Universidade Estadual de Londrina

BRUNA SILVESTRE RODRIGUES DA SILVA

ANÁLISE DA DIVERSIDADE GENÉTICA, FENOTÍPICA E MAPEAMENTO POR ASSOCIAÇÃO (GWAS) EM UM PAINEL DE *C. arabica*, INCLUINDO ACESSOS DO CENTRO DE ORIGEM

Londrina
2018

BRUNA SILVESTRE RODRIGUES DA SILVA

ANÁLISE DA DIVERSIDADE GENÉTICA, FENOTÍPICA E MAPEAMENTO POR ASSOCIAÇÃO (GWAS) EM UM PAINEL DE *C. arabica*, INCLUINDO ACESSOS DO CENTRO DE ORIGEM

Dissertação apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito para obtenção do título de Doutor.

Orientador: Dr. Luiz Filipe Protasio Pereira

Londrina
2018

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

da Silva, BRUNA Silvestre Rodrigues .
ANÁLISE DA DIVERSIDADE GENÉTICA, FENOTÍPICA E MAPEAMENTO POR ASSOCIAÇÃO (GWAS) EM UM PAINEL DE C. arabica, INCLUINDO ACESSOS DO CENTRO DE ORIGEM / BRUNA Silvestre Rodrigues da Silva. - Londrina, 2018.
131 f. : il.

Orientador: Luiz Filipe Protasio Pereira.
Coorientador: Juarez Pires Tomaz.
Coorientador: Diego Micheletti.
Tese (Doutorado em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Centro de Ciências Biológicas, , 2018.
Inclui bibliografia.

1. Melhoramento do Café Arábica - Tese. 2. Estudo de Associação - Tese. 3. Marcadores SSRs e SNPs - Tese. I. Pereira, Luiz Filipe Protasio. II. Pires Tomaz, Juarez. III. Universidade Estadual de Londrina. Centro de Ciências Biológicas. . IV. Título.

BRUNA SILVESTRE RODRIGUES DA SILVA

ANÁLISE DA DIVERSIDADE GENÉTICA, FENOTÍPICA E MAPEAMENTO POR ASSOCIAÇÃO (GWAS) EM UM PAINEL DE *C. arabica*, INCLUINDO ACESSOS DO CENTRO DE ORIGEM

Dissertação apresentada ao Programa de Pós – Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito para obtenção do título de Doutor.

BANCA EXAMINADORA

_____

Orientador Dr: Luiz Filipe Protasio Pereira
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA

_____

Dra. Francismar Corrêa Marcelino-Guimaraes
Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA

_____

Dr. Leandro Simoes Azeredo Gonçalves
Universidade Estadual de Londrina - UEL

_____

Dr. Luiz Gonzaga Esteves Vieira
Universidade do Oeste Paulista - UNOESTE

_____

Dr. Gustavo César Sant'Ana
Tropical Melhoramento e Genética - TMG

Londrina, 12 de Julho de 2018.

AGRADECIMENTOS

A Deus pela criação das ciências da vida, sem a qual não estaríamos aqui compreendendo a sua complexidade.

Aos meus pais Maria Eudália Rodrigues da Silva e Sidnei da Silva e meus irmãos Jaqueline Silvestre Rodrigues da Silva e Jeancarlos Rodrigues da Silva nos quais sinto muita saudade.

Aos meus segundos pais e exemplos de vida Olnei da Silva e Lécio Silva. Obrigada pelos ensinamentos, palavras de conforto e apoio financeiro fornecidos desde cedo. Com certeza não estaria aqui se não fosse por vocês.

Ao meu noivo e melhor amigo Victor Cavalcante pelo companheirismo, pela paciência, pelo incentivo e compreensão durante a realização desse projeto.

Ao meu orientador, Dr. Luiz Filipe Protasio Pereira pela sua amizade, paciência, competência e ética profissional. Todo meu agradecimento por todo conhecimento e ajuda compartilhada, além da oportunidade de estágio no IAPAR. Eu tenho muito orgulho de ter sido sua aluna e ter sido orientada por você.

Ao Dr. Diego Micheletti pela paciência e ajuda na finalização deste projeto e por todos os ensinamentos nas análises dos dados por bioinformática e programação.

À CAPES e CAPES PDSE pelo apoio financeiro para a realização desse trabalho e por toda a bagagem de conhecimento que obtive através desses recursos.

À Universidade Estadual de Londrina e ao curso de Pós-graduação em Genética e Biologia Molecular.

A todos do laboratório de Biotecnologia do IAPAR que de alguma forma contribuíram para o desenvolvimento desse projeto.

SILVA, Bruna Silvestre Rodrigues da. **Análise da diversidade genética, fenotípica e mapeamento por associação em um painel de *c. arabica*, incluindo acessos do centro de origem**. 2018. 129f. Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina. 2018.

RESUMO

A exploração da diversidade genética e fenotípica dos acessos do centro de origem de uma espécie é fundamental para estimar o potencial ganho genético e efetiva conservação de seus recursos genéticos. *Coffea arabica* é uma espécie recente e com pouca variabilidade genética, o que torna mais importante a caracterização destes materiais do centro de origem para busca de genótipos com características que possam ser incorporadas pelos programas de melhoramento. O objetivo deste trabalho foi analisar a diversidade e estrutura genética com base nos marcadores SSRs e SNPs de cafeeiros provenientes da Etiópia, como também fenotipar e realizar estudos de associação para características relacionadas à qualidade bioquímica dos grãos. Para o estudo de diversidade genética de *C. arabica* com base nos SSRs, 37 genótipos incluindo acessos da Etiópia, cultivares tradicionais e variedades, e seus dois parentais, *C. canephora* e *C. eugenioides* foram genotipados com 30 marcadores SSRs. Um total de 206 alelos foram polimórficos e vinte desses, com conteúdo de informação polimórfica (*PIC*) acima de 0.7. As estimativas de diversidade genética como a média da heterozigosidade esperada (*He*), heterozigosidade corrigida pelo tamanho amostral (*uHe*) e proporção de loci polimórficos (*P*%) demonstraram alta diversidade genética entre os acessos do lado Oeste do Vale do Rift em relação aos outros grupos genéticos estudados, como também 34 alelos privados foram observados nesse grupo. Duas subpopulações foram identificadas pelo estudo de estrutura populacional e Análise de Coordenadas Principais, uma correspondente aos acessos da Etiópia e o outro correspondente às cultivares tradicionais e variedades. A comparação da dissimilaridade genética entre os acessos de *C. arabica* com seus dois parentais diplóides demonstrou que as cultivares em estudo são geneticamente mais próximas de *C. eugenioides* do que os acessos etíopes. Todo o painel apresentou alta variabilidade quanto aos 8 compostos analisados por espectroscopia no infravermelho próximo (NIRS). Significativas correlações positivas entre ácidos clorogênicos (ACGs) e Lipídeos Totais (LT), Poteínas Totais (PT), Sacarose e Açúcares Totais (AT) bem como Cafeína e PT; Sacarose e AT; e LT e Caveol foram observadas. Significativa correlação negativa foi encontrada entre LT e Sacarose/AT e entre Cafestol e Caveol. A análise de agrupamento hierárquico identificou 3 grupos entre os acessos com características a serem exploradas para a qualidade da bebida e/ou para a produção de cafés com teores diferenciados de diterpenos. Maior diversidade fenotípica foi observada entre os acessos do lado Oeste do Vale do Rift. Foi realizada genotipagem por sequenciamento (GBS) de 159 genótipos incluindo acessos da histórica coleção Etíope da FAO, além de cultivares tradicionais e variedades de *C. arabica* e os dois parentais diploides da espécie. O *Pipeline* TASSEL-GBS foi realizado afim de montar os contigs, mapeá-los e identificar marcadores SNPs de *C. arabica* nos dois parentais diploides da espécie (*C. canephora* e *C. eugenioides).* Foram identificados um total de 1.719 e 2.949 SNPs em ambos subgenomas com uma cobertura média de 68X e 47X que subdividiu os

genótipos em 2 e 3 subpopulações. Para investigar as variações genotípicas subjacentes aos traços relacionados com a qualidade da bebida, GWAS foi realizado para identificar variantes SNPs e genes candidatos nos acessos de *C. arabica* da Etiópia com um total de 4.517 SNPs e 101 genótipos para 5 replicatas de fenotipagem (ano 2011, 2012, 2015, 2016 e média 2011/2015) e 8 modelos de associação. Um total de 33 SNPs de *C. arabica* foram significativamente associados aos compostos relacionados com a qualidade da bebida de café em ambos subgenomas. Dos 22 SNPs mapeados em *C. canephora* e significativamente associados, 17 são co-localizados a 13 genes candidatos envolvidos nas vias metabólicas dos compostos bioquímicos. Onze SNPs de *C. arabica* mapeados em *C. eugenioides* foram significativamente associados aos compostos bioquímicos. Nossos resultados confirmam que há uma grande riqueza alélica e fenotípica presente nos acessos da Etiópia, especialmente nos acessos originados de florestas do lado Oeste do grande Vale do Rift, ainda não explorados pelo melhoramento genético. Por fim, identificamos SNPs associados às caracteristicas fenotípicas que podem ser úteis no desenvolvimento de estratégias de seleção assistida objetivando melhorar a qualidade bioquímica de grãos de café verde, e genes candidatos para o melhoramento da qualidade da bebida de café.
Entre os genótipos analisados, foi observada uma grande variação fenotípica, com valores de 153 a 1.088 mg 100 g-1 para Cafestol e 303 a 1.144 mg 100 g-1 para Caveol.

Palavras-Chave: Café Arábica. Diversidade e estrutura genética. SSRs e SNPs. Diversidade fenotípica. Café Eugenioides. GWAS. QTLs.

Apoio Financeiro: Consórcio Pesquisa Café. CAPES. EMBRAPA CAFÉ. INCT CAFÉ. FINEP.

SILVA, Bruna Silvestre Rodrigues da. **Diversity genetic, phenotypic analysis and mapping by association (gwas) in *c. arabica* panel, including accessions from origin center**. 2018. 129p. Thesis (PhD in Genetics and Molecular Biology) – Universidade Estadual de Londrina, Londrina. 2018.

ABSTRACT

The exploration of genetic diversity and phenotypic of the accessions to the center of origin of a species is crucial to estimate the potential genetic gain and effective conservation of its genetic resources. *Coffea arabica* is a recent species with low genetic variability, which makes it more important the characterization of these materials from the center of origin to search genotypes with characteristics that can be incorporated by breeding programs. The aim of this work was analyze the diversity and genetic structure based on SSRs and SNPs markers of the coffee trees from Ethiopia, as well as to phenotype and carry out association studies for characteristics related to coffee drink quality. For the study of genetic diversity *of C. arabica* based on SSRs, 37 genotypes including accessions from Ethiopia, traditional cultivars and varieties, and their two parents, *C. canephora* and *C. eugenioides* were genotyped with 30 SSR markers. A total of 206 alleles showed polymorphic patterns and twenty of these, with polymorphic information content (PIC) above 0.7. The mean heterozygosity (He), unbiased expected heterozigosity (uHe) and proportion of polymorphic loci (P%) showed high levels of genetic diversity in the accessions on the Western side of Rift Valley in relation to the other genetic groups studied, as well as 34 private alleles were observed in this group. Two subpopulations were identified by the Population Structure and Principal Coordinate Analysis (PCoA) study, one corresponding to the Ethiopian accessions and the other corresponding to the traditional cultivars and varieties. The comparison of the genetic dissimilarity between the accessions of *C. arabica* and its two diploid parental showed that the cultivars under study are genetically closer to *C. eugenioides* than the Ethiopian accessions. It was also evaluated the phenotypic variability by Near Infrared Spectroscopy (NIRS) of eight biochemical compounds related to the quality of coffee drink in 68 accessions of *C. arabica* from Ethiopia. The all panel showed variability regarding the analyzed compounds, but high variability was observed between Cafestol and Caveol (35.96 and 22.14 mg 100 g^{-1}). Significant positive correlations between Chlorogenic Acids (ACGs) and Total Lipids (LT), Total Protein (PT), Sucrose and Total Sugars (AT) as well as Caffeine and PT; Sucrose and AT; and LT and Caveol were observed. Significant negative correlation was found between LT and Sucrose /AT and between Cafestol and Caveol. The hierarchical cluster analysis identified 3 groups between the accessions with characteristics to be explored for the quality of the beverage and /or for the production of coffees with different levels of diterpenes. Higher phenotypic diversity was observed between accessions on the Western Side of the Rift Valley. Genotyping by Sequence (GBS) was realized of 159 genotypes including accessions from the historical Ethiopian collection of FAO, as well as traditional cultivars and varieties of *C. arabica* and two diploid parental of the specie were carried out. A new Pipeline TASSEL-GBS was developed to assemble, align the *contigs* and identify SNPs of *C. arabica* in the two diploid parents of the specie (*C. canephora* and *C. eugenioides*), and identified a total of 1,719 and 2,949 SNPs with a coverage mean of 68X and 47X, in which the genotypes were subdivided into

2 and 3 subpopulations. To investigate the genotypic variations underlying the traits related to the quality of beverage, GWAS was performed to identify SNPs and candidate genes in the accessions of *C. arabica* from Ethiopia with a total of 4,517 SNPs and 101 genotypes for 5 replicates of phenotyping (year 2011, 2012, 2015, 2016 and 2011/2015 average) and 8 association models. A total of 33 SNPs from *C. arabica* were significantly associated with compounds related to the quality of coffee beverage in both subgenomes. Of the 22 SNPs mapped in *C. canephora* and significantly associated, 17 are co-localized to 13 candidate genes involved in the metabolic pathways of biochemical compounds. Eleven SNPs of *C. arabica* mapped on *C. eugenioides* were significantly associated with biochemical compounds. Our results confirm that there is a great allelic and phenotypic richness present in the accessions of Ethiopia, especially in the accessions originating from forests in the West Side of the Great Rift Valley, not yet explored by the genetic improvement. Finally, we identified SNPs associated with the phenotypic characteristics that may be helpful to develop assisted selection strategies aiming at improve the biochemical quality of green coffee beans and candidate genes for improve of coffee beverage quality.

Keywords: *Coffea arabica.* Diversity and genetic structure. SSRs and SNPs. Phenotypic diversity. *Coffea eugenioides*. GWAS. QTLs.

Financial support: Brazilian Coffee Research. CAPES. EMBRAPA COFFEE. INCT COFFEE. FINEP.

SUMÁRIO

1. INTRODUÇÃO

O café (*C. arabica* L.) é uma das principais commodities agroindustriais comercializadas mundialmente. O gênero pertence à família *Rubiaceae* e representa um total de 124 espécies, sendo a maioria diplóide (2n = 22X cromossomos) e autoincompatível, com exceção de *C. arabica*, um alotetraplóide com predominância de autocompatibilidade (DAVIS et al., 2011). Dentre as espécies de cafeeiros cultivadas, *Coffea arabica L.* e *Coffea canephora P.* (no Brasil, variedades conilon e robusta) são as de maior valor no mercado respondendo em 70 e 30% da produção mundial, diferindo consideravelmente no preço, qualidade e tipo de mercado (MEYER, 1968; GONZÁLEZ et al., 2001).

Uma maior compreensão da base molecular e fenotípica relacionada à qualidade da bebida de café é necessária para atender as demandas crescentes dos produtores e consumidores (CHENG et al., 2016). O melhoramento clássico de *C. arabica* tem demonstrado resultados expressivos, entretanto, sua eficiência é dificultada devido às características botânicas do gênero *Coffea* e à complexidade das características de interesse, como qualidade da bebida e tolerância a estresses bióticos e abióticos. Atualmente, a maioria das cultivares comerciais da espécie são susceptíveis a estresses bióticos e abióticos, isso devido a estreita base genética e diversidade molecular da espécie, o que limita as possibilidades de melhoramento (ANTHONY et al., 2001; CHAPARRO et al., 2004).

Dentre os vários tipos de marcadores moleculares disponíveis para o melhoramento de plantas, os mais utilizados em estudos de diversidade e mapeamento genético são os microssatélites (SSRs) e os polimorfismos de única base (SNPs). Estes possuem uma maior frequência dentro do genoma dos organismos e permitem a identificação e mapeamento de genes que controlam características de interesse agronômico, fator interessante para reduzir o tempo e o custo do melhoramento clássico de plantas (KAUR et al., 2015).

O conhecimento acerca da composição química dos grãos de café permite um melhor aproveitamento da variabilidade existente no café e é uma ferramenta útil para ampliar o mercado de exportação, melhorar a qualidade e consequentemente a competitividade (KITZBERGER et al., 2013). Um grande número de trabalhos tem se concentrado na determinação de compostos bioquímicos para que possam funcionar

como indicadores ou discriminadores, de forma isolada ou em conjunto (KURZROCK; SPEER, 2001; CAMPANHA, 2008; SCHOLZ et al., 2016).

A disponibilidade de dados moleculares e fenotípicos gerou um avanço no melhoramento de culturas principalmente nas com estreita base genética, pois a partir desses dados é possível inferir o quão associado um marcador pode estar em relação a um QTL. O Estudo de Associação Ampla do Genoma (GWAS) também conhecido por mapeamento por desequilíbrio de ligação, baseia-se em conceitos de genética de populações para identificar associações entre marcadores genéticos e características fenotípicas. Esta análise fornece marcadores SNPs que poderão auxiliar na construção de mapas mais saturados como também no desenvolvimento de estratégias de seleção assistida por marcadores (SAM) (GUPTA et al., 2005).

O GWAS utilizando populações naturais tem como vantagem maior resolução, uma vez que se aproveita dos eventos de mutação e recombinação acumulados ao longo do tempo na população para associação com um fenótipo de interesse. Sabe-se que o centro de diversidade, origem e dispersão de *C. arabica* se encontra nos altiplanos do sudoeste da Etiópia, precisamente região Oeste do Vale do Rift (SILVESTRINI et al., 2007; SOUZA et al., 2017).

Entre os anos 1964-1965 foram realizadas expedições pela FAO para coletar acessos de *C. arabica* na Etiópia (MEYER, 1968). As amostras colhidas foram encaminhadas para sete institutos (Índia, Tanzânia, Etiópia, Costa Rica, Peru, Portugal e Brasil). O Instituto Agronômico de Campinas (IAC) repassou 132 mudas de acessos da histórica coleção da FAO ao IAPAR que foram plantadas em 1976 e que são mantidas até os dias atuais. Existem estudos demonstrando maior variabilidade genética e fenotípica dos acessos de *C. arabica* da Etiópia em relação às cultivares (KURZROCK; SPEER, 2001; POT et al., 2008; SILVESTRINI et al., 2007; SCHOLZ et al., 2016) e até o o momento um estudo de GWAS para compostos relacionados com a qualidade da bebida nos acessos de *C. arabica* foi publicado (SANT'ANA et al., 2018). Porém os SNPs utilizados neste trabalho foram provenientes em sua maioria apenas do subgenoma de um dos parentais diploides, *C. canephora* (SANT'ANA et al., 2018).

Este trabalho foi realizado em 4 partes no qual na primeira foi feita uma revisão de literatura. O primeiro capítulo analisou a diversidade e estrutura genética de um painel de genótipos, incluindo acessos de *C. arabica* com marcadores SSRs, além de uma análise comparativa de distância genética entre os acessos e seus dois ancestrais

diploides *C. canephora* and *C. eugenioides.* O segundo capítulo aborda a caracterização e avaliação bioquímica dos acessos de *C. arabica* da safra 2016, explorando a composição bioquímica da espécie no que tange à qualidade da bebida. Por fim, visando identificar variantes SNPs subjacentes a traços relacionados com a qualidade da bebida de café, bem como genes candidatos, foi realizado o *Pipeline* TASSEL para montagem, mapeamento dos *tags* e identificação de *loci* SNPs de acessos de *C. arabica* nos subgenomas dos dois parentais diploides da espécie (*C. canephora* e *C. eugenioides*).

2 OBJETIVOS

2.1 OBJETIVO GERAL

Explorar os recursos genéticos de materiais selvagens de *Coffea arabica* visando estimar a diversidade e estrutura em seu centro de origem, além de identificar SNPs ligados às características fenotípicas e QTLs que influenciam a qualidade da bebida, para que possam ser incorporados pelos programas de melhoramento.

2.2 OBJETIVOS ESPECÍFICOS

- Estudar a diversidade genética e estrutura populacional entre genótipos de cafeeiros incluindo acessos de *C. arabica* da Etiopia com marcadores SSRs;
- Analisar a diversidade fenotípica dos acessos de *C. arabica* (safra 2016) quanto a compostos relacionados à qualidade da bebida pela aplicação da tecnologia NIRS;
- Montar e mapear *tags* obtidos pela tecnologia GBS de 159 genótipos de café nos genomas de referência dos dois parentais diploides de *C. arabica* (*C. canephora* e *C. eugenioides*) para a busca de variantes SNPs;
- Realizar o estudo de diversidade e estrutura genética nesses genótipos com base nos marcadores SNPs;
- Identificar SNPs associados a compostos relacionados com a qualidade da bebida de café e genes candidatos que possam ser incorporados nos programas de melhoramento do café.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 O CAFÉ E SUA IMPORTÂNCIA ECONÔMICA

O café é uma das bebidas mais populares mundialmente (JESKA-SHOWRON et al., 2016). A popularidade de sua bebida o tornou uma das mais importantes *commodities* agroindustriais do mundo, sendo produzido em mais de 80 países tropicais e subtropicais e é a fonte de subsistência para cerca de 100 milhões de cafeicultores (VEGA et al., 2003; CONAB, 2018).

O Brasil é o maior produtor e exportador mundial do café, responsável por cerca de um terço da produção e das exportações globais. Além disso, é o segundo maior consumidor da bebida no mundo, atrás apenas dos EUA, país que consome anualmente 24 milhões de sacas. Seu cultivo, processamento, comercialização, transporte e mercado proporcionam milhões de empregos em todo o mundo (CECAFÉ, 2018).

De acordo com o relatório mensal do CECAFÉ no período de abril de 2017 a março de 2018, as exportações dos Cafés do Brasil atingiram um volume equivalente a 30,58 milhões de sacas e o país arrecadou US$ 5,050 bilhões de receita cambial ao preço médio de US$ 165,07 a saca de 60kg. Desse volume, 27,14 milhões de sacas foram de café verde (26,79 milhões de café arábica e 345,32 mil de café robusta) e 3,44 milhões de café industrializado (3,42 milhões de café solúvel e 21,33 mil de café torrado e moído) (CECAFÉ, 2018).

Os principais estados produtores de café são por ordem Minas Gerais, Espírito Santo, São Paulo, Bahia, Rondônia, Paraná, Rio de Janeiro, Goiás, Mato Grosso e Amazonas; correspondendo cerca de 99,6% da produção nacional (CECAFÉ, 2018).

A diversidade de produtos de café oferecidos e muitos deles certificados pelo Programa de Qualidade do Café da ABIC tem mantido o interesse dos consumidores pela maior qualidade da bebida e aroma específicos. Além disso, pelo fato do Brasil oferecer uma ampla diversidade de cafés (grãos verdes, torrados, torrado moído, especiais, orgânicos, instantâneos e sustentáveis) confere-lhe uma vantagem competitiva sobre muitos outros países produtores e exportadores (ABIC, 2017).

3.2 A ESPÉCIE *C. ARABICA E SUAS CARACTERÍSTICAS*

C. arabica foi originada a partir da hibridização natural entre seus dois ancestrais diploides, *C. canephora* e *C. eugenioides*, e é considerada um alopoliploide do tipo

segmental, sendo a única espécie de café tetraploide (2n=2x=44), e que se multiplica preferencialmente por autofecundação. A espécie é nativa de uma região restrita e marginal às demais espécies do gênero, localizada no Sudoeste da Etiópia, Sudeste do Sudão e Norte do Quênia, entre 1.000 e 3.000 metros de altitude (CARVALHO, 1946). Estima-se que a especiação de *C. arabica* é um evento recente (≅ 665.000 anos) (YU et al., 2011).

