UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

HERISON VICTOR LIMA MUNIZ

ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA COM MODELOS SINGLE-LOCUS E MULTI-LOCUS PARA CARACTERÍSTICA DE LIPÍDIOS EM *Coffea arabica*

CORNÉLIO PROCÓPIO

2023

HERISON VICTOR LIMA MUNIZ

ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA COM MODELOS SINGLE-LOCUS E MULTI-LOCUS PARA CARACTERÍSTICA DE LIPÍDIOS EM *Coffea arabica*

Genome-wide association studies with single-locus and multi-locus models for lipid trait in *Coffea arabica*

Trabalho de conclusão de curso de Dissertação apresentada como requisito para obtenção do título de Mestre em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).

Orientador(a): Luiz Filipe Protasio Pereira
Coorientador(a): Alexandre Rossi Paschoal

CORNÉLIO PROCÓPIO

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

HERISON VICTOR LIMA MUNIZ

ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA COM MODELOS SINGLE-LOCUS E MULTI-LOCUS PARA CARACTERÍSTICA DE LIPÍDIOS EM COFFEA ARABICA

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 31 de Julho de 2023

Dr. Luiz Filipe Protasio Pereira, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Caroline Ariyoshi, Doutorado - Instituto de Desenvolvimento Rural do Paraná

(Idr-Paraná)Dr. Laurival Antonio Vilas Boas, Doutorado - Universidade Tecnológica

Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 31/07/2023.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por me abençoar em mais uma etapa incrível na minha vida.

Agradeço minha família que sempre foram minha base, principalmente meus pais Jose Raimundo Sousa Muniz e Josiana Santos Lima pela criação, educação, incentivo, e todo o apoio durante a minha vida.

Ao meu orientador Dr. Luiz Filipe Protasio Pereira por me receber como seu aluno, pela confiança, pelo incentivo ao meu desenvolvimento, e pelo apoio diário para a realização dessa dissertação.

Ao meu coorientador Dr. Alexandre Rossi Paschoal pelo apoio, incentivo, e por tornar minha adaptação melhor e mais rápida ao mestrado e na cidade de Londrina, Paraná.

Aos meus amigos e demais pessoas que, de alguma forma, foram essenciais para este momento importante, em especial a Dra. Caroline Ariyoshi, e a Dra. Rafaelle Vecchia Ferreira por contribuírem diretamente com a conclusão dessa dissertação.

Ao Instituto de Desenvolvimento Rural do Paraná (IDR-Paraná) por abrir as portas durante todo o processo de desenvolvimento no mestrado, onde pude atuar como colaborador na mesma.

Ao Consórcio Pesquisa Café pela Bolsa de Estudos durante a realização dessa dissertação.

À EMBRAPA Café pelos recursos que recebi durante a realização dessa dissertação, principalmente na utilização do servidor para trabalhar com os dados de Bioinformática.

À Universidade Tecnológica Federal do Paraná (UTFPR-CP) e todos os profissionais envolvidos na minha formação.

RESUMO

Lipídios são compostos que possuem um papel importante no desenvolvimento dos grãos de café, contribuindo para uma melhor qualidade da bebida. *Coffea arabica* é uma planta alotetraploide (4n = 2× = 44) que se originou da hibridização natural de duas espécies diploides, *Coffea canephora* e *Coffea eugenioides*. Devido as suas características de aroma e sabor, *C. arabica* é a espécie mais cultivada mundialmente, assim como no Brasil. Os estudos de associação genômica ampla (GWAS) são abordagens que permitem identificar a ligação entre uma variante genética e uma característica fenotípica de interesse. Com o objetivo de identificar nucleotídeos de traços quantitativos (QTNs) e regiões genômicas associadas ao metabolismo dos lipídios, foram realizados GWAS com dados de genotipagem de acessos de café alinhados ao genoma do *C. arabica* Et039. Foram utilizados dados de genotipagem por sequenciamento (GBS) e fenotipagem para o conteúdo lipídico total de um painel de 104 acessos de *C. arabica* provenientes da coleção FAO, originária da Etiópia (centro de origem da espécie), e as cultivares Mundo Novo 38 e *C. arabica* var. Typica. Para a GWAS, foram utilizados os métodos single-locus GLM e MLM, e os métodos multi-locus mrMLM, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, e FarmCPU. Como covariantes, para o ajuste dos métodos, foram utilizados os modelos matriz K, e matriz Q derivadas de PCA e de coeficiente de agrupamentos. As associações com os métodos de GWAS identificaram 13 QTNs com a correção com PCA + K, e 6 QTNs com a correção Q + K, sendo 5 QTNs convergentes com os diferentes modelos de correção. A correção da estrutura de população utilizando Q + K demonstrou um melhor ajuste nos métodos de GWAS. Os métodos multi-locus mrMLM e FarmCPU identificaram um maior número de QTNs associadas à lipídios. Dos QTNs identificados, 4 QTNs estão próximos a sete genes potencialmente envolvidos com o metabolismo dos lipídios. A frequência dos QTNs identificados nesse trabalho foi maior em acessos de *C.arabica* com maior conteúdo de lipídios, demonstrando o potencial para o desenvolvimento de marcadores visando auxiliar o melhoramento de cafeeiros.

Palavras-chave: Alotetraploide; Bioinfomática; Genótipos silvestres; Marcadores SNPs.

ABSTRACT

Lipids are compounds that play an important role in coffee bean development, contributing to the overall quality of the beverage. Coffea arabica is an allotetraploid plant (4n = 2× = 44) that originated from the natural hybridization of two diploid species, *Coffea canephora* and *Coffea eugenioides*. Due to its aromatic and flavorful characteristics, *C. arabica* is the most widely cultivated species globally, including in Brazil. Genome-wide association studies (GWAS) are approaches that enable the identification of links between genetic variants and phenotypic traits of interest. In order to identify quantitative trait nucleotides (QTNs) and genomic regions associated with lipid metabolism, GWAS were conducted using genotyping data from coffee accessions aligned to the *C. arabica* genome Et039. Genotyping by sequencing (GBS) data and phenotyping for total lipid content were used for a panel of 104 *C. arabica* accessions from the FAO collection, originating from Ethiopia (the species' center of origin), as well as the Mundo Novo 38 and *C. arabica* var. Typica cultivars. For GWAS, single-locus GLM and MLM methods, as well as multi-locus methods such as mrMLM, FASTmrMLM, FASTmrEMMA, ISIS EM-BLASSO, and FarmCPU were employed. Covariates for method adjustment included K matrix models and Q matrix models derived from PCA and clustering coefficients. Associations using GWAS methods identified 13 QTNs with PCA + K correction and 6 QTNs with Q + K correction, with 5 QTNs being consistent across different correction models. Correction of population structure using Q + K demonstrated better fitting in GWAS methods. Multi-locus methods such as mrMLM and FarmCPU identified a higher number of QTNs associated with lipids. Among the identified QTNs, 4 were located near seven genes potentially involved in lipid metabolism. The frequency of the identified QTNs was higher in *C. arabica* accessions with higher lipid content, demonstrating the potential for marker development in order to assist coffee plant breeding.

Keywords: Allotetraploid; Bioinformatics; SNP markers; Wild genotypes.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LD	Linkage Disequilibrium
EP	Estrutura de população
FarmCPU	Fixed and Random Model Circulating Probability Unification
FASTmrEMMA	Fast multi-locus random-SNP-effect EMMA
FASTmrMLM	A fast mrMLM multilocus mixed linear model
GBS	Genotyping by Sequencing
GISH	Genomic in situ hybridization
GLM	General Linear Model
GWAS	Genome-wide Association Studies
GWAS-ML	GWAS multi-locus
GWAS-SL	GWAS single-locus
IDR-Paraná	Instituto de Desenvolvimento Rural do Paraná
ISIS EM-BLASSO	Iterative modified-Sure Independence Screening EM-Bayesian LASSO
MAF	Minor Allele Frequency
ML	Multi-locus
MLM	Mixed Linear Model
MLMM	Multiple Loci Linear Mixed Model
mrMLM	Multi-locus random-SNP-effect mixed linear model
NGS	New Generation Sequencing
NGS	Next Generation Sequencing
PCA	Principal Component Analysis
QTN	Quantitative Trait Nucleotide
RFLP	Restriction fragment length polymorphism
rMLM	Random-SNP-effect MLM
SL	Single-locus
SNP	Single Nucleotide Polymorphism

SUMÁRIO

1 INTRODUÇÃO

O café é uma das bebidas mais populares e consumidas no mundo. O Brasil destaca-se por ser o maior produtor da *commodities*, a qual contribui diretamente na economia e na geração de empregos ao longo da sua produtividade no ano de 2020. As espécies *Coffea arabica* e *Coffea canephora* são as mais cultivadas (CONAB, 2023). *C. arabica* é a espécie mais importante, com participação de 69,3% da produção do café no Brasil (CONAB, 2023).

O gênero *Coffea* apresenta mais de 140 espécies identificadas (DAVIS *et al*., 2011; GUYOT *et al*., 2020). Todas as espécies de café são diploides (2n = 2× = 22), e em sua maioria auto-incompatíveis, com exceção de *C. arabica* que possui que possui um genoma tetraploide (4n = 2× = 44), sendo ainda uma planta autógama. Um estudo de cariograma permitiu a caracterização do genoma de *C. arabica*, evidenciando que essa espécie não é um alopolipoide segmentar, mas sim um alotretaploide verdadeiro (CLARINDO *et al*., 2008). Através da utilização da técnica *Restriction fragment length polymorphism* (RFLP) em combinação com *Genomic in situ hybridization* (GISH), foi possível identificar a origem do *C. arabica*, advindo da hibridização entre *C. canephora* (genitor paterno) e *C. eugenioides* (genitor materno) (LASHERMES *et al*., 1999)

Polimorfismos de nucleotídeo único (SNP - *single nucleotide polymorphism*), são os mais encontrados no genoma, sendo bialélicos e codominantes. Avanços na tecnologia de NGS (*Next Generation Sequencing*) reduziram significantemente o tempo e os recursos para identificação de SNPs. Esse grande número de polimorfismos vêm sendo utilizados cada vez mais em análises de seleção genômica, assim como estudos de associação genômica (GWAS - *Genome wide association studies*) (LIAO; LEE, 2018).

O GWAS é uma importante ferramenta que vem sendo utilizada para uma melhor compreensão de características complexas, com o objetivo de identificar loci/genes relacionados com uma determinada característica fenotípica. A partir de informações genômicas, de uma população composta por indivíduos não aparentados, permite a identificação de um grande número de marcadores moleculares. O desequilíbrio de ligação (DL) se baseia na recombinação e na proximidade entre os marcadores SNPs, assim, quanto mais próximos, maior é a probabilidade de estarem em DL. A partir dessa premissa, é possível fazer um mapeamento fino de SNPs e genes associados a variações fenotípicas complexas (HUANG; HAN, 2014; BARTOLI; ROUX, 2017; WEN *et al*., 2018).

