EM-BLASSO	0.51	4.60	5.97	0.49	NN
3e	2970187	Mg	D6	Yes	2017	mrMLM	0.09	4.05	25.24	0.39	CC
			D6	Yes	2017	FASTmrMLM	0.07	4.35	19.17	0.38	CC
			D6	Yes	2017	pLARmEB	0.07	3.96	16.46	0.38	CC
3e	9156601	K	D5	No	2017	mrMLM	0.85	4.60	20.92	0.13	AA
			D5	No	2017	FASTmrMLM	0.75	5.42	20.07	0.14	NN
3e	11127543	Mg	D8	Yes	2018	FASTmrMLM	0.07	3.05	9.31	0.18	AA
			D8	Yes	2018	pLARmEB	0.07	3.01	9.83	0.18	AA
			D8	Yes	2018	ISIS EM-BLASSO	0.09	4.55	13.52	0.18	AA
3e	30866116	Ca	D7	No	2018	ISIS EM-BLASSO	-0.05	3.42	8.59	0.41	GC
			D6	Yes	2017	mrMLM	0.14	3.65	17.86	0.07	AA
			D6	Yes	2017	FASTmrMLM	0.09	3.59	11.50	0.07	AA
4c	381534	K	D8	Yes	2018	mrMLM	-1.18	5.81	28.30	0.08	AA
			D8	Yes	2018	FASTmrMLM	-0.70	3.11	15.86	0.08	AA
			D8	Yes	2018	pLARmEB	-0.81	3.35	8.01	0.08	AA

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
4c	3072247	P	D5	No	2017	mrMLM	0.06	3.38	10.06	0.18	TT
			D5	No	2017	FASTmrMLM	0.06	3.31	13.78	0.17	TT
			D5	No	2017	ISIS EM-BLASSO	0.00	5.83	0.09	0.17	TT
4c	32173403	Ca	D6	Yes	2017	FASTmrEMMA	0.23	3.22	11.07	0.13	CC
			D9	Yes	2017/2018	mrMLM	0.08	3.57	1.63	0.12	C
4c	39262132	Mg	D9	Yes	2017/2018	FASTmrMLM	-0.05	9.40	8.57	0.15	TT
			D9	Yes	2017/2018	ISIS EM-BLASSO	-0.06	4.10	8.59	0.15	TT
4e	63587	K	D7	No	2018	mrMLM	-0.79	5.34	22.51	0.36	TT
			D7	No	2018	FASTmrMLM	-0.70	6.69	20.10	0.36	TT
			D7	No	2018	FASTmrEMMA	-1.55	4.77	11.91	0.36	TT
			D7	No	2018	pLARmEB	-0.82	6.29	18.45	0.36	TT
			D7	No	2018	ISIS EM-BLASSO	-0.71	5.86	22.46	0.36	TT
			D8	Yes	2018	mrMLM	-0.74	4.30	18.44	0.36	TT
			D8	Yes	2018	FASTmrMLM	-0.67	3.97	24.37	0.36	TT
			D8	Yes	2018	FASTmrEMMA	-1.47	3.45	10.66	0.36	TT
			D8	Yes	2018	pLARmEB	-0.69	4.19	9.54	0.36	TT
4e	493778	Mg	D9	Yes	2017/2018	FASTmrMLM	0.00	3.34	0.01	0.15	CC
			D5	No	2017	FASTmrMLM	0.00	3.47	0.00	0.23	TT
5c	41982008	K	D7	No	2018	mrMLM	0.75	3.83	19.41	0.15	CC
			D7	No	2018	FASTmrMLM	0.66	5.78	16.73	0.15	NN
			D7	No	2018	ISIS EM-BLASSO	0.60	3.79	15.04	0.15	NN
5c	45214650	K	D5	No	2017	mrMLM	1.02	5.03	17.46	0.48	AA
			D5	No	2017	FASTmrMLM	0.84	5.43	14.60	0.46	AA
			D5	No	2017	ISIS EM-BLASSO	0.49	4.06	4.59	0.46	AA
5e	36075183	Mg	D7	No	2018	mrMLM	0.12	3.60	17.59	0.10	GG
			D7	No	2018	FASTmrMLM	0.10	4.45	11.91	0.10	GG
			D7	No	2018	ISIS EM-BLASSO	0.11	3.61	13.50	0.10	GG
			D9	Yes	2017/2018	FASTmrEMMA	0.00	3.03	0.00	0.10	GG

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D9	Yes	2017/2018	ISIS EM-BLASSO	0.07	3.07	7.57	0.10	GG
6c	12684595	N	D6	Yes	2017	FASTmrMLM	1.39	6.06	22.48	0.09	AA
			D6	Yes	2017	ISIS EM-BLASSO	1.03	3.94	14.10	0.09	AA
6c	18861648	N	D7	No	2018	FASTmrMLM	0.68	3.46	7.74	0.16	NN
			D7	No	2018	ISIS EM-BLASSO	0.74	3.39	8.87	0.16	NN
			D9	Yes	2017/2018	pLARmEB	1.33	4.85	13.89	0.06	AA
			D9	Yes	2017/2018	ISIS EM-BLASSO	1.30	3.81	19.48	0.06	AA
6c	21514253	P	D8	Yes	2018	pLARmEB	0.07	3.44	19.23	0.17	TT
		Mg	D8	Yes	2018	pLARmEB	0.09	4.31	15.91	0.17	TT
			D8	Yes	2018	ISIS EM-BLASSO	0.09	4.07	14.57	0.17	TT
6e	1297121	K	D6	Yes	2017	FASTmrMLM	-0.56	4.04	11.12	0.14	TT
			D6	Yes	2017	pLARmEB	-0.52	3.36	9.78	0.14	TT
			D9	Yes	2017/2018	mrMLM	-0.74	4.74	27.06	0.11	T
			D9	Yes	2017/2018	FASTmrMLM	-0.51	3.68	15.74	0.11	TT
			D9	Yes	2017/2018	pLARmEB	-0.36	3.10	1.49	0.11	TT
6e	2219258	P	D5	No	2017	mrMLM	0.12	4.11	13.19	0.47	AA
			D5	No	2017	FASTmrMLM	0.11	3.49	16.74	0.46	AA
			D5	No	2017	pLARmEB	0.11	3.04	14.47	0.46	AA
6e	36057694	K	D7	No	2018	mrMLM	0.67	4.86	3.46	0.20	CC
			D7	No	2018	FASTmrEMMA	0.95	3.06	5.78	0.21	CC
			D7	No	2018	pLARmEB	0.62	4.37	2.27	0.21	CC
			D7	No	2018	ISIS EM-BLASSO	0.57	4.43	3.05	0.21	CC
6e	44825168	K	D5	Yes	2017	pLARmEB	0.39	3.02	7.13	0.22	GA
			D5	Yes	2017	ISIS EM-BLASSO	0.63	5.60	16.52	0.22	GA
7c	2399011	P	D5	No	2017	mrMLM	0.07	3.42	10.68	0.13	AA
			D5	No	2017	pLARmEB	0.07	4.82	16.56	0.13	AA
7c	3129645	N	D6	Yes	2017	FASTmrMLM	1.09	4.34	9.26	0.06	TT
			D6	Yes	2017	pLARmEB	0.93	3.09	7.24	0.06	TT