A base genética bastante estreita das cultivares de *C. arabica* é em grande parte consequência de seu histórico de dispersão e de sua natureza autopolinizadora (BERTHAUD e CHARRIER, 1988). Existem apenas três variedades amplamente cultivadas em todo mundo e que foram selecionadas a partir de duas variedades botânicas distintas, *Typica* e *Bourbon*, Caturra é um mutante do grupo *Bourbon*; Mundo Novo é um híbrido de *Bourbon* e *Typica* e Catuaí é um híbrido de Mundo Novo e Caturra. Estas variedades possuem alto rendimento e produzem uma bebida de alta qualidade, mas apresentam suscetibilidade a pragas e doenças (ANTHONY et al., 2002).

3.3 Marcadores Moleculares SSRs e SNPs e suas Utilidades em Análises Genéticas de Plantas

Com o aprimoramento das técnicas de biologia molecular na década de 80, os marcadores moleculares se tornaram o foco para uso no melhoramento de plantas, possibilitando a detecção de polimorfismos de DNA com comportamento mendeliano, passível de serem utilizados nas diferentes áreas da genética (FERREIRA; GRATTAPAGLIA, 1998).

Os marcadores moleculares são definidos como características de DNA que diferenciam dois ou mais indivíduos (PÍPOLO; GARCIA, 2006). Ao contrário dos marcadores fenotípicos, possuem a vantagem de serem independentes de variações do ambiente e podem ser detectados em qualquer tipo de tecido e fase de desenvolvimento da planta (FERREIRA; GRATTAPAGLIA, 1998).

Atualmente os distintos tipos de marcadores moleculares diferenciam-se pela tecnologia utilizada para revelar variabilidade no DNA, distribuição no genoma, e pelo custo de implementação e operação (GRIFFITHS et al., 2000). De acordo com as técnicas de detecção, os marcadores moleculares podem ser classificados em três classes: baseados em hibridização (RFLP, VNTR e DArt); em PCR (AFLP, RAPD, SCAR, STS, SSR e ISSR), e em sequenciamento (SNP) (BOSTEIN et al., 1980;

JEFFREYS et al., 1985; LITT; LUTTY, 1989; WILLIAMS, 1990; PARAN; MICHELMORE, 1993; VOS et al., 1995; BRITO et al., 2010; BORÉM; FRITSCHE-NETO, 2013).

Dentre os marcadores baseados em PCR, os SSRs (do inglês, Simple Sequence Repeat) são comumente utilizados em estudos de análises genéticas. Além de serem loco específicos, possuem várias vantagens incluindo alto grau de polimorfismo, repetibilidade, reprodutibilidade, codominância e multialelismo (BANERJEE et al., 2012; KAUR et al., 2015). Os SNPs também vêm ganhando notoriedade devido aos sequencimentos *De novo* e resequenciamentos cada vez mais precisos e baratos. A natureza dialélica dessa classe de marcador oferece baixa taxa de erro do alelo e eleva o nível de coerência entre os laboratórios (GIANCOLA, 2006).

Os SNPs são as formas mais frequentes de variação genética (90%) constituindo um número ilimitado de marcadores potenciais em estudos de análises genéticas. Normalmente ocorrem na frequência de 1/100-500 pb no genoma das plantas e variam dependendo da espécie/ variedade analisada, como por exemplo pode ocorrer na frequência de 1/ 490 pb em soja e 1/ 540 pb em ervilha (I.Y. CHOI et al., 2007; LEONFORTE et al., 2013). Portanto, os SNPs são atualmente marcadores-alvo devido a essas características e aliado ao fato de que apresentam facilidade e boa reprodutibilidade na genotipagem em larga escala (*"high-throughput"*) (GIANCOLA, 2006).

Para auxiliar o melhoramento, os marcadores possuem grande aplicação na proteção e discriminação de cultivares, espécies e variedades de forma precisa e acurada (SOUZA et al., 2017). Além disso, podem ser utilizados para estimativas de diversidade e estrututa genética, escolha de genitores, construção de mapas genéticos e mapas de QTLs, além de estudos de seleção genômica (SG) e de associação (GWAS). Esses estudos fornecem informações valiosas para a SAM (SALLES et al., 2003).

3.4 Aspectos Relacionados à Qualidade da Bebida de Café

A qualidade bioquímica da bebida do café é influenciada pelos diferentes teores de um grande número de compostos presentes nos grãos torrados. Assim, devido à grande importância econômica da espécie *C. arabica*, a composição bioquímica dos grãos tem sido extensivamente estudada em relação à implicação de determinados compostos envolvidos na formação de sabor e aroma da bebida (BELAY, 2008). Além disso, existe um crescente interesse do consumidor em compreender melhor sobre a

bebida, e com isso, exigência por cafés de qualidade (cafés *gourmet*) vem ganhando expressão nos últimos anos (FRIDELL, 2014).

A avaliação da qualidade da bebida é realizada de três formas: física (tamanho dos grãos), avaliação sensorial (qualidade da bebida) e análise bioquímica (compostos atribuídos à qualidade) (FRIDELL, 2014). A avaliação sensorial da bebida é feita organolepticamente por degustadores de café treinados utilizando terminologias para o aroma, sabor, corpo e acidez (CHENG et al., 2016).

A química da qualidade do café é altamente complexa. Seu efeito estimulante somado aos mais de 1.000 compostos voláteis formados após a torra e oriundos das transformações químicas de seus constituintes fixos como cafeína, trigonelina, ácidos clorogênicos, carboidratos, constituintes lipídicos, incluindo os diterpenos e ácidos graxos (livres e esterificados), acrescentam valor agregado ao café e aos resíduos gerados nas diversas etapas até alcançar seu principal objetivo, que é a bebida quente (DURÁN et al., 2017).

Sabe-se que o sabor de uma xícara de café recém preparada é a expressão final e o resultado de uma longa cadeia de transformações entre a semente e a bebida (JOËT et al., 2009). Os grãos de café verde submetidos à torrefação em temperaturas entre 100-200°C perdem água, expandem e tornam-se escuros e quebradiços (OESTREICH-JANZEN, 2010). Grãos maiores não são necessariamente responsáveis pelo gosto melhor da bebida, mas idealmente, a torrefação deve ser processada com grãos uniformes. Isso porque na torrefação de grãos irregulares, os menores tendem a queimar ou torrar, enquanto que os maiores demoram a torrar, o que afeta tanto a aparência visual dos grãos como a qualidade da bebida final (WINTGENS, 2012).

Após o processo de torrefação, o grão de café passa a ser constituído por Carboidratos (38-41,5%), Melanoidinas (23%), Proteínas Totais (10%), Lipídeos Totais (11-17%), Minerais (4,5-4,7%), Ácidos Clorogênicos - ACG (2,73,1%), Ácidos Alifáticos (2,4-2,5%), Cafeína (1,3-2,4%), Trigonelina e Niacina (0,7-1%), e Compostos Voláteis (0,1%) (BELITZ et al., 2009).

Fatores genéticos (espécie, variedade); edafoclimáticos (sombra, altitude, umidade, temperatura; agronômicos (condições de cultivo); tipo de processamento pós-colheita (secagem e armazenamento); e como já mencionado a torrefação, podem influenciar na composição química dos grãos de café cru e consequentemente na

qualidade da bebida final (SELMAR et al., 2006; KITZBERGER et al., 2013b; CHENG et al., 2016; SCHOLZ et al., 2016).

Alguns compostos-chave presente nas sementes maduras possuem efeitos significativos na qualidade do café. Dentre estes, alguns se mantém estáveis durante a torrefação e contribuem para os atributos de sabor e aroma da bebida, ou são degradados a precursores que dão um perfil característico à bebida (OESTREICH-JANZEN, 2010).

Esses compostos bioquímicos são potenciais ferramentas de identificação e discriminação entre genótipos pelo fato de estarem presentes em diferentes teores em espécies e variedades de café (CAMPANHA, 2008).

3.5 CONSTITUINTES QUÍMICOS DO CAFÉ

3.5.1 Cafeína

A cafeína quimicamente conhecida como 1,3,7 – trimetilxantina é um alcalóide de purina típico, portanto um composto nitrogenado não proteico e um metabólito secundário sintetizado em várias dicotiledôneas, incluindo chá (*Camellia sinensis*), cacao (*Theobroma cacao*) e café (*C. arabica* e *C. canephora*). Nas plantas de café a cafeína está presente em todas as partes e a biossíntese ocorre nas folhas, sementes e no pericarpo, a parte externa do fruto. Durante o desevolvimento do fruto a cafeína é translocada através das membranas e acumula-se no endosperma (OESTREICH-JANZEN, 2010).

Três principais papéis biológicos para a cafeína foram relatados para o café. Nas folhas, a cafeína possui propiedades inseticidas (NATHANSON, 1984), enquanto nos grãos induz a alelopatia, inibindo a germinação de sementes em espécies concorrentes (PACHECO et al., 2008). Nas flores, a cafeína pode encorajar o forrageamento eficiente e eficaz, auxiliando a memória das abelhas através de uma associação olfatória aprendida (WRIGHT et al., 2013).

Dentre os compostos do café a cafeína é aquela que recebe maior atenção, devido as suas conhecidas propriedades fisiológicas (reduz estresse oxidativo, ação protetora do sistema antioxidante e ação neuroprotetora); farmacológicas pelo efeito estimulante do sistema nervoso central, que melhora o estado de alerta no controle cognitivo, além de diminuir a fadiga (MAZZAFERA; YAMAOKA-YANO; VITORIA, 1996;

SIN et al., 2009; GLADE, 2010; ALI et al., 2012; OLEAGA et al., 2012); e sensoriais, sendo um dos mais importantes compostos que contribuem para os atributos do amargor e portanto, para a qualidade da bebida de café (OESTREICH-JANZEN, 2010). No entanto, quanto à dose que é ingerida, seu excesso de consumo pode resultar efeitos indesejados, como dores de cabeça, insônia, irritabilidade, doenças coronarianas, depressão e até mesmo dependência e risco de câncer (CHENG et al., 2016). Quanto aos efeitos ecológicos da cafeína, relatos na defesa contra a herbivoria de moluscos, insetos, fungos e bactérias tem sido discutido (BAUMANN, 2005).

Devido a esses estudos, a procura pelo café descafeinado tem aumentado consideravelmente, sendo até mesmo recomendado por médicos. Para a descafeinação do café é realizado um tratamento a vapor do café verde para suavizar os tecidos, seguido por extração com solventes orgânicos ou dióxido de carbono supercrítico (OESTREICH-JANZEN, 2010). Porém distintos genes N-methyltransferases que catalisam os 3 passos de metilação durante a biossíntese da cafeína já foram clonados em plantas de café, tornando possível a produção de cafés naturalmente descafeinados pela engenharia genética (UEFUJI et al., 2003).

A cafeína também possui variação entre espécies e cultivares. O café arábica é conhecido pelo seu baixo teor de cafeína comparado ao café robusta, com valores de 0.6 – 1.8% e 1.2 – 4.0%, respectivamente. Essa característica o torna um composto discriminante entre espécies, principalmente para a espécie *C. arabica*, que possui diversidade genética estreita (SCHOLZ et al., 2016). Já em outras espécies e genótipos de *C. arabica* foi relatado teor de cafeína quase que inexistente (SILVAROLLA, MASSAFERA; FAZUOLI, 2004; KY; BARRE; NOIROT, 2013). A espécie *C. arabica* var. Laurina possui a metade do conteúdo de cafeína encontrado nas outras variedades da espécie (0.6%) (CARVALHO et al., 1965).

Ao contrário dos outros compostos nitrogenados que podem sofrer alteração na estrutura química durante o processo de torra, a cafeína possui estabilidade no processamento pós-colheita, assim como no emprego de altas temperaturas (CAMPA et al., 2005; OESTREICH-JANZEN, 2010).

3.5.2 Ácidos Clorogênicos

A bebida de café é bem conhecida por seu efeito estimulante associado à cafeína, porém já se reconhece outros constituintes químicos presentes no grão de café que

possuem implicações na saúde humana, como os ácidos clorogênicos (ACG) e sua capacidade antioxidante (ALVES et al., 2006).

Os compostos fenólicos compreendem um grupo heterogêneo de substâncias, umas com estruturas químicas relativamente simples e outras complexas, e são os antioxidantes mais difundidos em algumas plantas cultivadas (UPADHYAY; MOHAN RAO, 2013).

Os compostos fenólicos estão entre os metabólitos secundários de plantas e muitos estão envolvidos na adaptação às condições ambientais. Os ácidos clorogênicos são uma família importante de compostos fenólicos. Os ACGs são uma família de ésteres formados entre ácidos hidroxicinâmicos (caféico, cumárico e ferúlico) e ácido quínico, sendo 71 diferentes de ACGs já relatados, e 30 identificados em grãos de café verde (UPADHYAY; MOHAN RAO, 2013).

São os mais abundantes polifenóis presentes no café e por serem encontrados em maior quantidade em grãos verdes, estão relacionados com a formação e maturação do grão, além de estimular o florescimento das plantas (JESZKA-SKOWRON et al., 2016). Esta característica peculiar das sementes de café em desenvolvimento é que seu acúmulo e mobilização em quantidades consideráveis são requeridos para o metabolismo de lignina das plântulas, pois estão relacionados com o fortalecimento da parede celular (AERTS; BAUMANN, 1994; UPADHYAY; MOHAN RAO, 2013).

Os ácidos clorogênicos totais em grãos de café verde de *C. canephora* foram encontrados a uma média de 3.5 – 14.0% (FARAH et al., 2006; MULLEN et al., 2013). Em contrapartida, a quantidade total de ACGs é relativamente baixa em grãos de café verde em *C. arabica* (aproximadamente de 3.4 a 4.8%) (UPADHYAY; MOHAN RAO, 2013). Além disso, são termicamente instáveis, e na mesma espécie a perda de ACGs após a torrefação, corresponde a 60.9% e 96.5% em torra leve a muito escura, respectivamente. Assim, embora a maioria das ACGs são perdidos durante a torrefação, um aumento acentuado da atividade antioxidante total na bebida é relatada sugerindo que os produtos de sua degradação são potenciais antioxidantes (UPADHYAY; MOHAN RAO, 2013).

Os ACGs dão pigmentação, aroma e sabor à bebida, além de conferir adstringência, amargor e contribui para a acidez final, possuindo efeitos negativos na qualidade da bebida quando em altas concentrações (CAMPA et al., 2005; FARAH; DONANGELO, 2006; UPADHYAY; MOHAN RAO, 2013).

Para fins terapêuticos, quando consumido regularmente foram associados com a redução de cálculos renais e baixo risco de doenças de pulmão e fígado, incluindo câncer (LEITZMANN, 2002; TVERDAL; SKURTVEIT, 2003).

3.5.3 Lipídeos Totais e Diterpenos

A fração lipídica (óleos e ceras) representa 7.7 – 16.0% da composição química de grãos de café cru, composta principalmente por triacilgliceróis, fosfolipídeos, esteróis e tocoferóis (SCHOLZ et al., 2016). Os lipídios desempenham um papel importante na qualidade da xícara, principalmente por causa da hidrólise dos triglicerídeos e da liberação de ácidos graxos que são então oxidados e inferem sabores característicos da bebida de café (SPEER, SEHAT e MONTAG, 1993). A oxidação dos ácidos graxos determina a formação de produtos termicamente induzidos em particular aldeídos, no qual reagem com intermediários da reação de *Maillard*, dando origem a compostos aromáticos adicionais (JOËT et al., 2009; 2010).

Os lipídeos no café servem como carreadores do aroma e vitaminas lipossolúveis e contribuem para textura (produção de espuma em café expresso), e sensação da bebida na boca, pois ficam retidos na fração lipídica do café (SPEER; KOÖLLING-SPEER, 2006; SCHOLZ et al., 2016). Durante o desenvolvimento dos frutos, os lipídeos se acumulam até 180 dias após a floração e depois diminuem (JOËT et al., 2009; 2010).

Adicionalmente, o óleo de café apresenta uma fração insaponificável que contém diterpenos, compostos formados um esqueleto de prenil Caureno com 20 carbonos que estão presentes na forma livre (baixo teor) e esterificada com diferentes ácidos graxos e são intimamente relacionados com o sabor da bebida de café (SELMAR et al., 2008). Eles representam 20% do conteúdo dos lipídeos totais em grãos de café (SPEER; KOLLING-SPEER, 2006).

Os diterpenos cafestol e caveol são dois furanoditerpenoides da família dos cauranos de grande interesse devido aos seus diferentes efeitos fisiológicos, sendo os compostos lipídicos mais importantes em café e estão presentes em grãos de café verde e torrado. Apesar da grande diversidade em estrutura e função, todos os isoprenoides (ou terpenoides) são derivados de dois compostos de cinco carbonos, o isopentenil difosfato (IPP) e o isômero difosfato de dimetilalilo (DMAPP), produzidos através de dois compartimentos intracelulares, via MVA (mevalonato citosólico) e MEP (2-C-metil-D-eritritol-4-fosfato ou não mevalonato plastidial) (PEREIRA E IVAMOTO, 2015). Dentre os

6 diterpenos exclusivos do gênero *Coffea* (16-O-metilcaveol, 16-O-metilcafestol, desidrocaveol e desidrocafestol), Cafestol e Caveol são os mais abundantes diterpenos e possuem uma estrutura química muito similar, com apenas uma diferença da ligação dupla em um hidrocarboneto aromático (DIAS et al., 2010; SCHOLZ et al., 2016).

Estes compostos assim como a maioria dos isoprenoides são metabólitos secundários do café e estão relacionados com a qualidade da bebida e tipo de preparo da bebida. No tipo expresso os teores de diterpenos podem ser de 5 a 15 vezes maior do que em bebidas filtradas, pois o filtro de papel tende a retê-los (SPEER; KOLLING-SPEER, 2006; NAIDOO et al., 2011).

Além disso, desempenham papéis-chave na mediação das interações das plantas com o meio ambiente, como tolerância das plantas a estresses bióticos pela atividade biopesticida e defesa contra herbivoria (PATERAKI et al., 2015). O cafestol também foi relacionado a compostos fenólicos voláteis liberados pelas flores que podem ser responsáveis pela atração de polinizadores (DEL TERRA et al., 2013).

O café também tem sido muito relatado quanto ao seu impacto na saúde, no qual os efeitos positivos dos diterpenos, como induzem a degradação de substâncias tóxicas e proteção hepatoprotora contra aflatoxina B1 e acroleína, atividade anti-inflamatória, antioxidante e anticarcinogênica aos consumidores da bebida estão sendo muito abordados. Somado ao potencial farmacêutico, os diterpenos têm atraído interesse substancial para sua utilização como nutracêuticos e fragrâncias na indústria farmacêutica e cosmética (CAVIN et al., 1998; BOEKSCHOTEN et al., 2004; ESQUIVEL; JIMENÉZ, 2012; LEE; CÁRDENAS et al., 2007; CHU et al., 2011; NAIDOO et al., 2011; WANG et al., 2012; DEL TERRA et al., 2013; PATERAKI et al., 2015).

Em contrapartida aos efeitos positivos desses compostos, o cafestol foi associado ao aumento do colesterol sérico (ação hipercolesterolêmica) quando ingerido em altas doses, que depende do método de preparação da bebida (KURZROCK; SPEER, 2001; NAIDOO et al., 2011). Assim, a fração de diterpenos pode ser levada em conta pelos programas de melhoramento, e cultivares com baixos níveis de cafestol combinado com altos níveis de caveol são recomendados (CHU et al., 2011).

A torrefação não altera a maioria dos lipídeos do café, no entanto, esses compostos são encontrados em baixos teores na bebida final. Em preparações filtradas normais, por exemplo, há menos de 0.2% de lipídeos na bebida; para expressos fortemente torrados, os lipídeos representam de 1 – 2% (RANHEIM; HALVORSEN,

2005). Já em grãos de café arábica verde esses compostos variam entre 12 – 14.40% (OESTREICH-JANZEN, 2010; KITZBERGER et al., 2013a).

O conteúdo desses compostos varia de acordo com espécies e cultivares, como por exemplo, o Caveol é específico de *C. arabica*, e raramente está presente na fração lipídica de *C. canephora,* de modo que ambos são interessantes para a discriminação das mesmas (SPEER; KOLLING-SPEER, 2006; DIAS et al., 2010; KITZBERGER et al., 2013b). Em contraste aos outros compostos bioquímicos presentes nos grãos de café verde, as concentrações desses compostos são estáveis quanto aos anos de cultivo e são ambientalmente independentes (SCHOLZ et al., 2013).

3.5.4 Proteínas Totais

Apesar de contribuir para a formação do sabor da bebida, as proteínas totais encontradas nos grãos de café são ainda pouco estudadas, sendo relacionadas com o acúmulo de compostos nitrogenados no grão (ACUÑA et al., 1999; OESTREICH-JANZEN, 2010).

O teor de proteína no café verde é de cerca de 14.5 – 17% do peso seco (SCHOLZ et al., 2011). As proteínas se acumulam durante o desenvolvimento da semente e após a germinação vão sendo degradadas em aminoácidos (SHIMIZU & MAZZAFERA, 2000). Além disso, as proteínas e os aminoácidos livres nos cafés verdes são em grande parte transformados após a torrefação. Durante a torra, as proteínas que ainda estão estocadas nas sementes sofrem desnaturação, e com a hidrólise das ligações peptídicas são liberadas aminas e carbonilas, que contribuem para o aroma da bebida (FERNANDES et al., 2001).

Sua concentração na bebida final é cerca de 6 – 7%, um valor relevante para cálculos de valor nutricional (LELOUP, 2006; OESTREICH-JANZEN, 2010). A composição de proteínas nos grãos verdes varia de acordo com a espécie e/ou variedade do café, além do tempo de maturação do fruto, e tudo isso implica no teor de proteínas encontrado nos grãos torrados (FERNANDES et al., 2001).

3.5.5 Carboidratos Solúveis (Sacarose, Açúcares Totais e Redutores)

Os carboidratos solúveis e insolúveis (celulose e hemicelulose) são os principais constituintes do grão de café cru. No caso dos carboidratos solúveis, o teor é de 6 – 12.5%, e corresponde a monossacarídeos (frutose, glicose, galactose, arabinose),

oligossacarídeos (sacarose, rafinose, estaquiose) e polissacarídeos (polímeros de galactose, manose e arabinose) (ARYA et al., 2007; BELITZ; GROSCH; SCHIEBERLE, 2009).

A qualidade da bebida está principalmente relacionada ao conteúdo de sólidos solúveis, no caso sacarose e açúcares totais, no qual compõem os precursores de aroma e sabor. O conteúdo de sacarose depende do genótipo de café analisado, mas em cultivares de *C. arabica*, varia entre 7.4 – 11.1% (OESTREICH-JANZEN, 2010; TRAN et al., 2016).

A sacarose é uma das principais fontes de açúcares redutores livres que participam da reação de *Maillard*, que ocorre durante a torrefação dos grãos de café. Esta reação gera um número significativo de produtos, incluindo produtos caramelizados, aromas do tipo doce e tipo queimado, e cores escuras, que são tipicamente associadas ao sabor do café (HOLSCHER; STEINHART, 1995). Outros açúcares no grão verde, como glicose e frutose, representam apenas 0,5% do total e podem se degradar ou reagir com aminoácidos, contribuindo para o aroma e sabor da bebida (OESTREICH-JANZEN, 2010).

Durante a torrefação a sacarose se degrada rapidamente em compostos voláteis e não-voláteis através das reações de Maillard. Assim, açúcares, proteínas e aminoácidos reagem conjuntamente para formar aromáticos típicos e compostos aromatizantes do café torrado (SELMAR et al., 2008). Em *C. arabica*, o acúmulo de sacarose é constante nos estádios iniciais da sua transição perisperma-endosperma aumentando drasticamente a partir do estágio intermediário (fruto verde para amarelo), e posteriormente o acúmulo estabiliza em 225 DAF (Dias Após a Florada), permancendo constante até o estágio cereja (PRIVAT et al., 2008).