De acordo com Wen *et al.* (2018) em seu estudo, a utilização de diferentes modelos de GWAS demonstraram melhores resultados, tendo uma melhor otimização de tempo, e redução

de viés nas análises. Os métodos que tratam o efeito do SNP como aleatórios e abordagens multi-locus melhoraram a eficiência, e houve um melhor ajuste para um modelo genético em comparação a outras metodologia de GWAS.

Sant'Ana e colaboradores (2018) publicaram o primeiro trabalho de GWAS em *C. arabica*, onde foi possivel identificar SNPs associados à caracteristicas bioquímicas do grão, como lipídios, e os diterpenos caveol e cafestol. No entanto, nesse trabalho foi utilizado somente o genoma de referência de um dos ancestrais diploides, *C. canephora*, limitando o número de SNPs identificados utilizados na associação.

Neste trabalho, com os mesmos dados do trabalho anterior, foram refeitos o GWAS, mas utilizando como referência dados de genotipagem mapeados no genoma de *C. arabica*, onde foram utilizados 11.136 marcadores, um número bem maior do que o trabalho anterior. Ainda, foram adicionamos os métodos single-locus GLM e MLM, e o método multi-locus FarmCPU para comparação com os metódos previamente utilizados no laboratório de Biotecnologia Vegetal (Instituto de Desenvolvimento Rural do Paraná, Londrina).

Diante do exposto, o presente estudo tem como objetivo a identificação de novas regiões genômicas para conteúdo de lipídios, a comparação de diferentes modelos para GWAS , e a identificação de qual(is) método(os) apresentam os melhores resultados na identificação de SNPs e regiões genômicas associadas às características de interesse agrônomico na espécie alotetraploide *C. arabica*.

2 OBJETIVOS

2.1 Objetivos geral

Realizar GWAS para conteúdo lipídico utilizando diferentes modelos de associação.

2.2 Objetivos específicos

Realizar estudos de associação com dados de GBS alinhados ao genoma de *C. arabica* para conteúdo de lipídios.

Realizar estudos de associação com modelos single-locus e multi-locus.

Avaliar quais métodos de associação apresentam melhor ajuste para o conteúdo de lipídios em *C. arabica*.

Identificar novas regiões genômicas ligadas ao metabolismo de lipídios em *C. arabica*.

3 REVISÃO BIBLIOGRÁFICA

3.1 Características botânicas e Aspectos econômicos de *Coffea arabica*

O gênero *Coffea* pertence a família *Rubiaceae*, é representado por 124 espécies, destacando-se *C. arabica* e *C. canephora* (DAVIS *et al*., 2011). *C. arabica* é uma espécie alotetraploide (4n = 2× = 44), e a única no gênero *Coffea* que possui essa característica em seu genoma. A origem do *C. arabica* ocorreu através da hibridização de duas espécies diploides, *C. canephora* e *C. eugenioides* (LASHERMES *et al*., 1999).

C. arabica tem sua origem nas regiões do sudoeste da Etiópia (ANTHONY, *et al*., 2002), mas sua dispersão ocorreu após a introdução de plantas no lêmen. O cultivo do café no Iêmen ocorre pelo menos há cinco séculos, e se espalhou pelo Sudeste Asiático por volta do ano de 1700. Duas bases genéticas provenientes do lêmen, que foram descritas como duas variedades distintas: *C. arabica* var. Typica e *C. arabica* var. Bourbon são responsáveis pela origem da maioria das cultivares comerciais de arábica pelo mundo (ANTHONY *et al*., 2002).

No Brasil, a cultura do café representa uma grande capacidade produtiva, tornado o país o maior produtor e exportador do mundo. O Brasil exportou 11,2 milhões milhões de sacas de 60 quilos de café nos quatro primeiros meses de 2023, e o pico de colheita para o ano de 2023 são nos meses junho e julho, onde estima-se a colheita de 31,4 milhões de sacas de café no bimestre (CONAB, 2023).

3.2 Polimorfismo de Nucleotídeo Único (SNP) e Genotipagem por Sequenciamento (GBS)

Os SNPs são alterações que ocorrem em um ponto específico no genoma. Essas alterações podem estar presentes em qualquer lugar no genoma, em regiões intergênicas, assim como em regiões intragênicas. A utilização desses marcadores moleculares pode aumentar o poder de identificação genótipos superiores, reduzindo tempo e custo em programas de melhoramento (KUMAR *et al*., 2012; ARIYOSHI *et al*., 2022). Nos genomas de plantas, assim como de outras espécies, os SNPs possuem uma ocorência de alta densidade (MANIVANNAN *et al*., 2018).

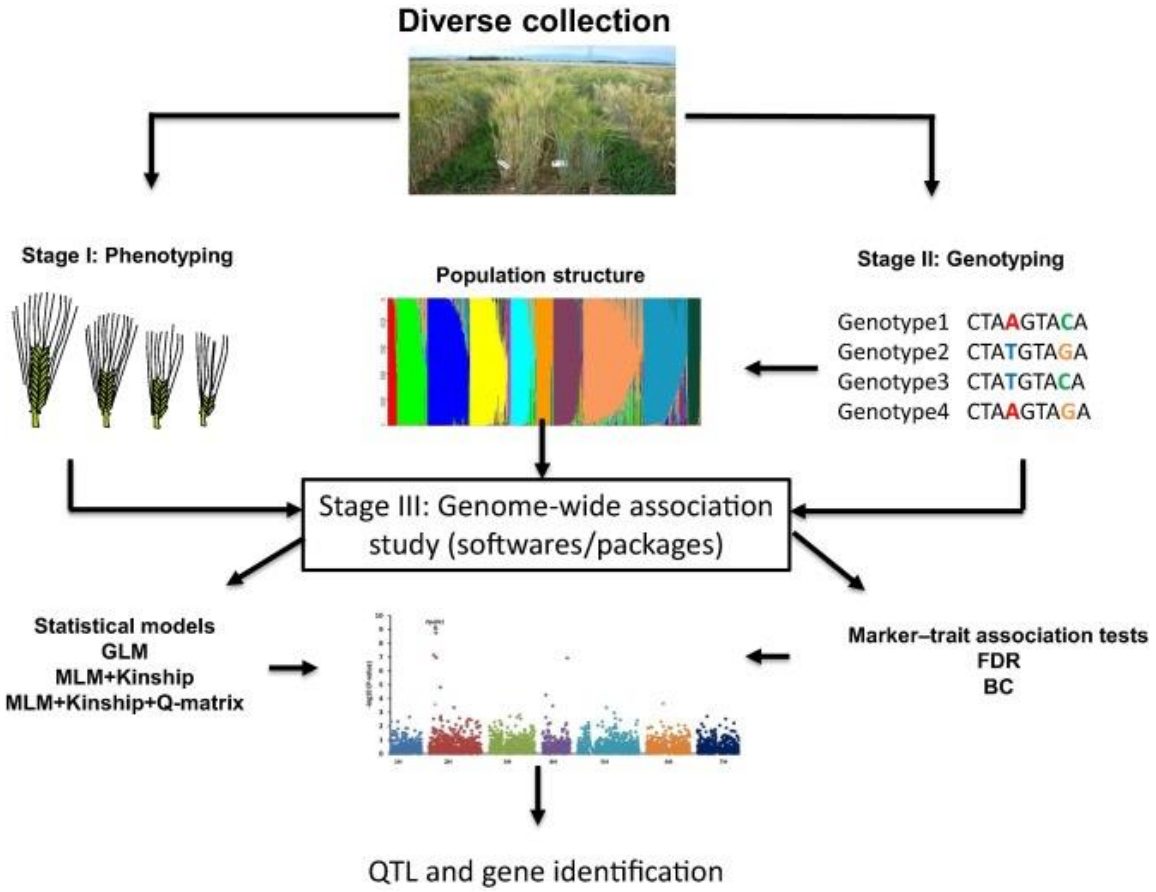
Além disso, o avanço nas técnicas de sequenciamento de nova geração (NGS) contribuiu significantemente na redução de custos do sequenciamento de DNA, seja completo ou reduzida, como por exemplo, a genotipagem por sequenciamento (GBS). GBS consiste em um sistema para criação e sequencimento de bibliotecas representativas de forma reduzida, capaz

de gerar uma quantidade grande de SNPs para análises de genotipagem (ELSHIRE *et al*., 2011; HE *et al*., 2014). Diferentes métodos computacionais com o GBS, permitiram a detecção e validação de maneira eficiente desses SNPs, possibilitando seu uso em estudos de associação genômica, para descoberta de polimorfismos, regiões genômicas, e genes relacionados a uma característica de interesse (MANIVANNAN *et al*., 2018).

3.3 Estudos de Associação Genômica Ampla (GWAS)

Estudos de Associação Genômica Ampla (GWAS – Genome-Wide Association Studies) permitem a identificação de loci/genes relacionados às características de interesse. A partir disso, pode-se mapear genes associados a variações fenotípicas complexas (HUANG; HAN, 2014; BARTOLI; ROUX, 2017; WEN *et al*., 2018). Uma das vantagens da GWAS em relação as técnicas tradicionais de mapeamento genético, é a exploração da diversidade genética existente, preferencialmente de genótipos não aparentados, eliminando a necessidade da produção de populações de cruzamento para o trabalho. Para realizar um experimento de GWAS com sucesso, há três estágios importantes (Figura 1).

Figura 1. Os três estágios mais importantes para realizar um experimento de GWAS.



Fonte: Alqudah *et al*. (2020)

O estágio I, envolve a fenotipagem de todos os genótipos para uma determinada característica ou conjunto de características, de acordo com os objetivos do estudo. A precisão da fenotipagem é extremamente importante para identificar associações entre genótipo e fenótipo. O estágio II, refere-se à genotipagem dos mesmos indivíduos previamente fenotipados por meio de marcadores moleculares de DNA. O método mais comumente empregado nesse estágio é o GBS, devido a sua capacidade de gerar um grande número de marcadores SNPs de baixo custo, que abrangem principalmente as regiões de não metiladas do genomas. Por fim, no estágio III, os dados fenotípicos e genotípicos são integrados utilizando um software adequado, como o TASSEL, por meio do qual é possível identificar os alelos associados a uma característica específica após a seleção do modelo de GWAS (ALQUDAH *et al*., 2020).

O GWAS leva em conta o DL, onde os loci que são fisicamente próximos são separados no cromossomo por recombinação com menor frequência em relação aos loci mais distantes uns dos outros, e essa associação não aleatória entre dois loci é conhecida como DL. Desta forma, o GWAS pode detectar essas associações através da proximidade dos SNPs aos loci causadores se houver o alto DL com os polimorfismos funcionais, assim, associados ao fenótipo de interesse (LANDER; SCHORK, 1994; LIPKA *et al*., 2015; VISSCHER *et al*., 2017; XIAO *et al*., 2017).

Os métodos de GWAS mais comuns eram os single-locus, a exemplo do *General Linear Model* (GLM) (PRICE *et al*., 2006) e o *Mixed Linear Model* (MLM) (YU *et al*., 2006). Esses modelos testam um SNP por vez, dessa forma, múltiplos testes necessitam da correção de Bonferroni. A correção de Bonferroni é frequentemente utilizada para controlar falsos positivos, e por ser altamente restritiva, pode resultar na exclusão de loci importantes relacionados às características de estudo (DING *et al*., 2019; ZHANG; JIA; DUNWELL, 2019).