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D6	Yes	2017	ISIS EM-BLASSO	1.40	5.88	17.39	0.06	TT
7c	6206055	Mg	D5	No	2017	mrMLM	-0.09	5.21	21.02	0.42	AA
			D5	No	2017	FASTmrMLM	-0.05	3.14	9.87	0.42	AA
			D5	No	2017	pLARmEB	-0.07	7.74	17.13	0.42	AA
			D6	Yes	2017	mrMLM	-0.07	4.02	12.77	0.39	AA
			D6	Yes	2017	FASTmrEMMA	-0.11	3.98	8.93	0.39	AA
			D6	Yes	2017	pLARmEB	-0.06	7.54	12.32	0.39	AA
			D6	Yes	2017	ISIS EM-BLASSO	-0.06	5.05	13.92	0.39	AA
			D9	Yes	2017/2018	mrMLM	-0.05	3.59	7.20	0.40	A
7c	8016140	N	D6	Yes	2017	FASTmrEMMA	-1.63	3.87	13.41	0.26	CA
			D6	Yes	2017	pLARmEB	-0.59	5.03	4.92	0.26	CA
			D6	Yes	2017	ISIS EM-BLASSO	-0.51	3.73	3.83	0.26	CA
		Mg	D8	Yes	2018	pLARmEB	-0.07	7.07	7.40	0.18	CC
			D8	Yes	2018	ISIS EM-BLASSO	-0.06	4.14	5.48	0.18	CC
7c	17671324	K	D7	No	2018	mrMLM	-0.59	5.84	9.89	0.17	CC
			D7	No	2018	FASTmrMLM	-0.45	5.62	6.34	0.17	CC
			D7	No	2018	pLARmEB	-0.50	5.02	5.40	0.17	CC
			D7	No	2018	ISIS EM-BLASSO	-0.49	6.03	8.39	0.17	CC
7e	45587	K	D9	Yes	2017/2018	pLARmEB	-0.49	4.24	2.69	0.09	GG
			D9	Yes	2017/2018	ISIS EM-BLASSO	-0.75	6.37	28.49	0.09	GG
7e	1384928	P	D6	Yes	2017	pLARmEB	-0.08	5.78	4.56	0.09	AA
			D6	Yes	2017	ISIS EM-BLASSO	-0.09	5.22	23.88	0.09	AA
			D9	Yes	2017/2018	mrMLM	-0.12	7.69	35.50	0.07	A
			D9	Yes	2017/2018	FASTmrMLM	-0.10	6.31	30.35	0.08	AA
			D9	Yes	2017/2018	FASTmrEMMA	0.00	3.13	0.00	0.08	AA
			D9	Yes	2017/2018	pLARmEB	-0.09	5.37	7.36	0.08	AA
7e	2651208	N	D7	No	2018	mrMLM	1.43	3.72	21.89	0.11	AA
			D7	No	2018	pLARmEB	1.19	4.35	7.90	0.10	AA

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D7	No	2018	ISIS EM-BLASSO	1.17	3.79	17.67	0.10	AA
			D8	Yes	2018	ISIS EM-BLASSO	0.99	3.04	12.79	0.09	AA
7e	6942055	K	D5	No	2017	pLARmEB	0.76	6.86	26.07	0.19	NN
			D5	No	2017	ISIS EM-BLASSO	0.81	7.71	26.07	0.19	NN
7e	17047040	P	D5	No	2017	mrMLM	-0.11	5.58	18.68	0.07	TT
			D5	No	2017	FASTmrMLM	-0.10	3.72	20.74	0.07	TT
			D5	No	2017	pLARmEB	-0.10	5.94	21.03	0.07	TT
			D5	No	2017	ISIS EM-BLASSO	-0.08	4.83	18.28	0.07	TT
			D6	Yes	2017	mrMLM	-0.10	3.41	18.63	0.06	TT
8c	33347309	N	D8	Yes	2018	mrMLM	-2.77	6.15	56.19	0.43	AA
			D9	Yes	2017/2018	FASTmrMLM	-1.90	3.80	49.64	0.42	TA
			D9	Yes	2017/2018	pLARmEB	-1.41	3.99	21.24	0.42	TA
8c	36771722	K	D9	Yes	2017/2018	mrMLM	-0.70	5.18	36.73	0.16	T
			D9	Yes	2017/2018	FASTmrEMMA	0.00	3.06	0.00	0.17	TT
			D9	Yes	2017/2018	pLARmEB	-0.46	4.99	3.55	0.17	TT
8c	38942595	N	D5	No	2017	FASTmrMLM	-0.77	4.04	11.25	0.20	GG
			D9	Yes	2017/2018	ISIS EM-BLASSO	-0.64	3.98	9.84	0.20	GG
		Ca	D9	Yes	2017/2018	pLARmEB	0.04	3.39	2.98	0.20	GG
8e	34004590	N	D7	No	2018	FASTmrMLM	-0.90	3.09	10.43	0.41	NN
			D7	No	2018	pLARmEB	-0.98	4.14	5.20	0.41	NN
9c	9131195	P	D7	No	2018	mrMLM	-0.11	4.08	19.31	0.49	AA
			D7	No	2018	FASTmrMLM	-0.04	3.25	4.35	0.49	NN
			D7	No	2018	pLARmEB	-0.06	4.28	9.73	0.49	NN
			D7	No	2018	ISIS EM-BLASSO	-0.09	5.04	15.36	0.49	NN
			D8	Yes	2018	mrMLM	0.08	3.01	17.44	0.48	GG
			D8	Yes	2018	pLARmEB	0.06	3.12	7.94	0.48	GG
			D8	Yes	2018	ISIS EM-BLASSO	0.06	3.49	10.19	0.48	GG
9e	2042013	N	D5	No	2017	FASTmrEMMA	-2.73	5.45	14.11	0.14	NN