3.6 Fenotipagem por Espectroscopia no Infravermelho Próximo

O Infravermelho Próximo (NIR, do inglês, Near Infrared) é a denominação dada à região do espectro eletromagnético logo após a região visível, abrangendo a ampla região espectral (PASQUINI., 2003). Quando a radiação no infravermelho próximo incide em uma molécula, faz seus átomos vibrarem em maior amplitude. Os espectros NIR contêm informações relacionadas com a diferença entre as forças das ligações químicas, espécies químicas presentes e eletronegatividade. No entanto, somente

vibrações que resultam em uma mudança no momento de dipolo da molécula podem absorver radiação no NIR (DE BEER et al., 2011).

Assim o NIRS é uma alternativa de análise de compostos químicos de forma rápida e de baixo custo, quando se possui um modelo de calibração confiável, sendo obtida por meio de comparação com um método de referência (MORGANO et al., 2007).

A análise por NIRS é muito utilizada na discriminação entre as espécies de café, verificação de amostras puras e misturas, na definição do grau de torrefação dos grãos, avaliação das propriedades sensoriais, e para quantificar compostos químicos com base nos modelos de calibração (ESTEBAN-DIEZ; GONZÁLEZ-SÁIZ; PIZARRO, 2004; HUCK; GUGGENBICHLER; BONN, G.K, 2005; MORGANO et al., 2007; SCHOLZ et al., 2014).

3.7 GBS (GENOTIPAGEM POR SEQUENCIAMENTO)

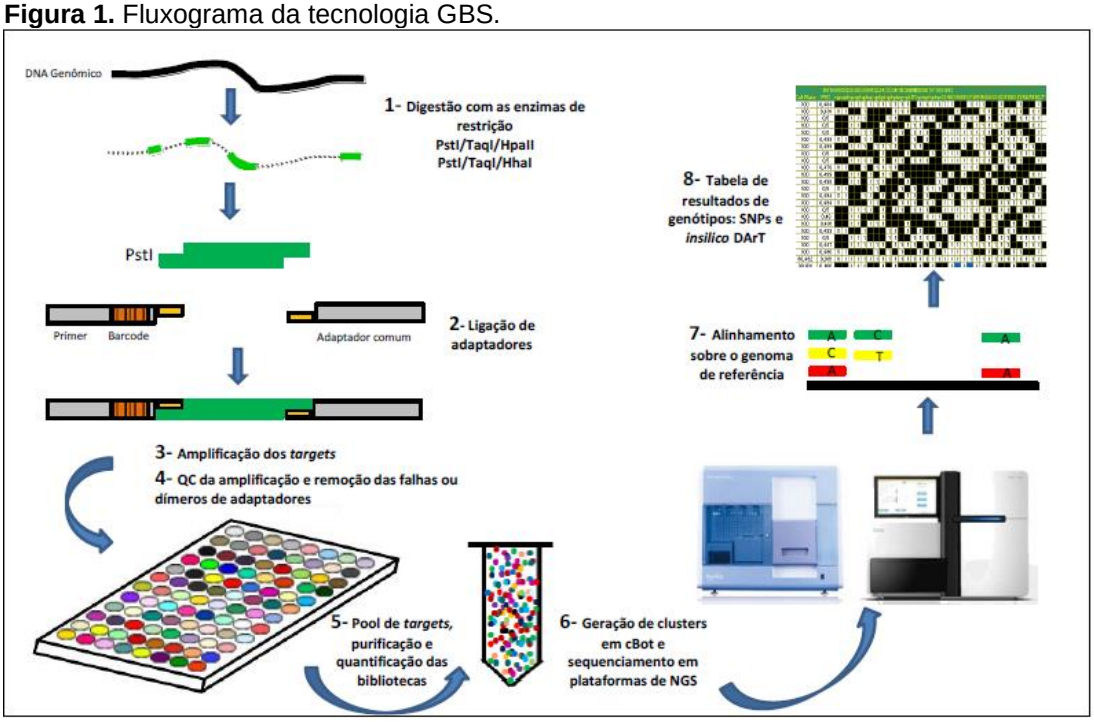
Os avanços nas tecnologias de nova geração (NGS) permitiram o resequenciamento do genoma de muitas espécies permitindo a descoberta e a caracterização de milhares de polimorfismos. Com isso houve um aumento na descoberta de marcadores do tipo SNPs em plantas-modelo e não-modelo, tais como arroz, milho, soja, feijão café e sorgo (POLAND et al., 2012).

A Genotipagem por Sequenciamento (do inglês, Genotyping by Sequence - GBS) foi desenvolvida como uma ferramenta para estudos de associação genômica ampla (GWAS) e permite sequenciamento e ao mesmo tempo genotipagem de um grande número de indivíduos incluindo aqueles com genomas complexos. POLAND et al. (2012) demonstraram para cevada e trigo a robustez da tecnologia GBS, que são espécies com grandes e complexos genomas.

A técnica GBS visa reduzir a complexidade de grandes genomas mas como um método simples e robusto (ELSHIRE et al., 2011). Essa redução da complexidade genômica consiste na utilização de enzimas sensíveis à metilação (regiões não expressas) e por isso, a tecnologia enriquece o genoma para regiões expressas permitindo uma representação reduzida do genoma. O desenvolvimento de bibliotecas GBS é bastante simplificado, pois requer pouca quantidade de DNA, evita corte aleatório e seleção por tamanho, e é concluído em apenas dois passos nas placas, seguido por amplificação por reação em cadeia da polimerase (PCR) da biblioteca (SANSALONI., 2012).

As principais etapas da metodologia GBS são apresentadas na Figura 1. Nesta imagem é possível observar a redução da complexidade genômica pelas enzimas sensíveis à metilação seguido da ligação dos adaptadores e códigos barras (*barcoding*) às sequências complementares ao sítio de reconhecimento da enzima *PstI*, amplificação dos *targets* por PCR, agrupamento (*pooling*) das amostras, sequenciamento dos *clusters* em plataformas NGS e finalmente a análise de dados (SANSALONI., 2012).

Figura 1. Fluxograma da tecnologia GBS.



Sansaloni., 2012.

3.8 O *PIPELINE* TASSEL

Com os avanços das tecnologias de sequenciamento de nova geração houve uma revolução na biologia molecular impulsionado por uma onda de seqüências com dados brutos. O *pipeline* TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) é uma ferramenta de bioinformática que é especificamente adaptada para o protocolo GBS de ELSHIRE et al. (2011) e POLAND et al. (2012) e permite o processamento eficiente de dados brutos das sequências GBS. Com esse *pipeline* centenas de milhares ou mesmo milhões de marcadores SNPs podem ser anotados em até 100.000 indivíduos. Além disso, é uma ferramenta robusta para estudo de diversidade genômica e possui habilidade para rodar em máquinas de pesquisa com 8 a 16 GB de RAM (GLAUBITZ et al., 2014).

Embora o *Pipeline* TASSEL-GBS tenha sido desenvolvido para espécies para os quais o genoma de referência seja necessário, no entanto, é também possível o uso de genomas incompletos consistindo de inúmeros *contigs* como uma “pseudo-referência” (GLAUBITZ et al., 2014).

Além disso, não se limita às enzimas de restrição específicas utilizadas em protocolos preliminares. Atualmente são aceitas 15 enzimas de restrição únicas e 15 pares de enzimas de restrição, como também novas enzimas são facilmente adicionadas. O principal objetivo do *Pipeline* é usar os dados das sequências cumulativas de todas as amostras geradas em uma população de melhoramento para descobrir marcadores SNPs (GLAUBITZ et al., 2014).

O número potencialmente grande de marcadores disponíveis pela tecnologia GBS tornou o GWAS viável em populações onde o DL se estende o suficiente para que os polimorfismos causadores tenham chances razoáveis de estarem em DL com um ou mais marcadores (GLAUBITZ et al., 2014).

3.9 Estudo de Associação Genômica Ampla (GWAS)

Durante anos o mapeamento de QTLs usando populações bi-parentais e multi-parentais tem sido empregado para a identificação de genes responsáveis pela variação genética natural de características de interesse. No entanto, o mapeamento de QTLs possui baixa resolução e requer muito tempo e recursos para construção de uma população segregante (WEIGEL, 2012).

Associação genética é um estudo multidisciplinar, que envolve componentes da genética, estatística, biologia molecular e bioinformática, formando a base para a identificação das regiões genômicas associadas com a variação (ORAGUZIE et al., 2007). O mapeamento por associação tem sido amplamente utilizado na dissecação de características complexas em todos os sistemas e se tornou uma das mais efetivas abordagens para a mineração de alelos favoráveis (SU et al., 2016).

O conceito de desequilíbrio de ligação refere-se à associação não aleatória entre alelos de diferentes *loci* ou entre um QTL e um locus marcador. O GWAS também referido como mapeamento por desequilíbrio de ligação (DL) é babseada no uso do DL quando *loci* no genoma são herdados conjuntamente como resultado da história compartilhada de mutações e recombinações (alelos derivados de múltiplos fundadores), para entender os efeitos genéticos sobre fenótipos específicos (ZHU, 2008).

O GWAS supera várias desvantagens do mapeamento de QTLs (mapeamento de ligação) oferecendo maior resolução e considerando uma diversidade alélica maior, é menos demorado e requer menos recursos (ZHU et al. 2008, KORTE e FARLOW, 2013). No entanto o GWAS possui algumas limitações, como por exemplo, é uma abordagem efetiva de associação genótipo-fenótipo quando a informação sobre a estrutura populacional e o DL está disponível; requer grandes dimensões populacionais; pode gerar uma inflação de falsos positivos como resultado da estrutura populacional e parentesco; possui baixo poder estatístico para identificar alelos raros; e tem dificuldades na dissecação de características complexas (variantes raras de grandes efeitos ou variantes comuns de pequenos efeitos) (ZHU et al., 2008; KORTE et al., 2012). Porém, ambas estratégias relatadas se complementam levando a um maior poder na detecção de uma variação genética causal (KORTE et al., 2012).

O GWAS foi introduzido em plantas, especificamente no milho no início do século 21 (THORNSBERRY et al., 2001) e a partir daí tem sido aplicado com sucesso em inúmeros estudos em culturas de importância econômica como soja, arroz, trigo, sorgo e cevada (BELO´ et al., 2008; ZHOU et al., 2015; UPADHYAYA et al., 2015; LIU et al., 2016; YANO et al.,2016; LI et al., 2016; MUQADDASI et al., 2016; MAURER et al., 2016). Em café, nosso grupo realizou o primeiro trabalho de GWAS para o conteúdo de Lipídeos Totais, Caveol, Cafestol e Razão Caf/CAv, onde foram encontrados 21 SNPs associados com o conteúdo desses compostos químicos e 9 genes candidatos relacionados às vias metabólicas dos lipídeos e diterpenos (SANT'ANA et al., 2018).

Os SNPs associados com as caractísticas de interesse obtidos por GWAS podem acelerar o processo de melhoramento através da seleção assistida por marcadores (SAM) ou podem ser incorporados em estratégias de seleção genômica (SG). Além disso, os SNPs significativamente associados podem auxiliar no conhecimento sobre a função biológica dos polimorfismos e como se relacionam com determinada característica fenotípica (GUPTA et al., 2005).

4 REFERÊNCIAS BIBLIOGRÁFICAS

ASSOCIAÇÃO BRASILEIRA DA INDÚSTRIA DE CAFÉ, 2017. < http://abic.com.br/consumo-de-cafe-especial-aumentou-em-ate-15-em-2017- diz associacao-brasileira-de-cafes-especiais/>. Acesso em 16 de junho de 2018.

AERTS, R.J.; BAUMANN, T.W. Distribution and utilization of chlorogenic acid in Coffea seedlings. **Journal of Experimental Botany**, 45: 497-503, 1994.

ALI, M. M. et al. Determination of caffeine in some Sudanese beverages by High Performance Liquid Chromatography. **Pakistan Journal of Nutrition**, v. 11, n. 4, p. 336-342, 2012.

ALVES, R. C.; CASAL, S.; OLIVEIRA, B. Benefícios do café na saúde: Mito ou realidade?. **Química Nova**, 32, 2169, 2009.

ANTHONY, F.; COMBES, M.C.; ASTORGA, C.; BERTRAND, B.; GRAZIOSI, G.; LASHERMES, P. The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. **Theoretical and Applied Genetics**, v.104, p.894-900, 2002.

ARYA, M.; RAO, J.M. **An impression of coffe carbohydrates**. Critical Reviews in Food Science and Nutrition, 47, 51, 2007.

BANERJEE, S.; DAS, M.; MIR, R et al. Assessment of genetic diversity and population structure in a selected germplasm collection of 292 jute genotypes by microsatellite (SSR) markers. **Mol. Plant. Breed**, 3:11–25, 2012.

BELO´, A.; ZHENG, P.; LUCK, S.; SHEN, B., MEYER, D.; LI, B.; TINGEY, S AND RAFALSKI, A. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. **Mol. Genet. Genomics**, 279:1–10,2, 2008.

BRITO, G.G.; TEIXEIRA, E.; GALLINA, A.P.; ZAMBOLIM, E.M.; ZAMBOLIM, L.; DIOLA, V.; LOUREIRO, M.E. Inheritance of the coffee leaf rust resistance and identification of AFLP markers linked to the resistance gene. **Euphytica**, 173:255-264, 2010.

BAUMANN, T. W. Biochemical Ecology. **In Espresso Coffee**: The Science of Quality; Illy, A., Viani, R., Eds, Elsevier: Amsterdam, p. 58, 2005.

BELAY, A.; TURE, K.; REDI, M.; ASFAW, A. Measurement of caffeine in coffee beans with UV/vis spectrometer. **Food Chemistry**, 108, 310–315, 2008.

BELITZ, H.D.; GROSCH, W.; SCHIEBERLE, P. **Food Chemistry**, 4a. ed., Springer: Berlin, 2009.

BELÓ A.; ZHENG P.; LUCK S.; SHEN B.; MEYER D., et al. Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. **Mol. Genet. Genomics**, 279: 1–10, 2008.

BOEKSCHOTEN, M. V.; SCHOUTEN, E. G.; KATAN, M. B. Coffee bean extracts rich and poor in kahweol both give rise to elevation of liver enzymes in healthy volunteers. v. 8, p. 1–8, 2004.

BORÉM, A.; FRITSHE-NETO, R. **Biotecnologia aplicada ao melhoramento de plantas**. Viçosa: UFV, 2013.

BOSTEIN, D.; WHITE, R.L.; SKOLNICK, M.; DAVIS, R.W. Construction of a genetic linkage map in man using restriction fragment lenght polymorphisms. **American. Journal. Human. Genetics**, Chicago, v.32, n.3, p.314-331, 1980.

CAMPA, C.; DOULBEAU, S.; DUSSERT, S.; HAMON, S.; NOIROT, M. Qualitative relationship between caffeine and chlorogenic acid contents among wild *Coffea* species. **Food. chemistry** 93(1):135-139, 2005.

CAMPANHA, F. G. D**iscriminação de espécies de café (Coffea arabica e Coffea canephora) pela composição de diterpenos**. 91 f. Dissertação (Mestrado em Ciência de Alimentos) – Universidade Estadual de Londrina, Londrina. 2008.

CARVALHO, A.; TANGO, J.S.; MONACO, L.C. Genetic control of caffeine in coffee. **Nature**, 205: 314, 1965.

CAVIN, C. et al. The coffee-specific diterpenes cafestol and kahweol protect against aflatoxin B1-induced genotoxicity through a dual mechanism. **Carcinogenesis**, v. 19, n. 8, p. 1369-1375, 1998.

CECAFÉ. Consumo de café especial aumentou em até 15% em 2017. **Relatório mensal março de 2018**. Disponível <http://www.cecafe.com.br/> Acesso em 15 de abril de 2018.

CHAPARRO, A. P.; CRISTANCHO, M. A.; CORTINA, H. A AND GAITAN, A. L. Genetic variability of Coffea arabica L. accessions from Ethiopia evaluated with RAPDs. **Genetic. Resources. Crop. Evolution**, 51:291–297, 2004.

CHENG, B.; FURTADO, A.; SMYTH, H. E AND HENRY, R. J. Influence of genotype and environment on coffee quality. **Trends. Food. Sci. Technol***.* 57, 20–30, 2016.

CHU, Y. F.; CHEN, Y.; BROWN, P. H.; LYLE, B. J.; BLACK, R. M.; CHENG, I. H & PRIOR, R. L. Bioactivities of crude caffeine: Antioxidant activity, cyclooxygenase-2 inhibition, and enhanced glucose uptake. **Food. Chemistry**, 131(2), 564-568, 2011.

CONAB – Companhia Nacional de Abastecimento. Acompanhamento da safra brasileira - safra 2018 - N. 8. Levantamento março, 2018.

DAVIS, A.P.; TOSH, J.; RUCH, N.; FAY, M.F. Growing coffee: Psilanthusb (*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, 2011.

DE BEER, T.; BURGGRAEVE, A.; FONTEYNE, M.; SAERENS, L.; REMON, J.P.; VERVAET, C. Near infrared and Raman spectroscopy for the in-process monitoring of pharmaceutical production processes. **Int. J. Pharm**, 417, 32–47, 2011.

DEL TERRA, L.; LONZARICH, V.; ASQUINI, E.; NAVARINI, L.; GRAZIOSI, G.; SUGGI LIVERANI, F & PALLAVICINI, A. Functional characterization of three *Coffea Arabica* L. monoterpene synthases: Insights into the enzymatic machinery of coffee aroma. **Phytochemistry**, *89*, 6-14, 2013.

DIAS, R. C. E.; CAMPANHA, F. G.; VIEIRA, L. G. E.; FERREIRA, L. P. F.; POT, D.; MARRACCINI, P.; BENASSI, M. T. Evaluation of kahweol and cafestol in coffee tissues and roasted coffee by a new high-performance liquid chromatography methodology. **Journal of Agricultural and Food Chemistry**, Easton, v. 58, n. 1, p. 88-93, 2010.

DURÁN, C. A. A. et al. **Café: aspectos gerais e seu aproveitamento para além da bebida**. Revista Virtual de Química. Rio de Janeiro, RJ. v. 9, n. 1, p. 107-134, Nov. 2016. Disponível em: Acesso em 24 de maio de 2018.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS One,** 6, e1937910, 2011.

ESQUIVEL, P.; JIMÉNEZ, V. M. Functional properties of coffee and coffee by-products. **Food Research International**, v. 46, n. 2, p. 488–495, 2012.

ESTEBAN-DÍEZ, I.; GONZÁLEZ-SÁIZ, J. M.; PIZARRO, C. **Anal. Chim. Acta**, *514*, 57, 2004.

FARAH, A.; DONANGELO, C.M. Phenolic compounds in coffee. **Brazilian Journal of Plant Physiology**, v.18, p.23-36, 2006.

FERREIRA, M.E.; GRATTAPAGLIA, D. **Introdução ao Uso de Marcadores Moleculares em Análises Genéticas**. 2nd edition. Brasília: Embrapa-Cnargen, 1998.

FRIDELL, G. **Coffee**. John Wiley & Sons, 2014.

GLADE, M. **Caffeine**. Not just a stimulant. Nutrition, 26, 932–938, 2010.

GLAUBITZ, J. C. et al. TASSEL-GBS: A high capacity Genotyping-by-Sequencing analysis pipeline. **PLoS One**, 9, e90346, 2014.

GIANCOLA, S; MCKHANN, H.I; BERARD, A et al. Utilization of the three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and TaqMan, in diploid and polyploid plants. **Theoretical and Applied Genetics**, 112, 1115–1124, 2006.

GRFFITHS, W.C.; GREGORY, M.P.; MILLER, J.H.; SUZUKI, D.T.; LEWONTIN, R.C.; GELBART, W.M. **An introduction to Genetic Analysis**. 7 ed. New York: W.H Freeman & Co. 2000.

GONZÁLEZ, A. G.; PABLOS, F.; MARTÍN, M. J.; LEÓNCAMACHO, M.; VALDENEBRO, M. S. HPLC analysis of tocopherols and triglycerides in coffee and their use as authentication parameters. **Food. Chemistry**, 73, 93-101., 2001.

GUPTA, P.K.; RUSTGI, S.; KULWAL, P.L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. **Plant. Mol. Biol**, 57: 461–485, 2005.

HOLSCHER, W.; STEINHART, H. Aroma compounds in green coffee. In: Charalambous G, ed. Food flavors: generation, analysis and process influence. Amsterdam, the Netherlands: **Elsevier Science**, 785–803, 1995.

HUCK, C.W.; GUGGENBICHLER, W.; BONN, G.K. Analysis of caffeine, theobromine and theophylline in coffee by near infrared spectroscopy (NIRS) compared to high-performance liquid chromatography (HPLC) coupled to mass spectrometry" **Anal. Chim. Acta**, 538, 195, 2005.

I.Y. CHOI; D.L. HYTEN.; L.K. MATUKUMALLI.; Q. SONG.; J.M. CHAKY.; C.V. QUIGLEY.; K. CHASE.; K.G. LARK.; R.S. REITER.; M.S. YOON, et al. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis, **Genetics**, 176 (1); 685–696, 2007.

JEFFREYS, A.J.; WILSON, V.; THEIN, S.L. Hypervariable 'minisatelite' regions in human DNA. **Nature**, 316:76-79, 1985.

JESZKA-SKOWRON, M.; SENTKOWSKA, A.; PYRZYNSKA, K.; MARIA PAZ DE PEÑA, M. Chlorogenic acids, caffeine content and antioxidant properties of green coffee extracts: influence of green coffee bean preparation. **Eur. Food. Res. Technol**, 242:1403–1409, 2016.

JOËT, T.; LAFFARGUE, A.; SALMONA, J.; DOULBEAU, S.; DESCROIX, F.; BERTRAND, B.; DE KOCHKO, A.; DUSSERT, S. Metabolic pathways in tropical dicotyledonous albuminous seeds: Coffea arabica as a case study. **New. Phytol**, 182:146–162, 2009.

JOËT, T.; LAFFARGUE, A.; DESCROIX F.; DOULBEAU, S.; BERTRAND, B.; KOCHKO, A.; DUSSERT, S. Influence of environmental factors, wet processing and their interactions on the biochemical composition of green coffee beans. **Food. Chem**, 118:693–701, 2010.

KAUR, S.; PANESAR, P.S.; BERA, M.B.; KAUR, V. Simple sequence repeat markers in genetic divergence and marker-Assisted selection of rice cultivars: a review. **Crit. Ver. Food. Sci. Nutr**, 55:41–49, 2015.

KITZBERGER, C.S.G.; SCHOLZ, M.B.S.; PEREIRA, L.F.P.; BENASSI, M.T. Composição química de cafés arábica de cultivares tradicionais e modernas. **Pesquisa Agropecuária Brasileira**, 48:1498–1506, 2013a.

KITZBERGER, C.S.G.; SCHOLZ, M.B.S.; PEREIRA, L.F.P.; VIEIRA, L.G.E.; SERA, T.; SILVA, J.B.G.D.; BENASSI, M.T. Diterpenes in green and roasted coffee of Coffea arabica cultivars growing in the same edapho-climatic conditions. **J. Food. Compos. Anal**, 2013b.

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

KY, C.L.; BARRE, P.; NOIROT, M. Genetic investigations on the caffeine and chlorogenic acid relationship in an interspecific cross between *Coffea liberica dewevrei* and *C. pseudozanguebariae*. **Tree Genet Genomes**, 9:1043–1049, 2013.

KURZROCK, T.; SPEER, K. **Diterpenes and diterpenes esters in coffee**. Food Reviews International, New York, v. 17, n. 4, p. 433-450, 2001.

LEE, K. J.; JEONG, H. G. Protective effects of kahweol and cafestol against hydrogen peroxide-induced oxidative stress and DNA damage. **Toxicology letters**, v. 173, n. 2, p. 80-87, 2007.