Os modelos multi-locus apresentam vantagens em relação aos single-locus, como por exemplo, uma quantidade menor da taxa de falsos positivos, permitindo a utilização de um limiar de associação menos estringente que o Bonferroni (LI *et al*., 2018). Assim, é possível evitar que haja perda de loci importantes com o uso de um nível de significância menos estringente. Outra vantagem é maior capacidade de identificação de SNPs associados a QTLs de menor efeito (ZHANG; JIA; DUNWELL, 2019).

3.4 Modelos de Estudos de Associação Genômica Ampla (GWAS)

O modelo MLM é uma abordagem utilizada devido a sua eficácia na correção de diversos pequenos efeitos genéticos e controlar viés de estratificação populacional no modelo, com a utilização de análise de componentes principais (PCA) e matriz de parentesco. O método General Linear Mixed (GLM), diferente do MLM, utiliza apenas a PCA para corrigir a estrutura populacional (EP) (YU *et al*., 2005; PRICE *et al*., 2006; WEN, *et al*., 2016).

O modelo mrMLM usa o *random-SNP-effect MLM* (rMLM) e o multi-locus RMLM (MRMLM) para o GWAS. O rMLM trata o efeito-SNP como aleatório, mas permite uma modificação no cálculo de correção de Bonferroni para calcular o limiar p-valor para testes de significância (WANG *et al*., 2016). O mrMLM é um modelo multi-locus que inclui marcas selecionadas do método rMLM que utiliza uma menor estringência como critério de seleção (WANG *et al*., 2016). Devido a natureza multi- locus, não são necessárias correções para múltiplos testes. Os resultados das análises de dados reais e estudos de simulação mostraram que o mrMLM possui um maior poder de detecção QTN, um mínimo *bias* na estimação do efeito de um QTN, e uma maior robustez, comparado com o RMLM e o EMMA.

O FASTmrMLM é uma versão mais rápida e eficiente do mrMLM, que utiliza transformações matriciais (TAMBA; ZHANG, 2018). No método FASTmrEMMA, uma nova matriz de transformação é construída para se obter um novo modelo genético que inclui somente variação de QTN e erro residual normal; permitindo aos números de autovalores diferentes de zero serem um e corrigindo a variância da taxa poligênica-para-residual para aumentar a velocidade de computação. Os resultados das análises, tanto dos simulados quanto dos dados reais, mostraram que FASTmrEMMA é: mais vantajoso na detecção de QTN, possui um modelo mais robusto e ajustado, possui menos *bias* na estimação do efeito do QTN, e necessita de menor tempo de corrida do que as atuais metodologias single e multi-locus (E-BAYES, SUPER, EMMA, CMLM e ECMLM) (WEN *et al*., 2017).

O método ISIS-EM-BLASSO usa uma triagem interativa de segurança modificada independente (interative modified-sure independende screening – ISIS), focada em reduzir o número dos SNPs para uma quantidade moderada. Maximização de expectativa (*expection-maximization –EM*), mínimo encolhimento absoluto Bayesiano e operador de seleção (*Bayesian least absolute shrinkage and selection operator* – BLASSO) são usados para estimar todos os SNPs de efeito selecionados para a detecção de QTN (TAMBA; NI; ZANG, 2017).

Em estudo realizado por Ariyoshi (2021), a partir de um modelo MLM single-locus

e quatro modelos multi-locus: mrMLM, FASTmrMLM, FASTmrEMMA e ISIS EM-BLASSO, foram identificados 13 QTNs associados com a resposta de *C. arabica* a mancha aureolada do cafeeiro. O modelo MLM single-locus identificou quatro SNPs. Os modelos multi-locus conseguiram identificar como associados as mesmas regiões genômicas identificadas no modelo single-locus, e identificaram outras regiões de menor efeito para a característica.

O modelo FarmCPU (LIU *et al*., 2016), que unifica as vantagens do modelo linear misto e da regressão passo a passo (efeito fixo) usando-os iterativamente. O FarmCPU substitui o parentesco por um conjunto de marcadores associados aos genes causais, dessa maneira, o método consegue eliminar a confusão entre o modelo misto e genes subjacentes à característica de interesse. Os marcadores associados são ajustados como modelo fixo para testar um marcador por vez no genoma, e o conjunto desses marcadores associados são otimizados pelo método de máxima verossimilhança.

Tabela 1. Ferramentas de GWAS.

Nome	Abordagem	Autores	Ano
General Linear Model (GLM)	Single-locus	Price et al.	2006
Mixed Linear Model (MLM)	Single-locus	Yu et al.	2006
Multi-locus RMLM (mrMLM)	Multi-locus	Wang et al.	2016
Fixed and random model Circulating Probability Unification (FarmCPU)	Multi-locus	Liu et al.	2016
Fast multi-locus random-SNP-effect EMMA (FASTmrEMMA)	Multi-locus	Wen et al.	2017
Integrative sure independence screening EM-Bayesian LASSO (ISIS EM-BLASSO)	Multi-locus	Tamba; Ni; Zhang.	2017
A fast mrMLM algorithm for multi-locus genome-wide association studies (FASTmrMLM)	Multi-locus	Tamba; Zhang.	2018

Fonte: Autoria própria (2023)

4 MATERIAL E MÉTODOS

4.1 Material vegetal

Foram utilizados dados de GBS realizados por Sant'Ana *et al.* (2018), onde foram previamente alinhados e submetidos à chamada de SNPs pelo pipeline Tassel 5 GBS v2 (GLAUBITZ *et al*., 2014), no genoma de *C. arabica*, acesso Et039, fornecido pelo Arabica Coffee Genome Consortium, no qual 159.000 marcadores SNPs foram identificados em 159 acessos de café (FELICIO; RODRIGUES, 2018, dados não publicados). Desses acessos, foram utilizados 104 acessos silvestres da Etiópia (centro de origem da espécie), provenientes da coleção FAO, e as cultivares *C. arabica* var. Typica e Mundo Novo 38, os quais foram fenotipados para o conteúdo lipídico total (Sant'Ana *et al.* 2018), utilizando os métodos descritos na Associação de Químicos Analíticos Oficiais (AOAC), utilizando o éter de petróleo como solvente.

4.2 Filtragem dos SNPs

Os 159.000 SNPs do alinhamento foram filtrados utilizando o software TASSEL versão 5.2.89 descrito por Bradbury *et al.* (2007). A filtragem foi feita com os parâmetros de frequência do alelo menor (MAF > 0.05), e call rate > 0.8. Para a imputação dos dados foram utilizados os softwares Beagle v4.1 (BROWNING; BROWNING, 2016), e o LD-kNNi baseado no método K-vizinho mais próximo (MONEY *et al*., 2015).

4.3 Estrutura de população

Para estrutura de população foram utilizados os cinco primeiros componentes da PCA e kinship (K) calculados pelo TASSEL 5.2.53 para a GWAS single-locus (GWAS-SL), e para a GWAS multi-locus (GWAS-ML) os cinco primeiros componentes da PCA fornecida pelo TASSEL 5.2.53, matriz Q (Q) pelo software STRUCTURE v2.3.4 (PRITCHARD *et al*., 2000), e matriz K (K) fornecido pelo pacote mrMLM. Na estrutura genética estimada pelo Structure, as frequências alélicas de cada cluster K (2 a 10) foram estimadas. Foi utilizado 1.000 (burn-in period) e 1.000 iterações, esses parâmetros resultaram em 10 corridas por cada valor de K. Foi utilizado o critério de ΔK (EVANNO; REGNAUT; GOUDET, 2005) no software Structure Harvester (EARL; VONHOLDT, 2012) para estimar o nível mais alto da estrutura de população.

4.4 GWAS single-locus (GWAS-SL) e multi-locus (GWAS-SL)

A GWAS single-locus (GWAS-SL) foi realizada pelo TASSEL 5.2.53 com dois

métodos: GLM (PRICE *et al*., 2006) e MLM (YU *et al*., 2006), e a GWAS multi-locus (GWAS-ML) foram realizadas com cinco métodos: mrMLM (WANG *et al*., 2016), FASTmrMLM (TAMBA; ZHANG, 2018), FASTmrEMMA (WEN *et al*., 2017) e ISIS EM-BLASSO (TAMBA; NI; ZHANG, 2017) no pacote mrMLM (ZHANG *et al*., 2020), e o FarmCPU (LIU *et al*., 2016) no pacote GAPIT3 (WANG; ZHANG, 2021), ambos os pacotes no software R.

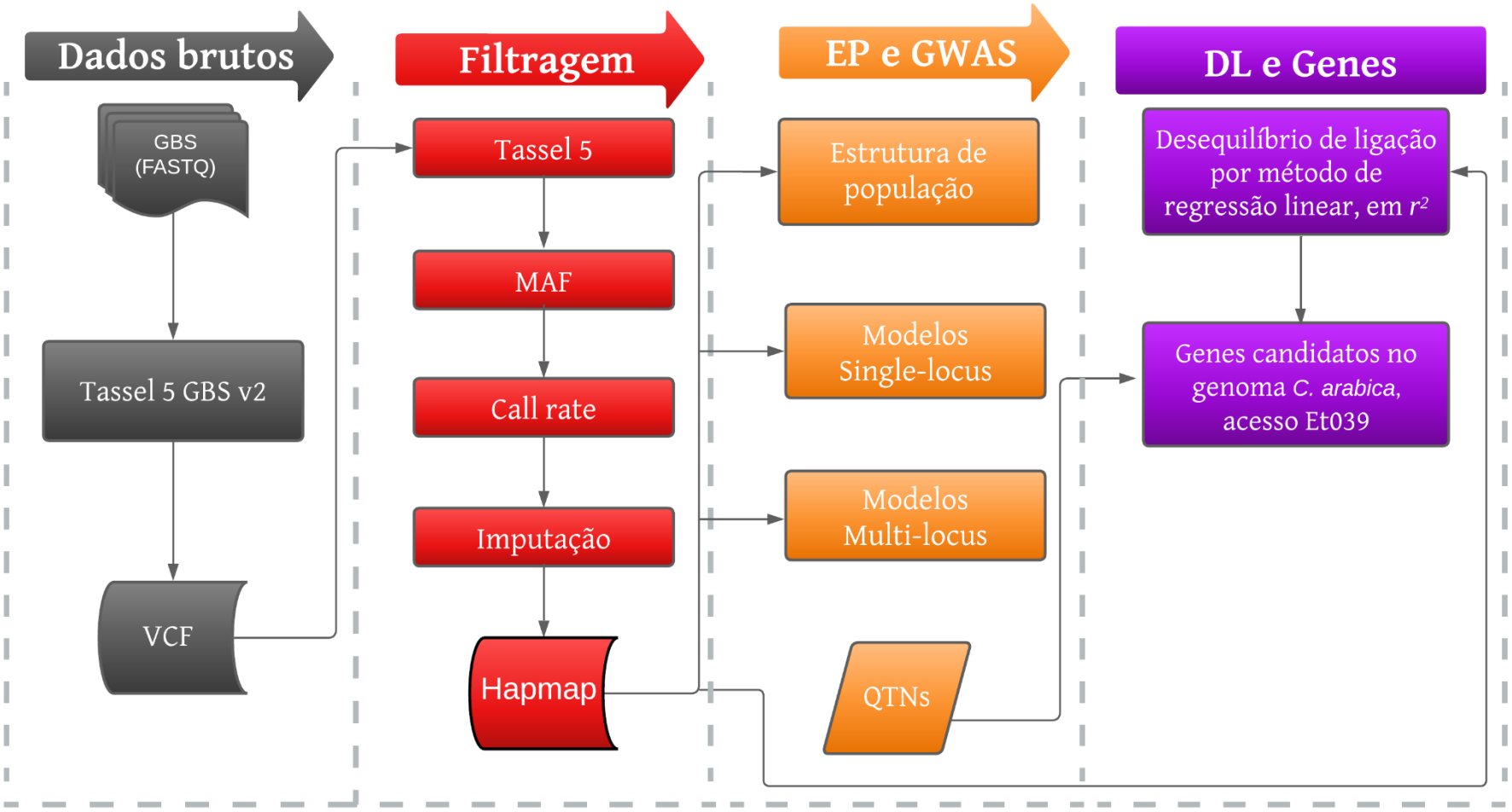
O limiar de associação para o GWAS-SL foi $p \leq 0.05/n$, onde n é o número de marcadores. Para a GWAS-ML no pacote mrMLM, na primeira etapa, foram utilizados os valores críticos $p \leq 0.01$, 0.01, 0.005, e 0.01 para os métodos mrMLM, FASTmrMLM, FASTmrEMMA, e ISIS EM-BLASSO para resultar um resultado intermediário. Na segunda etapa, todos os SNPs selecionados na primeira etapa foram filtrados pelos modelos multi-locus, e os marcadores com maiores efeitos que ultrapassaram o limiar LOD score foram considerados como SNPs potencialmente associados. O limite crítico da pontuação LOD foi definido como 3 para SNPs na fase final. Para o FarmCPU, foi utilizado o critério $p \leq 0.0005$.