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D5	No	2017	ISIS EM-BLASSO	-1.01	3.02	2.32	0.14	NN
10e	3347245	P	D6	Yes	2017	pLARmEB	-0.11	4.84	7.38	0.07	AA
			D6	Yes	2017	ISIS EM-BLASSO	-0.10	3.96	27.71	0.07	AA
			D9	Yes	2017/2018	mrMLM	-0.13	6.52	40.85	0.08	A
			D9	Yes	2017/2018	FASTmrMLM	-0.11	6.03	36.97	0.08	AA
			D9	Yes	2017/2018	pLARmEB	-0.10	4.54	8.68	0.08	AA
10e	8674437	N	D9	Yes	2017/2018	mrMLM	-0.76	3.82	13.61	0.26	G
			D9	Yes	2017/2018	FASTmrEMMA	-1.67	5.26	18.80	0.25	GG
			D9	Yes	2017/2018	pLARmEB	-0.65	4.07	7.13	0.25	GG
10e	9082830	Ca	D5	No	2017	mrMLM	-0.09	3.31	8.05	0.12	GG
			D5	No	2017	FASTmrMLM	-0.07	3.61	5.99	0.12	NN
			D5	No	2017	pLARmEB	-0.10	5.79	16.98	0.12	NN
			D5	No	2017	ISIS EM-BLASSO	-0.11	6.24	17.57	0.12	NN
			D6	Yes	2017	FASTmrMLM	-0.12	4.67	12.11	0.07	GG
			D6	Yes	2017	pLARmEB	-0.09	3.19	2.45	0.07	GG
			D6	Yes	2017	ISIS EM-BLASSO	-0.09	3.11	5.88	0.07	GG
			D9	Yes	2017/2018	FASTmrMLM	0.00	3.62	0.00	0.07	GG
10e	25772478	P	D5	No	2017	mrMLM	-0.08	4.53	14.80	0.11	TT
			D5	No	2017	pLARmEB	-0.05	3.77	7.62	0.12	NN
11c	23616508	N	D7	No	2018	mrMLM	1.18	7.87	19.47	0.24	GG
			D7	No	2018	FASTmrMLM	1.14	3.69	22.15	0.24	GG
			D7	No	2018	pLARmEB	1.39	14.26	14.21	0.24	GG
			D7	No	2018	ISIS EM-BLASSO	1.12	8.86	21.18	0.24	GG
11e	7373080	K	D6	Yes	2017	mrMLM	-0.83	3.47	30.64	0.22	GG
			D6	Yes	2017	ISIS EM-BLASSO	-0.63	4.36	19.49	0.21	GG
11e	24925979	P	D5	No	2017	mrMLM	0.07	4.39	12.04	0.21	AA
			D5	No	2017	pLARmEB	0.05	3.54	9.72	0.21	NN
11e	39282028	Ca	D6	Yes	2017	mrMLM	0.11	4.11	4.25	0.10	GG

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D6	Yes	2017	FASTmrMLM	0.00	3.04	0.00	0.12	GG

Table S7 List of SNPs identified in association with the macronutrient content in coffee grains. The SNPs were identified by GBS data aligned to the Et39 reference genome. GWAS analyses were performed using phenotypic and genotypic data from 110 *C. arabica* genotypes. The SNPs listed were identified by more than two GWAS methods or in two environments (2017, 2018, 2017/2018).

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
1c	1045971	P	D10	No	2017	FASTmrMLM	0.01	5.00	0.07	0.22	NN
			D10	No	2017	pLARmEB	0.04	3.05	1.40	0.22	NN
			D11	Yes	2017	FASTmrMLM	-0.05	3.68	9.89	0.14	CC
			D11	Yes	2017	pLARmEB	-0.05	4.49	4.25	0.14	CC
1c	22254476	N	D12	No	2018	mrMLM	-1.21	3.80	14.84	0.14	GG
			D12	No	2018	FASTmrMLM	-0.98	3.87	12.83	0.13	GG
			D12	No	2018	pLARmEB	-0.96	4.04	9.36	0.13	GG
			D13	Yes	2018	mrMLM	-1.61	5.51	26.35	0.13	GG
1c	25976280	Mg	D13	Yes	2018	pLARmEB	0.08	3.37	8.80	0.11	TT
			D13	Yes	2018	ISIS EM-BLASSO	0.08	4.15	11.80	0.11	TT
1c	29501607	P	D12	No	2018	FASTmrMLM	0.00	3.05	0.00	0.10	AA
			D12	No	2018	ISIS EM-BLASSO	0.00	3.67	0.00	0.10	AA
1c	30481936	Ca	D10	No	2017	FASTmrMLM	0.07	3.69	11.61	0.09	TT
			D10	No	2017	ISIS EM-BLASSO	0.10	5.53	15.17	0.09	TT
1c	31229877	Ca	D12	No	2018	FASTmrMLM	0.05	3.31	4.96	0.26	CT
			D12	No	2018	FASTmrEMMA	0.08	3.09	3.54	0.26	CT
			D12	No	2018	pLARmEB	0.07	6.67	5.62	0.26	CT
			D12	No	2018	ISIS EM-BLASSO	0.05	6.91	4.56	0.26	CT
		Mg	D13	Yes	2018	mrMLM	0.08	3.93	12.60	0.27	TT
			D13	Yes	2018	pLARmEB	0.07	3.01	6.94	0.27	TC
			D13	Yes	2018	ISIS EM-BLASSO	0.07	5.34	9.27	0.27	TC
			D14	Yes	2017/2018	mrMLM	0.09	6.66	23.63	0.28	T

