CROP BREEDING AND APPLIED BIOTECHNOLOGY

ARTICLE

A new set of quantitative trait loci linked to lipid content in *Coffea arabica*

Herison Victor Lima Muniz[1,2], Caroline Ariyoshi[2], Rafaelle Vecchia Ferreira[2], Mariane Silva Felicio[3] and Luiz Filipe Protasio Pereira[2,4*]

Crop Breeding and Applied Biotechnology
24(2): e478824212, 2024
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332024v24n2a25

Abstract: *Lipids are compounds that play an important role in coffee bean development, contributing to beverage quality. Genome-wide association studies (GWAS) were conducted to pinpoint quantitative trait nucleotides (QTNs) linked to lipid metabolism in Coffea arabica. Genotyping by sequencing (GBS) and phenotyping data from 104 wild C. arabica accessions, Mundo Novo cultivar, and C. arabica var. Typica were utilized. GBS data were aligned to C. arabica Et039 reference genome, and both single-locus and multi-locus GWAS methods were employed. Methods were adjusted for kinship matrix, population structure, and principal component analysis. Of the 19 QTNs identified, 5 showed consistency across different population structure adjustments. The multi-locus methods mrMLM and FarmCPU proved more effective in identifying QTNs associated with lipid content. Four QTNs were situated near seven genes potentially involved in lipid metabolism. Higher frequencies of identified QTNs in accessions with elevated lipid content suggest their utility as markers for coffee plant breeding.*

Keywords: *Allotetraploid, coffee, GWAS, SNP markers*

INTRODUCTION

Coffee is one of the most popular beverages in the world, and its commercial production is based on two species: *Coffea arabica* and *Coffea canephora*. Brazil stands out for being the largest producer of this commodity, which contributes directly to the economy and job creation (CONAB 2023). *Coffea arabica* has higher production, mainly due to its superior cup quality, which adds greater economic value. Cup quality in coffee is due to the composition and combination of several compounds, including chlorogenic acids, caffeine, sugars, diterpenes, and lipids (Scholz et al. 2016). Lipids are key coffee compounds that play an important role in coffee bean development, contributing to the flavor and aroma of coffee beverages (Sant'Ana et al. 2018).

Genome-wide association studies (GWAS), together with next-generation sequencing technology, have emerged as powerful tools for identifying molecular markers associated with agronomic traits of interest. GWAS can overcome the limitations of traditional genetic linkage mapping, including maps with little refinement and limited parental diversity (Bartoli and Roux 2017). GWAS can explore the genetic diversity found in wild crop relatives, offering a higher mapping resolution in comparison with biparental quantitative trait loci (QTLs) experiments and is considered a cost-effective way to detect associations between molecular markers and traits of interest (Korte and Farlow 2013, Su et al. 2016).

***Corresponding author:**
E-mail: filipe.pereira@embrapa.br
ORCID: 0000-0002-4872-6607

Received: 18 December 2023
Accepted: 07 March 2024
Published: 10 April 2024

[1] Universidade Tecnológica Federal do Paraná, Avenida Alberto Carazzai, 1640, Centro, 86300-000, Cornélio Procópio, PR, Brazil
[2] Instituto de Desenvolvimento Rural do Paraná, Rodovia Celso Garcia Cid, km 375, Vivendas do Arvoredo, 86047-902, Londrina, PR, Brazil
[3] Universidade Estadual Paulista "Júlio de Mesquita Filho", Rua Professor Doutor Antônio Celso Wagner Zanin, 250, Distrito de Rubião Junior, 18618-689, Botucatu, SP, Brazil
[4] Embrapa Café, Parque Estação Biológica, PQEB W3, 70770-901, Brasília, DF, Brazil

Sant'Ana et al. (2018) published the first GWAS in *C. arabica* and identified single nucleotide polymorphisms (SNPs) associated with the biochemical characteristics of the grain, such as lipids, and the diterpenes cafestol and kahweol. However, it was used only in the reference genome of *C. canephora* (one of the diploid ancestors of the allotetraploid *C. arabica*), limiting the number of SNPs used in the association.

In this study, our objectives were to use the same genotyping by sequencing (GBS) and lipid phenotype data from previous work but with the GBS data aligned to the complete *C. arabica* genome in the association study, aiming to identify novel genomic regions linked to lipid metabolism in *C. arabica*.

MATERIAL AND METHODS

Plant material

A *Coffea arabica* collection of 104 wild genotypes from Ethiopia, *C. arabica* Mundo Novo 38, and *C. arabica* var. Typica was used. GBS and phenotyping for total lipid content in green grains from this population were previously carried out by Sant'Ana et al. (2018). The GBS data were aligned to *C. arabica* reference genome Et39 (Arabica Coffee Genome Consortium, Salojärvi et al. 2023) and subjected to SNP calling using the Tassel 5 GBS v2 pipeline, resulting in a panel of 159,000 SNPs (Glaubitz et al. 2014).