Para comparação dos métodos foi levado em consideração o número de QTNs associadas.

4.5 Desequilíbro de ligação e busca dos genes associados aos QTNs

Os coeficientes de correlação ao quadrado (r^2) foram calculados usando janelas de 50 SNPs adjacentes, no TASSEL versão 5.2.53, usado para avaliar o decaimento de DL. O valor da distância do DL (em bp) foi avaliada por um método de regressão não linear, em $r^2 = 0.2$, no software R. Os genes em DL com os QTNs associados ao conteúdo de lipídios foram extraídos da anotação do genoma de *C. arabica* Et039.

Figura 2. Fluxograma da metodologia.



EP - Estrutura de população
DL - Desequilíbrio de ligação

Fonte: Autoria própria (2023)

5 RESULTADOS E DISCUSSÃO

5.1 Filtragem dos SNPs

Após os filtros de qualidade e imputação, a partir de 159.000 SNPs identificados na chamada de SNPs, um total de 11.136 SNPs foram utilizados para a estrutura de população e GWAS. Desse total, 5.032 foram identificados no subgenoma do *C. canephora*, 4.870 no subgenoma do *C. eugenioides*, e 1.234 no cromossomo zero. O cromossomo zero corresponde aos *scaffolds* que não têm uma posição definida durante a montagem do genoma.

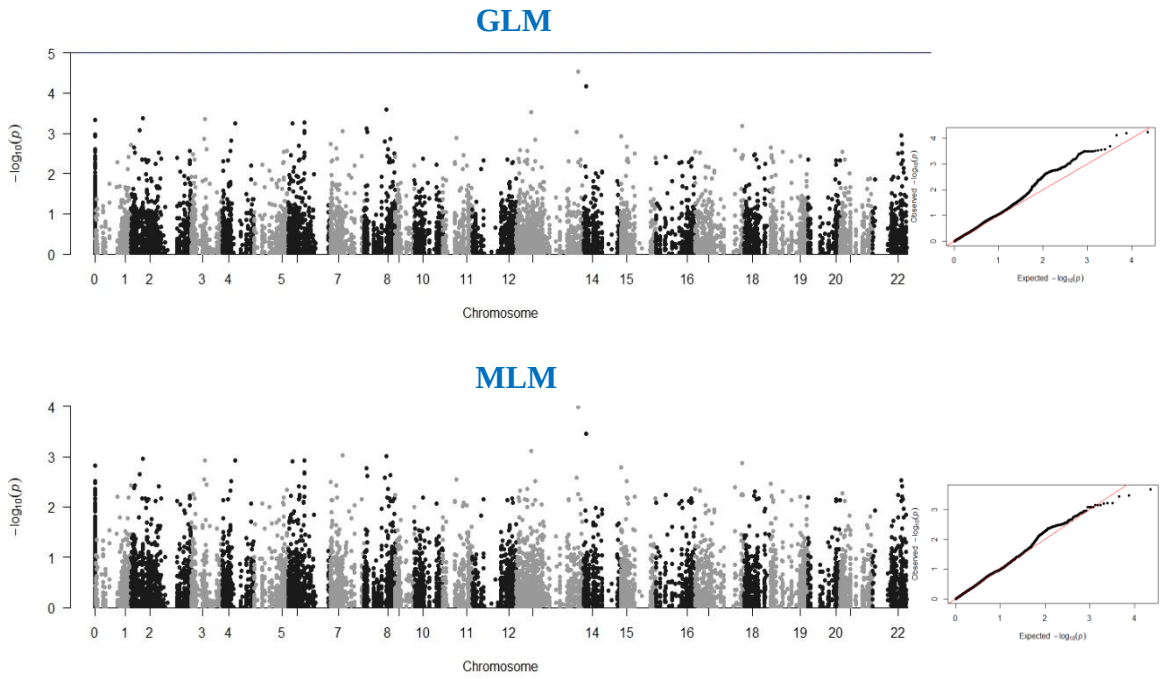
5.2 Estrutura de população

As análises de estrutura de população com o painel de acessos utilizando PCA (apêndice A), estão de acordo com o trabalho de Ariyoshi *et al.* (2022). O resultado do Structure (apêndice B) foi similiar aos trabalhos de Ariyoshi *et al.* (2022) e Sant'Ana *et al*. (2018), baseado em três grupos (K = 3), que foi o nível mais alto da população de acordo com o critério de Evanno (ΔK) (apêndice C).

5.3 GWAS nos métodos single-locus (SL)

Os métodos SL não foram capazes de detectar QTNs (Figura 3). Uma característica comum de GWAS por métodos SL é a varredura unidimensional do genoma, onde cada marcador é testado por vez. No entanto, essa abordagem não facilita boas estimativas dos efeitos dos marcadores controlados por múltiplos loci, o qual acontece na maioria dos traços complexos (WANG *et al*., 2016). Outro problema com o método, é a necessidade da correção de múltiplos testes para o valor limiar do teste de significância. Devido a correção de Bonferroni, que normalmente é muito conservadora, métodos como GLM e MLM não foram eficientes para detectar pequenos loci efetivos de um traço complexo (WANG *et al*., 2016). Por outro lado, utilizando o modelo PCA como correção de estrutura de população, os métodos SL obtiveram um ótimo ajuste (figura 3).

Figura 3. Gráficos de manhattan plots com os métodos single-locus MLM e GLM. A coluna X representa os cromossomos, e a coluna Y representa o -log10(p-value). Gráficos quantil-quantil (Q-Q plots) com os métodos single-locus GLM e MLM.



Fonte: Autoria própria (2023)

5.4 GWAS nos métodos multi-locus (ML)

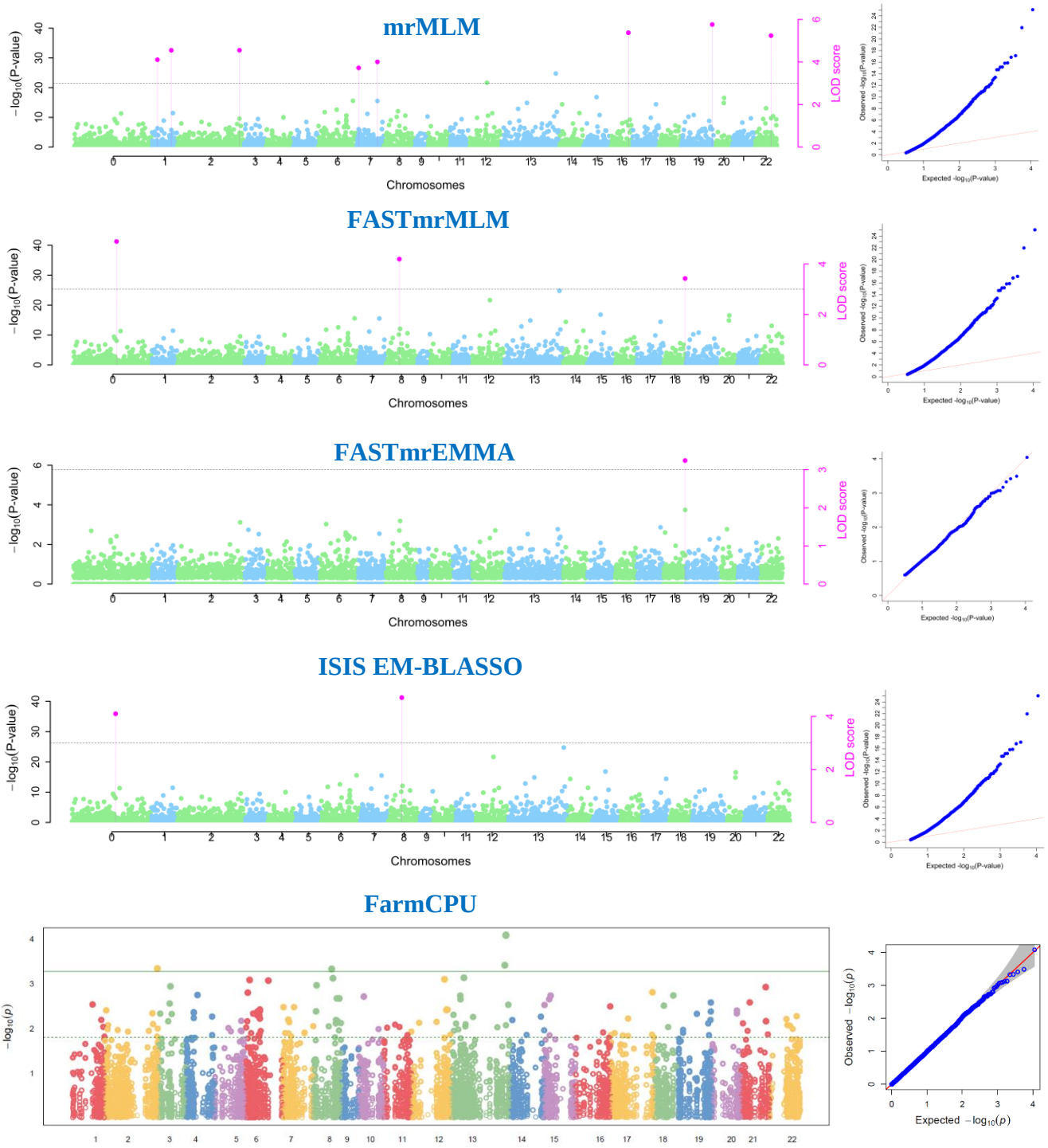
Neste estudo, apenas os métodos ML detectaram QTNs relacionadas com o conteúdo de lipídios. Devido as limitações dos métodos SL, algumas metodologias ML foram implementadas recentemente, como o mrMLM (WANG *et al*., 2016), FASTmrEMMA (WEN *et al*., 2017) , FASTmrMLM (TAMBA; ZHANG, 2018), e FarmCPU (LIU *et al*., 2016).