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D14	Yes	2017/2018	FASTmrMLM	0.08	5.09	21.22	0.28	TC
			D14	Yes	2017/2018	FASTmrEMMA	0.15	3.03	19.43	0.28	TC
			D14	Yes	2017/2018	pLARmEB	0.08	7.55	8.72	0.28	TC
1c	35028505	N	D10	No	2017	mrMLM	1.50	6.91	19.35	0.10	AA
			D10	No	2017	FASTmrMLM	1.32	7.82	18.15	0.11	NN
			D10	No	2017	pLARmEB	1.32	7.78	12.01	0.11	NN
			D10	No	2017	ISIS EM-BLASSO	1.29	8.85	20.12	0.11	NN
			D11	Yes	2017	pLARmEB	1.53	4.68	19.86	0.06	AA
			D11	Yes	2017	ISIS EM-BLASSO	1.53	3.50	23.79	0.06	AA
1e	29539506	Mg	D10	No	2017	mrMLM	-0.11	5.85	37.78	0.25	AA
			D10	No	2017	pLARmEB	-0.07	4.05	12.73	0.26	AA
			D10	No	2017	ISIS EM-BLASSO	-0.07	3.12	11.08	0.26	AA
			D11	Yes	2017	mrMLM	-0.09	4.49	17.27	0.25	AA
			D11	Yes	2017	FASTmrEMMA	-0.20	5.39	23.34	0.26	AA
			D11	Yes	2017	pLARmEB	-0.08	4.53	5.88	0.26	AA
			D11	Yes	2017	ISIS EM-BLASSO	-0.08	4.53	15.45	0.26	AA
2c	1009617	K	D12	No	2018	FASTmrMLM	0.55	4.81	12.06	0.19	TT
			D12	No	2018	pLARmEB	0.56	5.30	10.75	0.19	TT
			D12	No	2018	ISIS EM-BLASSO	0.59	6.85	11.85	0.19	TT
2c	2489490	Ca	D12	No	2018	mrMLM	-0.15	3.34	28.64	0.12	TT
			D13	Yes	2018	ISIS EM-BLASSO	-0.16	7.94	31.67	0.10	CT
			D13	Yes	2018	mrMLM	-0.17	5.64	27.00	0.09	TT
			D13	Yes	2018	FASTmrMLM	-0.15	5.01	26.67	0.10	CT
			D13	Yes	2018	pLARmEB	-0.16	8.12	33.20	0.10	CT
		Mg	D12	No	2018	ISIS EM-BLASSO	-0.10	4.67	12.82	0.13	CT
2c	3036194	K	D11	Yes	2017	mrMLM	0.89	3.31	23.40	0.07	AA
			D11	Yes	2017	pLARmEB	0.68	3.34	18.76	0.08	AG
			D11	Yes	2017	mrMLM	-0.77	7.51	20.07	0.14	GG

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D11	Yes	2017	FASTmrMLM	-0.70	6.44	21.36	0.15	GG
			D11	Yes	2017	FASTmrEMMA	-1.05	4.43	9.85	0.15	GG
			D11	Yes	2017	pLARmEB	-0.46	5.20	9.89	0.15	GG
2c	13968839	Mg	D10	No	2017	pLARmEB	0.02	3.40	0.65	0.38	GG
			D10	No	2017	ISIS EM-BLASSO	0.04	5.24	1.74	0.38	GG
2c	46800088	P	D12	No	2018	pLARmEB	0.06	3.27	5.94	0.12	TT
			D13	Yes	2018	pLARmEB	0.08	3.09	12.54	0.08	TT
2c	52520849	N	D12	No	2018	mrMLM	0.86	3.05	9.63	0.24	GG
			D12	No	2018	ISIS EM-BLASSO	0.99	3.95	15.53	0.24	NN
2e	10671091	K	D10	No	2017	pLARmEB	-0.46	5.87	9.48	0.19	NN
			D10	No	2017	ISIS EM-BLASSO	-0.44	4.28	10.15	0.19	NN
2e	15411452	K	D10	No	2017	mrMLM	0.86	4.15	24.25	0.12	TT
			D10	No	2017	FASTmrMLM	0.71	4.34	19.36	0.11	TT
			D10	No	2017	pLARmEB	0.54	3.83	9.86	0.11	TT
			D10	No	2017	ISIS EM-BLASSO	0.59	3.17	13.65	0.11	TT
2e	21305731	N	D10	No	2017	FASTmrMLM	0.57	5.28	5.37	0.48	CC
			D10	No	2017	pLARmEB	0.57	6.42	3.56	0.48	CC
			D10	No	2017	ISIS EM-BLASSO	0.52	6.09	5.18	0.48	CC
2e	25615669	Ca	D12	No	2018	FASTmrMLM	0.11	5.56	23.02	0.15	TT
			D12	No	2018	pLARmEB	0.11	4.55	11.41	0.15	TT
			D12	No	2018	ISIS EM-BLASSO	0.13	5.30	27.44	0.15	TT
			D13	Yes	2018	mrMLM	0.14	3.24	15.19	0.08	TT
2e	32049748	Ca	D12	No	2018	FASTmrMLM	-0.07	4.51	8.93	0.15	NN
			D12	No	2018	pLARmEB	-0.10	4.96	8.97	0.15	NN
			D12	No	2018	ISIS EM-BLASSO	-0.09	5.61	11.62	0.15	NN
3c	2252158	K	D10	No	2017	pLARmEB	0.44	3.20	6.26	0.11	NN
			D10	No	2017	ISIS EM-BLASSO	0.48	3.28	8.73	0.11	NN
3e	36725563	Mg	D12	No	2018	mrMLM	-0.13	4.83	16.22	0.11	TT