SNP filtering

The 159,000 SNPs aligned to the *C. arabica* genome were filtered using TASSEL software version 5.2.89, as described by Bradbury et al. (2007). Filtering was performed with the parameters of minor allele frequency (MAF > 0.05) and call rate > 0.8. For data imputation, Beagle v4.1 software (Browning and Browning 2016) and LD-kNNi based on the K-nearest neighbor method (Money et al. 2015) were used.

Population structure

The first five principal component analyses (PCA) and kinship matrix (K) were calculated using TASSEL 5.2.53 and used for single-locus GWAS (SL-GWAS) and multi-locus GWAS (ML-GWAS). The Q matrix (Q) was produced using Structure v2.3.4 software (Pritchard et al. 2000). For the Q matrix, the allele frequencies of each K cluster (2-10) were estimated, with 1000 runs as the burn-in period, 1000 runs for the Markov chain Monte Carlo (MCMC), and 10 runs for each K value. The ΔK criterion (Evanno et al. 2005) was used in Structure Harvester software (Earl and Vonholdt 2012) to estimate the uppermost level of the population structure.

Single-locus and multi-locus GWAS

Single-locus GWAS (SL-GWAS) was performed by TASSEL 5.2.53 with two methods: a general linear model (GLM, Price et al. 2006) and a mixed linear model (MLM, Yu et al. 2006). The association threshold for SL-GWAS was $p \leq 0.05/n$, where n is the number of markers. Multi-locus GWAS (ML-GWAS) were performed using five methods: mrMLM (Wang et al. 2016), FASTmrMLM (Tamba and Zhang 2018), FASTmrEMMA (Wen et al. 2018), ISIS EM-BLASSO (Tamba et al. 2017), and FarmCPU (Liu et al. 2016) using R software (R Core Team 2023). For mrMLM, FASTmrMLM, FASTmrEMMA, and ISIS EM-BLASSO methods, in the first step, the critical values $p \leq 0.01$, 0.01, 0.005, and 0.01 were used, respectively, for the intermediate result. In the second step, all SNPs selected in the first step were filtered by the multi-locus methods, and the markers with the largest effects that exceeded the LOD score threshold were considered potentially associated SNPs. The critical LOD score threshold was set to 3 for SNPs in the final phase. For FarmCPU, the criterion $p \leq 0.0005$ was used.

Linkage disequilibrium analysis and identification of candidate genes

Squared correlation coefficients (r^2) were calculated on sliding windows with 50 adjacent SNPs in TASSEL version 5.2.53, which was used to evaluate linkage disequilibrium (LD) decay. The LD distance value in base pairs (bp) was evaluated by a non-linear regression method, at $r^2 = 0.2$, using R software (R Core Team 2023). The search for candidate genes was performed using the *C. arabica* Et039 genome functional annotation according to the LD threshold upstream and downstream of the associated QTN positions.

RESULTS AND DISCUSSION

SNP filtering and population structure

After quality filters and imputation, of the 159,000 SNPs identified in the SNP calling, 11,136 SNPs were used for population structure adjustments and GWAS. Of this total, 5032 were identified in the *canephora* subgenome, 4870 in the *eugenioides* subgenome, and 1234 in chromosome zero. Chromosome zero corresponds to scaffolds that do not have a defined position during genome assembly. This represents a significant improvement concerning our first GWAS work, in which the GBS data were aligned only in the *C. canephora* genome, and GWAS was performed with only 2587 SNPs (Sant'Ana et al. 2018).

PCA-based population structure analysis (Supplementary Figure S1) and a Q matrix from Structure K=3 (the higher delta K) (Supplementary Figure S2) were used for population structure adjustments to the GWAS methods. In Sant'Ana et al. (2018), the higher delta K obtained was K=2 but, interestingly, the second highest value K=3 showed a population structure very similar to the present work. Both the PCA and structure analysis (Supplementary Figure 3) presented here have similar results to those of Ariyoshi et al. (2022), who used a similar plant collection and genotyped data mapped in the *C. arabica* genome.

Single-locus GWAS

Single-locus methods were unable to detect QTNs (Figure 1). A common feature of SL-GWAS is a one-dimensional genome scan, in which each marker is tested in turn. However, this approach does not facilitate good estimates of the effects of markers controlled by multiple loci, which occur in most complex traits (Wang et al. 2016). Another problem with the method is the issue of multiple test corrections for the threshold value of the significance test. Due to the Bonferroni correction, which is normally very conservative, methods such as GLM and MLM are not efficient in detecting small effective loci of a complex trait (Wang et al. 2016).