LEONFORTE, A.; SUDHEESH.; COGAN, N.; SALISBURY, P.; NICOLAS, M; MATERNE, M.; FORSTER, J.; KAUR, J. SNP marker discovery, linkage map construction and identification of QTLs for enhanced salinity tolerance in field pea (Pisum sativum L.). **BMC. Plant. Biol**, 13 (1); 1–14, 2013.

LEITZMANN, M.F. Coffee intake is associated with lower risk of symptomatic gallstone disease in women. **Gastroenterology**, 123:1832-1830, 2002.

LI, Y et al. Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. **Plant. J***.* 86, 391–402, 2016.

LITT, M.; LUTY, J. A. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. **Am. J. Hum. Genet**, 44, 397-401, 1989.

LIU H.; CHE Z.; ZENG X.; ZHANG G.; WANG H.; YU D. Identification of single nucleotide polymorphisms in soybean associated with resistance to common cutworm (*Spodoptera litura* Fabricius). **Euphytica**, 209, 49–62, 2016.

MAURER, A.; DARBA, V.; PILLEN, K. Genomic dissection of plant development and its impact on thousand grain weight in barley through nested association mapping. **J. Exp. Bot**, 67, 2507–2518, 2016.

MAZZAFERA, P.; YAMAOKA-YANO, D.M.; VITÓRIA, A.P. Para que serve a cafeína em planta. **Revista Brasileira de Fisiologia Vegetal**, Brasília, v.8, n.1, p.67-74, 1996.

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

MORGANO, M. A.; FARIA, C. G.; FERRÃO, M. F.; FERREIRA, M. M. C. Determinação de açúcar total em café cru por espectroscopia no infravermelho próximo e regressão por mínimos quadrados parciais. **Química Nova**, v. 30, n. 2, p. 346-350, 2007.

MULLEN, W.; NEMZER, B.; STALMACH, A.; ALI, S.; COMBET, E. Polyphenolic and Hydroxycinnamate Contents of Whole Coffee Fruits from China, India, and Mexico Phenolic compounds. **Journal of Agricultural and Food Chemistry**, 61, 5298, 2013.

MUQADDASI, Q. H. et al. Genome-wide association mapping of anther extrusion in hexaploid spring wheat. **PLoS One**, 11, 2016.

NAIDOO, N.; CHEN, C.; REBELLO, S. A.; SPEER, K.; TAI, E. S.; LEE, J & VAN DAM, R. M. Cholesterol-raising diterpenes in types of coffee commonly consumed in Singapore, Indonesia and India and associations with blood lipids: A survey and cross sectional study. **Nutr. J**, *10*, 48, 2011.

NATHANSON, J.A. Caffeine and related methylxanthines – possible naturally- occurring pesticides. **Science**, 226, 184–187, 1984.

OESTREICH-JANZEN, S. **Chemistry of coffee**. Comprehensive Natural Products, II, 1085e1117, 2010.

ORAGUZIE, N.C.; RIKKERINK E.H.A.; GARDINER S.E.; DE SILVA H.N. **Association mapping in plants**. Springer, New York, p. 277, 2007.

PACHECO, A.; POHLAN, J.; SCHULZ, M. Allelopathic effects of aromatic species intercropped with coffee: investigation of their growth stimulation capacity and potential of caffeine uptake in Puebla, Mexico. **Allelopathy. J**, 21, 39–56, 2008.

PARAN, I.; MICHELMORE, R.W. Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. **Theor. Appl. Genet**, 85:985-993, 1993.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. Campinas-SP: **Journal of the Brazilian Chemical Society**, v. 14, p.198-219, 2003.

PATERAKI, I.; HESKES, A. M.; HAMBERGER, B. Cytochromes P450 for Terpene Functionalisation and Metabolic Engineering. **Advances in biochemical engineering/ biotechnology**, v. 123, n. July 2015, p. 127–141, 2015.

PEREIRA, L. F. P & IVAMOTO, S. T. Chapter 6: **Characterization of coffee genes involved in isoprenoid and diterpene metabolic pathways**. In: *Coffee in Health and Disease Prevention* (Preedy, R. V. Ed.). London: Academic Press, 45-51, 2015.

PÍPOLO, V.C.; GARCIA, J.C. **Biotecnologia aplicada ao melhoramento genético**. In: Biotecnologia na agricultura: aplicações e biossegurança. Cascavel: COODETEC, 2006.

POT, D.; SCHOLZ M.B.S.; L, S. D.; DEL GROSSI, L.; PERREIRA, L.F.P.; VIEIRA, L.G.; SERA, T. **Phenotypic analysis of *Coffea arabica* accessions from Ethiopia: Contribution to the undestanding of *Coffea arabica* diversity**. 2008. In: 22nd International Conference on Coffee Science, 14-19 september 2008, Campinas, Brasil. ASIC. Campinas: ASIC, 1 p.

PRIVAT, I.; FOUCRIER, S.; PRINS, A.; EPALLE, T.; EYCHENNE, M.; KANDALAFT, L.; CAILLET, V.; LIN, C.; TANKSLEY, S.; FOYER ,C et al. Differential regulation of grain sucrose accumulation and metabolism in *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta) revealed through gene expression and enzyme activity analysis. **New Phytologist**, 178: 781–797, 2008.

POLAND, J.A.; BROWN, P.J.; SORRELLS, M.E.; JANNINK, J.L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **Plos One**, 7(2): e32253, 2012.

RANHEIM, T.; HALVORSEN, B. Coffee consumption and human health - beneficial or detrimental? - Mechanisms for effects of coffee consumption on different risk factors for cardiovascular disease and type 2 diabetes mellitus. **Molecular Nutrition & Food Research**, v. 49, p. 274-284, 2005.

SALLES, G. et al. Marcadores Microssatélites em Espécies Vegetais, 2003.

SANSALONI, Carolina Paola. **Desenvolvimento e aplicações de DArT (Diversity Arrays Technology) e genotipagem por sequenciamento (Genotyping-by-Sequencing) para análise genética em eucalyptus**. 2012. xiii, 100 f., il. Tese (Doutorado em Biologia Molecular)—Universidade de Brasília, Brasília, 2012.

SANT'ANA, GUSTAVO C.; PEREIRA, LUIZ F. P.; POT, DAVID.; IVAMOTO, SUZANA T.; DOMINGUES, DOUGLAS S.; FERREIRA, RAFAELLE V.; PAGIATTO, NATALIA F.; DA SILVA, BRUNA S. R.; NOGUEIRA, LÍVIA M.; KITZBERGER, CINTIA S. G.; SCHOLZ, MARIA B. S.; DE OLIVEIRA, FERNANDA F.; SERA, GUSTAVO H.; PADILHA, LILIAN.; LABOUISSE, JEAN-PIERRE; GUYOT, ROMAIN; CHARMETANT, PIERRE; LEROY, THIERRY. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica L*. **Sci. Reports**, v. 8, p. 465, 2018.

SCHOLZ, M.B.S.; KITZBERGER, C.S.G.; PEREIRA, L.F.P.; DAVRIEUX, F.; POT, D; CHARMETANT, P; LEROY, T. Application of near infrared spectroscopy for green coffee biochemical phenotyping. **JNear Infrared Spectrosc**, 22:411–421, 2014.

SCHOLZ, M. B. S.; SILVA, J. V. N.; FIGUEIREDO, V. R. G.; KITZBERGER, C. S. G. **Atributos sensoriais e características físico-químicas de bebida de cultivares de café do IAPAR**. Coffee Science, Lavras, v. 8, n. 1, p. 6-16, 2013.

SCHOLZ, M. B. S.; GOOD-KITZBERGER, C. S.; PAGIATTO, N. F.; PROTASIO, P. L. F.; DAVRIEUX, F.; POT, D.; CHARMETANT, P.; LEROY, T. Chemical composition in wild Ethiopian Arabica coffee accessions. **Euphytica**, 10 p, 2016.

SELMAR, D.; BYTOF, G.; KNOPP, S. The storage of green coffee (*Coffea arabica*): decrease of viability and changes of potential aroma precursors. **Ann Bot**, 101:31–38, 2008 SILVAROLLA, M.B; MAZZAFERA, P; LIMA, M.M.A. Caffeine content of Ethiopian *Coffee arabica* beans. **Genet. Mol. Biol**, 23:213–215, 2000.

SELMAR, D.; BYTOF, G & KNOPP, S. E. The storage of green coffee (*Coffea arabica* L.): Decrease of viability and changes of potential aroma precursors. *Ann. Bot.* 101, 31–38, 2008.

SILVAROLLA, M.B.; MAZZAFERA, P.; FAZUOLI, L.C. Plant biochemistry: a naturally decaffeinated Arabica coffee. **Nature**, 429:826, 2004.

SILVESTRINI, S.; JUNQUEIRA, M.G.; FAVARIN, A.C.; GUERREIRO-FILHO, O.; MALUF, M.P.; SILVAROLLA, M.B.; COLOMBO C.A. Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. **Genet. Resour. Crop. Evol**, 10722-006-9122-4, 2007.

SIN, C. W. M.; HO, J. S. C.; CHUNG, J. W. Y. Systematic review on the effectiveness of caffeine abstinence on the quality of sleep. **Journal of Clinical Nursing**, v. 18, n. 01, p. 1321, 2009.

SOUZA, T.V.; CAIXETA, E.T.; ALKIMIM, E.R.; OLIVEIRA, A.C.B.; PEREIRA, A.A.; ZAMBOLIM, L.; SAKIYAMA, N. S. Molecular markers useful to discriminate *Coffea arabica* cultivars with high genetic similarity. **Euphytica**, 213:75, 2017.

SPEER, K.; KÖLLING-SPEER, I. The lipid fraction of the coffee bean—mini-review. Braz **J. Plant. Physiol**, 18:201–216, 2006.

SPEER, K.; SEHAT, N.; MONTAG, A. Fatty acids in coffee. In: Proceedings of the 15th ASIC Colloquium, Montpellier. Association for Science and Information on Coffee (ASIC); pp. 583–92, 1993.

SU, J. et al. Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. **BMC Genomics**, 17, 687,2016.

THORNSBERRY, J.; GOODMAN, M., DOEBLEY, J.; KRESOVICH, S.; NIELSEN, D.; AND BUCKLER, E. Dwarf8 polymorphisms associate with variation in flowering time. **Nat. Genet**. 28:286–289, 2001.

TRAN, H.; SLADE LEE, L.; FURTADO, A.; SMYTH, H.; HENRY, R. **Advances in genomics for the improvement of quality in Coffee.** Journal of the Science of Food and Agriculture, 96, p. 3310–3312, 2016.

TVERDAL, A.; SKUTVEIT, S. Coffee intake and mortality from liver cirrhosis. **Ann. Epidemiology**, 13(6):419-423, 2003.

UEFUJI, H.; YAMAGUCHI, Y.; HOIZUMU, N.; SANO, H. Molecular cloning and functional characterization of three distinct N-metiltransferases involved in the caffeine biosynthetic pathway in coffee plants. **Plant Physiology**. V. 132, p. 372-380, 2003

UPADHYAY, L.J.; MOHAN RAO. **An outlook on chlorogenic acids—occurrence, chemistry, technology, and biological activities.** Critical Reviews in Food Science and Nutrition, 53 (9), p. 968–984, 2013.

VEGA, F.E.; ROSENQUIST, E.; COLLINS, W. Global project needed to tackle coffee crisis. **Nature**, 425. 343-343, 2003.

VOS, P.; HOGERS, R.; BLEEKER, M.; REIJANS, M.; VAN DE LEE, T.; HORNES, M.; FRIJTERS, A.; POT, J.; PELEMAN, J.; KUIPER, M.; ZABEAU, M. AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Research**, v.23, p.4407-4414, 1995.

WANG, S. et al. Antiangiogenic properties of cafestol, a coffee diterpene , in human umbilical vein endothelial cells. **Biochemical and Biophysical Research Communications**, v. 421, n. 3, p. 567–571, 2012.

WEIGEL D. Natural variation in Arabidopsis: from molecular genetics to ecological genomics. **Plant. Physiol**, 158:2–22, 2012.

WILLIAMS, J.G.K. Polimorphisms amplified by arbitrary primeres are usefull as genetic markers. **Nucleic Acids Research**, Oxford, v.18, p.6531-6535, 1990.

WINTGENS, J. N. COFFEE: **Growing, processing, sustainable production**. A guidebook for growers, processors, traders and researchers (2nd ed). Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2012.

WRIGHT, G.A.; BAKER, D.D.; PALMER, M.J.; STABLER, D.; MUSTARD, J.A.; POWER, E.F.; BORLAND, A.M.; STEVENSON, P.C. Caffeine in floral nectar enhances a pollinator's memory of reward. **Science**, 339, 1202–1204, 2013.

YANO, K. et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. **Nat. Genet**. 48, 927–934, 2016.

YU, Q.; GUYOT. R.; DE KOCHKO. A.; BYERS. A.; NAVAJAS-PÉREZ, R.; LANGSTON, B.J.; DUBREUIL-TRANCHANT, C.; PATERSON, A.H.; PONCET, V.; NAGAI, C.; MING, R. Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*). **Plant. J**. 67:305–317, 2011.

ZHOU, Z.; JIANG, Y.; WANG, Z.; GOU, Z.; LYU, J.; LI, W et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. **Nat. Biotechnol**. 33, 408–414, 2015.

5 CAPÍTULO 1: Population Structure and Genetics Relationships between Ethiopian and Brazilian *Coffea arabica* Genotypes Revealed by SSRs.

ABSTRACT

Information about population structure and genetic relationships within and among wild and cultivate *Coffea arabica* L. genotypes is highly relevant to optimize the use of genetic resources for breeding purposes. In this study, we evaluated genetic diversity and population structure in 34 genotypes of *C. arabica* and of three diploid *Coffea* species (*C. canephora*, *C. eugenioides* and *C. racemosa*) using 30 SSR markers. A total of 206 alleles were identified. The set of SSR markers was able to discriminate all genotypes and revealed that Ethiopian accessions presented higher genetic diversity than commercial varieties. Population structure analysis indicated two genetic groups, one corresponding to Ethiopian Arabica accessions and another corresponding predominantly to commercial cultivars. Thirty private alleles were detected in the group of accessions collected from West side of Great Rift Valley. We observed a lower average genetic distance of *C. arabica* genotypes in relation to *C. eugenioides* than to *C. canephora*. Interestingly, commercial cultivars were genetically closer to *C. eugenioides* than *C. canephora* and *C. racemosa*. The great allelic richness observed in Ethiopian Arabica coffee, especially in West group showed that these accessions may be potential source of new alleles to be explored by coffee breeding programs.

Key words: Coffea spp, *Coffea arabica*, SSR markers, Population structure, cultivated and wild gene pools.

6 INTRODUCTION

Coffee is one of the most important agricultural commodities in tropical countries. More than 90% of its production occurs in developing countries providing an income for millions of families around the world that are dependent of coffee for their subsistence (Tran et al., 2016). *Coffea L*. genus belong to the *Rubiaceae* family and encompass 124 species, but only 10 are cultivated (Davis et al., 2011). *Coffea arabica* L. and *C. canephora* P. are the two most commercially relevant with approximately 70% and 30% of global production, respectively (Meyer, 1968). *C. arabica* produces a high-quality beverage, with pleasant aroma and flavor, but diseases and pests as well as abiotic stresses often affect its yield (Tran et al., 2016).

Coffee breeding programs invested intense efforts to develop cultivars with high productivity, biotic and abiotic stresses tolerance, and high biochemical quality of the beans (Tran et al. 2016). However, several factors are limiting the genetic gains in breeding programs (Pestana et al., 2015; Vieira et al., 2010). Commercial coffee plants originate from a limited number of cultivars, mainly Typica and Bourbon types and as a consequence only narrow genetic base is available to support breeding programs. In addition, the reproductive behavior of *Coffea arabica* (i.e. autogamy) also contributes to the narrow genetic diversity available in this species (Anthony et al., 2002). As expected, several studies based on molecular markers demonstrated the low genetic variability available among commercial *C. arabica* varieties (Silvestrini et al., 2007; Setotaw et al., 2013; Pestana et al., 2015; Vieira et al., 2010).

The origin center of *C. arabica* is located in the highlands of southwestern Ethiopia (Meyer, 1968). Studies report wide agronomic diversity of Arabica coffee accessions collected in this region regarding leaf size, height, biotic and abiotic stresses tolerance, and productivity (Bertrand et al., 2005; Tran et al., 2017). In addition, studies using

molecular markers indicated the presence of higher genetic variability of Ethiopian accessions in comparison with cultivars, demonstrating the potential of these accessions for breeding purposes (Silvestrini et al., 2007; López-Gartner et al., 2009; Teressa et al., 2010; Aerts et al., 2013; Sant'Ana et al., 2018). These accessions also showed great variability for caffeine, chlorogenic acids, lipids, sucrose and diterpenes contents of coffee beans (Scholz et al., 2016; Sant'Ana et al., 2018).

The knowledge about population structure and genetic relationships of these Ethiopian accessions, among themselves and in relation to traditional cultivars, is fundamental for efficient use of this material in Arabica coffee breeding programs. Simple Sequence Repeats (SSR) markers have been widely used to analyze genetic relationships among cultivated and wild populations (Lashermes et al., 1995;1999; Missio et al., 2011; Chaparro et al., 2004; Silvestrini et al., 2007; Teressa et al., 2010; Aerts et al., 2013; Motta et al., 2014), due to the technique simplicity, speed, great resolving power, high levels of polymorphism and codominance. In addition, they are evenly dispersed across the genomes enabling accurate discrimination even between genetically related individuals (Vieira et al., 2010).

We analyzed the population structure and genetic relationships of a *C. arabica* panel, including wild genotypes from the primary center of origin of the species (Ethiopia) and commercial varieties in order to evaluate the allelic richness of Ethiopian accessions for breeding purposes. In addition, a comparative analysis of genetic distances among *C. arabica* accessions and their two ancestral diploids, *C. canephora* and *C. eugenioides*, was carried out.

7 MATERIAL AND METHODS

7.1 PLANT MATERIAL

A total of 37 *Coffea* genotypes were analyzed, including 34 *C. arabica* genotypes (twenty-six from the Ethiopian collection, four cultivars (Typica, Bourbon, Iapar 59 and Icatu x Catuaí) and four lines developed by the breeding programs of IAPAR (L1C1, L3C3, 90-3-1 and 90-8-1); and 3 genotypes from different diploid *Coffea* species (*C. canephora*, *C. eugenioides*, and *C. racemosa*) (Table I).

All plants are cultivated at the Londrina experimental station at Agronomic Institute of Parana (IAPAR), Brazil (23°23’00”S and 51°11’30”W). The FAO collection at IAPAR comes from open pollinated seeds from the original collection at CATIE (Costa Rica) introduced in Brazil in 1976, and transferred from Instituto Agronômico de Campinas (IAC) to IAPAR. These accessions were collected in different Ethiopian regions. Twenty accessions were collected from the West side of the Great Rift Valley (Kaffa, Kama, Illubador, Gojjam and Shoa provinces), and six from the East side of the Great Rift Valley (Sidamo and Harar provinces).

Table I – List of 37 *Coffea* genotypes analyzed in this study.

Genetic Materials	Country	Region	Accession origin [a]	Specie
E044	Ethiopia	Kaffa	West	*Coffea arabica*
E516	Ethiopia	Kaffa	West	*Coffea arabica*
E130	Ethiopia	Kaffa	West	*Coffea arabica*
E131	Ethiopia	Kaffa	West	*Coffea arabica*
E335	Ethiopia	Kaffa	West	*Coffea arabica*
E332	Ethiopia	Kaffa	West	*Coffea arabica*
E272	Ethiopia	Kaffa	West	*Coffea arabica*
E383	Ethiopia	Kaffa	West	*Coffea arabica*
E148	Ethiopia	Illubador	West	*Coffea arabica*
E208	Ethiopia	Illubador	West	*Coffea arabica*
E370	Ethiopia	Illubador	West	*Coffea arabica*
E363	Ethiopia	Illubador	West	*Coffea arabica*
E196	Ethiopia	Illubador	West	*Coffea arabica*
E454	Ethiopia	Illubador	West	*Coffea arabica*
E087	Ethiopia	Illubador	West	*Coffea arabica*
E464	Ethiopia	Illubador	West	*Coffea arabica*
E123A	Ethiopia	Kama	West	*Coffea arabica*
E123B	Ethiopia	Kama	West	*Coffea arabica*
E565	Ethiopia	Gojjam	West	*Coffea arabica*
E017	Ethiopia	Sidamo	East	*Coffea arabica*
E018	Ethiopia	Sidamo	East	*Coffea arabica*
E022	Ethiopia	Sidamo	East	*Coffea arabica*
E021	Ethiopia	Sidamo	East	*Coffea arabica*
E037	Ethiopia	Shoa	East	*Coffea arabica*
E237	Ethiopia	Sidamo	East	*Coffea arabica*
E238	Ethiopia	Sidamo	East	*Coffea arabica*
Typica	Amsterdam Gardens	x	Cutivar	*Coffea arabica*
Bourbon	La Réunion (Bourbon Island)	x	Cutivar	*Coffea arabica*
Iapar 59	Brazil	Paraná	Inbred line	*Coffea arabica**
Icatu x Catuai	Brazil	Paraná	Inbred line	*Coffea arabica***
H9733(90-3-1)	Brazil	Paraná	Inbred line	*Coffea arabica****
H9733(90-8-1)	Brazil	Paraná	Inbred line	*Coffea arabica****
L1C1	Brazil	Paraná	Inbred line	*Coffea arabica*****
IAPAR 88480 (L3C3)	Brazil	Paraná	Inbred line	*Coffea arabica***
Coffea canephora	West/Central Africa/South Asia	x	Other Coffea spp	*Coffea canephora*
Coffea eugenioides	East/Central Africa	Kenya	Other Coffea spp	*Coffea eugenioides*
Coffea_racemosa	East Africa	Mozambique	Other Coffea spp	*Coffea racemosa*

[a] Geographical origin of the Ethiopian accessions and breeding status
x Genotypes that are not from Rift Valley region
* Villa Sarchi CIFC 971/10 x Hybrid of Timor 832/2 (Introgression of *C. canephora*)
** Introgression of *C. canephora*
*** IAPAR 88480-8 L3C3 x (IAPAR 88480-8 L3C3 x C1195-5-6-2). C1195-5-6-2 = [(*C. arabica* x *C. racemosa*) x *C. arabica*] x *C. arabica*
****Crossing of the accession E335 x Catuai

7.2 DNA Extraction and Genotyping

Genomic DNA was isolated from leaves by CTAB method (Doyle & Doyle, 1990) and diluted to a final concentration of 5ng/µL. DNA quality was verified in 0.8% agarose gel stained with ethidium bromide, and quantification was estimated using spectrophotometry with absorbance at 260 and 280 nm, using Nanodrop™.

For genotyping 30 SSR markers previously described as being polymorphic in *C. arabica* (Table II) were used (Da Silva et al., 2013). Amplification reactions of the DNA were performed using *Promega® Go Taq Green Master Mix* Kit with a final volume of 10 µL on *GeneAmp*®-PCR System 9700 (*Applied Biosystems*) thermocycler, with the following parameters: 95 °C for 2 minutes, followed by 30 cycles of 95 °C for 50 seconds, 50 °C for 1 minute, 72 °C for 30 seconds, and final extension of 72 °C for 5 minutes.

Table II – List of the 30 SSR markers used in this study.