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D12	No	2018	FASTmrMLM	-0.11	4.78	17.32	0.11	NN
			D12	No	2018	FASTmrEMMA	0.00	3.49	0.00	0.11	NN
			D12	No	2018	pLARmEB	-0.10	4.62	14.12	0.11	NN
			D12	No	2018	ISIS EM-BLASSO	-0.12	4.85	16.04	0.11	NN
4c	398594	K	D12	No	2018	mrMLM	-1.01	3.06	24.45	0.08	TT
			D12	No	2018	ISIS EM-BLASSO	-0.84	5.72	13.51	0.08	TT
			D13	Yes	2018	mrMLM	-0.93	3.69	13.04	0.06	TT
4c	763740	N	D10	No	2017	mrMLM	-1.45	6.48	24.22	0.15	AA
			D10	No	2017	FASTmrMLM	-1.23	7.07	21.43	0.15	NN
			D10	No	2017	pLARmEB	-1.23	7.79	14.17	0.15	NN
			D10	No	2017	ISIS EM-BLASSO	-1.08	6.57	19.09	0.15	NN
4c	763805	Ca	D11	Yes	2017	mrMLM	-0.15	4.56	26.56	0.07	TT
			D11	Yes	2017	pLARmEB	-0.15	4.10	10.66	0.08	TT
			D11	Yes	2017	ISIS EM-BLASSO	-0.13	3.36	27.54	0.08	TT
			D14	Yes	2017/2018	mrMLM	-0.14	3.76	41.50	0.07	T
			D14	Yes	2017/2018	FASTmrMLM	-0.11	3.98	32.90	0.08	TT
4e	1616535	Mg	D10	No	2017	pLARmEB	0.03	3.15	2.77	0.30	CT
			D10	No	2017	ISIS EM-BLASSO	0.06	8.95	7.88	0.30	CT
4e	28869207	P	D12	No	2018	pLARmEB	0.04	3.42	3.26	0.37	CT
			D12	No	2018	ISIS EM-BLASSO	0.04	3.44	4.73	0.37	CT
5c	14986549	K	D12	No	2018	pLARmEB	-0.67	4.06	11.97	0.12	TT
			D11	Yes	2017	ISIS EM-BLASSO	-0.67	3.76	14.99	0.11	TT
5c	24419353	Ca	D12	No	2018	FASTmrMLM	0.07	6.43	11.80	0.19	AA
			D12	No	2018	pLARmEB	0.07	3.41	5.07	0.19	AA
			D12	No	2018	ISIS EM-BLASSO	0.08	5.20	10.90	0.19	AA
5c	27867264	Ca	D10	No	2017	FASTmrEMMA	0.00	3.12	0.00	0.15	NN
			D10	No	2017	pLARmEB	-0.14	7.49	30.58	0.15	NN
5c	28812383	Ca	D13	Yes	2018	pLARmEB	-0.07	3.14	7.63	0.12	AA

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D13	Yes	2018	ISIS EM-BLASSO	-0.08	3.27	7.22	0.12	AA
5c	29553723	K	D12	No	2018	FASTmrMLM	0.68	4.55	15.50	0.13	TT
			D12	No	2018	pLARmEB	0.69	5.47	13.57	0.13	TT
			D12	No	2018	ISIS EM-BLASSO	0.85	7.42	20.12	0.13	TT
5c	29867171	P	D11	Yes	2017	FASTmrMLM	0.07	4.88	1.83	0.16	GG
			D11	Yes	2017	pLARmEB	0.04	7.01	0.27	0.16	GG
			D11	Yes	2017	ISIS EM-BLASSO	0.07	3.19	1.91	0.16	GG
5e	4484667	N	D13	Yes	2018	FASTmrMLM	-0.94	3.10	9.70	0.09	TT
			D13	Yes	2018	pLARmEB	-1.03	3.63	10.87	0.09	TT
			D13	Yes	2018	ISIS EM-BLASSO	-0.99	3.28	10.29	0.09	TT
5e	23073079	P	D11	Yes	2017	mrMLM	-0.15	4.65	51.49	0.07	GG
			D11	Yes	2017	FASTmrMLM	-0.08	3.13	19.90	0.08	GG
			D11	Yes	2017	ISIS EM-BLASSO	-0.08	3.18	21.77	0.08	GG
5e	36481790	Mg	D12	No	2018	mrMLM	0.13	5.72	18.68	0.12	GG
			D12	No	2018	FASTmrMLM	0.11	6.44	16.38	0.12	NN
			D12	No	2018	FASTmrEMMA	0.00	3.32	0.00	0.12	NN
			D12	No	2018	ISIS EM-BLASSO	0.11	5.15	14.36	0.12	NN
5e	36481803	K	D13	Yes	2018	FASTmrMLM	-0.57	3.72	14.56	0.14	TT
			D13	Yes	2018	pLARmEB	-0.45	3.08	9.48	0.14	TT
			D13	Yes	2018	ISIS EM-BLASSO	-0.57	3.29	16.11	0.14	TT
			D14	Yes	2017/2018	ISIS EM-BLASSO	-0.44	4.52	14.44	0.16	TT
6c	3018182	Mg	D10	No	2017	FASTmrMLM	0.00	3.18	0.01	0.25	NN
			D10	No	2017	pLARmEB	-0.01	4.89	0.12	0.25	NN
			D10	No	2017	ISIS EM-BLASSO	-0.06	4.41	11.65	0.25	NN
6c	7442606	P	D14	Yes	2017/2018	mrMLM	-0.12	5.66	36.42	0.07	G
			D14	Yes	2017/2018	pLARmEB	-0.10	5.21	20.93	0.07	GG
6e	767196	K	D10	No	2017	mrMLM	-0.63	4.78	11.92	0.15	TT
			D10	No	2017	FASTmrMLM	-0.55	5.00	10.27	0.14	TT