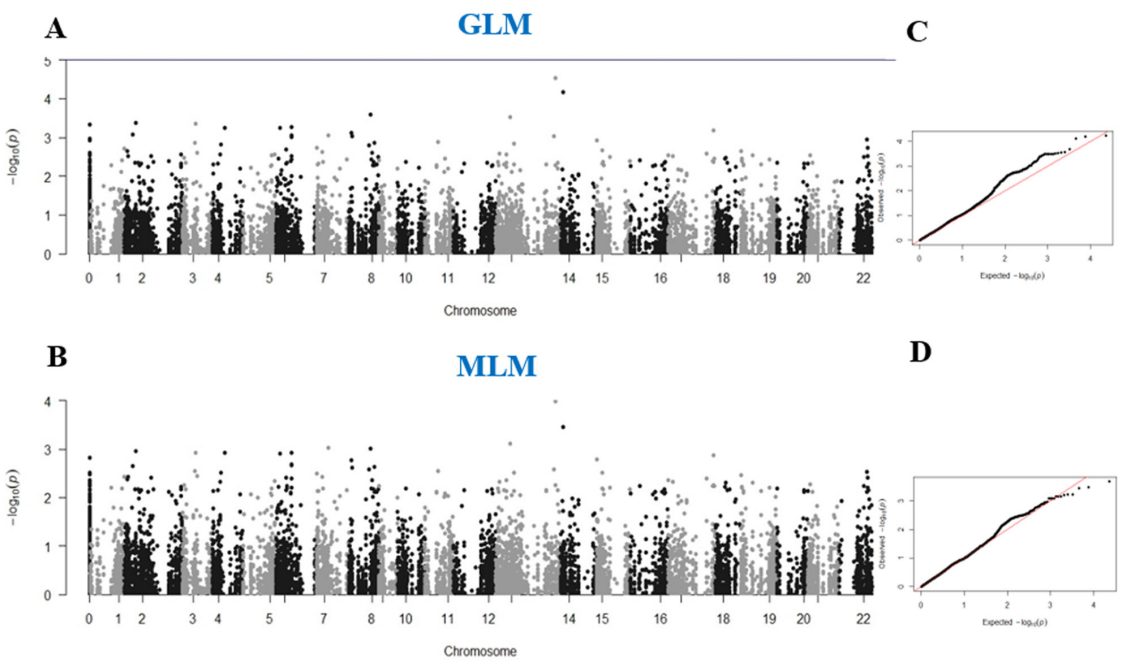

Figure 1. Manhattan plots (A–B) of the single-locus genome-wide association study for lipid content using the GLM and MLM methods. The *x*-axis represents the chromosomes, and the *y*-axis represents the $-\log_{10}$(*p*-value). The solid line indicates the thresholds for lipid content in *C. arabica*. Quantile–quantile (Q–Q) plots (C–D) of a single-locus genome-wide association study for lipid content in *C. arabica*. Q–Q plots show the observed vs. expected negative $\log_{10}$ *p*-values.

Multi-locus GWAS

In this study, only ML-GWAS detected QTNs related to lipid content. The advantages of ML-GWAS over SL-GWAS have already been described in studies of other plant species, such as cotton, corn, tobacco, and coffee (Li et al. 2018, Su et al. 2018, Xu et al. 2018, Ariyoshi et al. 2022, Ikram et al. 2022). This is because multi-locus methods are characterized by a multidimensional genome scanning approach in which the effects of all markers are estimated simultaneously (Cui et al. 2018).