As vantagens dos métodos ML em relação aos SL já foram descritas em estudos com outras espécies de plantas, como o algodão, milho, tabaco, e também o café (LI *et al*., 2018; SU *et al*., 2018; XU *et al*., 2018; IKRAM *et al*., 2022; ARIYOSHI *et al*., 2022). Isso ocorre porque os métodos ML são caracterizados por uma abordagem multidimensional de varredura do genoma, em que os efeitos de todos os marcadores são estimados de forma simultânea (CUI *et al*., 2018).

Nas análises com correção da estrutura populacional PCA + K (figura 4), foram identificadas os QTNs. Desses QTNs, cada modelo identificou: (8) mrMLM, (3) FASTmrMLM, (1) FASTmrEMMA, (2) ISIS EM-BLASSO, e (5) FarmCPU (Tabela 2). Podemos observar que com exceção dos métodos FASTmrEMMA e FarmCPU, os gráficos de Q-Q Plot (Figura 4) apresentaram p-valores observados muito abaixo do esperado desde o início

da plotagem com os modelos PCA + K.

Figura 4. Gráficos de manhattan plots com os métodos multi-locus mrMLM, FASTmrMLM, FASTmrEMMA, ISIS-EM-BLASSO e FarmCPU com os modelos PCA + K. A coluna X representa os cromossomos, e a coluna Y representa o -log10(p-value). Gráficos de quantil-quantil (Q-Q plots) com os métodos multi-locus mrMLM, FASTmrMLM, FASTmrEMMA, ISIS-EM-BLASSO e FarmCPU com os modelos PCA + K.



Fonte: Autoria própria (2023)

Tabela 2. QTNs associadas nos métodos multi-locus mrMLM, FASTmrMLM, FASTmrEMMA ISIS-EM-BLASSO e FarmCPU ajustados com os modelos PCA + K.

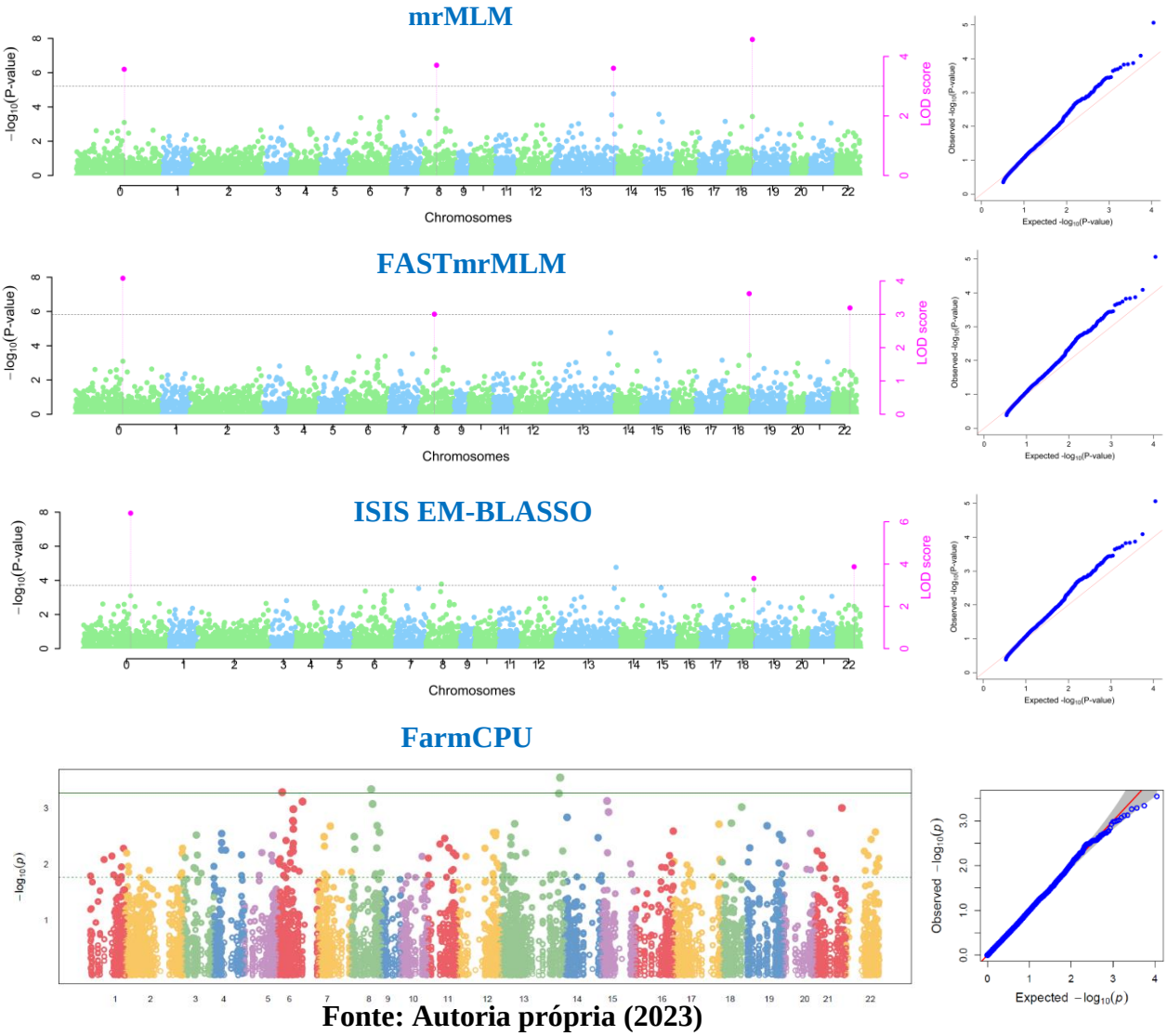
		mrMLM		FASTmrMLM		FASTmrEMMA		ISIS EM-BLASSO		FarmCPU
Posição no genoma de *C. arabica* Et039	**Chr[1]**	**-Log10 (p-valor)**	**LOD SCORE**	**-Log10 (p-valor)**	**LOD SCORE**	**-Log10 (p-valor)**	**LOD SCORE**	**-Log10 (p-valor)**	**LOD SCORE**	**-Log10 (p-valor)**
	Sca_4634HRSCAF=4635									3.488117
Chr_0_4634_168274				5.6818	4.8901			4.8566	4.1001	
Chr_1_sg_C_22364613	1 sg C	4.8608	4.1041							
Chr_1_sg_C_31802172	1 sg C	5.3237	4.5466							
Chr_2_sg_C_56859771	2 sg C	5.3267	4.5495							3.335358
Chr_7_sg_C_11717962	7 sg C	4.7552	4.0035							
Chr_7_sg_C_409622	7 sg C	4.4578	3.7205							
Chr_8_sg_C_19869722	8 sg C			4.9522	4.1914			5.4941	4.71	3.326979
	2 sg E									3.411168
Chr_2_sg_E_ 58841892										
Chr_2_sg_E_ 60056603	2 sg E									4.082494
Chr_5_sg_E_35714509	5 sg E	6.1815	5.3709							
Chr_7_sg_E_20849219	7 sg E			4.1472	3.4261	3.9487	3.2386			
Chr_8_sg_E_34967946	8 sg E	6.5827	5.758							
Chr_11_sg_E_29236376	11 sg E	6.0433	5.2378							

[1]Identificação do cromossomo no genoma de *C. arabica* Et039.
sg = subgenoma

Fonte: Autoria própria (2023)

Com o ajuste dos modelos multi-locus, utilizando dados de coeficiente de agrupamento obtidos pelo software STRUCTURE e matriz K (Q + K) foram identificados QTNs (Figura 5). Desses QTNs, cada modelo identificou: (4) mrMLM, (4) FASTmrMLM, (3) ISIS EM-BLASSO, e (3) FarmCPU (Tabela 3). Com essa correção da estrutura populacional, os métodos apresentaram um melhor ajuste, como podemos observar pelos gráficos de Q-Q plot (Figura 5), com exceção para o método FASTmrEMMA, no qual a correção por PCA apresentou um melhor ajuste dos dados de p-valor, mas não associou QTN com a correção pelos modelos Q + K.

Figura 5. Gráficos de manhattan plots com os métodos multi-locus mrMLM, FASTmrMLM, ISIS-EM-BLASSO e FarmCPU com os modelos Q + K . A coluna X representa os cromossomos, e a coluna Y representa o -log10(p-value). Gráficos de quantil-quantil (Q-Q plots) com os métodos multi-locus mrMLM, FASTmrMLM, ISIS-EM-BLASSO e FarmCPU com os modelos Q + K.



Fonte: Autoria própria (2023)

Tabela 3. QTNs associadas nos métodos multi-locus mrMLM, FASTmrMLM, ISIS-EM-BLASSO e FarmCPU ajustados com os modelos Q + K.