SSR loci	Primer Forward	Primer Reverse	Repeat motif	Nature of the repeat	Reference/Contig position
CM2	TGTGATGCCATTAGCCTAGC	TCCAACATGTGCTGGTGATT	$(AC)_{10}$ $(AT)_{9}$	Di	Baruah et al., 2003
CFGA792b[2]	GATCAGAACTTTGAGCTCAGCA	AATGTGGCACGCTAGAAGTG	$(AG)_{12}$	Di	Cristancho et al., 2008
CFCA281[2]	GCGTCCACGTGTTAAGTCTT	TCAAGTGGCAGACATGTCAC	$(AC)_{13}$	Di	Cristancho et al., 2008
CFCA331[2]	TGATGGACAGGAGTTGATGG	CACTCATTTTGCCAATCTACC	$(CT)_{17}(AC)_{18}$	Di	Cristancho et al., 2008
CFCA360[2]	TTAAGACATCGGTGCATTCA	TGTGTACTGGGTTTTTTGATGT	$(AC)_{15}$	Di	Cristancho et al., 2008
CaM03[b]	CGCGCTTGCTCCCTCTGTCTCT	TGGGGGAGGGGCGGTGTT	AAC	Di	Geleta et al., 2012
M24	GGCTCGAGATATCTGTTTAG	TTTAATGGGCATAGGGTCC	$(CA)_{15}(CG)_{4}CA$	Di	Combes et al., 2000
M47	TGATGGACAGGAGGTGATGG	TGCCAATCTACCTACCCCTT	$(CT)_{9}(CA)_{8}/(CT)_{4}/(CA)_{5}$	Di	Combes et al., 2000
SSRCa 002	CTGTCCCACCAACCAAAA	CTTCAACCCCCAACACAC	$(TTCC)_{3}.(GT)_{17}$	Tetra/Di	Missio et al., 2009
SSRCa 052	GATGGAAACCCAGAAAGTTG	TAGAAGGGCTTTGACTGGAC	$(TTG)_{7}$	Tri	Missio et al., 2009
SSRCa 081	ACCGTTGTTGGATATCTTTG	GGTTGAACCTAGACCTTATTT	$(CT)_{38}$	Di	Missio et al., 2009
SSRCa 085	ATGTGAAAATGGGAAGGATG	CACAGGAAAGTGACACGAAG	$(TC)_{24}$	Di	Missio et al., 2009
SSRCa 091	CGTCTCGTATCACGCTCTC	TGTTCCTCGTTCCTCTCTCT	$(GT)_{8}(GA)_{10}$	Di	Missio et al., 2009
LEG11	CACTGAAGGCCTGGAAGAAT	AGCATCTGCAGCCTCCATAG	TGG	Tri	Pereira et al., 2011
LEG12	CACCATAGCAACTTCAAACACG	CACATCCAGGAACCTTGCTC	TC	Di	Pereira et al., 2011
LEG13	GAAGAGGAAGAAGGGGCAAG	GTGGTGGAGGAAAGGGATTC	GAA	Tri	Pereira et al., 2011
LEG32	GGGTGATGGAAAAGCAAATG	CCAGCATCAGCAAGTAAAAGG	AGA	Tri	Pereira et al., 2011
M32	AACTCTCCATTCCCGCATTC	CTGGGTTTTCTGTGTTCTGC	$(CA)_{3}/(CA)_{3}/(CA)_{18}$	Di	Combes et al., 2000
No Identification	CTCTCCCTCAGTCAATTCCA	CTTGGTCTCCCTCCTTTTTC	$(ATC)_{14}$	Tri	Silvestrini et al., 2007
AJ250253	CTTGTTTGAGTCTGTCGCTG	TTTCCCTCCCAATGTCTGTA	$(GA)_{5}$ $(GT)_{8}TT(GT)_{4}TT(GT)_{7}/(GA)_{11}/(TC)_{2}/(CT)_{3}GT$	Di	Silvestrini et al., 2007
AJ250256	AGGAGGGAGGTGTGGGTGAAG	AGGGGAGTGGATAAGAAGG	$(GT)_{11}$	Di	Maluf et al., 2005
DCM01	TTTTTGGGAAATGAAGGTGC	TGCACTTCAAGATCCCCTTT	$(AG)_{15}$	Di	Aggarwal et al., 2007
CaM41	CATCGTCTCCATCGTTGCTCTATC	CCCTCCCCCTCTTTCCTATCTAAT	$(TAAA)_{5}$	Tetra	Hendre et al., 2008
CFGA249	TAAGAAGCCACGTGACAAGTAAGG	TATGGCCCTTCTCGCTTTAGTT	$(AG)_{13}$	Di	Moncada and Couch, 2004
IAPAR 14	GCGGATCTAACCAAGTAGCC	ATGATGCCGGTGATGTTTAT	$(TTC)_{4}$	Tri	size37248
CHT03	GTCTCTCCGCTTTTTCTTCC	CTTGGTTGCCTGTTTCCTAA	$(CT)_{8}$	Di	scaffold8|size27678
CHT12	CCGAGCATTGTGACTCGTAT	CAGGAAAAACCAGAGACGAA	$(AT)_{9}$	Di	scaffold19|size38401
CHT25	CCTGTCTTGGCTCTACCTGA	TCTGTTGATCCGTGTTGATG	$(CTAT)_{3}$	Tetra	scaffold59|size4646
CHT28	CCGACGGGTCTCTTCTTTAT	TTCTTTACGGGATTGCTCTG	$(AG)_{5}$	Di	scaffold121|size2154
CHT29	AAACCCAACCTGGCTTTTT	CATCGCCTCTCTTTCTCATC	$(TTCC)_{3}$	Tetra	scaffold121|size2154

PCR products were submitted to 10% polyacrylamide gel electrophoresis and stained by ethidium bromide. Gels were visualized under ultraviolet light and captured by the Kodak ® 120 digital system. Molecular size of the amplified products was estimated using a 50 bp DNA ladder (*Ludwig biotec* ®).

8 DATA ANALYSES

8.1 DIVERSITY ANALYSES

Due the allotetraploid genome of *C. arabica*, it is impossible to distinguish between the triallelic combinations of SSRs loci. Therefore, although microsatellites are co-dominant markers, data analysis was based on presence/absence (1/0) of each allele, as performed by Aggarwal et al. (2007) and Silvestrini et al. (2007). Was also estimated the mean number of alleles for all 30 SSRs by the arithmetic mean of all loci.

The analyzed genotypes were allocated to 4 groups: 1) Cultivars and inbreed lines developed at IAPAR; 2) Eastern accessions (i.e. accessions from the East side of the Great Rift Valley); 3) Western accessions (i.e. accessions from the West side of the Great Rift Valley); and 4) others *Coffea* diploid species (diploid *Coffea* species *C. canephora*, *C. eugenioides*, and *C. racemosa*).

The genetic diversity structure was estimated using the following parameters for each group: unbiased expected heterozigosity (uHe), private alleles number, proportion of polymorphic loci *(P)* and Shannon's genetic index (*H'*). These analyses were performed using GenAlex software version 6.5 (Peakall & Smouse, 2006).

Analysis of Molecular Variance (AMOVA) was performed to estimate variation within and among genetics groups using the SSRs polymorphic loci. The analysis was performed using Arlequim 3.11 software (Excoffier et al., 2005) based on the Weir & Cockerham method with 10,000 permutations for the genetic distances, 100,000 steps of Markov chain for the exact population differentiation test and 10,000 dememorisation steps, with a significance level of 0.01 (Weir & Cockerham, 1984) in which the fixation index (Fst) was also estimated (Wright, 1965).

8.2 Population Structure Analysis

The binary matrix representing the SSR profile for each genotype was used to perform the principal coordinate analysis (PCoA). This analysis was performed using GenAlex software version 6.5 (Peakall & Smouse, 2006).

Population structure was estimated using Bayesian clustering method implemented in STRUCTURE software version 2.3.4 (Pritchard et al., 2000). Allele frequencies in each of the K groups (from 2 to 10) were estimated. We used a 10^5 burn-in period and 10^5 interactions MCMC (Markov Chain Monte Carlo), as these parameters resulted in relative stability of the results with 10 runs per K value. The genome composition (genome plot) of each accession was represented for each K. The most probable number of populations was estimated based on ΔK values (Evanno et al., 2005) using Structure Harvester software (Earl & Bridgett, 2012). The level of membership that we considered to assign the genotypes to the different groups was 0.6.

9 RESULTS

9.1 Genetic Diversity and Population Differentiation

We observed a total of 206 alleles across the 37 *Coffea* genotypes. The mean number of allele frequencies over all loci was 6.0 ranging from 3 to 16 per locus and the PIC values ranged from 0.39 to 1.00.

Among Eastern, Western, Cultivars/inbred lines and species groups (*C. canephora, C. eugenioides* and *C. racemosa*) the western group show the highest proportion of polymorphic markers (*P)* and Shannon's index *(H')* (69.95% e 0.281), and the cultivars/inbred lines group presented the lowest ones (32.86% e 0.153). The number of private alleles in Western group was also higher than in the others groups, confirming the higher genetic diversity in this group of accessions of *C. arabica*. Regarding the uHe values, which measures the genetic diversity weighted by the sample size of each group,

the species group presented the highest value (0.2), which can be explained by the fact that this group is formed by three individuals from distinct species. Considering only the *C. arabica's* groups, the Western group also had a higher value of uHe (0.183) than Eastern and cultivars/inbreed's groups (0.177 and 0.106, repectively) (Table III).

Table III **–** Proportion of polymorphic loci (*P*), Shannon index of all loci (*H'*), Unbiased expected heterozigosity (uHe) and Private alleles for each genetic group over all loci.

Genetic Group	N° of genotypes	*P* (%)	*H'*	uHe	Private alleles
East	7	52.58	0.252	0.177	10
West	19	69.95	0.281	0.183	34
Cultivars and inbred lines	8	32.86	0.153	0.106	7
Species	3	45.07	0.249	0.200	18

The AMOVA shows that most of the genetic variance originates from the within-group level (75%; p = 0.05), with 25% of the total variance distributed between groups. The Fst value, which measures the magnitude of genetic differentiation between populations, was 0.25 for whole sample (Table IV).

Table IV **–** Analysis of Molecular Variance (AMOVA) among and within *Coffea* genetic groups.

Source of Variation	d.f	Sum of squares	Variance Component	Variation (%)
Among groups	3	232.95	7.11	25.38
Within groups	33	689.65	20.89	74.62

Fst for whole genetic groups = 0.25; Fst between West and East Ethiopian group = 0.05; Fst between Ethiopian and Cultivars/inbred lines group = 0.32; d.f = degrees of freedom.
p = 0.05.

9.2 Population Genetic Structure

According to Evanno criterion (Evanno et al., 2005) the most likely number of groups (K) was 2. The structure with 3 groups (K = 3) also presented a high ΔK value (Fig 1A, 1B and 1C). With K = 2, Brazilian cultivars/inbred lines, eastern accessions E017 and E037 and the diploid species, *C. canephora*, *C. eugenioides* and *C. racemosa* were allocated to the Q1 group (grey). The Q2 group (black) was formed by all accessions from the west and east side of Rift Valley. In the structure with K = 3 the Q3 group was formed by *C. canephora* and *C. racemosa* species, demonstrating that these species are more genetically distant in relation to the other groups analyzed (Figure I).

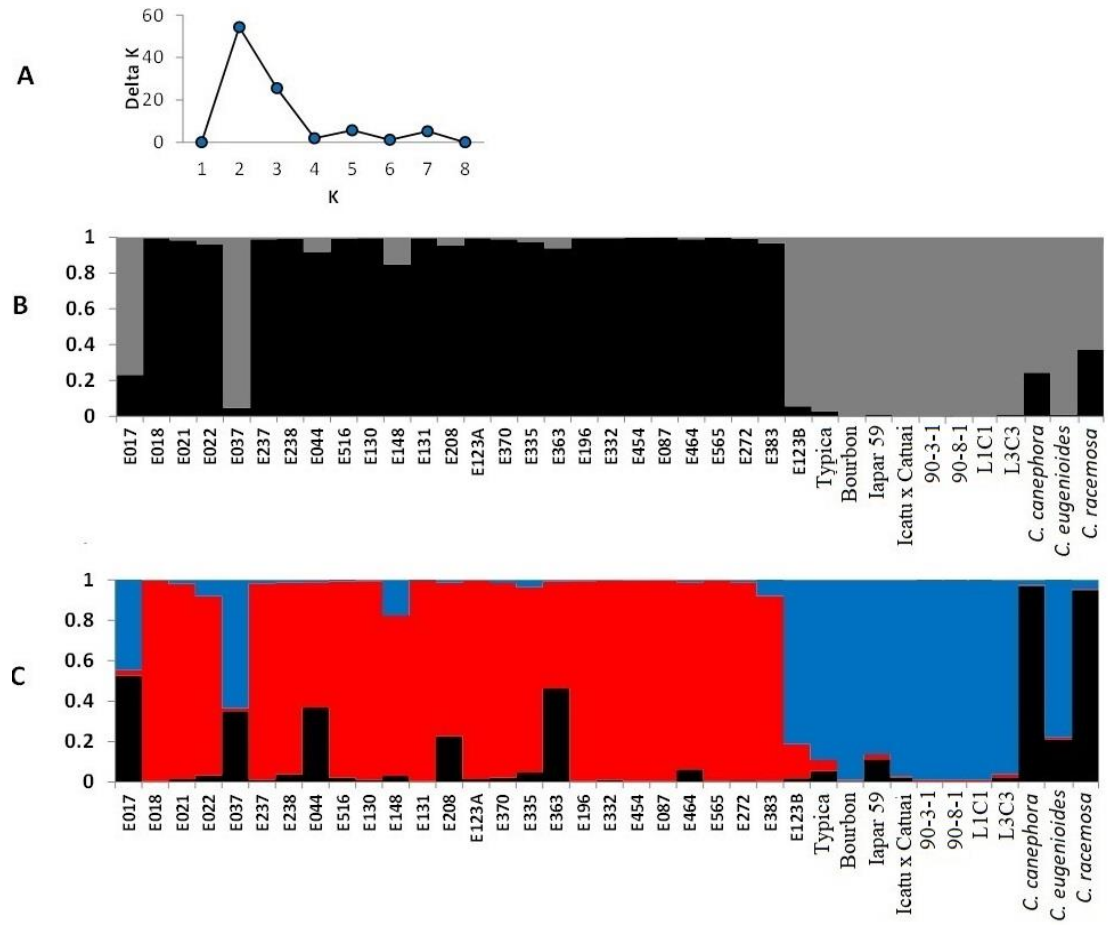


Figure I. Bar plot of Structure software used to study diversity of wild genotypes of *Coffea arabica*, cultivars/inbred lines and species from data of 30 SSR markers.
ΔK values in function of subgroups number (A) obtained from Bayesian clustering analysis considering *K* = 2 (B) and *K* = 3 (C) by STRUCTURE software version 2.3.4 from 30 microsatellite loci in 37 *Coffea* samples.

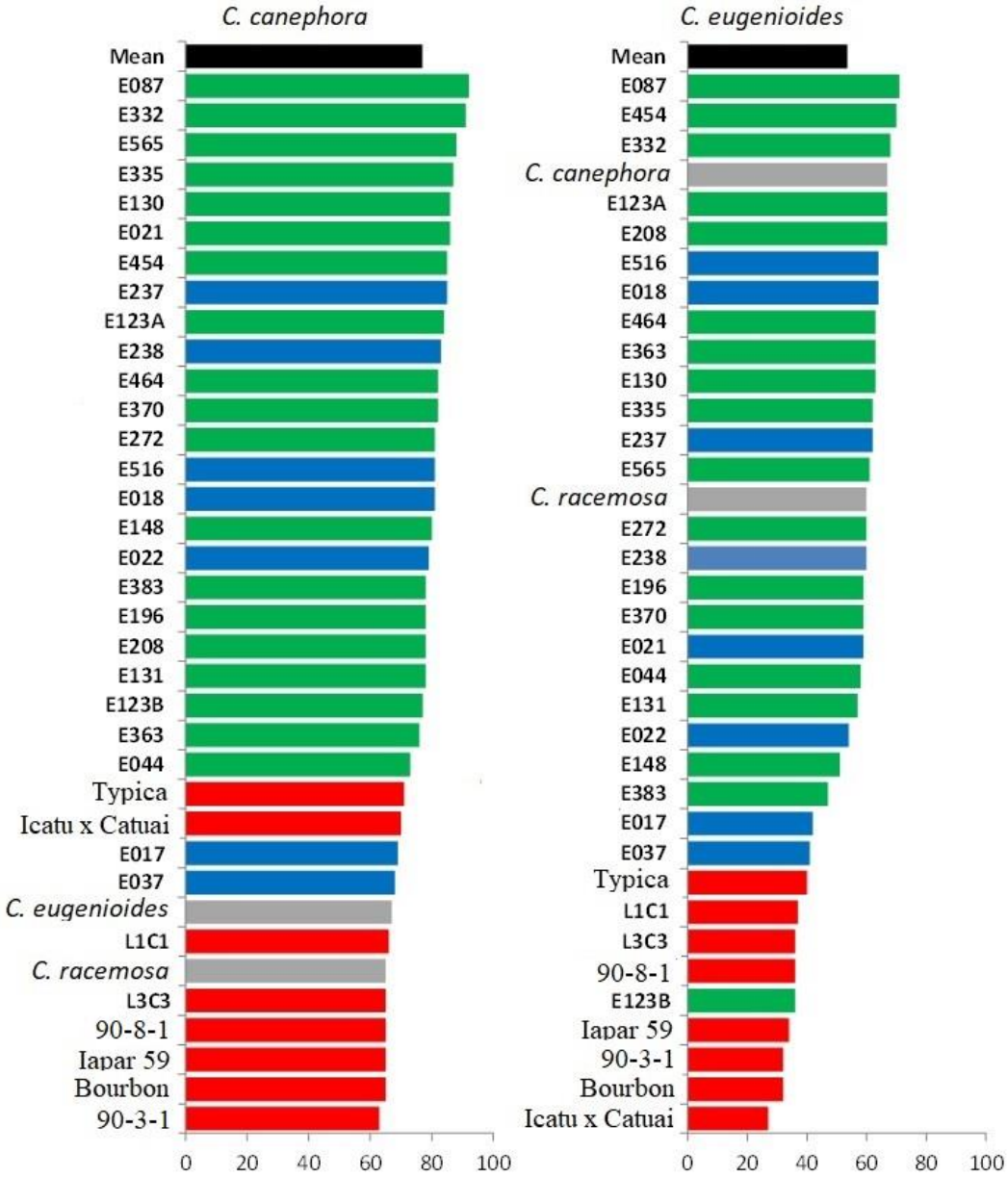
Principal coordinate analysis based on binary genetic distance matrix was consistent with Bayesian STRUCTURE analysis and explained 91.19% of the total genetic variation of the panel (PCoA1 – 50.25% and PCoA2 – 40.94%). There was a clear division of whole panel in two groups, one (Q1) formed exclusively by accessions from Ethiopia, and another (Q2) formed by the cultivars/inbred lines, two accessions from east side of Rift Valley (E017 and E037) and three others coffee species (i.e. *C. canephora*, *C. eugenioides* and *C. racemosa*). This second group presented a subdivision in two groups, in which one was formed specifically by genotypes of *C. canephora*, *C. racemosa* plus E017 accession and another one formed by other

genotypes. As in the bayesian analysis, the genotype of *C. eugenioides* presented great genetic proximity to the cultivars and breeding lines (Figure II).

Figure II. Principal coordinate analysis based on genetic binary distance among the *Coffea* genotypes analyzed (n=37). The dots are colored according to the colors of STRUCTURE results using K = 3.

The comparison of binary genetic distance of *C. canephora* and *C. eugenioides* in relation to each one of *C. arabica* genotypes (Figure III) indicate that all *C. arabica* genotypes are genetically closer to *C. eugenioides* than to *C. canephora*. In addition, cultivars showed markedly lower genetic distances in relation to *C. eugenioides* compared to Ethiopian accessions.

Figure III. Binary genetic distances between each *C. arabica* genotype in relation to *C. canephora* and *C. eugenioides* individuals. The bars representing genotypes from Cultivars/breeding lines, Western and Eastern accessions from Rift Valley and Species groups are colored in red, green, blue and grey respectively.



10 DISCUSSION

In the present study, 30 highly polymorphic SSR markers were used to genotype 37 Coffea spp genotypes. An average of 6.0 alleles and an average PIC of 0.73 were observed. Larger values were obtained for the different genetic parameters analyzed than previous studies. Anthony et al. (2002) reported an average number of 4.7 alleles per SSRs using six SSRs in *C. arabica* sample containing four Typica, five Bourbon and

10 subspontaneous derived accessions. Using 34 SSRs, Moncada & McCouch (2004) reported an average of 2.5 and 1.9 amplified alleles for SSRs in 11 wild and 12 cultivated *C. arabica* genotypes, respectively, with the number of alleles ranging from one to eight. Maluf et al. (2005) also reported an average number of 2.87 alleles in 28 cultivated *C. arabica* lines using 23 SSRs. One reason for such differences could be due to smaller sample size and the coffee genotypes (Ethiopian vs Cultivated) used in the previous studies, mainly to the enrichment of Ethiopian *C. arabica* genotypes, as compared to the present study.

On the other hand, similar results were obtained by Teressa et al. (2010) analyzing *C. arabica* collection of 133 genotypes (78 accessions from Ethiopia and 55 cultivars) with 32 SSRs. They detected 209 alleles, with the number of alleles per marker ranging from 2 to 14, with an average of 6.5 alleles for each marker. Aerts et al. (2013) in a based on populations from two coffee production systems (forest coffee and semi-forest coffee), identified a total of 159 alleles across 703 wild accessions collected in forests from Ethiopia with 24 SSR markers. The number of alleles ranged from 2 to 19 per locus.

Our results indicated that accessions collected by the FAO mission in 1968 at the primary origin center of the specie (Ethiopia) presented a high genetic diversity. However, among the cultivars a low level of diversity was verified. This result is in agreement with the early history of *C. arabica* distribution when the commercial cultivars have undergone successive genetic reductions (Anthony et al., 2002). Historical data indicated that the *C. arabica* populations in major producing countries were derived from few plants and/or seeds originated from Ethiopia. This could be the main factor for the low allelic richness and low polymorphism of the commercial cultivars. However, relatively high level of SSR polymorphism and genetic richness were also reported in *C.*

arabica from Ethiopia (Anthony et al., 2002; Moncada & McCouch, 2004; Silvestrini et al., 2007; Aerts et al., 2013).

The proportion of polymorphic loci and Shannon's index values estimated in our study (33% to 70% and 0.1 to 0.3) were similar to analyzes performed in natural populations of *Coffea arabica* from Ethiopia (López-Gartner et al., 2009). These authors determined that the *P%* and *H'* values ranged from 37% to 73% and 0.2 to 0.4 respectively. Comparing the unbiased expected heterozigosity diversity within each genetic group (Table III), the variability was higher in the Western group than Eastern and cultivars/inbred lines what is consistent with *H'* and the number of private alleles.

The high genetic richness of *Coffea arabica* Ethiopian accessions suggests that it can be used as a complementary source of diversity in breeding programs. In the West group 34 private alleles were identified, suggesting that particular efforts should be targeted towards the introduction of this genetic group in *C. arabica* breeding programs. The lower number of private alleles was found in Brazilian cultivars/inbred lines (7 alleles), corroborating the low diversity observed among cultivars in other studies (Setotaw et al., 2013; Vieira et al., 2010). Genetic structure analyses (genetic distance and bayeasian based approaches) indicated the presence of two main subgroups in our samples of accessions, which clearly distinguished Ethiopian Coffea arabica accessions from the others *Coffea* genotypes. Similar result was observed in previous studies comparing the genetic diversity among wild and cultivated genotypes of *C. arabica* (Silvestrini et al., 2007; Lópes-Gartner et al., 2009; Teressa et al., 2010).

Interestingly, in the present study the genetic distance analyzes as well the Structure and PCoA results, demonstrated a closer proximity of the *C. eugenioides* genotype in relation to the cultivars group. We also observe that *C. canephora* and *C. racemosa* demonstrated high genetic dissimilarity in relation to *C. eugenioides*.

Lashermes et al. (1995) studying the evolutionary history of *C. arabica* and their genetic relationships with other *Coffea* species also reported that *C. eugenioides*, followed by *C. canephora* and *C. racemosa* were the most related to *C. arabica*. Our data indicates that selection performed during the genetic improvement of *C arabica* may have led to a decrease in genetic divergence of the breeding cultivars in relation to its diploid ancestor *C. eugenioides*. *C. eugenioides* is the female ancestral parent of *C. arabica* and probably it is the main source of genes related to beverage quality (Medina Filho et al., 2007; 2012).