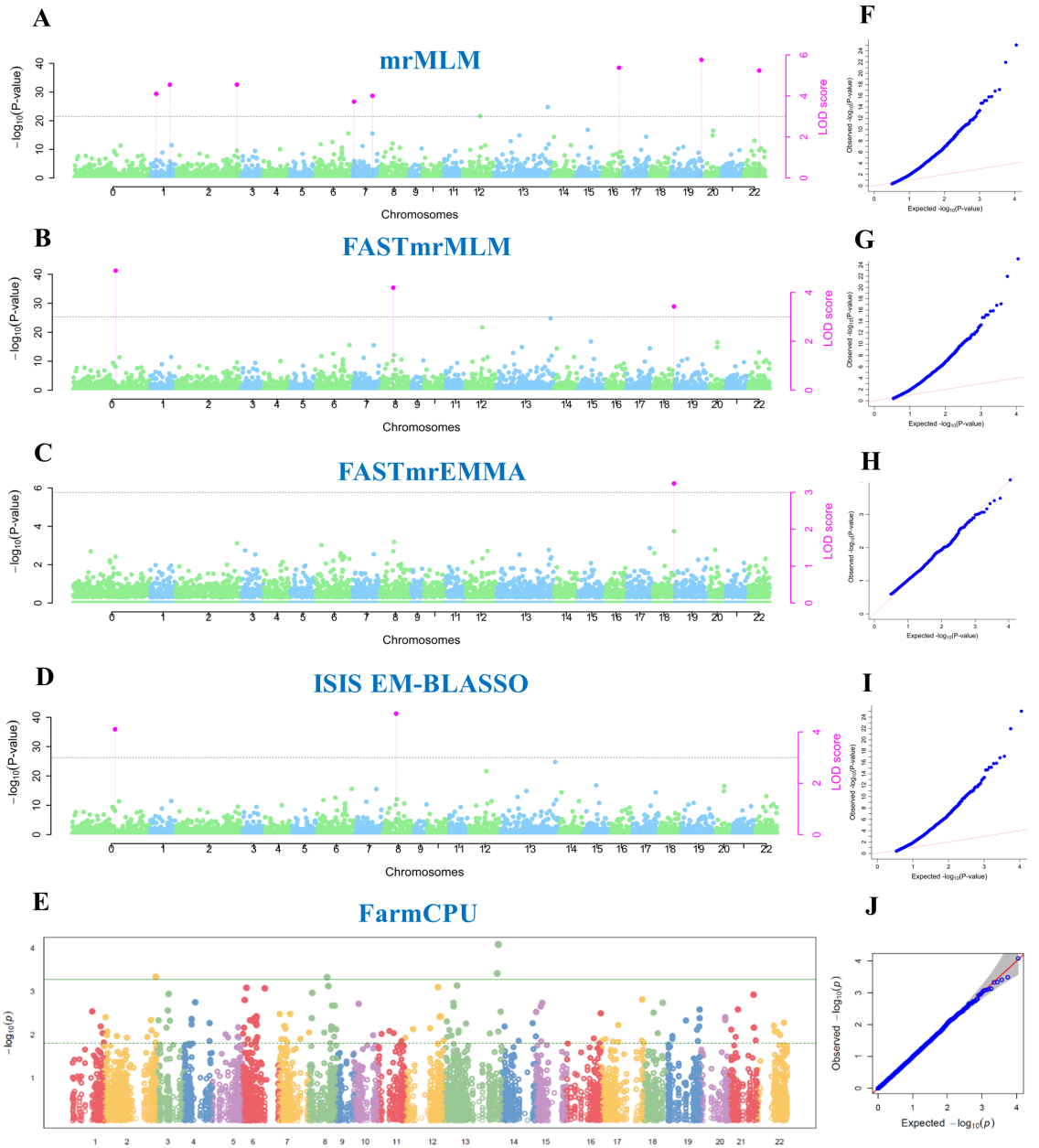

Figure 2. Manhattan plots (A–E) of the multi-locus genome-wide association study for lipid content with the methods mrMLM, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, and FarmCPU with the PCA + K models. The *x*-axis represents the chromosomes, and the *y*-axis represents the $-\log_{10}$(*p*-value). The dashed lines indicate the LOD score threshold and the solid line indicates the $-\log_{10}$(*p*-value) threshold for lipid content in *C. arabica*. Quantile–quantile (Q–Q) plots (F–J) of a multi-locus genome-wide association study for lipid content in *C. arabica* with the PCA + K models. Q–Q plots show the observed vs. expected negative $\log_{10}$ *p*-values.

In the population structure correction by PCA + K (Figure 2), 13 QTNs were identified (Table 1). However, except for the FASTmrEMMA and FarmCPU, the Q-Q plot graphs (Figure 2) depicted observed *p*-values consistently lower than expected throughout the plot with the PCA + K models. In those methods, the inclusion of PCA as a population structure correction resulted in unfavorable adjustments. In a study developed by Elhaik (2022), 12 common test cases were analyzed in human population data, in which the adjustment with PCA demonstrated unfavorable results in association studies for the studied trait.

In the adjustment of multi-locus methods, using clustering coefficient data obtained by the Structure software and K matrix (Q + K) (Figure 3), 6 QTNs were identified (Table 2). The methods presented a better adjustment, as shown in the Q-Q plot graphs (Figure 3), except for the FASTmrEMMA method, in which the PCA correction presented a better adjustment of the *p*-value data but did not associate QTN with the correction by the Q + K models. The inclusion of the Q matrix in the GWAS methods decreased the number of QTNs by approximately three times. Yang et al. (2011) evaluated complex traits in maize using the PCA + K and Q + K models as population structure corrections, in which Q + K showed a better reduction in false positives.

GWAS using PCA + K resulted in 13 QTNs, and in 3 of them (Chr_7_sg_C_409622, Chr_8_sg_C_19869722, and Chr_5_sg_E_35714509), we observed nearby genes related to lipid biosynthesis and/or metabolism. Meanwhile, the GWAS using Q + K resulted in 6 QTNs, 2 of which (Chr_6_sg_C_4567047 and Chr_8_sg_C_19869722) had nearby genes related to lipid genes. The different population structure correction models identified five convergent QTNs (Chr_0_4634_168274, Chr_8_sg_C_19869722, Chr_2_sg_E_60056603, Chr_7_sg_E_20849219, and Chr_11_sg_E_29236360).