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D10	No	2017	pLARmEB	-0.45	5.88	6.16	0.14	TT
			D10	No	2017	ISIS EM-BLASSO	-0.52	5.54	9.63	0.14	TT
			D11	Yes	2017	mrMLM	-0.61	4.42	12.64	0.14	TT
			D11	Yes	2017	FASTmrMLM	-0.61	4.67	15.96	0.14	TT
			D11	Yes	2017	FASTmrEMMA	-1.10	3.58	9.13	0.14	TT
			D11	Yes	2017	pLARmEB	-0.50	4.93	11.63	0.14	TT
			D11	Yes	2017	ISIS EM-BLASSO	-0.70	6.30	14.73	0.14	TT
			D14	Yes	2017/2018	mrMLM	-0.70	5.06	21.95	0.11	T
			D14	Yes	2017/2018	FASTmrMLM	-0.56	4.24	17.22	0.11	TT
			D14	Yes	2017/2018	FASTmrEMMA	0.00	5.19	0.00	0.11	TT
			D14	Yes	2017/2018	pLARmEB	-0.60	6.22	8.79	0.11	TT
			D14	Yes	2017/2018	ISIS EM-BLASSO	-0.50	3.29	11.83	0.11	TT
6e	7581121	Ca	D11	Yes	2017	FASTmrMLM	-0.06	3.17	8.58	0.17	AA
			D11	Yes	2017	pLARmEB	-0.08	3.22	4.25	0.17	AA
			D11	Yes	2017	ISIS EM-BLASSO	-0.08	3.59	15.55	0.17	AA
			D14	Yes	2017/2018	mrMLM	-0.08	3.05	24.71	0.18	A
			D14	Yes	2017/2018	FASTmrMLM	-0.06	3.27	15.69	0.18	AA
6e	12969697	K	D10	No	2017	mrMLM	1.18	8.04	43.65	0.40	AA
			D10	No	2017	FASTmrMLM	1.09	7.99	42.87	0.41	AT
			D10	No	2017	FASTmrEMMA	1.77	4.44	13.33	0.41	AT
			D10	No	2017	pLARmEB	0.99	11.31	32.18	0.41	AT
			D10	No	2017	ISIS EM-BLASSO	0.97	7.11	35.21	0.41	AT
			D11	Yes	2017	pLARmEB	0.64	4.88	22.97	0.39	AT
			D11	Yes	2017	ISIS EM-BLASSO	0.69	4.84	17.44	0.39	AT
			D14	Yes	2017/2018	ISIS EM-BLASSO	0.31	3.49	6.61	0.39	AT
7c	777009	Mg	D10	No	2017	FASTmrMLM	-0.01	3.12	0.31	0.12	AG
			D10	No	2017	pLARmEB	-0.09	6.41	15.98	0.12	AG
			D10	No	2017	ISIS EM-BLASSO	-0.09	4.93	18.84	0.12	AG

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D14	Yes	2017/2018	FASTmrEMMA	0.00	3.02	0.00	0.12	AG
			D14	Yes	2017/2018	ISIS EM-BLASSO	-0.10	4.39	22.84	0.12	AG
7c	1131684	P	D10	No	2017	FASTmrMLM	0.07	3.38	28.95	0.16	AA
			D10	No	2017	pLARmEB	0.07	3.21	20.87	0.16	AA
7c	5878838	Mg	D12	No	2018	pLARmEB	-0.07	3.48	8.29	0.16	GG
			D12	No	2018	ISIS EM-BLASSO	-0.07	3.56	7.25	0.16	GG
7c	7474388	N	D12	No	2018	mrMLM	1.45	7.12	27.34	0.22	TT
			D12	No	2018	FASTmrMLM	1.28	4.45	27.51	0.23	TC
			D12	No	2018	pLARmEB	1.13	6.68	16.43	0.23	TC
			D12	No	2018	ISIS EM-BLASSO	1.21	8.56	23.00	0.23	TC
			D13	Yes	2018	mrMLM	1.48	5.86	28.60	0.22	TT
			D13	Yes	2018	FASTmrMLM	1.09	4.88	23.94	0.22	CT
			D13	Yes	2018	FASTmrEMMA	0.00	4.04	0.00	0.22	CT
			D13	Yes	2018	pLARmEB	1.14	4.31	24.36	0.22	CT
			D13	Yes	2018	ISIS EM-BLASSO	1.13	4.25	24.40	0.22	CT
7c	9532384	P	D11	Yes	2017	FASTmrMLM	-0.06	5.00	12.53	0.49	GG
			D11	Yes	2017	pLARmEB	-0.01	5.00	0.08	0.49	GG
7c	12758930	N	D10	No	2017	mrMLM	0.97	4.32	12.14	0.19	AA
			D10	No	2017	FASTmrMLM	0.69	4.42	7.37	0.19	NN
			D10	No	2017	pLARmEB	0.69	3.65	4.87	0.19	NN
			D10	No	2017	ISIS EM-BLASSO	0.72	5.33	9.35	0.19	NN
7c	23065519	N	D10	No	2017	mrMLM	1.58	6.49	26.97	0.13	AA
			D10	No	2017	FASTmrMLM	1.48	7.25	28.70	0.13	NN
			D10	No	2017	pLARmEB	1.48	6.89	18.98	0.13	NN
			D10	No	2017	ISIS EM-BLASSO	1.28	6.97	25.23	0.13	NN
7e	2558936	Ca	D13	Yes	2018	pLARmEB	-0.05	3.97	3.43	0.44	TT
			D13	Yes	2018	ISIS EM-BLASSO	-0.04	3.14	2.41	0.44	TT
7e	6318625	P	D12	No	2018	mrMLM	-0.08	3.83	18.89	0.14	AA