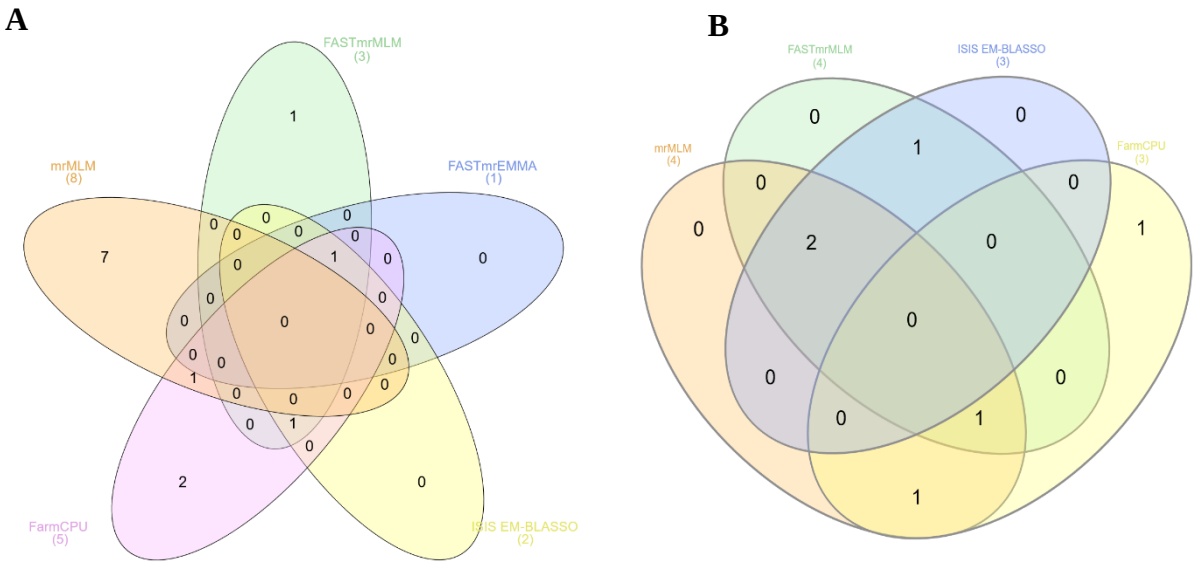
		mrMLM		FASTmrMLM		ISIS EM-BLASSO		FarmCPU
Posição no genoma de *C. arabica* Et039	**Chr[1]**	**-Log10 (p-value)**	**LOD SCORE**	**-Log10 (p-value)**	**LOD SCORE**	**-Log10 (p-value)**	**LOD SCORE**	**-Log10 (p-value)**
Chr_0_4634_168274	Sca_4634HRSCAF=4635	4.2953	3.5664	4.8379	4.0823	7.2502	6.4039	
Chr_6_sg_C_4567047	6 sg C							3.286509
Chr_8_sg_C_19869722	8 sg C	4.4391	3.7028	3.6986	3.0031			3.338187
Chr_2_sg_E_60056603	2 sg E	4.3333	3.6024					3.546682
Chr_7_sg_E_20849219	7 sg E	5.3459	4.5679	4.353	3.6211	4.0384	3.3233	
Chr_11_sg_E_29236360	11 sg E			3.8991	3.1918	4.6147	3.8697	

[1]Identificação do cromossomo no genoma de *C. arabica* Et039.
sg = subgenoma

Fonte: Autoria própria (2023)

Conforme o diagrama de Venn, os cinco métodos ML em conjunto associaram 13 QTNs utilizando PCA + K (Figura 6A). Quatro métodos de GWAS em conjunto associaram 6 QTNs utilizando Q + K (Figura 6B), enquanto o FASTmrEMMA não detectou associações para a característica estudada.

Figura 6. Diagramas dos QTNs identificadas pelos métodos multi-locus com diferentes correções de estrutura de população.



PCA + K **Q + K**

Fonte: Autoria própria (2023)

As associações com a GWAS (PCA + K) resultaram um total de 10 QTNs sem genes ligados ao metabolismo de lipídios e 3 QTNs com genes relacionados (Chr_7_sg_C_409622, Chr_8_sg_C_19869722, e Chr_5_sg_E_35714509) nos cinco métodos testados. A inclusão da PCA como correção da estrutura de população nos métodos de GWAS demonstraram ajustes desfavoráveis. No estudo desenvolvido por Elhaik (2022), foram analisadas doze casos de testes comuns em dados de população humana, em que o ajuste com PCA demonstrou resultados desfavoráveis em estudos de associações para a característica estudada.

O GWAS (Q + K) resultaram em 4 QTNs sem novas regiões genômicas associadas, e 2 QTNs com regiões genômicas associadas (Chr_6_sg_C_4567047 e Chr_8_sg_C_19869722). A inclusão da matriz Q nos métodos de GWAS diminuiu em aproximadamente três vezes a quantidade de QTNs sem associações com o conteúdo de lipídios. No trabalho de Yang *et al.* (2010), foram realizados estudos de características complexas em milho utilizando os modelos

PCA + K e Q + K como estrutura de população, onde Q + K apresentaram melhor redução dos falsos positivos.

Diferentes modelos de correção de estrutura de população identificaram 5 QTNs convergentes (Chr_0_4634_168274, Chr_8_sg_C_19869722, Chr_2_sg_E_60056603, Chr_7_sg_E_20849219, e Chr_11_sg_E_29236360).

5.5 Comparação dos diferentes métodos de GWAS

O mrMLM foi o método ML que mais identificou QTNs em ambos correções de estrutura da população, mas por outro lado, resultou em uma quantidade menor de QTNs sem genes ligados ao conteúdo de lipídios. Quando o número de QTNs putativos excede consideravelmente o tamanho da amostra, o modelo multi-locus nesse método pode apresentar superajuste (WANG *et al*., 2016).

FASTmrMLM identificou 1 QTN com gene ligado ao metabolismo dos lipídios nos dois modelos de EP, como também uma quantidade maior de QTNs sem genes ligados ao conteúdo lipídico. Nesse método, é implementado o *algoritmo Least Angle Regression* (LARS) entre a varredura single-locus e a estimação *EM-Emperical Bayes* na identificação de QTNs verdadeiras na segunda etapa (TAMBA; ZHANG, 2018).

FASTmrEMMA identificou apenas 1 QTN (sem genes ligados ao conteúdo de lipídios) com a correção PCA + K, por outro lado, o ajuste com Q + K não identificou QTN. O método mostrou-se bem ajustado quando utilizado PCA + K. Nesse método, é empregado um algoritmo de aproximação que utiliza uma transformação de matriz para branquear a matriz de covariância da matriz poligênica K e o ruído ambiental. Essa técnica acelera o processo computacional. Os efeitos de todos os SNPs selecionados na primeira etapa são colocados no modelo multi-locus, onde os efeitos dos SNPs são estimados por *expectation–maximization empirical Bayes*, e efeitos diferentes de zero foram posteriormente detectados pelo *likelihood ratio test* para identificação de QTN verdadeira (WEN *et al*., 2017).

ISIS EM-BLASSO identificou 1 QTN com um gene ligado ao metabolismo de lipídios nos modelos PCA + K, e QTNs sem genes ligados ao conteúdo de lipídios nos modelos Q + K. Na primeira etapa deste método, é utilizado o *iterative-modified sure independence screening* com o objetivo de diminuir o número de SNPs para um tamanho moderado, após isso, *Expectation-Maximization (EM)-Bayesian least absolute shrinkage and selection operator* é usado para estimar todos os efeitos dos SNPs para detecção de QTNs verdadeiras (TAMBA; NI; ZHANG, 2017).

FarmCPU identificou 1 QTN com gene ligado ao metabolismo de lipídios e QTNs sem

genes ligados ao conteúdo de lipídios quando utilizado os modelos PCA + K. O método mostrou-se eficiente nesse estudo, quando ajustado aos modelos Q + K, identificando a maior quantidade de QTNs com genes ligados ao metabolismo de lipídios, e menos QTNs sem genes ligados ao conteúdo de lipídios em relação aos outros métodos ML. Esse é método ML que divide o *Multiple Loci Linear Mixed Model* (MLMM) em um modelo de efeito fixo e um modelo de efeito aleatório, usando-os iterativamente (LIU *et al*., 2016).

Para fins de comparação, a Tabela 4 mostra a quantidade de QTNs identificadas com regiões genômicas anotadas para a conteúdo de lipídios, de acordo com a anotação do genoma de *C. arabica***,** pelos métodos de GWAS-SL e GWAS-ML, com os modelos de correção de estutura de população PCA + K e Q + K.

Tabela 4. Comparação da identificação de QTNs associadas a lipídios por métodos de GWAS.

Métodos de GWAS	Nº de QTNs identificadas	Regiões genômicas anotadas
GLM	0	0
MLM	0	0
mrMLM	3	5
FASTmrMLM	2	1
FASTmrEMMA	0	0
ISIS EM-BLASSO	1	1
FarmCPU	3	2

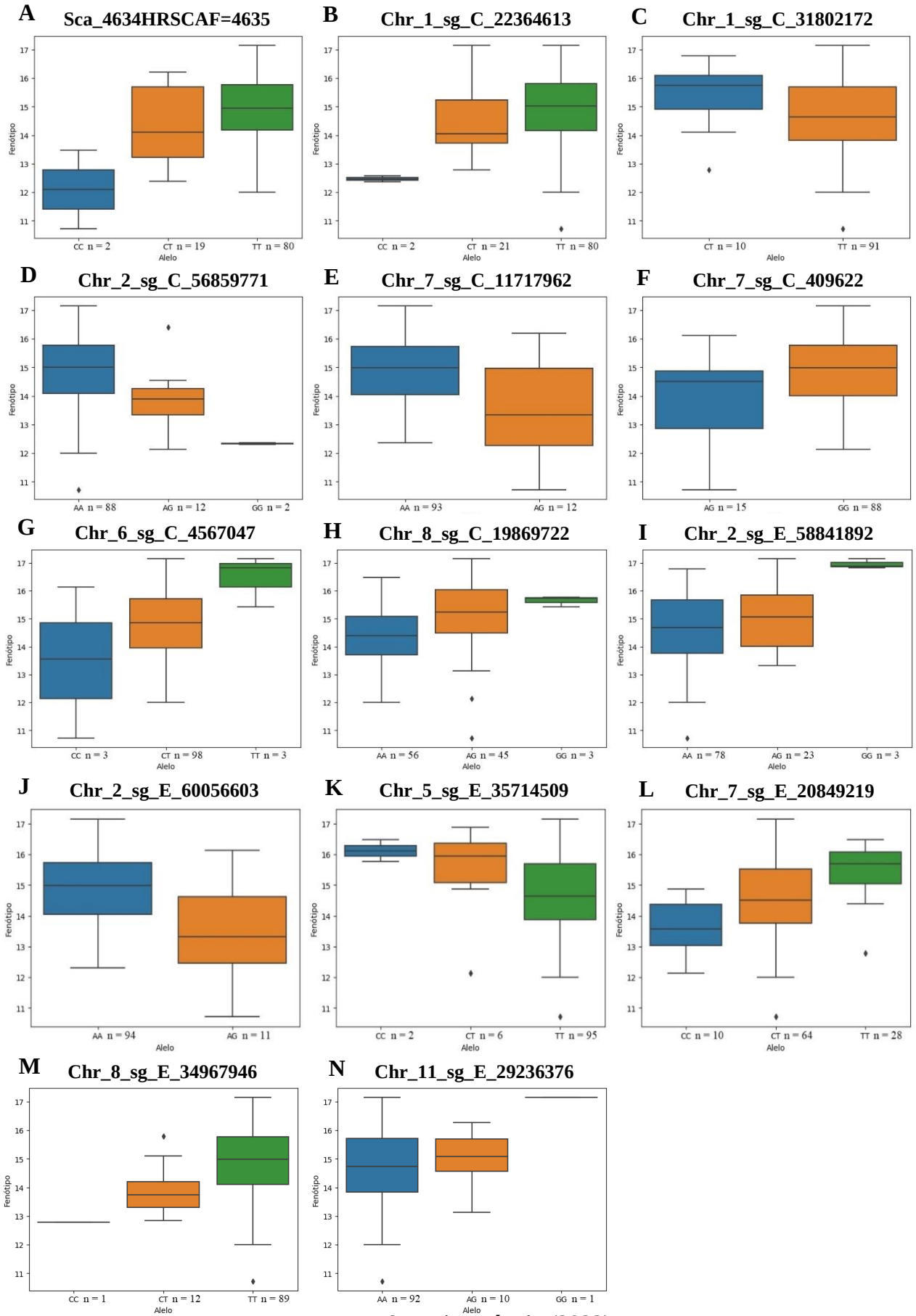
Fonte: Autoria própria (2023)