Our previous GWAS study related to lipid content identified 5 QTNs (Sant'Ana et al. 2018). In this work, 14 QTNs were identified by the GWAS methods using the PCA + K and Q + K models, a number approximately three times greater than our initial results. Of the 6 QTNs of the *canephora* subgenome, 4 were on the same pseudo-chromosome as our original work (Sant'Ana et al. 2018).

Identification of candidate genes related to lipid content

In the four QTNs, genes involved in lipid metabolism and/or fatty acid biosynthesis were identified. From the functional annotation of *C. arabica* Et039, seven genes were identified close to the QTNs associated with the lipid trait (Table 3). To consider a gene linked to a QTN, the approximate distance based on the LD decay result was used. The LD decay r^2 = 0.2 was 158,774 bp (Supplementary Figure S4).

Table 1. Associated QTNs in the multi-locus mrMLM, FASTmrMLM, FASTmrEMMA ISIS-EM-BLASSO, and FarmCPU methods adjusted by PCA + K models

		mrMLM		FASTmrMLM		FASTmrEMMA		ISIS EM-BLASSO		FarmCPU
Position of QTN in the genome of *C. arabica* Et039*	**Allele change**	**−log10 (*p*-value)**	**LOD SCORE**	**−log10 (*p*-value)**	**LOD SCORE**	**−log10 (*p*-value)**	**LOD SCORE**	**−log10 (*p*-value)**	**LOD SCORE**	**−log10 (*p*-value)**
Chr_0_4635_168274	T/C			5.6818	4.8901			4.8566	4.1001	3.488117
Chr_1_sg_C_22364613	T/C	4.8608	4.1041							
Chr_1_sg_C_31802172	T/C	5.3237	4.5466							
Chr_2_sg_C_56859771	A/G	5.3267	4.5495							3.335358
Chr_7_sg_C_11717962	A/G	4.7552	4.0035							
Chr_7_sg_C_409622	G/A	4.4578	3.7205							
Chr_8_sg_C_19869722	A/G			4.9522	4.1914			5.4941	4.71	3.326979
Chr_2_sg_E_ 58841892	A/G									3.411168
Chr_2_sg_E_ 60056603	A/G									4.082494
Chr_5_sg_E_35714509	T/C	6.1815	5.3709							
Chr_7_sg_E_20849219	T/C			4.1472	3.4261	3.9487	3.2386			
Chr_8_sg_E_34967946	T/C	6.5827	5.758							
Chr_11_sg_E_29236376	A/G	6.0433	5.2378							

* Chr: chromosome, Sg: subgenome

QTN Chr_6_sg_C_4567047 is close to the *g15.102* gene, with its functional annotation for 4-phosphopantetheinyl transferase isoform X1. For Chr_5_sg_E_35714509, the nearby gene is *g119.153*, with functional annotation for the electron transfer flavo mitochondrial subunit. For these two QTNs, no descriptions were found in the literature for a better understanding of their functions.

QTN Chr_7_sg_C_409622 is close to the *g10.29* gene with functional annotation for the protein triacylglycerol lipase-like 1. This protein is involved in acyl-lipid metabolism in *Arabidopsis thaliana*, as well as in other plant species. The acyl lipid has several functions, including providing the central membrane diffusion barrier that separates cells and subcellular organelles. This function alone encompasses more than 10 classes of membrane lipids, such as phospholipids, galactolipids, and sphingolipids (Li-beisson et al. 2013). In addition to this QTN, in the annotation of *C. arabica* Et039, genes *g10.11* and *g10.8* were also identified, with participation in lipid metabolism, with functional annotations for patatin 3 and patatin 1, respectively.