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D12	No	2018	FASTmrMLM	-0.06	8.18	12.47	0.14	NN
			D12	No	2018	pLARmEB	-0.06	5.36	6.24	0.14	NN
			D12	No	2018	ISIS EM-BLASSO	-0.05	7.59	9.24	0.14	NN
7e	10898466	N	D12	No	2018	mrMLM	0.89	3.69	10.25	0.29	TT
			D12	No	2018	FASTmrMLM	0.74	3.87	9.17	0.29	TT
			D12	No	2018	ISIS EM-BLASSO	0.67	4.06	7.05	0.29	TT
8c	1741546	K	D12	No	2018	mrMLM	1.03	4.64	30.77	0.10	AA
			D13	Yes	2018	mrMLM	1.24	6.63	24.95	0.07	AA
			D13	Yes	2018	pLARmEB	0.85	3.17	19.14	0.07	AA
8c	9060394	Ca	D13	Yes	2018	mrMLM	0.15	3.85	17.64	0.08	TT
			D13	Yes	2018	FASTmrMLM	0.12	3.37	14.31	0.08	TT
			D13	Yes	2018	pLARmEB	0.11	5.05	13.06	0.08	TT
			D13	Yes	2018	ISIS EM-BLASSO	0.11	4.30	13.10	0.08	TT
8c	25790868	Ca	D10	No	2017	FASTmrEMMA	0.00	4.13	0.00	0.12	GG
			D10	No	2017	pLARmEB	0.10	4.04	14.05	0.12	GG
8c	28326873	Mg	D12	No	2018	mrMLM	-0.10	3.34	7.67	0.12	TT
			D12	No	2018	FASTmrMLM	-0.07	3.36	4.50	0.12	TT
8c	29217952	Mg	D10	No	2017	pLARmEB	0.07	3.58	10.53	0.17	NN
			D11	Yes	2017	FASTmrMLM	0.10	5.05	22.46	0.13	AA
8c	29801841	Mg	D12	No	2018	mrMLM	0.12	4.38	17.61	0.14	AA
			D12	No	2018	FASTmrMLM	0.08	3.06	10.22	0.14	AA
			D12	No	2018	pLARmEB	0.07	4.96	8.22	0.14	AA
			D12	No	2018	ISIS EM-BLASSO	0.10	4.47	12.46	0.14	AA
8c	30292619	N	D10	No	2017	FASTmrEMMA	-1.66	3.12	9.54	0.20	GG
			D10	No	2017	ISIS EM-BLASSO	-0.39	3.01	2.45	0.20	GG
			D11	Yes	2017	pLARmEB	-0.87	3.53	13.58	0.19	GG
			D11	Yes	2017	ISIS EM-BLASSO	-0.87	3.52	16.27	0.19	GG
8e	250072	Mg	D12	No	2018	pLARmEB	-0.08	3.06	8.72	0.11	TT

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D13	Yes	2018	mrMLM	-0.17	4.98	38.49	0.11	TT
			D13	Yes	2018	FASTmrMLM	-0.14	4.81	31.88	0.10	TT
			D13	Yes	2018	pLARmEB	-0.11	3.12	15.94	0.10	TT
			D14	Yes	2017/2018	mrMLM	-0.14	3.78	43.85	0.11	T
			D14	Yes	2017/2018	pLARmEB	-0.11	3.90	13.88	0.11	TT
			D14	Yes	2017/2018	ISIS EM-BLASSO	-0.11	5.75	26.08	0.11	TT
8e	408365	K	D10	No	2017	mrMLM	-0.72	3.98	1.35	0.33	GG
			D10	No	2017	FASTmrMLM	-0.63	4.68	1.22	0.34	GG
			D10	No	2017	pLARmEB	-0.47	3.97	0.60	0.34	GG
			D10	No	2017	ISIS EM-BLASSO	-0.60	5.26	1.13	0.34	GG
			D11	Yes	2017	FASTmrEMMA	-1.21	3.37	7.95	0.31	GG
			D11	Yes	2017	pLARmEB	-0.49	3.12	1.03	0.31	GG
8e	28487464	N	D10	No	2017	FASTmrMLM	-0.74	3.82	3.84	0.46	AA
			D10	No	2017	pLARmEB	-0.74	4.56	2.54	0.46	AA
			D10	No	2017	ISIS EM-BLASSO	-0.61	3.24	3.04	0.46	AA
			D11	Yes	2017	mrMLM	-1.51	5.52	18.02	0.43	AA
			D11	Yes	2017	FASTmrMLM	-1.26	4.32	19.20	0.44	AA
			D11	Yes	2017	FASTmrEMMA	-2.06	3.27	10.41	0.44	AA
			D11	Yes	2017	pLARmEB	-1.02	4.15	10.05	0.44	AA
			D11	Yes	2017	ISIS EM-BLASSO	-1.02	5.45	12.04	0.44	AA
8e	33206159	K	D13	Yes	2018	FASTmrMLM	-0.57	3.33	16.73	0.20	TT
			D13	Yes	2018	FASTmrEMMA	-1.24	3.70	9.67	0.20	TT
			D14	Yes	2017/2018	mrMLM	-0.75	4.91	40.48	0.15	T
			D14	Yes	2017/2018	FASTmrMLM	-0.63	4.63	35.30	0.16	TT
			D14	Yes	2017/2018	FASTmrEMMA	0.00	4.20	0.00	0.16	TT
			D14	Yes	2017/2018	pLARmEB	-0.57	4.82	12.68	0.16	TT
			D14	Yes	2017/2018	ISIS EM-BLASSO	-0.70	6.66	36.36	0.16	TT
		P	D14	Yes	2017/2018	mrMLM	-0.08	4.22	33.74	0.15	T