5.6 Distribuição fenotípica dos QTNs identificadas pelos métodos de GWAS

A partir dos dados de distribuição fenotípica e de todos os QTNs identificados, foi possível relacionar a concentração de lipídios presentes nos acessos de *C. arabica* de acordo com a forma alélica dos genótipos (Figura 7). Nos QTNs Sca_4634HRSCAF=4635 (Figura 7A), Chr_1_sg_C_22364613 (Figura 7B), Chr_2_sg_C_56859771 (Figura 7D), Chr_7_sg_C_11717962 (Figura 7E), Chr_7_sg_C_409622 (Figura 7F), Chr_6_sg_C_4567047 (Figura 7G), Chr_2_sg_E_58841892 (Figura 7I), Chr_2_sg_E_60056603 (Figura 7J), Chr_7_sg_E_20849219 (Figura 7L), Chr_8_sg_E_34967946 (Figura 7M), e Chr_11_sg_E_29236376 (Figura 7N), o maior teor de lipídios estão presentes em indivíduos com genótipos homozigotos para o alelo associado à característica de conteúdo de lipídios. Por

outro lado, os QTNs Chr_1_sg_C_31802172 (Figura 7C), Chr_8_sg_C_19869722 (Figura 7H), e Chr_5_sg_E_35714509 (Figura 7K), a maior concentração lipídica está presente em indivíduos com genótipos heterozigotos para o alelo associado à característica de conteúdo de lipídios.

Figura 7. Distribuição de dados fenotípicos de conteúdo de lipídios em acessos de *C. arabica*.



Fonte: Autoria própria (2023)

5.7 Identificação de genes candidatos

A partir da anotação funcional do *C. arabica* Et039, sete genes foram identificados próximos às QTNs associadas com a característica de lipídios (Tabela 5). Para considerar um gene ligado a uma QTN, foi utilizado a distância aproximada baseada no resultado de decaimento do DL (apêndice D). O decaimento do DL $r^2 = 0.2$ foi 158.774 pares de bases.

Em estudo prévio de Sant'Ana *et al.* (2018), dados de GBS foram utilizados para estudos de GWAS com objetivo de decifrar a base genética de lipídios. Devido a falta de um genoma de referência de *C. arabica*, os dados foram alinhados ao genoma de *C. canephora*. Com os dados mapeados, foram utilizados 2.587 SNPs para a estrutura de população e para o GWAS. A partir disso, o presente estudo utilizou os mesmos dados de sequenciamento, o qual foram alinhados no genoma de *C. arabica*, acesso Et039. A partir do mapeamento no genoma completo do *C. arabica*, 11.136 SNPs foram utilizados para a estrutura de população e para o GWAS, um número superior ao estudo anterior. O uso dos dados de GBS alinhados ao genoma de *C. arabica* aumentou a identificação de QTNs nos acessos de café. Foram identificados quatro QTNs próximos a genes envolvidos no metabolismo dos lipídios e /ou biossíntese de ácidos graxos, um número quatro vezes maior do que o observado no estudo de Sant'Ana *et al.* (2018).

A QTN Chr_6_sg_C_4567047 está próximo ao gene *g15.102*, com sua anotação funcional para 4 -phosphopantetheinyl transferase isoform X1. Para o Chr_5_sg_E_35714509, o gene próximo é *g119.153*, com anotação funcional para electron transfer flavo subunit mitochondrial. Para essas duas QTNs, não foram encontradas descrições na literatura para uma melhor compreensão de suas funções.

A QTN Chr_7_sg_C_409622 está próximo ao gene *g10.29*, com anotação funcional para a proteína triacylglycerol lipase-like 1. Essa proteína está envolvida com o metabolismo acil-lipídio em *Arabidopsis thaliana*, assim como em outras espécies de plantas. O acil-lipídio possui diversas funções, incluindo o fornecimento da barreira de difusão central da membrana que separam as células e as organelas subcelulares. Essa função sozinha engloba mais de 10 classes de lipídios de membrana, como fosfolipídios, galactolipídeos e esfingolipídios. (LI-BEISSON *et al.*, 2013). Próximos a esse QTN, na anotação do *C. arabica* Et039, também foram identificados os genes *g10.11* e *g10.8*, com participação no metabolismo dos lipídios, com anotação funcional para patatin 3 e patatin 1, respectivamente.

A QTN Chr_8_sg_C_19869722 está próximo ao gene *g66.10*, o qual possui anotação funcional para 3-ketoacyl- synthase 10. Essa proteína contribui para a biossíntese de cera

cuticular e suberina (LOLLE, *et al*., 1997). Como precursores de compostos de cera, os ácidos graxos de cadeia muito longa participam da limitação da perda de água não estomática e da prevenção de ataques de patógenos. Eles também servem como armazenamento de energia em sementes e como blocos de construção de membranas. 21 genes 3-ketoacyl-CoA synthase foram identificados no genoma de *A. thaliana* expressos em sementes, flores, e folhas (JOUBÈS *et al*., 2008).

Tabela 5. Anotação funcional dos genes próximos às QTNs associadas a lipídios.

Gene[1]	QTN	Posição do gene[2]	Anotação funcional[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]Nome dos genes recuperados da anotação funcional do genoma de *C. arabica* Et039.
[2]Posição (+) upstream ou (-) downstream em pares de bases.
[3]Anotação funcional dos genes recuperados da anotação funcional do genoma de *C. arabica* Et039.

Fonte: Autoria própria (2023)

6 CONCLUSÃO

O estudo com os métodos de associação genômica ampla utilizando dados de GBS alinhados ao genoma de *C. arabica* possibilitou identificar uma quantidade maior de SNPs em relação ao trabalho publicado anteriormente. As análises com diferentes modelos para a correção de estrutura de população mostraram que os métodos de GWAS podem se ajustar melhor, dependendo dos dados genotípicos e fenotípicos, e dos modelos aplicados para correção. Nesse estudo os modelos Q + K se ajustaram melhor, reduzindo a quantidade de SNPs sem regiões genômicas associadas, de acordo com a anotação funcional do *C. arabica*. Os métodos multi-locus mrMLM e FarmCPU demonstraram melhor desempenho na busca por QTNs associados com a característica de lipídios na anotação do genoma de *C. arabica*. Quatro QTNs próximos a sete genes potencialmente envolvidos com o metabolismo de lipídios foram encontrados.

REFERÊNCIAS

ALQUDAH, A. M. et al. GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley–A review. **Journal of advanced research**, v. 22, p. 119-135, 2020.

ANTHONY, F et al. The origin of cultivated Coffea arabica L. varieties revealed by AFLP and SSR markers. **Theoretical and Applied Genetics**, v. 104, n. 5, p. 894-900, 2002.

ARIYOSHI, C. **Associação Genômica Ampla e Seleção Assistida para resistência a *Pseudomonas syringae* pv. *garcae* em *Coffea arabica*.** 2021, 109 p. Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina, 2021.

ARIYOSHI, C. et al. Genome-wide association study for resistance to Pseudomonas syringae pv. garcae in Coffea arabica. **Frontiers in Plant Science**, v. 13, 18 out. 2022.

ARIYOSHI, C. et al. Development and Validation of an Allele-Specific Marker for Resistance to Bacterial Halo Blight in Coffea arabica. Agronomy, v. 12, n. 12, p. 3178, 1 dez. 2022.

BARTOLI, C.; ROUX, F. Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. **Frontiers in Plant Science**, v. 8, n. May, p. 763, 2017.

BRADBURY, P. J. et al. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics, v. 23, n. 19, p. 2633–2635, 2007.

BROWNING, B. L.; BROWNING, Sharon R. Genotype imputation with millions of reference samples. **The American Journal of Human Genetics**, v. 98, n. 1, p. 116-126, 2016.

CLARINDO, W. R.; CARVALHO, C. R. First Coffea arabica karyogram showing that this species is a true allotetraploid. **Plant Systematics and Evolution**, v. 274, n. 3, p. 237-241, 2008.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da safra brasileira de café**, Brasília, DF, v. 10, n. 2 segundo levantamento, maio. 2023.

CUI, Y.; ZHANG, F.; ZHOU, Y. The application of multi-locus GWAS for the detection of salt-tolerance loci in rice. **Frontiers in plant science**, v. 9, p. 1464, 2018.

DAVIS, A. P. et al. Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. **Botanical Journal of the Linnean Society**, v. 167, n. 4, p. 357-377, 2011.

DERIBE, H. Review on Factors which Affect Coffee (Coffea Arabica L.) Quality in South Western, Ethiopia. **International Journal of Forestry and Horticulture**, v. 5, n. 1, p. 12-19, 2019.

DING, R. et al. Single-locus and multi-locus genome-wide association studies for intramuscular fat in Duroc pigs. **Frontiers in genetics**, p. 619, 2019.

EARL, D. A.; VONHOLDT, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation genetics resources**, v. 4, p. 359-361, 2012.

ELHAIK, E. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. **Scientific Reports**, v. 12, n. 1, p. 14683, 2022.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular ecology**, v. 14, n. 8, p. 2611-2620, 2005.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PloS one**, v. 6, n. 5, p. e19379, 2011.

FERRÃO, L. F. V. et al. Insights into the genetic basis of blueberry fruit-related traits using diploid and polyploid models in a GWAS context. **Frontiers in Ecology and Evolution**, v. 6, p. 107, 2018.

GLAUBITZ, J. C. et al. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. **PLoS ONE**, v. 9, n. 2, 2014.

GUYOT, R. al (2020). WCSdb: a database of wild Coffea species. Database 00:1–6

HE, J. et al. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. **Frontiers in plant science**, v. 5, p. 484, 2014.

HUANG, Xuehui; HAN, Bin. Natural variations and genome-wide association studies in crop plants. **Annual review of plant biology**, v. 65, p. 531-551, 2014.

KORTE, A.; FARLOW, A. The advantages and limitations of trait analysis with GWAS: A review. **Plant Methods**, v. 9, p. 1-9, 2013.

JOUBÈS, J. et al. The VLCFA elongase gene family in Arabidopsis thaliana: phylogenetic analysis, 3D modelling and expression profiling. **Plant Molecular Biology**, v. 67, n. 5, p. 547–566, 2008.

KUMAR, S.; BANKS, T. W.; CLOUTIER, S. SNP discovery through next-generation sequencing and its applications. **International Journal of Plant Genomics**, 2012.

LANDER, E. S.; SCHORK, N. J. Genetic dissection of complex traits. **Science**, v. 265, n. 5181, p. 2037-2048, 1994.

LASHERMES, P. et al. Molecular characterisation and origin of the Coffea arabica L. genome. **Molecular and General Genetics MGG**, v. 261, n. 2, p. 259-266, 1999.