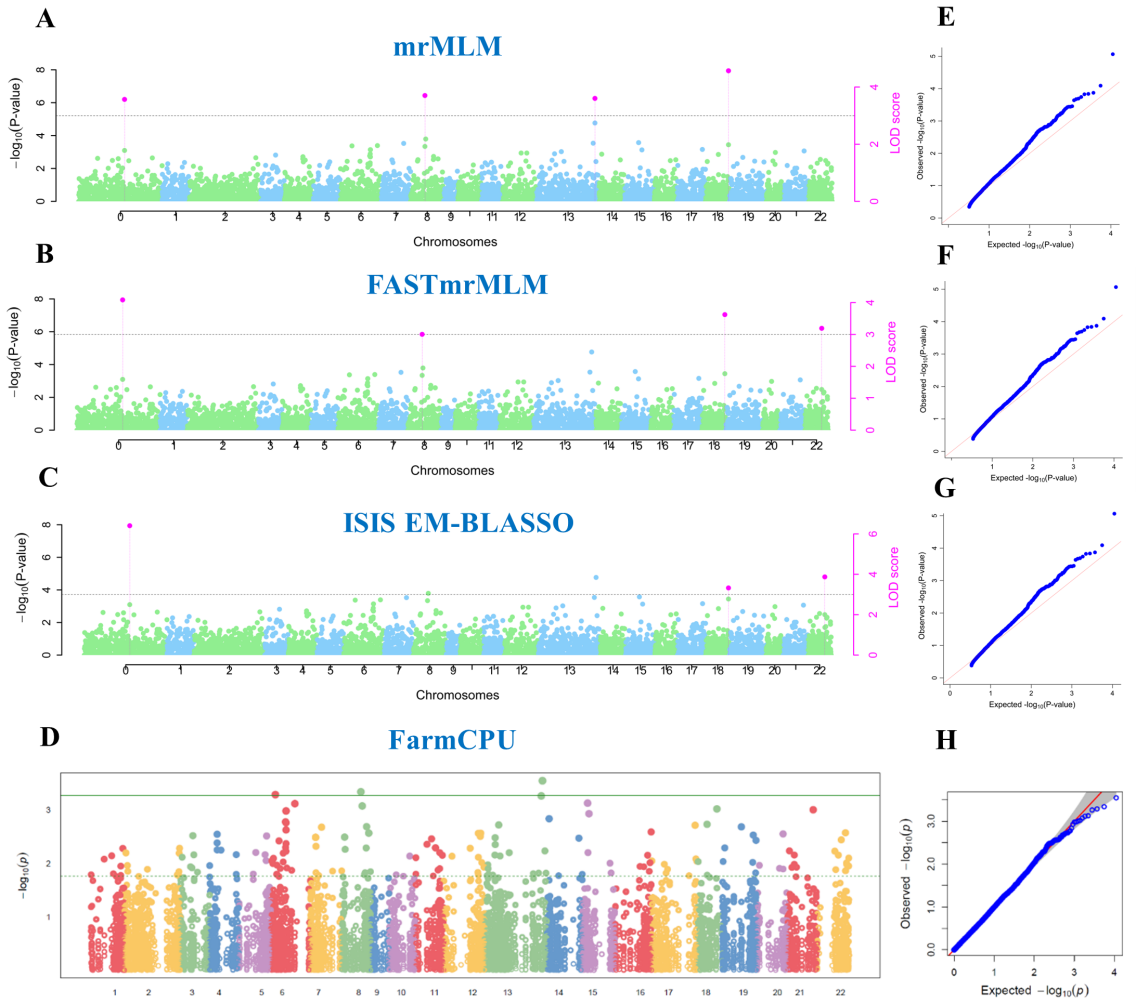

Figure 3. Manhattan plots (A–D) of the multi-locus genome-wide association study for lipid content with the methods mrMLM, FASTmrMLM, ISIS EM-BLASSO, and FarmCPU with the Q + K models. The *x*-axis represents the chromosomes, and the *y*-axis represents the $-\log_{10}(p\text{-value})$. The dashed lines indicate the LOD score threshold, and the solid line indicates the $-\log_{10}(p\text{-value})$ threshold, for lipid content in *C. arabica*. Quantile–quantile (Q–Q) plots (E–H) of a multi-locus genome-wide association study for lipid content in *C. arabica* with the Q + K models. Q–Q plots show the observed vs. expected negative $\log_{10}$ *p*-values.

Table 2. Associated QTNs in the multi-locus mrMLM, FASTmrMLM, ISIS-EM-BLASSO, and FarmCPU methods adjusted by Q + K models

		mrMLM		FASTmrMLM		ISIS EM-BLASSO		FarmCPU
Position of QTN in the genome of *C. arabica* Et039*	**Allele change**	**–log10 (p-value)**	**LOD SCORE**	**–log10 (p-value)**	**LOD SCORE**	**–log10 (p-value)**	**LOD SCORE**	**–log10 (p-value)**
Chr_0_4635_168274	T/C	4.2953	3.5664	4.8379	4.0823	7.2502	6.4039	
Chr_6_sg_C_4567047	C/T							3.286509
Chr_8_sg_C_19869722	A/G	4.4391	3.7028	3.6986	3.0031			3.338187
Chr_2_sg_E_60056603	A/G	4.3333	3.6024					3.546682
Chr_7_sg_E_20849219	T/C	5.3459	4.5679	4.353	3.6211	4.0384	3.3233	
Chr_11_sg_E_29236360	A/G			3.8991	3.1918	4.6147	3.8697	

*Chr: chromosome, Sg: subgenome

Table 3. Functional annotation of genes close to QTNs associated with lipids in *Coffea arabica*