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D14	Yes	2017/2018	FASTmrMLM	-0.07	3.95	34.13	0.16	TT
			D14	Yes	2017/2018	FASTmrEMMA	0.00	3.94	0.00	0.16	TT
			D14	Yes	2017/2018	ISIS EM-BLASSO	-0.07	4.78	23.64	0.16	TT
9c	304428	Mg	D10	No	2017	FASTmrMLM	0.00	3.14	0.00	0.09	TT
			D10	No	2017	ISIS EM-BLASSO	0.00	3.15	0.00	0.09	TT
9c	375837	Mg	D11	Yes	2017	mrMLM	-0.25	4.63	49.96	0.04	AA
			D11	Yes	2017	pLARmEB	-0.22	4.37	16.88	0.04	GA
			D11	Yes	2017	ISIS EM-BLASSO	-0.22	4.37	44.36	0.04	GA
9c	3500675	Ca	D11	Yes	2017	mrMLM	-0.10	4.04	19.49	0.22	TT
			D11	Yes	2017	FASTmrMLM	-0.10	5.71	25.76	0.22	CT
			D11	Yes	2017	FASTmrEMMA	0.00	3.48	0.00	0.22	CT
			D11	Yes	2017	pLARmEB	-0.08	3.13	5.12	0.22	CT
10c	4582941	K	D12	No	2018	pLARmEB	0.66	4.28	14.79	0.18	NN
			D12	No	2018	ISIS EM-BLASSO	0.59	3.91	11.71	0.18	NN
10c	6503365	K	D14	Yes	2017/2018	mrMLM	-0.80	3.10	9.79	0.48	A
			D14	Yes	2017/2018	pLARmEB	-0.62	3.19	3.16	0.48	GA
10e	3044516	P	D12	No	2018	FASTmrMLM	0.00	4.46	0.00	0.11	TT
			D12	No	2018	ISIS EM-BLASSO	-0.04	3.59	5.78	0.11	TT
			D13	Yes	2018	mrMLM	-0.12	3.63	29.26	0.08	TT
			D13	Yes	2018	FASTmrMLM	-0.09	3.89	22.50	0.08	TT
			D13	Yes	2018	pLARmEB	-0.08	3.06	12.06	0.08	TT
10e	3763094	Ca	D12	No	2018	FASTmrMLM	-0.02	3.38	0.69	0.15	NN
			D12	No	2018	pLARmEB	-0.06	4.16	3.75	0.15	NN
10e	3763101	Ca	D12	No	2018	mrMLM	-0.10	4.15	14.77	0.28	AA
			D13	Yes	2018	mrMLM	-0.11	4.35	16.47	0.24	AA
			D13	Yes	2018	FASTmrMLM	-0.09	3.84	13.21	0.24	AA
			D13	Yes	2018	pLARmEB	-0.06	3.57	8.28	0.24	AA
10e	9090235	Ca	D10	No	2017	mrMLM	-0.12	3.31	29.72	0.12	GG

Chromosome	Position	Trait	Dataset	Imputed	Environment	Method	QTN effect	LOD score	r^2	MAF	Genotype
			D10	No	2017	FASTmrMLM	0.00	3.06	0.05	0.12	NN
			D10	No	2017	pLARmEB	-0.06	3.70	4.66	0.12	NN
			D10	No	2017	ISIS EM-BLASSO	-0.10	4.75	13.74	0.12	NN
11c	2494757	K	D11	Yes	2017	mrMLM	-0.49	5.08	13.10	0.37	GG
			D11	Yes	2017	FASTmrMLM	-0.46	5.72	15.00	0.36	GG
11c	17308057	N	D12	No	2018	mrMLM	1.08	7.53	13.03	0.24	GG
			D12	No	2018	FASTmrEMMA	1.50	3.05	8.54	0.24	GG
			D12	No	2018	pLARmEB	0.79	5.03	6.92	0.24	GG
			D12	No	2018	ISIS EM-BLASSO	1.11	8.60	16.77	0.24	GG
			D13	Yes	2018	mrMLM	1.17	9.21	14.20	0.21	GG
			D13	Yes	2018	FASTmrMLM	1.07	7.00	18.20	0.22	GG
			D13	Yes	2018	FASTmrEMMA	2.33	6.52	20.58	0.22	GG
			D13	Yes	2018	pLARmEB	1.17	6.42	20.55	0.22	GG
			D13	Yes	2018	ISIS EM-BLASSO	1.14	6.46	19.87	0.22	GG
			D14	Yes	2017/2018	mrMLM	1.10	4.63	26.74	0.21	G
11c	27532072	N	D12	No	2018	mrMLM	-1.03	3.68	6.77	0.48	AA
			D12	No	2018	FASTmrMLM	0.00	4.00	0.00	0.48	NN
			D12	No	2018	ISIS EM-BLASSO	-0.73	4.59	4.18	0.48	NN
			D13	Yes	2018	mrMLM	-1.06	3.17	7.76	0.48	AA
11e	23756404	K	D12	No	2018	pLARmEB	-0.52	4.34	7.75	0.13	AA
			D12	No	2018	ISIS EM-BLASSO	-0.56	4.72	8.91	0.13	AA
11e	29533307	Ca	D11	Yes	2017	mrMLM	0.14	8.01	9.04	0.20	GG
			D11	Yes	2017	FASTmrMLM	0.12	9.54	7.15	0.21	GG
			D11	Yes	2017	FASTmrEMMA	0.22	3.66	17.22	0.21	GG
			D11	Yes	2017	pLARmEB	0.11	7.11	2.11	0.21	GG
			D11	Yes	2017	ISIS EM-BLASSO	0.12	6.32	9.35	0.21	GG