LIAO, Pei-Yu; LEE, K. H. From SNPs to functional polymorphism: The insight into biotechnology applications. **Biochemical Engineering Journal**, v. 49, n. 2, p. 149-158, 2010.

LI, C. et al. 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**, v. 9, p. 1083, 2018.

LI-BEISSON, Y. et al. Acyl-Lipid Metabolism. **The Arabidopsis Book**, v. 11, p. e0161, 2013.

LIPKA, A. E. et al. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. **Current Opinion in Plant Biology**, v. 24, p. 110-118, 2015.

LIU, X. et al. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. **PLoS genetics**, v. 12, n. 2, p. e1005767, 2016.

LOLLE, S. J. et al. Developmental Regulation of Cell Interactions in the Arabidopsis fiddlehead-1Mutant: A Role for the Epidermal Cell Wall and Cuticle. **Developmental biology**, v. 189, n. 2, p. 311-321, 1997.

MANIVANNAN, A. et al. Next-generation sequencing approaches in genome-wide discovery of single nucleotide polymorphism markers associated with pungency and disease resistance in pepper. **BioMed research international**, v. 2018, 2018.

MEYER. **Coffee Mission to Ethiopia**, 1964–65. FAO, Rome, Italy, 1968.

MONEY, D. et al. LinkImpute: fast and accurate genotype imputation for nonmodel organisms. **G3: Genes, Genomes, Genetics**, v. 5, n. 11, p. 2383-2390, 2015.

PRICE, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. **Nature genetics**, v. 38, n. 8, p. 904-909, 2006.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945-959, 2000.

PORCU, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. **Nature communications**, v. 10, n. 1, p. 1-12, 2019.

SANT'ANA, G. C. et al. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. **Scientific reports**, v. 8, n. 1, p. 1-12, 2018.

SU, J. et al. Multi-Locus Genome-Wide Association Studies of Fiber-Quality Related Traits in Chinese Early-Maturity Upland Cotton. **Frontiers in Plant Science**, v. 9, 16 ago. 2018.

TAMBA, C. L.; NI, Y. Li.; ZHANG, Y. M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. **PLoS computational biology**, v. 13, n. 1, p. e1005357, 2017.

TAMBA, C. L; ZHANG, Y. M. A fast mrMLM algorithm for multi-locus genome-wide association studies. **biorxiv**, p. 341784, 2018.

TRAN, H. T. M et al. SNP in the Coffea arabica genome associated with coffee quality. **Tree Genetics & Genomes**, v. 14, n. 5, p. 1-15, 2018.

VISSCHER, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. **The American Journal of Human Genetics**, v. 101, n. 1, p. 5-22, 2017.

WANG, J.; ZHANG, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. **Genomics, Proteomics & Bioinformatics**, 2021.

WANG, S. B. et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. **Scientific reports**, v. 6, n. 1, p. 1-10, 2016.

WEN, Y. J. et al. A fast multi-locus random-SNP-effect EMMA for genome-wide association studies. **bioRxiv**, p. 077404, 2016.

WEN, Y. J. et al. Methodological implementation of mixed linear models in multi-locus genome-wide association studies. **Briefings in bioinformatics**, v. 19, n. 4, p. 700-712, 2018.

XIAO, Y. et al. Genome-wide association studies in maize: praise and stargaze. **Molecular plant**, v. 10, n. 3, p. 359-374, 2017.

XU, Y. et al. Genome-Wide Association Mapping of Starch Pasting Properties in Maize Using Single-Locus and Multi-Locus Models. **Frontiers in Plant Science**, v. 9, 5 set. 2018.

YANG, X. et al. Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. **Molecular Breeding**, v. 28, p. 511-526, 2011.

YU, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature genetics**, v. 38, n. 2, p. 203-208, 2006.

ZHANG, C. et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. **Bioinformatics**, v. 35, n. 10, p. 1786-1788, 2019.

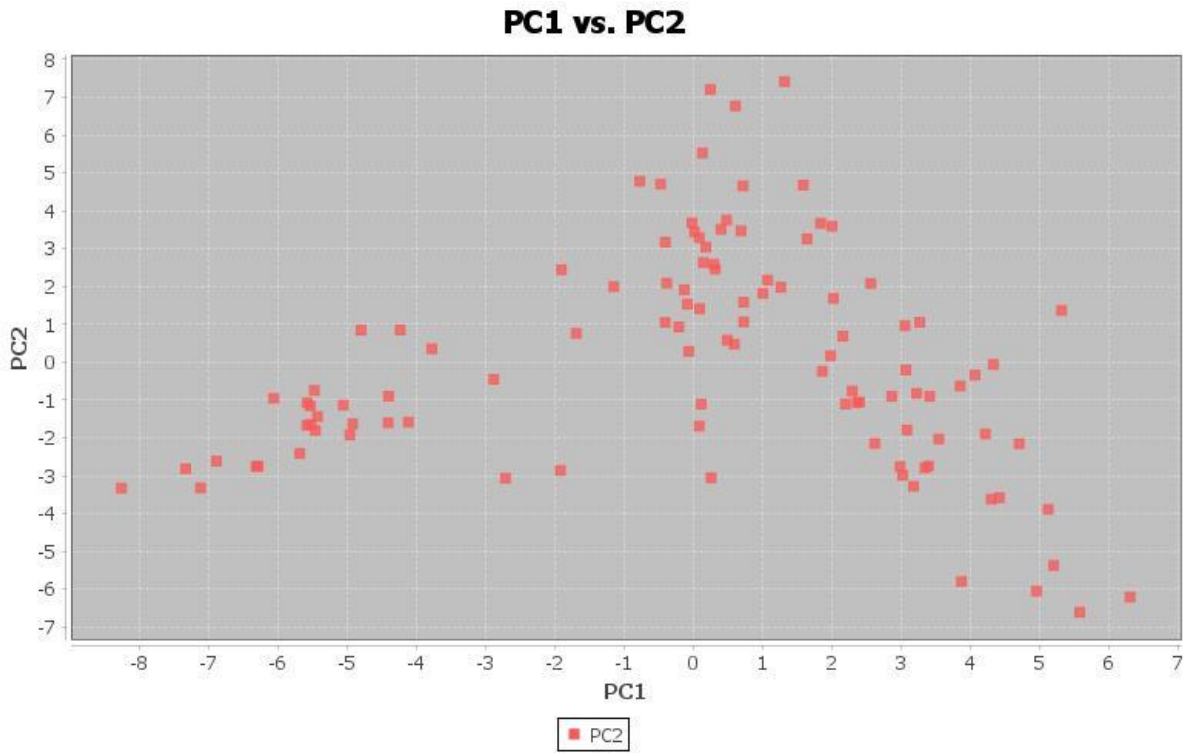
ZHANG, Y. M.; JIA, Z.; DUNWELL, J. M. The applications of new multi-locus GWAS methodologies in the genetic dissection of complex traits. **Frontiers in plant science**, p. 100, 2019.

.

APÊNDICE A – Análise de componentes principais

ANÁLISE DE COMPONENTES PRINCIPAIS

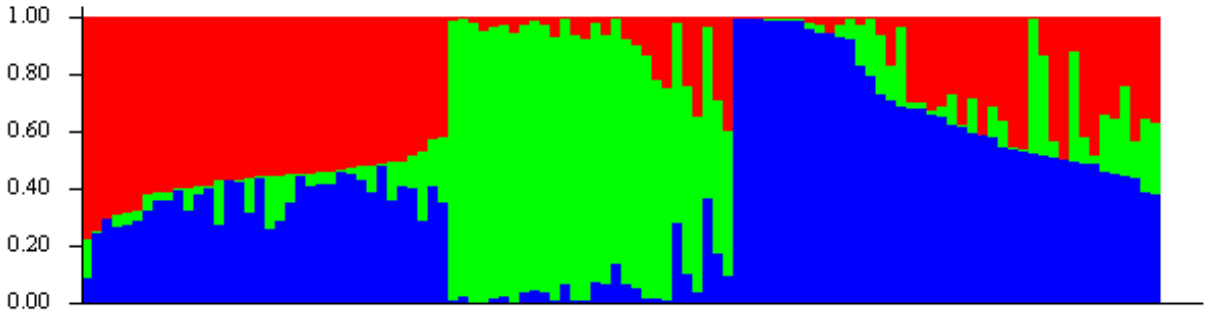
Figura 8. Análise de componentes principais dos 106 acessos de *Coffea arabica*.



APÊNDICE B – STRUCTURE

STRUCTURE

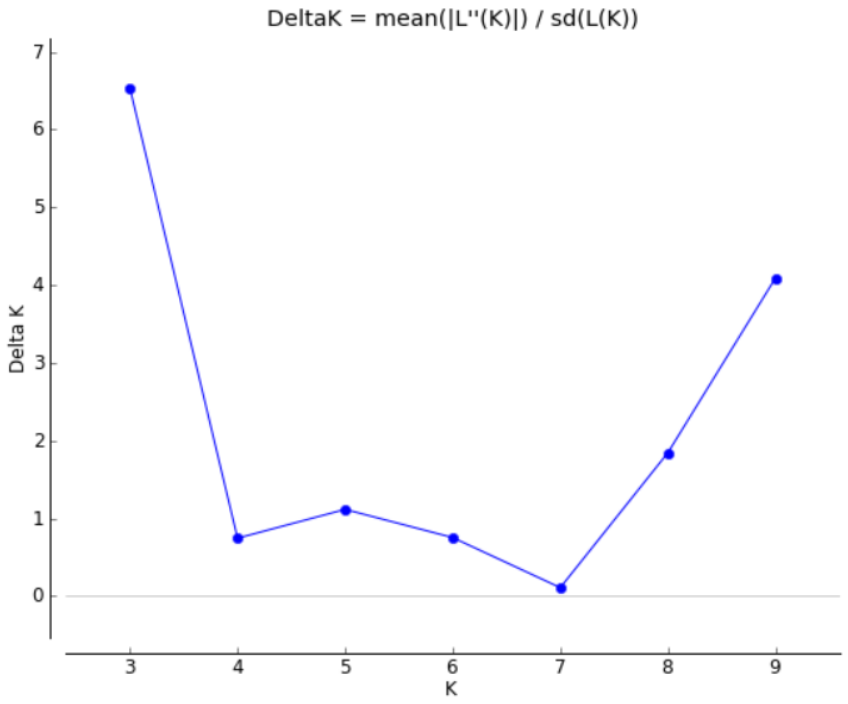
Figura 9. Barplot da estimativa do coeficiente de adesão estimado (Q) dos 106 acessos diferentes baseados em 11.136 SNPs para o K = 3.



APÊNDICE C – Structure Harvester

STRUCTURE HARVESTER

Figura 10. Evolução dos valores ΔK (eixo y) de acordo com o número de grupos genéticos (eixo x).



APÊNDICE D – Decaimento do desequilíbrio de ligação

DECAIMENTO DO DESEQUILÍBRIO DE LIGAÇÃO

Figura 11. Análise do decaimento do desequilíbrio de ligação dos 106 acessos de *Coffea arabica* mensurado por r^2 em função da distância genética entre os marcadores SNPs.