Ashihara & Crozier, 1999, reported that this low caffeine content of the parental *C. eugenioides* is due to the reduction of caffeine biosynthesis along with the rapid catabolism that is regulated by specific genes. On the other side, *C. canephora* contains higher levels of the caffeine and chlorogenic acids (CGA), compounds directly related to both coffee bitterness and astringency, affecting its quality (Charrier & Berthaud, 1988; Jeszka-Skowron et al., 2016).

11 CONCLUSION

Our results indicate the presence of a high allelic richness in accessions from Ethiopia, especially in those collected in the West side of the Great Rift Valley, and this reinforces the importance of conserving and using germplasm of the primary center of origin of this important species. Interestingly, our results indicate that *C. arabica* cultivars are genetically closer to its diploid ancestor *C. eugenioides* than wild Arabica accessions. Overall, information about genetic relationships of *Coffea* accessions estimated using SSR markers are valuable for conservation strategies and utilization of this germoplasm in breeding programs.

Further analyses, including genomic comparisons, of a higher number of *C. eugenioides* and *C. canephora* genotypes in comparison with wild type and *C. arabica* cultivars should provide a better understanding of the influence of the two diploid sub genomes in the domestication process of *C. arabica*.

12 REFERENCES

AERTS, R.; BERECHA, G.; GIJBELS, P.; HUNDERA, K.; VAN GLABEKE ,S.; VANDEPITTE, K, et al. Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian montane rainforests. **Evol. Appl**. 6(2):243–52, 2013.

AGA, E.; BRYNGELSSON, T.; BEKELE, E.; SALOMON, B. Genetic diversity of forest arabica coffee (*Coffea arabica L*.) in Ethiopia as revealed by random amplified polymorphic DNA (RAPD) analysis. **Hereditas**. 138: 36-46, 2003.

AGGARWAL, R.K.; HENDRE, P.S.; VARSHNEY, R.K.; BHAT, P.R.; KRISHNAKUMAR, V.; SINGH, L. Identification , characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. **Theor. Appl. Genet**. 114:359–72, 2007.

ANTHONY, F.; COMBS, C.; ASTORGA, C.; BERTRAND, B.; GRAZIOSI; LASHERMES, P. The origin of cultivated *Coffea arabica L*. varieties revealed by AFLP and SSR markers. **Theor. Appl. Genet**. 104: 894–900, 2002.

ASHIHARA, H & CROZIER, A. Biosynthesis and catabolism of caffeine in low- caffeine-containing species of *Coffea*. **J. Agric. Food. Chem**. 47(8), 3425e3431, 1999.

BARUAH, A.; NAIK, V.; HENDRE, P.S.; RAJKUMAR, R.; RAJENDRAKUMAR, P and AGGARWAL, R.K. Isolation and characterization of nine microsatellite markers from Coffea Arabica L., showing wide cross-species amplifications. **Mol. Ecol**. Notes 3: 647–650, 2003.

BERTRAND, B.; ETIENNE, H.; CILAS, C.; CHARRIER, A.; BARADAT, P. *Coffea arabica* hybrid performance for yield, fertility and bean weight. **Euphytica**. 141(3): 255-262, 2005.

CHAPARRO, A. P.; CRISTANCHO, M. A.; CORTINA, H. A AND GAITAN, A. L. Genetic variability of *Coffea arabica L*. accessions from Ethiopia evaluated with RAPDs. **Genet. Resour. Crop. Evol,** 51:291–297, 2004.

CHARRIER, A.; BERTHAUD, J. **Principles and methods in coffee plant breeding**: *Coffea canephora Pierre*. In: Clarke RJ, Macrae R (eds), Coffee, Vol. 4: Agronomy, pp.167- 198. Elsevier Applied Science, London and New York, Great Britain, 1988.

COMBES, M.C.; ANDRZEJEWSKI, S.; ANTHONY, F.; BERTRAND, B.; ROVELLI, P.; GRAZIOSI, G.; LASHERMES, P. Characterization of microsatellites loci in *Coffea arabica* and related coffee species. **Mol. Ecol**. v.9, p.1171-1193, 2000.

CRISTANCHO, M & ESCOBAR, C. Transferability of SSR markers from related Uredinales species to the coffee rust Hemileia vastatrix. **Genet. Mol. Res**. 7, 1186–1192, 2008.

DA SILVA, B. S.R.; CAÇÃO, S.B.; IVAMOTO, S.T.; SILVA, J.C.; DOMINGUES, D.S.; PEREIRA, L.F.P. Identificação e Caracterização de Microssatélites de *Coffea arabica* a partir de dados de sequenciamento de RNA e de BACs. **BBR**. 2,3:186-190, 2013.

DAVIS, A.P.; TOSH, J.; RUCH, N.; FAY, M.F. 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(4):357–77, 2011.

DOYLE, J.J.; DOYLE, J.L. **Isolation of plant DNA from fresh tissue.** Focus 12: 13 15, 1990.

EARL, D.A.; BRIDGETT, M STRUCTURE HARVESTER : a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conserv. Genet**. 4:359–61, 2012.

EVANNO, G.; REGNAUT, S.; GOUDET J. Detecting the number of clusters of individuals using the software structure: a simulation study. **Mol. Ecol**. 14: 2611-2620, 2005.

EXCOFFIER, L.; LAVAL, G.; SCHNEIDER, S. Arlequin (version 3.0): An integrated software package for population genetics data analysis. **Evol. Bioinform. Online**. 47-50, 2005.

GELETA, M.; HERRERA, I.; MONZÓN, A AND BRYNGELSSON. Genetic diversity of arabica coffee (*Coffea arabica L*) in Nicaragua as estimated by Simple Sequence Repeat Markers. **Sci. World. J**. 2012:1-11, 2012.

HENDRE, P.S.; PHANINDRANATH, R.; ANNAPURNA, V.; LALREMRUATA, A AND AGGARWAL, K. Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex A. Froehner) showing broad cross-species transferability and utility in genetic studies. **BMC. Plant. Biol**. 8:51-70, 2008.

JESZKA-SKOWRON, M.; SENTKOWSKA, A.; PYRZYŃSKA, K.; PAZ DE PEŃA, M. Chlorogenic acids, caffeine content and antioxidant properties. **Eur. Food. Res. Technol**. 242:1403–1409, 2016.

LASHERMES, P.; COMBES, M.C.; CROS, J.; TROUSLOT, P.; ANTHONY, F.; CHARRIER, A. Origin and genetic diversity of *Coffea arabica L*. based on DNA molecular markers. **Agronomie**. 528–36, 1995.

LASHERMES, P.; COMBES, M.C.; ROBERT, J.; TROUSLOT, P.; D'HONT, A.; ANTHONY F, et al. Molecular characterisation and origin of the *Coffea arabica L*. genome. **Mol. Gen. Genet**. Mar;261(2):259–66, 1999.

LÓPEZ-GARTNER, G.; CORTINA, H.; MC COUCH.; SUSAN, R.; MONCADA, M.D.P. Analysis of genetic structure in a sample of coffee (*Coffea arabica* L.) using fluorescent SSR markers. **Tree Genetics and Genomes**. 5:435-446, 2009.

MALUF, M.P.; SILVESTRINI, M.; RUGGIERO, L.M.C.; GUERREIRO, FILHO, O.; COLOMBO, C. Genetic diversity of cultivated *Coffea arabica* inbred lines assessed by RAPD, AFLP and SSR marker systems. **Sci. Agric**. 62(4):366-373, 2005.

MEDINA-FILHO, H.P.; MALUF, M.P.; BORDIGNON, R.; GUERREIRO FILHO, O.; FAZUOLI, L.C. Traditional breeding and modern genomics: a summary of tools and

developments to exploit biodiversity for the benefit of the coffee agroindustrial chain. **Acta Horticulturae**. 745:351-368, 2007.

MEDINA-FILHO, H.P.; BORDIGNON, R.; SOUZA, F.F.; TEIXEIRA, A.L.; DIOCLECIANO, J.M.; FERRO, G.O. **Arabica selections with *Coffea eugenioides* and *C.canephora* introgressions for Rondônia state in Brazilian Amazon**: Proceedings of the 24th International Conference on Coffee Science. Costa Rica. Anais, pp.1303-1307, 2012.

MEYER, F.G.; FERNIE, L.M.; NARASIMHASWAMY, R.L.; MONACO, L.C.; GREATHEAD, D.J. **Coffee Mission to Ethiopia**. In: Food and agriculture organization of the United Nations. p. 1964–5, 1968.

MISSIO, R.F. et al. Development and validation of SSR markers for *Coffea arabica L*. **Crop. Breed. Appl. Biotechnol**. 9: 361-371, 2009.

MISSIO, R.F.; CAIXETA, E.T.; ZAMBOLIM, E.M.; PENA, G.F.; ZAMBOLIM, L.; DIAS, L.A.S AND SAKIYAMA, N.S. Genetic characterization of an elite coffee germplasm assessed by gSSR and EST-SSR markers. **Genet. Mol. Res**. 10 (4): 2366-2381, 2011.

MONCADA, P.; MCCOUCH, S. Simple sequence repeat diversity in diploid and tetraploid Coffea species. **Genome**. 47: 501-509, 2004.

MOTTA, L.B.; SOARES, T.C.B.; FERRAO, M.A.G et al. Molecular characterization of arabica and conilon coffee plants genotypes by SSR and ISSR markers. **Braz. Arch. Biol. Technol**. 57:728–735, 2014.

PEAKALL, R.; SMOUSE, P.E. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. **Mol. Ecol**. 6:288–95, 2006.

PEREIRA, G.S.; PADILHA, L.; PINHO, E. V. R. V.; TEIXEIRA, R. K. S.; CARVALHO, C. H. S. de; MALUF, M. P.; CARVALHO, B. L. Microsatellite markers in analysis of resistance to coffee leaf miner in Arabica coffee. **Pesq. Agropec. Bras**. 46(12), 1650-1656, 2011.

PESTANA, K.N.; CAPUCHO, A.S.; CAIXETA, E.T et al. Inheritance study and linkage mapping of resistance loci to Hemileia vastatrix in Híbrido de Timor UFV 443-03. **Tree. Genet. Genomes**. 11:72, 2015.

POT, D.; SCHOLZ, M. B. S.; LANNES, S. D.; DEL GROSSI, L.; PEREIRA, L. F. P.; VIEIRA, L. G.; SERA, T. **Phenotypic analysis of *Coffea arabica* acessions from Ethiopia: contribution to the understanding of *Coffea arabica* diversity**. In: 22nd International Conference on Coffee Science, Campinas. *Anais*… Campinas, p.165. 2008.

PRITCHARD, J.K.; STEPHENS, M.; DONNELLY, P. Inference of Population Structure Using Multilocus Genotype Data. **Genet. Soc. Am. Inference**.155:945–59, 2000.

SCHOLZ, M.B.S.; KITZBERGER, C.S.G.; PAGIATTO, N.F.; PEREIRA, L.F.P.; DAVRIEUX, F.; POT, D et al. Chemical composition in wild ethiopian Arabica coffee accessions. **Euphytica**. 209(2):429–38, 2016.

SANT'ANA, GUSTAVO C.; PEREIRA, LUIZ F. P.; POT, DAVID.; IVAMOTO, SUZANA T.; DOMINGUES, DOUGLAS S.; FERREIRA, RAFAELLE V.; PAGIATTO, NATALIA F.; DA SILVA, BRUNA S. R.; NOGUEIRA, LÍVIA M.; KITZBERGER, CINTIA S. G.; SCHOLZ, MARIA B. S.; DE OLIVEIRA, FERNANDA F.; SERA, GUSTAVO H.; PADILHA, LILIAN.; LABOUISSE, JEAN-PIERRE.; GUYOT, ROMAIN.; CHARMETANT, PIERRE.; LEROY, THIERRY. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. **Sci. Rep**. 8:465, 2018.

SETOTAW, T.A.; CAIXETA, E.T.; PEREIRA, A.A.; OLIVEIRA, A.C.B.; DE CRUZ C.D.; ZAMBOLIM, E.M et al. Coefficient of Parentage in *Coffea arabica L*. Cultivars Grown in Brazil. **Crop. Sci**. 53(4):1237–47, 2013.

SILVESTRINI, S.; JUNQUEIRA, M.G.; FAVARIN, A.C.; GUERREIRO-FILHO, O.; MALUF, M.P.; SILVAROLLA, M.B.; COLOMBO C.A. Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. **Genet. Resour. Crop. Evol**. 54(6): 1367-1379, 2007.

SOUSA, T.V.; CAIXETA, E.T.; ALKIMIM, E.R.; DE OLIVEIRA, A.C.B.; PEREIRA, A A.; ZAMBOLIM, L.; SAKIYAMA, N.S. Molecular markers useful to discriminate *Coffea arabica* cultivars with high genetic similarity. **Euphytica**. 213:75, 2017.

TERESSA, A.; CROUZILLAT, D.; PETIARD, V AND BROUHAN, P. Genetic diversity of Arabica coffee (*Coffea arabica L*.) Collections. **EJAST**. 1(1): 63-79, 2010.

TRAN, H.T.; LEE, L.S.; FURTADO, A.; SMYTH, H.; HENRY, R.J. Advances in genomics for the improvement of quality in coffee. **J. Sci. Food. Agric**. 96:3300–12, 2016.

UEFUJI, H.; S. OGITA.; Y. YAMAGUCHI.; N. KOIZUMI AND H. SANO: Molecular cloning and functional characterization of three distinct Nmethyltransferases involved in the caffeine biosynthetic pathway in coffee plants. **Plant. Physiol**. 132, 372-380, 2003.

WEIR, B.S.; COCKERHAM, C.C. Estimating F-statistics for the analysis of population structure. **Evolution**. 1358–70, 1984.

WRIGHT, S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. **Soc. Study. Evol**. 19(3):395, 1965.

VIEIRA, E.S.N.; VON PINHO, E.V.R.; CARVALHO, M.G.; ESSELINK, G.D.; VOSMAN, B. Development of Microsatellite Markers for the Identification of Brazilian Coffea arabica Varieties. **Genet. Mol. Biol**. 33:507–14, 2010.

13 CAPÍTULO 2: Estudo de Diversidade Fenotípica de acessos de *Coffea arabica* oriundos da Etiópia

RESUMO

A exploração de recursos genéticos de genótipos do centro de origem de uma espécie é uma opção para os programas de melhoramento que possuem cultivares com baixa variabilidade genética. Visando seleção de genótipos para o melhoramento e estudos de associação genômica, este trabalho objetivou avaliar a variabilidade fenotípica para oito compostos bioquímicos relacionados à qualidade da bebida em 68 acessos de *C. arabica* oriundos da Etiópia. Todos os compostos mostraram alta variabilidade demonstrando grandes amplitudes entre os valores. Significantes correlações positivas foram observadas entre Lipídeos Totais (LT) e Caveol; Ácidos clorogênicos (ACG) e LT, Açúcares Totais (AT), Sacarose e Proteínas Totais (PT); bem como Cafeína e PT. Significante correlação negativa entre LT e Sacarose/ Açúcares Totais; e entre Cafestol e Caveol também foram observadas. A análise de agrupamento hierárquico identificou 4 grupos entre os genótipos. O Grupo 1 apresentou maiores teores de Caveol, LT, Sacarose, AT, e baixos de Cafeína. No grupo 3 estão os acessos com altos teores de Caveol, LT e PT e AT, e baixos teores de Cafestol e ACGs. Assim ambos os grupos 1 e 3 são indicados para fins de melhoramento da espécie, e visando benefícios à saúde humana quanto aos baixos conteúdos de Cafestol, o grupo 3 é o mais indicado. Foi observada maior diversidade entre os acessos do lado oeste do Vale do Rift, mas sem clara separação entre os acessos dessas regiões. Esse trabalho permitiu a seleção de potenciais genótipos para a produção de uma bebida de melhor qualidade. Os resultados obtidos serão utilizados em estudos de associação genética com marcadores moleculares SNPs.

Palavras-Chave: *Coffea arabica.* Cafés selvagens. Composição Bioquímica. Diversidade fenotípica.

14 INTRODUÇÃO

O café é comercializado há séculos e consumido por mais de 800 milhões de pessoas que apreciam do seu sabor, aroma, benefícios à saúde bem como suas propriedades estimulantes (PEREIRA e IVAMOTO, 2015). Dentre as 10 espécies cultivadas mundialmente, C. arabica e *C. canephora* (variedade Robusta) são as de maior cotação no mercado internacional (DAVIS et al., 2011; CONAB et al., 2016).

A espécie *Coffea arabica* é responsável por bebidas de alta qualidade devido às características bioquímicas da bebida (bebida mais aromática), e pelo baixo teor de cafeína, porém é susceptível a doenças e vulnerável a estresses bióticos e abióticos (SILVAROLLA; MAZZAFERA; FAZUOLI, 2004). Isso devido ao gargalo genético gerado pelo seu histórico de dispersão, assim como às características botânicas da espécie. Em contrapartida, *C. canephora* é mais resistente a estresses ambientais (bióticos e abióticos), porém possui características sensoriais inferiores e é utilizado principalmente para a produção de cafés solúveis e em *blends* comerciais com *C. arabica*, sendo raramente utilizado sozinho (ANTHONY et al., 2001; SOUZA; BENASSI, 2012; CONAB, 2016).

O valor econômico da diversidade genética disponível em um germoplasma depende dos potencias benefícios que ele apresenta. Uma vez que características agronômicas de interesse são identificadas para o café, tais como tolerância a pragas, aumento da produtividade, baixa cafeína (SILVAROLLA et al., 2000), baixo cafestol e alto caveol (SCHOLZ et al., 2014; 2016) ou cafés com baixo teor de cafeína (SILVAROLLA et al., 2004), essas informações podem ser incluídas no melhoramento para a seleção de novas variedades.

A crescente conscientização da qualidade da bebida do café fez com que a demanda por cafés de alta qualidade aumentasse cada vez mais, e devido a isso, a composição bioquímica dos grãos de *C. arabica* tem sido extensivamente estudada (UPADHYAY, MOHAN RAO, 2013; KITZBERGER et al., 2013 SCHOLZ et al., 2016). Isso porque a concentração de compostos químicos presentes nos grãos de café, possuem papel primordial na formação do aroma e sabor da bebida durante o processo de torrefação dos grãos, sendo

responsáveis então pela qualidade da bebida. Isso é resultante da presença combinada de vários constituintes químicos voláteis e não voláteis que influenciam nas características sensoriais comercialmente importantes (CHENG et al., 2016).

Além disso, o café como um alimento funcional possui propriedades antioxidantes que reduz o risco de incidência a várias doenças crônicas, como inflamação, diabetes, doenças cardiovasculares, hepáticas, além de proteger contra a doença de Parkinson e câncer (BHUPATHIRAJU et al., 2013; CANO-MARQUINA et al., 2013). Estas propriedades estão ligadas a compostos bioativos como cafeína, teofilina, teobromina, cafestol, caveol, tocoferóis, ácidos clorogênicos e seus derivados (PERRONE et al., 2008).

Em espécies com estreita base genética como *C. arabica*, as cultivares são intimamente relacionadas e fenotipicamente similares. Já foi demonstrado que os acessos de *C. arabica* da Etiópia possuem variabilidade quanto a conteúdos bioquímicos, moleculares, características morfológicas e agronômicas, assim como tolerância e resistência a estresses. Essas características fazem da exploração de acessos originados da Etiópia uma opção para aumentar a base genética das variedades de café (BERTRAND et al., 2003, TESSEMA et al., 2011; SCHOLZ et al., 2014; 2016; POT et al., 2010).

O conhecimento da composição química entre genótipos é um pré-requisito para estudos de associação entre marcadores moleculares e características da bebida para melhorar a eficiência e velocidade do melhoramento. Este trabalho teve como objetivo caracterizar e explorar a composição bioquímica relacionada à qualidade da bebida em acessos etíopes da safra 2016 através da Espectroscopia no Infravermelho Próximo (NIRS). Esse trabalho ajudará no mapeamento de marcas de interesse relacionadas a essas características bioquímicas do grão, além de fornecer um estudo de diversidade, relação genética e seleção de genitores pelos programas de melhoramento do café visando melhorar a qualidade da bebida de café.

15 MATERIAIS E MÉTODOS

15.1 MATERIAL VEGETAL

Foram utilizados 68 genótipos que compõem o germoplasma de *C. arabica* da Etiópia, da safra de 2016 que são mantidos no IAPAR (FAO, 1968) e cultivados nas mesmas condições edafoclimáticas. Do total dos acessos, 62 são do lado oeste do Vale do Rift e 6 do lado leste do vale do Rift (E012/226, E016/456, E018/494, E022/163, E237/071 e E238/022). Amostras de frutos maduros estágio cereja com aproximadamente 225 DAF (Dias Após a Florada), foram coletados e estes secos naturalmente em caixas com fundo de malha, até atingir 12% de umidade. Após secagem, os frutos foram beneficiados pela remoção da pele e pergaminho, e então armazenados em sacos de papel em local seco e escuro.

Para caracterizar o tamanho dos grãos de café verde beneficiados, os mesmos foram passados peneiras com tela 14 (0,56 mm), 15 (0,59 mm), 16 (0,63 mm), 17 (0,67 mm) e 18 (0,71 mm). Os grãos considerados ardidos (marrons), pretos e brocados foram removidos nessa etapa por não serem características próprias dos genótipos, e sim gerados pelo ambiente. Posteriormente os grãos de café foram mergulhados em nitrogênio líquido (-196°C), moídos em moinho de disco (PERTEN 3600, Kungens Kurva, Suécia), e passados em peneiras com malha 0,5 mm para controlar o tamanho das partículas. As amostras moídas foram mantidas congeladas a -18 C até posterior análise.

15.2 ANALISE FENOTÍPICA DOS COMPOSTOS QUÍMICOS

A composição bioquímica dos acessos foi analisada quanto ao conteúdo de Ácidos Clorogênicos (ACG), Cafeína, Lipídeos Totais (LT), Proteínas Totais (PT), Sacarose, Açúcares Totais (AT) e dois terpenóides específicos do café; Caveol e Cafestol.

Para determinar os compostos químicos foi utilizado o espectrofotômetro NIRSystem 6500 (Foss NIRSystems, Silver Spring, Maryland - USA) com reflectância difusa e região espectral de 400 a 2500 nm com modelos de predição previamente desenvolvidos para cada composto (SCHOLZ et al., 2014

a, b). Foram realizadas varreduras com comprimentos de onda variando de 1100 a 2500 nm em intervalos de 2 nm, utilizando cubeta retangular contendo 6 g de café moído em temperatura ambiente. Todos os cálculos foram feitos usando os programas ISIscan e WinISI 4.5 (*Infrasoft Internacional*, Porto Matilda, Pennsylvania - EUA).

15.3 ANÁLISES ESTATÍSTICAS

Todas as análises estatísticas foram realizadas com auxílio do *software* XLSTAT (ADDINSOFT, 2010). A Análise de Componentes Principais (PCA) e Análise de Agrupamento Hierárquico (AAH) foram realizadas a partir de uma matriz de covariância gerada a partir dos genótipos. Os dados utilizados para AAH foram primeiramente normalizados, e em seguida, aplicado o algoritmo de *Ward's* com base na distância Euclidiana em conjunto para definir os grupos. A correlação de Pearson foi realizada com uma significância de $\alpha = 0,05$, visando observar a correlação bilateral entre os compostos químicos.