Gene[1]	**QTN**	**Gene position in relation to the QTN[2]**	**Functional annotation[3]**
g15.102	Chr_6_sg_C_4567047	(+) 27.242	4 -phosphopantetheinyl transferase isoform X1
g1.69	Chr_7_sg_C_409622	(-) 172.382	senescence-associated carboxylesterase 101-like isoform X1
g10.11	Chr_7_sg_C_409622	(+) 46.483	patatin 3
g10.29	Chr_7_sg_C_409622	(+) 87.019	triacylglycerol lipase-like 1
g10.8	Chr_7_sg_C_409622	(+) 138.414	patatin 1
g66.10	Chr_8_sg_C_19869722	(+) 46.073	3-ketoacyl- synthase 10
g119.153	Chr_5_sg_E_35714509	(+) 19.826	electron transfer flavo subunit mitochondrial

[1] The name of the genes was retrieved from the *C. arabica* Et039 genome functional annotation. [2] (+) upstream or (-) downstream position in bp. [3] The genes functional annotation was retrieved from the *C. arabica* Et039 genome functional annotation.

QTN Chr_8_sg_C_19869722 is close to the *g66.10* gene, which has a functional annotation for 3-ketoacyl-synthase 10. This protein contributes to the biosynthesis of cuticular wax and suberin (Lolle et al. 1997). As precursors of wax compounds, very long-chain fatty acids participate in limiting non-stomatal water loss and preventing pathogen attacks. They are also used as energy storage in seeds and as building blocks for membranes. Twenty-one 3-ketoacyl-CoA synthase genes were identified in the *Arabidopsis thaliana* genome and expressed in seeds, flowers, and leaves (Joubès et al. 2008).

CONCLUSION

Using the *C. arabica* genome as a reference for the GWAS study for lipids represented a significant improvement from our previous work, where only the *C. canephora* genome was available. This allowed us to recover a higher number of SNPs, increase the number of QTNs, and identify candidate genes involved with lipids and/or fatty acid biosynthesis. The information generated is also important for the development of markers for breeding and for providing candidate genes for transcriptome analysis to depict lipid biosynthesis in *C. arabica*.

ACKNOWLEDGMENTS

We acknowledge the Consórcio Pesquisa Café for financial support and the Scholarship for H.V.L.M. (Grant 10.18.20.027.00). INCT Café provided fellowships for C.A., R.F.V., and M.S.F. L.F.P.P. acknowledges CNPq for the research fellowship. Supplementary Figures and Tables may be requested from the corresponding author.

REFERENCES

Ariyoshi C, Sant'ana GC, Felicio MS, Sera GH, Nogueira LM, Rodrigues LMR, Ferreira RV, Silva BSR, Resende MLV, Deste'fano SAL, Domingues DS and Pereira LFP (2022) Genome-wide association study for resistance to *Pseudomonas syringae* pv. *garcae* in *Coffea arabica*. **Frontiers in Plant Science 13**: 989847.

Bartoli C and Roux F (2017) Genome-Wide Association studies in plant pathosystems: Toward an ecological genomics approach. **Frontiers in Plant Science 8**: 763.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y and Buckler ES (2007) TASSEL: Software for association mapping of complex traits in diverse samples. **Bioinformatics 23**: 2633-2635.

Browning BL and Browning SR (2016) Genotype imputation with millions of reference samples. **The American Journal of Human Genetics 98**: 116-12.

CONAB - Companhia Nacional de Abastecimento (2023) Acompanhamento da safra brasileira de café. Available at <https://www.conab.gov.br/info-agro/safras/cafe/boletim-da-safra-de-cafe>. Accessed on April 23, 2023.

Cui Y, Zhang F and Zhou Y (2018) The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. **Frontiers in Plant Science 9**: 1464.

Earl DA and Vonholdt BM (2012) Structure Harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources 4**: 359-361.

Elhaik E (2022) Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. **Scientific Reports 12**: 14683.

Evanno G, Regnaut S and Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular Ecology 14**: 2611-2620.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q and Buckler ES (2014) TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. **PLoS One 9**: e90346.

Ikram M, Xiao J, Li R, Xia Y, Zhao W, Yuan Q, Siddique KHM and Guo P (2022) Identification of superior haplotypes and candidate genes for yield-related traits in tobacco (*Nicotiana tabacum* L.) using association mapping. **Industrial Crops and Products 189**: 115886.

Joubès J, Raffaele S, Bourdenx B, Garcia C, Laroche-Traineau J, Moreau P, Domergue F and Lessire R (2008) The VLCFA elongase gene family in *Arabidopsis thaliana*: phylogenetic analysis, 3D modelling and expression profiling. **Plant Molecular Biology 67**: 547-566.

Korte A and Farlow A (2013) The advantages and limitations of trait analysis with GWAS: A review. **Plant Methods 9**: 29.

Li C, Fu Y, Sun R, Wang Y and Wang Q (2018) Single-locus and multi-locus genome-wide association studies in the genetic dissection of fiber quality traits in upland cotton (*Gossypium hirsutum* L.). **Frontiers in Plant Science 9**: 1083.

Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, Bates PD, Baud S, Bird D, DeBono A, Durrett TP, Franke RB, Graham IA, Katayama K, Kelly AA, Larson T, Markham JE, Miquel M, Molina I, Nishida I, Rowland O, Samuels L, Schmid KM, Wada H, Welti R, Xu C, Zallot R and Ohlrogge J (2013) Acyl-lipid metabolism. **The Arabidopsis Book 11**: e0133.

Liu X, Huang M, Fan B, Buckler ES and Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. **PLoS Genetics 12**: e1005957.

Lolle SJ, Berlyn GP, Engstrom EM, Krolikowski KA, Reiter WD and Pruitt RE (1997) Developmental regulation of cell interactions in the Arabidopsis fiddlehead-1mutant: A role for the epidermal cell wall and cuticle. **Developmental Biology 189**: 311-321.

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong GY and Myles S (2015) LinkImpute: fast and accurate genotype imputation for nonmodel organisms. **G3: Genes, Genomes, Genetics 5**: 2383-2390.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics 38**: 904-909.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. **Genetics 155**: 945-959.

R Core Team (2023) **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Available at < https://www.r-project.org/>. Accessed on July, 20, 2023.

Salojärvi J, Rambani A, Yu Z, Guyot R, Strickler S, Lepelley M, Wang C, Rajaraman S, Rastas P, Zheng C, Muñoz DS, Meidanis J, Paschoal AR, Bawin Y, Krabbenhoft T, Wang ZQ, Fleck S, Aussel R, Bellanger L, Charpagne A, Fournier C, Kassam M, Lefebvre G, Métairon S, Moine D, Rigoreau M, Stolte J, Hamon P, Couturon E, Tranchant-Dubreuil C, Mukherjee M, Lan T, Engelhardt J, Stadler P, De Lemos SMC, Suzuki SI, Sumirat U, Man WC, Dauchot N, Orozco-Arias S, Garavito A, Kiwuka C, Musoli P, Nalukenge A, Guichoux E, Reinout H, Smit M, Carretero-Paulet L, Filho OG, Braghini MT, Padilha L, Sera GH, Ruttink T, Henry R, Marraccini P, Van de Peer Y, Andrade A, Domingues D, Giuliano G, Mueller L, Pereira LF, Plaisance S, Poncet V, Rombauts S, Sankoff D, Albert VA, Crouzillat D, de Kochko A and Descombes P (2023) The genome and population genomics of allopolyploid *Coffea arabica* reveal the diversification history of modern coffee cultivars. **BioRxiv** 2023.09.06.556570.

Sant'Ana GC, Pereira LFP, Pot D, Ivamoto ST, Domingues DS, Ferreira RV, Pagiatto NF, Silva BSR, Nogueira LM, Kitzberger CSG, Scholz MBS, Oliveira FF, Sera GH, Padilha L, Labouisse JP, Guyot R, Charmetant P and Leroy T (2018) Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica* L. **Scientific Reports 8**: 465.

Scholz MB, Kitzberger CS, Pagiatto NF, Pereira LF, Davrieux F, Pot D, Charmetant P and Leroy T (2016) Chemical composition in wild Ethiopian Arabica coffee accessions. **Euphytica 209**: 429.

Su J, Ma Q, Li M, Hao F and Wang C (2018) Multi-Locus Genome-Wide Association studies of fiber-quality related traits in Chinese early-maturity upland cotton. **Frontiers in Plant Science 9**: 1169.

Su J, Pang C, Wei H, Li L, Liang B, Wang C, Song M, Wang H, Zhao S, Jia X, Mao G, Huang L, Geng D, Wang C, Fan S and Yu S (2016) Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. **BMC Genomics 17**: 687.

Tamba CL, Ni YL and Zhang YM (2017) Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. **PLoS Computational Biology 13**: e1005357.

Tamba CL and Zhang YM (2018) A fast mrMLM algorithm for multi-locus genome-wide association studies. Available at <https://api.semanticscholar.org/CorpusID:90293082>. Accessed on April 15, 2023.

Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, Zhang J, Dunwell

JM, Xu S and Zhang YM (2016) Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. **Scientific Reports 6**: 19444.

Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY, Wang SB, Dunwell JM, Zhang YM and Wu R (2018) Methodological implementation of mixed linear models in multi-locus genome-wide association studies. **Briefings in Bioinformatics 19**: 700-712.

Xu Y, Yang T, Zhou Y, Yin S, Li P, Liu J, Xu S, Yang Z and Xu C (2018) Genome-Wide Association mapping of starch pasting properties in maize using single-locus and multi-locus models. **Frontiers in Plant Science 9**: 1311.

Yang X, Gao S, Xu S, Zhang Z, Prasanna BM, Li L, Li J and Yan J (2011) Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. **Molecular Breeding 28**: 511-526.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics 38**: 203-208.