16 RESULTADOS E DISCUSSÃO

16.1 VARIAÇÃO DOS COMPOSTOS

Os acessos de *C. arabica* da Etiópia apresentaram grande variabilidade no conteúdo dos compostos avaliados. Os valores mínimo e máximo dos compostos analisados apresentaram amplitudes elevadas (Tabela 1), demonstrando alta variabilidade desses compostos nos acessos selvagens. Além do conteúdo lipídico, a composição da fração diterpênica deve ser levada em conta pelos programas de melhoramento de café. Recomendam-se baixos níveis de Cafestol combinados com altos níveis de Caveol devido ao respectivo aumento negativo do colesterol sérico, quando ingeridos em altas quantidades (MURIEL; ARAUZ, 2010). Dentre os genótipos analisados, foi observada uma grande variação fenotípica, com valores de 153 a 1.088 mg 100 g^{-1} para Cafestol e 303 a 1.144 mg 100 g^{-1} para Caveol. Como o conteúdo destes compostos varia de acordo com espécie e cultivar (KITZBERGER et al., 2013b), estes resultados indicam que são potenciais compostos para seleção e estudo de

diversidade, devido ao potencial discriminante de ambos (Tabela 1) (SCHOLZ et al., 2016).

Em relação aos acessos da Etiópia cultivados no IAC, um apresentou baixos níveis de Cafeína (SILVAROLLA et al., 2004). Foram também relatados teores de Cafeína variando de 0.76 (considerado descafeinado) a 1.65 g.100 g^{-1}; 1.00 a 1.50 g .100 g^{-1} e 0.72 a 1.23 g.100 g^{-1} para acessos etíopes cultivados Brasil, respectivamente (SILVAROLLA et al., 2000, TESSEMA et al., 2011; SCHOLZ et al., 2016).

No presente estudo, foram encontrados teores de Cafeína variando de 0.94 a 1.66 g.100 g^{-1} para os acessos E478/408 e E144/447, respectivamente. Esses resultados sugerem alta variabilidade nos níveis de cafeína em *C. arabica*. Porém, por ser um componente derivado de compostos nitrogenados, seu aumento pode ser inlfuenciado pelos processos de adubações aplicados a partir de 2011.

Os teores de PT variaram de 12.04 a 16.89 g.100^{-1}. Em um estudo com variedades comerciais e cafés etíopes, foram encontradas valores de 14.5 a 17.0 g.100^{-1} e 12.33 a 16.78 g.100 g^{-1} em grãos de café verde (SCHOLZ et al., 2011; SCHOLZ et al., 2016). Altos teores de PT foram encontrados em cultivares modernas, segundo Kitzberger et al., 2013a (16.10 a 18.0 g.100^{-1}). Esses resultados demonstram que o teor de proteína é altamente variável, e pode estar relacionado com técnicas culturais, como o tipo e nível de adubação; diferentes condições climáticas e também pela característica genética dos genótipos.

Tabela 1. Valores dos compostos químicos quanto à média, máximo, mínimo e variação fenotípica dos acessos de *C. arabica* na safra 2016 (n = 68).

Variável	Mínimo	Máximo	Média	Desvio Padrão	Variância
ACG[a]	3,82	9,30	6,56	0,944	0.89
Cafeína[a]	0,94	1,66	1,30	0,148	0.02
Lipídeos Totais[a]	11,76	16,06	13,91	0,827	0.68
Proteínas Totais[a]	12,04	16,89	14,46	0,999	0.99
Sacarose[a]	4,64	8,38	6,51	0,949	0.90
Cafestol[b]	152,91	1088,40	620,66	201,112	40.44
Caveol[b]	303,15	1144,10	723,63	183,324	33.60
Açúcares Totais[a]	5,04	8,90	6,97	0,952	0.91

ACG – Ácidos Clorogênicos
[a] Expresso em g.100 g^{-1}
[b] Expresso em mg.100g^{1}

Os ACGs são encontrados em maior quantidade em grãos de café verde representando 5% da massa seca no perisperma (75 DAF) e se acumulam por um curto período tempo. Esses compostos decaem durante o desenvolvimento do endosperma (180 DAF) e, portanto, estão envolvidos na maturação dos grãos (JOËT et al., 2009; JESZKA‐SKOWRON et al., 2016).

O conteúdo dos ACGs em cultivares de café varia de 6 a 12 g.100 g^{-1} (FARAH et al., 2006). SCHOLZ et al. (2016) encontraram teores de ACGs variando de 5.63 a 10.11 g.100 g^{-1}. Neste estudo o conteúdo de ACGs também apresentou valores variáveis de 3.82 a 9.30 g.100 g^{-1}, apresentando acessos com valores extremos desse composto.

Sabe-se que bebidas com teor de Sacarose mais elevado são preferidas, pois a sacarose é um importante precursor do aroma e sabor. Isso porque durante a torrefação o açúcar, as proteínas e aminoácidos reagem para formar típicos compostos do aroma e sabor no café torrado (SELMAR et al., 2008). Foram relatados conteúdos de Sacarose em acessos de *C. arabica* da Etiópia variando de 6.40 a 9.06 g.100^{-1} e 6. 36 a 10.42 g.100^{-1} (TESSEMA et al., 2011; SCHOLZ et al., 2016). Valores de 6.19 a 9.27 g.100 g^{-1} foram encontrados em variedades de café obtidos nas mesmas condições edafoclimáticas (KITZBERGER et al., 2013b). Nos acessos avaliados neste trabalho, os valores de sacarose variaram de 4.64 a 8.38 g. 100^{-1}, demonstrando que os acessos etíopes possuem grande variabilidade quanto a esse composto.

Os lipídeos são responsáveis pela retenção do aroma e pela produção de espuma no café expresso. A maioria dos compostos formados durante a

torrefação são solúveis em lipídeos e, portanto, ficam retidos na fase lipídica, o que confere aroma à bebida (SCHOLZ et al., 2016). A espécie *C. arabica* apresenta um maior teor lipídico, principalmente em relação ao seu parental *C. canephora*. Para os genótipos desse estudo o teor lipídico variou de 11.76 a 16.06 g.100^{-1}, sendo valores próximos aos encontrados anteriormente (13.58 a 16.58 g.100^{-1}) e superiores aos obtidos em cultivares comerciais e modernas (12.0 a 14.40 g.100^{-1}) (SCHOLZ et al., 2011; SCHOLZ et al., 2016; KITZBERGER et al., 2013a).

16.2 CORRELAÇÃO ENTRE AS VARIÁVEIS

Foram observadas correlações positivas significativas entre Cafeína com PT; LT com Caveol; ACG com LT, PT, AT, Sacarose; e Sacarose com AT. Correlação negativa significativa foi encontrada entre LT com Sacarose/AT e Caveol com Cafestol.

Tabela 2. Correlação de Pearson entre compostos em acessos de *C. arabica* da Etiópia safra 2016 (n = 68).

Variáveis	ACG	Cafeína	LT	PT	Sacarose	Cafestol	Caveol	AT[a]
ACG[a]	**1.0**							
Cafeína[a]	0.17	**1.0**						
LT[a]	**0.25**	-0.21	**1.0**					
PT[a]	**0.31**	**0.65**	0.01	**1.0**				
Sacarose[a]	**0.27**	0.06	**-0.42**	-0.12	**1.0**			
Cafestol[b]	0.15	-0.10	0.01	-0.17	-0.04	**1.0**		
Caveol[b]	-0.06	-0.12	**0.31**	0.03	-0.13	**-0.59**	**1.0**	
AT[a]	**0.27**	0.02	**-0.39**	-0.14	**0.99**	-0.04	-0.09	**1.0**

Os valores em negrito são diferentes de 0 com níveis de significância α = 0,05. ACG – Ácidos Clorogênicos; [a]Expresso em g.100 g^{-1}; [b]Expresso em mg.100g^{-1}.

Durante o desenvolvimento do grão de café, há um acúmulo contínuo de Sacarose/AT juntamente com Lipídeos por volta de 105 DAF, sendo que para Lipídeos o acúmulo é até 180 e diminui gradativamente. Podemos observar esta relação sendo significativa e negativa (r = -0.42, -0.39, p-value < 0.05), já que o acúmulo de Lipídeos diminui enquanto que o de Sacarose continua a acumular até o término do desenvolvimento do fruto (JOËT et al., 2009; 2010; SCHOLZ et al., 2016).

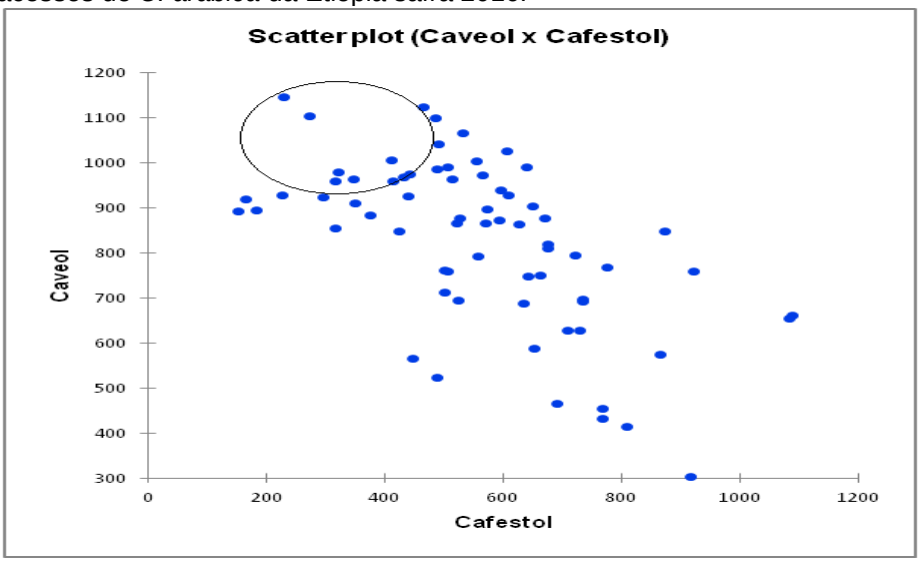
Nós observamos uma negativa correlação entre as concentrações dos compostos Caveol e Cafestol (r = - 0.59, p-value < 0.05), demonstrando que no

estágio final de maturação do fruto há uma relação inversa entre o acúmulo dos mesmos (Figura 1). Scholz et al., 2016 e Sant'Ana et al., 2018 também observaram correlação negativa entre ambos compostos ($r = -0.55$, p-value < 0.05/ $r = -0.30$, p-value < 0.005). O conteúdo de Caveol também mostrou correlação com o conteúdo de LT ($r = 0.31$, p-value < 0.05), enquanto que para Cafestol não houve uma correlação significativa. Sant'Ana et al. (2018) obtiveram um resultado semelhante ($r = 0.29$, p-value < 0.005).

O efeito dos diterpenos Caveol e Cafestol na saúde humana já foi relatado em outras espécies e cultivares de café (KITZBERGER et al., 2010; BUTT; SULTAN., 2011). No entanto, estudos do perfil desses compostos nos acessos da Etiópia estão sendo conduzidos desde 2012 (PAGIATTO et al., 2012; SCHOLZ et al., 2014b; SCHOLZ et al., 2016), e a partir desse conjunto de dados, será possível selecionar genótipos de interesse.

O gráfico de dispersão selecionou genótipos-alvo com maiores teores de Caveol e baixos de Cafestol. Neste estudo 9 acessos da Etiópia apresentaram valores de Caveol variando de 957.46 a 1.144.10 $mg.100g^{-1}$ e de Cafestol variando de 228.84 a 442.62 $mg.100g^{-1}$ (E151/574, E302/083, E159/180, E458/097, E146/717, E081/041, E068/014, E261/053 e E196/117) (Figura 1).

Figura 1. Variação do teor de Cafestol em função do teor de Caveol de 68 acessos de *C. arabica* da Etiópia safra 2016.



Embora existam estudos sobre a caracterização de putativos genes relacionados à biossíntese de Caveol e Cafestol em café, pouca informação sobre a formação desses diterpenos está disponível, principalmente dos genes envolvidos nas etapas finais de biossíntese. Porém os resultados obtidos até o momento sugerem que a síntese de um composto pode ser dependente da inibição do outro na via (WILLE et al., 2004; ROBERTS, 2007; WANG et al., 2012; IVAMOTO et al., 2016).

Tanto as Proteínas quanto a Cafeína são compostos nitrogenados que se acumulam durante a formação do fruto e ambas acumulam no grão em função da disponibilidade de nitrogênio. Foi observada correlação significativa positiva ($r = 0.65$, p-value < 0.05) entre esses compostos (SCHOLZ et al., 2016). A correlação positiva entre Cafeína e ACG ($r = 0.31$, p-value < 0.05) provavelmente é pelo fato do ligeiro aumento de ACGs que ocorre por volta dos 225 DAF onde as proteínas estariam acumulando (JOËT et al., 2009; 2010).

A correlação positiva entre ACGs e Sacarose/AT ($r = 0.27$, pvalue < 0.05) pode ser explicada pelo fato dos frutos terem sido colhidos por volta de 225 DAF, e é após esse estágio de maturação que os ACGs decaem abruptamente em relação à Sacarose, ao passo que posteriormente passam a se acumular conjuntamente no fruto. O mesmo raciocínio vale para a correlação positiva entre ACGs e LT ($r = 0.25$, pvalue < 0.05), pois os ACGs decaem entorno de 180 DAF e voltam a se acumular em torno de 225 DAF (JOËT et al., 2009).

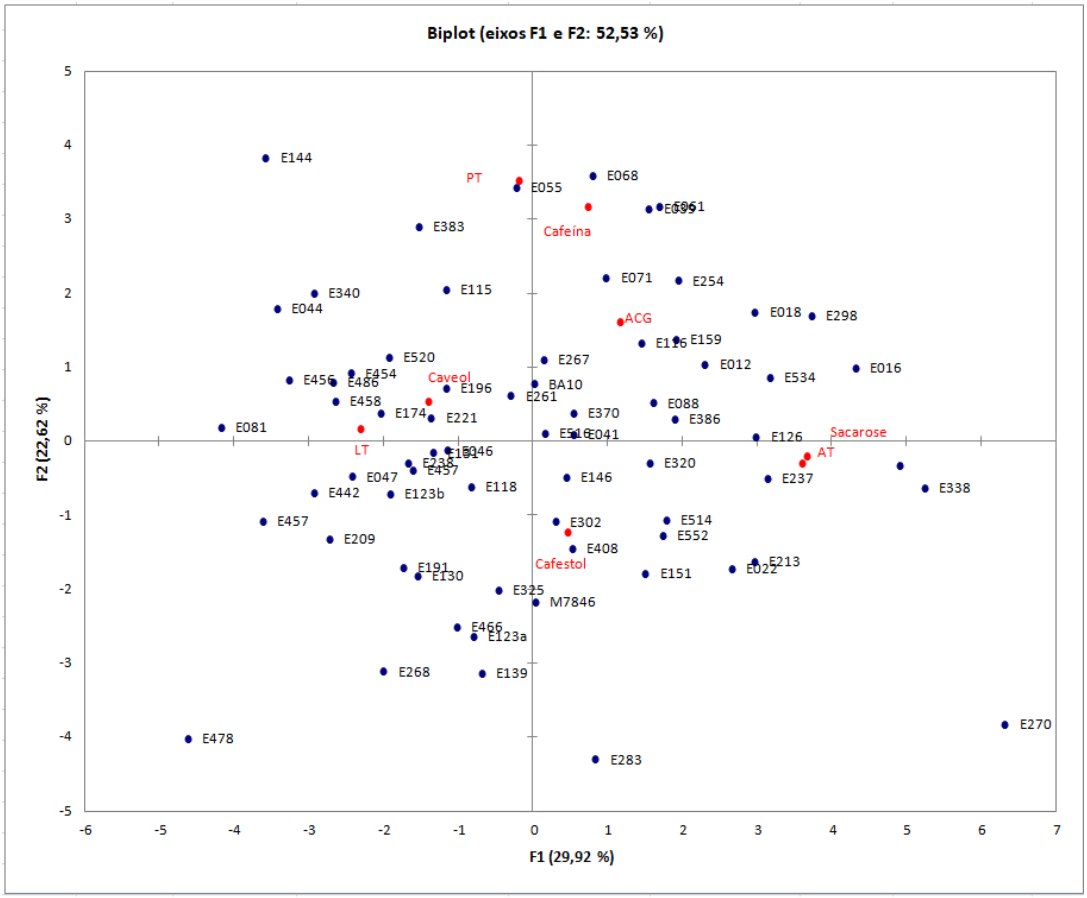
16.3 ANÁLISE MULTIVARIADA

A aplicação de análises multivariadas a uma matriz de dados é uma forma eficiente de discriminar genótipos em grupos e associá-los com variáveis discriminantes. A análise de componentes principais (ACP) reduz as dimensões de um conjunto de dados e explica sua estrutura pela variância de cada composto. Esse tipo de análise é essencial para identificar possíveis grupos em uma população com base na composição química (BERTRAND et al., 2012). A projeção dos acessos e as variáveis podem ser observadas no biplot da Figura 2, onde os dois primeiros eixos explicam 52,53% da variação total (29,92% para CP1 e 22,62% para CP2, respectivamente).

Sacarose/AT, Caveol e LT foram os compostos que mais contribuíram para a formação do primeiro componente, enquanto que Cafestol, Cafeína e PT contribuíram para a formação do segundo componente.

Os acessos que estão entre o CP1 < 0 e CP2 > 0 são caracterizados por apresentarem altos níveis de Caveol, LT, PT e menores valores de Cafestol. Essas características fazem desses acessos, alvos para a produção de novas variedades com melhor qualidade da bebida, pois esses compostos são responsáveis pela obtenção do aroma e sabor na bebida final. Vale ressaltar que para os acessos desse grupo, os níveis de Sacarose foram relativamente dentro dos limites para a produção de uma bebida de qualidade (Figura 2). Os acessos que estão entre o CP1 > 0 e CP2 < 0 se destacaram pelos altos teores de Scarose e AT, com destaque para o melhoramento de café visando a qualidade da bebida, os acessos E270 e E338.

Figura 2. ACP dos compostos bioquímicos de 68 acessos de café da Etiópia da safra 2016.



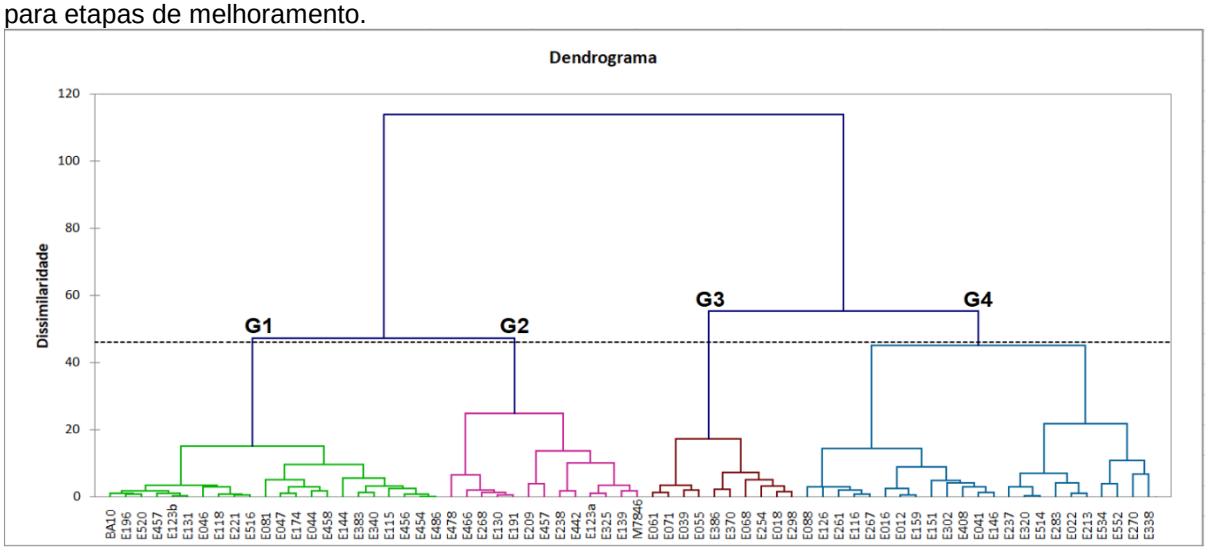
Os acessos que estão entre CP1 < 0 e CP2 < 0 não demonstraram padrões de interesse ao melhoramento e à saúde humana, pelo fato de possuírem altos teores de Cafestol e baixos de Caveol, Sacarose e AT.

A análise de agrupamento hierárquico (AAH) (Tabela 3) classificou os grupos com base em uma matriz de dissimilaridade entre os genótipos formando um dendrograma apresentado na Figura 3. A AAH é uma análise multivariada que classifica e agrupa genótipos baseado em uma matriz de dissimilaridade de dados de composição química. A AAH entre os 68 acessos indicou a formação de 4 grupos distintos conforme observados na Tabela e Figura 3.

Tabela 3. Classificação de agrupamento hierárquico baseada nos 68 acessos de café da Etiópia.

Classe	G1	G2	G3	G4
Número de genótipos	23	10	22	13
	E012	E018	E044	E123a
	E016	E039	E046	E130
	E022	E055	E047	E139
	E041	E061	E081	E191
	E088	E068	E144	E209
	E116	E071	E115	E238
	E126	E254	E118	E268
	E146	E298	E123b	E325
	E151	E386	E131	E442
	E159	E370	E174	E457
	E213		E196	E466
	E237		E221	E478
	E261		E383	M7846
	E267		E340	
	E270		E454	
	E283		E456	
	E302		E457	
	E320		E458	
	E338		E486	
	E408		E516	
	E514		E520	
	E534		BA10	
	E552			

Figura 3. Dendrograma com base na análise de agrupamento hierárquico (AAH) para os compostos químicos de 68 acessos de *C. arabica* safra 2016. Destaque para o Grupo 1 e 3, selecionando cafés para etapas de melhoramento.



Na tabela 4 estão a média, mínimo e máximo dos 8 compostos químicos de grãos de café verde a partir dos 4 grupos definidos pela AAH (Figura 3) de 68 acessos da Etiópia. Os acessos do grupo 1 apresentaram alto conteúdo de Caveol, LT, PT, Sacarose/AT, bem como baixos conteúdos de Cafeína e Cafestol.

No grupo 2 estão os acessos que apresentaram as maiores concentrações de ACGs e Cafeína, bem como concentrações equivalentes de Caveol e Cafestol, e no grupo 3 estão os acessos com altos teores de PT, LT, AT, e baixos de Cafestol e ACGs. Já no grupo 4 estão os acessos com concentrações equivalentes de Caveol e Cafestol, além de baixos teores de PT e Sacarose.

Para que um café seja considerado de qualidade superior é necessário que apresente baixos teores de ACGs e Cafeína, devido à interferência destes no sabor final da bebida. Além disso, se for possível agregar a essa bebida de café um conteúdo naturalmente maior de açúcar e lipídeos, a bebida poderá apresentar um maior padrão de qualidade. Assim, para fins de melhoramento da qualidade da bebida, os cafés contidos no grupo 1 e 3 são os mais indicados. Porém se o interesse é um aumento de benefícios à saúde humana, o grupo 3 seria o mais indicado.

Tabela 4. Valores médios para oito compostos químicos entre os quatro grupos definidos pela AAH e entre 68 acessos de *C. arabica* da Etiópia, safra 2016.

	Grupos	**ACG**	**Cafeína**	**LT**	**PT**	**Sacarose**	**AT**	**Cafestol**	**Caveol**
Máximo		7.55	1.60	15.25	15.74	8.38	8.90	1.082,75	1.144,10
Média	**G1**	6.16	1.31	13.63	14.42	7.31	7.75	518.39	844.59
Mínimo		4.40	1.13	11.76	12,04	6,10	6,62	152,91	303,15
Máximo		9.30	1.65	15.16	16.89	8.01	8.54	922.76	958.81
Média	**G2**	7.77	1.46	14.40	15.86	6.92	7.28	610.85	707.48
Mínimo		6.25	1.29	13.64	14.53	5.91	6.05	414.30	522.28
Máximo		7.28	1.66	16.06	16.78	6.79	7.25	675.34	1.122,70
Média	**G3**	6.36	1.31	14.67	15.28	5.76	6.21	495.49	948.12
Mínimo		4.87	1.17	14.04	14.18	4.98	5.26	228.10	761.02
Máximo		7.23	1.34	15.43	15.04	6.61	7.25	1.088,40	1.039,99
Média	**G4**	5.95	1.14	14.62	14.00	5.67	6.12	689.76	715.59
Mínimo		3.82	0.94	13.27	12.59	4.64	5.04	491.62	414.27

ACG - Ácidos clorogênicos
[a] Expresso em g.100 g^{-1}
[b] Expresso em mg.100g^{-1}

17 CONCLUSÃO

A análise da composição química dos acessos da Etiópia cultivados no IAPAR revelou alta variabilidade neste material. Além disso foi possível identificar genótipos e grupos com potencial para produção de uma bebida de alta qualidade e/ou com benefícios à saúde humana.

18 REFERÊNCIAS

ADDINSOFT XLSTAT: **software for statistical analysis**. Versão 2008.4.02 (2008). Paris. 1 CD-ROM, 2010.

ANTHONY, F.; BERTRAND, B.; QUIROS, O.; WILCHES, A.; LASHERMES, P.; BERTHAUD, J.; CHARRIER, A. Genetic diversity of wild coffee (*Coffea arabica L*.) using molecular markers. **Euphytica**. v.118, p.53-65, 2001.

BERTRAND, B.; GUYOT, B.; ANTHONY, F.; LASHERMES, P. Impact of the *Coffea canephora* gene introgression on beverage quality of *C. arabica*. **Theor. Appl. Genet**, 107:387–394, 2003.

BERTRAND, B.; BOULANGER, R.; DUSSERT, S.; RIBEYRE, F.; BERTHIOT, L.; DESCROIX, F.; JOËT, T. Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality. **Food. Chem**. 135:2575–2583, 2012.

BHUPATHIRAJU, S.N.; PAN, A.; MALIK, V.S.; MANSON, J.E.; WILLETT, W.C.; VAN DAM, R.M.; HU, F.B. Caffeinated and caffeine-free beverages and risk of type 2 diabetes. **Am .J. Clin. Nutr.** 97:155–166, 2013.

CANO-MARQUINA, A.; TARÍN, J.J.; CANO, A. **The impact of coffee on health**. Maturitas 75:7–21, 2013.

SILVAROLLA, M.B.; MAZZAFERA, P.; FAZUOLI, L.C. Plant biochemistry: a naturally decaffeinated arabica coffee. **Nature**. 429(June):826, 2004.

CHENG, B.; FURTADO, A.; SMYTH, H.E.; HENRY, R.J. Influence of genotype and environment on coffee quality. **Trends. Food. Sci. Technol**. 57:20–30, 2016.

COMPANHIA NACIONAL DE ABASTECIMENTO (Brasil). Acompanhamento da safra brasileira de café: Segundo levantamento – Safra 2016. Brasília: CONAB, 2016. Disponível em: <http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16_06_10_15_13_24_boletim_cafe_-_maio_2016.pdf> Acesso em: 09 de maio de 2016.

DAVIS, A.P.; TOSH, J.; RUCH, N.; FAY, M.F. Growing coffee: Psilanthusb (*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, 2011.

FARAH, A.; DONANGELO, C.M. Phenolic compounds in coffee. **Braz. J. Plant. Physiol.** 18:23–26, 2006.

IVAMOTO, S.T.; POT, D.; CHARMETANT, P.; MARRACCINI, P.; FERREIRA, L.P.; DOMINGUES, D.S.; VIEIRA, L.G.E.; PEREIRA, L.F.P. Diterpenes in *Coffea arabica*: biochemical aspects and transcriptional analysis of candidate CYPS involved in cafestol and kahweol biosynthesis. **Plant. Physiol. Biochem**. 111, 340-347. 2016.

JESZKA‐SKOWRON, M.; SENTKOWSKA, A.; PYRZYNSKA, K.; MARIA PAZ DE PEÑA, M. Chlorogenic acids, caffeine content and antioxidant properties of green coffee extracts: influence of green coffee bean preparation. **Eur. Food. Res. Technol**. 242:1403–1409, 2016.

JOËT, T.; LAFFARGUE, A.; SALMONA, J.; DOULBEAU, S.; DESCROIX, F.; BERTRAND, B.; DE KOCHKO, A.; DUSSERT, S. Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study. **New. Phytol**. 182:146–162, 2009.

JOËT, T.; LAFFARGUE, A.; DESCROIX F.; DOULBEAU, S.; BERTRAND, B.; KOCHKO, A.; DUSSERT, S. Influence of environmental factors, wet processing and their interactions on the biochemical composition of green coffee beans. **Food. Chem**. 118:693–701, 2010.

KITZBERGER, C.S.G.; SCHOLZ, M.B.S.; PEREIRA, L.F.P.; VIEIRA, L.G.E.; SERA, T.; SILVA, J.B.G.D.; BENASSI, M.T. **Analysis of diterpenes in green and roasted coffee of *Coffea arabica* cultivars growing in the same edapho-climatic conditions**. In: 23 International conference on coffee Science, vol 1, Bali, pp 110–117, 2010.

KITZBERGER, C.S.G; SCHOLZ, M.B.S; PEREIRA, L.F.P; BENASSI, M.T. Composição química de cafés arábica de cultivares tradicionais e modernas. **Pesquisa Agropecuária Brasileira**. 48:1498–1506, 2013a.

KITZBERGER, C.S.G.; SCHOLZ, M.B.S.; PEREIRA, L.F.P.; VIEIRA, L.G.E.; SERA, T.; SILVA, J.B.G.D.; BENASSI, M.T. Diterpenes in green and roasted coffee of Coffea arabica cultivars growing in the same edapho-climatic conditions. **J. Food. Compos. Anal**. 2013b.

LEROY, Thierry et al. **Genetics of coffee quality.** Braz. J. Plant. Physiol. vol.18, n.1, p.229-242, 2006.

MURIEL, P; ARAUZ J**. Coffee and liver diseases**. Fitoterapia, 8:297–305, 2010.

PAGIATTO, N.F.; IVAMOTO, S.; SERA, T.; KITZBERGER, C.S.G.; SCHOLZ, M.B.S.; CHARMETANT, P.; LERROY, T.; DOMINGUES, D.; PEREIRA, L.F.P. **Evaluation of kahweol and cafestol levels in Ethiopian *Coffea arabica L*. accessions grown in Brazil**. In: 24 International conference on coffee science, San Jose´, p 217, 2012.

PEREIRA, L. F. P & IVAMOTO, S. T. Chapter 6: Characterization of coffee genes involved in isoprenoid and diterpene metabolic pathways. In: *Coffee in Health and Disease Prevention* (Preedy, R. V. Ed.). London: Academic Press, 45-51, 2015.

PERRONE, D.; FARAH, A.; DONANGELO, C.M.; DE PAULIS, T.; MARTIN, P.R. Comprehensive analysis of major and minor chlorogenic acids and lactones in economically relevant Brazilian coffee cultivars. **Food. Chem**.106:859–867, 2008.

POT, D.; SCHOLZ, M. B. S.; LANNES, S. D.; DEL GROSSI, L.; PEREIRA, L. F. P.; VIEIRA, L. G.; SERA, T. Phenotypic analysis of *Coffea arabica* acessions from Ethiopia: contribution to the understanding of *Coffea arabica* diversity. In: 22nd INTERNATIONAL CONFERENCE ON COFFEE SCIENCE, Campinas. ***Anais*…** Campinas, p.165, 2010.

SANT'ANA, GUSTAVO C.; PEREIRA, LUIZ F. P.; POT, DAVID.; IVAMOTO, SUZANA T.; DOMINGUES, DOUGLAS S.; FERREIRA, RAFAELLE V.; PAGIATTO, NATALIA F.; DA SILVA, BRUNA S. R.; NOGUEIRA, LÍVIA M.; KITZBERGER, CINTIA S. G.; SCHOLZ, MARIA B. S.; DE OLIVEIRA, FERNANDA F.; SERA, GUSTAVO H.; PADILHA, LILIAN.; LABOUISSE, JEAN-PIERRE.; GUYOT, ROMAIN.; CHARMETANT, PIERRE.; LEROY, THIERRY. Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. **Sci. Rep**. 8:465, 2018.

SCHOLZ, M.B.S.; FIGUEIREDO, V.R.G.; SILVA, J.V.N.; KITZBERGER, C.S.G. Características físico-químicas de grãos verdes e torrados de cultivares de café do Iapar. **Coffee. Sci**. 6:245–255, 2011.

SCHOLZ, M.B.S; KITZBERGER, C.S.G; PEREIRA, L.F.P; DAVRIEUX, F; POT, D; CHARMETANT, P; LEROY, T. Application of near infrared spectroscopy for green coffee biochemical phenotyping. **J.Near. Infrared. Spectrosc**. 22:411–421, 2014b.

SCHOLZ, M. B. S.; GOOD-KITZBERGER, C. S.; PAGIATTO, N. F.; PROTASIO, P. L. F.; DAVRIEUX, F.; POT, D.; CHARMETANT, P.; LEROY, T. Chemical composition in wild Ethiopian Arabica coffee accessions. **Euphytica**. 10 p, 2016.

SELMAR, D.; BYTOF, G.; KNOPP, S. The storage of green coffee (*Coffea arabica*): decrease of viability and changes of potential aroma precursors. **Ann. Bot**. 101:31–38, 2008.

SILVAROLLA, M.B.; MAZZAFERA, P.; LIMA, M.M.A. Caffeine content of Ethiopian coffee arabica beans. **Genet. Mol. Biol**, 23:213–215, 2000.

SILVAROLLA, M.B.; MAZZAFERA, P.; FAZUOLI, L.C. Plant biochemistry: a naturally decaffeinated Arabica coffee. **Nature**, 429:826, 2004.

SOUZA, R.M.N.; BENASSI, M.T. Discrimination of commercial roasted and ground coffees according to chemical composition. **Journal of the Brazilian Chemical Society**, v.23, p.1347-1354, 2012

TESSEMA, A.; ALAMEREW, S.; KUFA, T.; GAREDEW, W. Variability and association of quality biochemical attributes in some promising *Coffea arabica* germplasm collections in Southwestern Ethiopia. **Int. J. Plant .Breed. Genet**. 5:302–316. 2011.

UPADHYAY, L.J.; MOHAN RAO. **An outlook on chlorogenic acids—occurrence, chemistry, technology, and biological activities.** Critical Reviews in Food Science and Nutrition, 53 (9), pp. 968–984, 2013.

WANG, Q.; HILLWIG, M. L.; WU, Y.; PETERS, R. J. CYP701A8: A rice ent-kaurene oxidase paralog diverted to more specialized diterpenoid metabolism. **Plant Physiology**. v.158, n.1, 2012.

WILLE, A.; ZIMMERMANN, P.; VRANDÓVA, E.; FÜRHOLZ, A.; LAULE, O.; BLEULER, S. HENNIG, L.; PRELIC, A.; ROHT, P. V.; THIELE, L.; ZITZLER, E.; GRUISSEM, W; BÜHLMANN, P. Sparse Graphical Gaussin Modeling of the Isoprenoid Gene Network in *Arabidopsis thaliana*. **Genome Biology**. V. 5, n. 1, 2004.

19 MATERIAL SUPLEMENTAR

Suplementar 1 – Conteúdo de Ácidos Clorogênicos (ACG), Cafeína, Lipídeos Totais, Proteínas Totais, Sacarose e os Diterpenos Cafestol e Caveol encontrados nos 68 genótipos de *C. arabica* da Etiópia na safra 2016.

Identificação dos acessos	ACG[a]	Cafeína[a]	Lipídeos Totais[a]	Proteínas Totais[a]	Sacarose[a]	Cafestol[b]	Caveol[b]
E012/226	6,64	1,37	13,88	14,94	7,59	182,93	894,77
E016/456	6,16	1,42	12,85	14,93	8,37	152,91	892,61
E018/494	8,98	1,34	14,25	15,48	7,44	523,36	694,58
E022/163	6,93	1,37	14,17	12,88	7,52	776,90	767,22
E039/435	7,59	1,61	14,39	15,95	6,70	489,44	522,28
E041/079	5,37	1,46	14,03	14,29	6,87	424,74	848,02
E044/122	6,13	1,40	15,37	15,44	5,29	228,10	926,52
E046/021	6,86	1,31	14,19	14,28	5,75	501,76	761,02
E047/267	5,43	1,28	14,06	14,39	5,52	375,06	883,87
E055/005	7,37	1,49	13,92	16,88	5,91	447,79	564,24
E061/128	6,25	1,65	13,64	16,52	6,92	642,86	746,86
E068/014	8,45	1,54	15,16	15,94	6,97	414,30	958,81
E071/258	6,74	1,55	13,89	15,63	6,53	501,90	712,53
E081/041	6,85	1,25	16,06	14,46	5,28	412,77	1003,93
E088/034	6,38	1,46	12,78	14,06	6,75	297,30	922,11
E144/447	5,89	1,66	14,99	16,78	4,98	572,78	895,33
E115/099	6,32	1,45	14,04	16,34	5,69	675,34	817,61
E116/286	6,55	1,38	13,11	15,74	6,75	527,09	876,22
E118/213	5,86	1,17	14,64	15,20	6,45	596,31	939,04
E123B/121	6,76	1,22	14,98	14,18	5,97	607,56	1024,17
E123A/231	5,69	1,09	14,30	14,29	5,90	809,98	414,27
E126/359	6,11	1,28	12,33	15,35	7,27	506,23	759,54
E130/169	6,64	1,15	14,98	13,64	5,96	627,60	861,93
E131/378	6,46	1,27	14,74	14,76	6,08	641,22	989,17
E139/054	5,67	1,12	14,53	13,20	6,18	653,75	587,93
E146/717	5,65	1,27	14,28	14,46	6,97	346,56	963,83
E151/574	5,43	1,18	12,81	13,48	7,28	228,84	1144,10
E159/180	6,65	1,34	13,81	15,63	7,44	322,75	977,40
E174/164	4,87	1,41	14,46	14,80	5,76	317,17	853,70
E191/091	6,07	1,13	15,43	14,02	6,16	522,56	864,84
E196/117	6,71	1,32	14,70	14,96	6,10	430,89	967,62
E209/031	4,80	1,24	14,12	15,01	4,93	767,55	453,20
E213/021	7,13	1,23	13,86	13,76	7,50	734,74	691,97
E221/330	6,46	1,21	14,39	15,65	5,94	609,33	927,38
E237/071	7,55	1,16	14,23	14,66	7,99	515,45	961,85
E238/022	6,70	1,33	14,93	15,04	5,46	1088,40	660,05
E254/284	7,68	1,33	14,79	16,89	7,23	735,29	696,23
E261/053	5,32	1,39	13,00	15,18	6,10	442,62	974,75
E267/090	6,81	1,31	13,10	15,56	6,19	440,64	925,63
E268/067	6,69	0,99	14,95	13,09	5,84	594,25	871,92
E270/263	5,67	1,15	11,76	13,43	8,29	917,41	303,15
E283/096	5,91	1,13	14,59	12,04	6,98	873,29	846,34
E298/382	9,30	1,29	14,89	15,87	8,01	722,41	792,87
E302/083	6,06	1,14	15,25	14,19	7,44	274,30	1101,74
E320/145	6,49	1,32	14,94	14,71	7,38	674,58	808,81
E325/523	6,66	1,17	14,73	13,91	5,97	767,59	431,45
E383/142	7,28	1,43	14,60	16,23	5,74	350,84	910,27

E386/131	7,77	1,38	14,91	14,53	7,19	708,36	627,21
E338/218	6,98	1,35	13,24	14,19	8,38	634,21	688,39
E340/179	7,25	1,38	14,22	15,69	5,07	489,29	984,93
E370/314	7,52	1,38	14,17	14,92	6,33	922,76	759,15
E408/192	4,40	1,18	14,22	14,34	7,40	166,12	918,85
E442/279	7,23	1,19	15,19	14,81	4,88	865,80	573,49
E457/073	3,82	1,34	13,27	14,47	4,64	557,96	791,45
E454/107	5,99	1,32	14,60	15,59	5,57	506,99	989,49
E456/062	6,38	1,34	14,97	15,38	5,10	650,38	903,40
E457/477	6,30	1,18	14,93	14,76	6,26	466,68	1122,70
E458/097	5,65	1,21	14,66	15,70	5,63	315,98	957,46
E466/125	6,40	1,02	15,29	13,54	6,61	491,62	1039,99
E478/408	5,65	0,94	14,53	12,59	4,67	556,52	1002,61
E486/189	6,12	1,31	14,69	15,49	5,49	565,69	972,59
E514/129	6,33	1,23	14,64	14,44	7,58	670,29	876,10
E516/069	6,41	1,25	14,05	15,30	6,54	569,53	865,88
E520/04	6,91	1,26	14,60	15,65	5,81	486,19	1098,43
E534/202	5,22	1,60	13,20	15,19	7,40	730,41	627,91
E552/323	5,88	1,41	13,33	14,18	6,67	1082,75	654,37
M7846/064	5,38	1,11	13,86	14,38	6,48	663,29	749,55
BA10/57	7,05	1,30	14,75	15,16	6,79	531,01	1064,07

[a]Expresso em g.100 g^{-1}; [b]Expresso em mg.100g^{-1}.

20 CAPÍTULO 3: Diversidade, estrutura populacional e associação genômica ampla (GWAS) em *Coffea arabica* para características relacionadas com a qualidade da bebida de café.

RESUMO

Os derivados de constituintes nitrogenados, lipídicos e fenólicos, bem como a sacarose e açúcares totais presentes em grãos de café verde contribuem para atributos do sabor e aroma da bebida de café, e portanto, para a qualidade da bebida. Nesse estudo foi realizada genotipagem por sequenciamento (GBS) de 159 genótipos de *C. arabica,* incluindo acessos provenientes do centro de origem, cultivares tradicionais e variedades. A partir dos arquivos *fastq* obtidos do sequenciamento GBS, o *Pipeline* TASSEL v5 foi realizado para uma nova montagem, mapeamento dos *tags* e busca de putativos SNPs utilizando os dois parentais diploides da espécie (*C. canephora* e *C. eugenioides*) como referência. Após o controle de qualidade dos dados (Heterozigosidade < 0.5, MAF > 0.05, *Call Rate* 0.8 e retirada de multialelos), foram obtidos 1.719 e 2.949 SNPs, com uma cobertura média de 68X e 47X, para ambos subgenomas *C. canephora* e *C. eugenioides*, respectivamente. Análise de estrutura populacional sugeriu a presença de 2 e 3 grupos (K = 2 and K = 3) correspondentes aos acessos do lado Leste do Vale do Rift e acessos do lado Oeste do Vale do Rift. Um grupo com maior diversidade foi formado por acessos selvagens coletados em florestas e parques de reservas. A fim de identificar marcadores e regiões genômicas associadas a características bioquímicas relacionadas com a qualidade da bebida, o GWAS foi conduzido com oito modelos de associação para nove compostos químicos, além de Diterpenos Totais, Média dos Diterpenos e Razão Caf/Cav entre 4 anos de coleta (2011, 2012, 2015 e 2016) e média (2011/2015). Cento e um acessos de *C. arabica* da histórica coleção da FAO com um total de 4.517 SNPs gerados pela união dos dois subgenomas foram utilizados para essa análise. Um total de 33 SNPs de *C. arabica* foram significativamente associados a compostos relacionados com a qualidade da bebida de café. Dos 22 SNPs de *C. arabica* mapeados em *C. canephora e* significativamente associados, nós identificamos 17 SNPs co-localizados a 13 genes candidatos potencialmente envolvidos nas vias metabólicas dos compostos bioquímicos (1 SNP associado com Cafeína, 4 com Proteínas Totais, 4 com Cafestol, 4 com Caveol e 4 com Razão Caf/Cav). Onze SNPs de *C. arabica* mapeados em *C. eugenioides* foram significativamente associados aos compostos bioquímicos, sendo 1 SNP associado com o conteúdo de Lipídeos Totais, 5 com Caveol, 4 com Cafestol e 1 com Razão Caf/Cav. Porém ainda não há anotação genômica para esse parental diploide. A partir da distribuição dos valores fenotípicos em relação aos SNPs significativamente associados, foi possível identificar padrões genotípicos contrastantes para cada composto analisado, inclusive 2 blocos de haplótipos significativamente associados a Caveol, Cafestol e Razão Caf/Cav que correspondem a regiões similares em ambos subgenomas. Nosso estudo ressaltou a importância de se conservar as plantas originadas de florestas e parques de reservas florestais oriundos do lado Oeste do Grande Vale do Rift presente em nosso banco de germoplasma do IAPAR. Também identificou SNPs para utilização em seleção assistida visando melhorar a qualidade bioquímica de grãos de café verde, assim como genes candidatos relacionados com a qualidade da bebida de café para posterior validação da expressão gênica.

Palavras-chave: Parentais diplóides. Marcadores SNPs. GWAS. Qualidade de Bebida.

21 INTRODUÇÃO

O gênero *Coffea* compreende um total de 124 espécies sendo que 10 são cultivadas em todo mundo (DAVIS et al. 2011). No Brasil, cerca de 70% da produção é de *Coffea arabica*, espécie mais importante economicamente (CONAB, 2018). O café arábica é preferido entre as demais espécies por apresentar uma bebida de maior qualidade (DE CASTRO e MARRACCINI, 2006).

A qualidade da bebida de café é uma característica complexa que é geneticamente controlada por múltiplos QTLs, sendo, em sua maioria, resultante da ação conjunta de vários genes e o melhoramento para essa característica é demorado e depende do conhecimento da genética e de seus componentes (LEROY et al., 2011). A composição química dos grãos do café cru faz parte dos fatores que determinam a qualidade da bebida e a identificação desses componentes é importante para indicar as características desejáveis (FERRÃO et al., 2003).

Durante os últimos anos, estudos visando mapeamento de QTLs (mapeamento de ligação) para o café foram conduzidos e muitos destes foram realizados com populações biparentais (PESTANA et al., 2010; LEROY et al., 2011; MÉROT-L'ANTHOËNE et al., 2014; MONCADA et al., 2016). Porém para o melhor entendimento dos mecanismos genéticos de características complexas envolvidas com a qualidade da bebida, mais genes relacionados podem estar relacionados. Isso porque no mapeamento de populações biparentais há uma limitação de diversidade genética e com isso, possui menor resolução de mapeamento, sendo então um empecilho a descoberta de novos genes. Em uma população natural existem fenótipos distintos que são influenciados por vários QTLs comuns, além de alguns fenótipos serem determinados por um ou poucos QTLs específicos. Assim, buscar genes responsáveis por determinado fenótipo é uma tarefa de grande relevância (PRADHAN et al., 2016).

O sistema reprodutivo de *C. arabica* é predominantemente autógamo (PRAKASH et al., 2002) e junto à sua origem recente e ao seu limitado histórico de dispersão (LASHERMES et al., 1999; VIDAL et al., 2010), confere a essa, baixa variabilidade genética. No entanto, como já relatado, uma maior

variabilidade é encontrada nos altiplanos do Sudoeste da Etiópia (ANTHONY et al., 2001; SILVESTRINI et al., 2007).

O Estudo de Associação Ampla do Genoma (GWAS) fornece uma estratégia promissora para mapear QTLs devido a um grande número de eventos históricos de recombinação que levam ao rápido decaimento do desequilíbrio de ligação (DL) (FLINT-GARCIA et al., 2003). Esse tipo de análise é considerada um ensaio natural, onde os dados genotípicos e fenotípicos são coletados de uma população em que parentesco não é controlado pelo pesquisador. A ocorrência não-aleatória entre alelos dentro de um cromossomo é chamada de desequilíbrio de ligação. Assim o GWAS provou ser muito útil para dissecar características quantitativas complexas com base na abordagem de mapeamento por desequilíbrio de ligação (HUANG et al., 2010).

Existem duas abordagens alternativas em estudos de mapeamento de associação: estudo de associação genômica ampla e mapeamento de associação do gene candidato. A associação genômica ampla ("*Genome Wide Association Study – GWAS"*) refere-se a uma varredura do genoma que busca a variação genética causal usando um grande número de marcadores em todo o genoma (SHEHZAD et al., 2009). Alternativamente, o mapeamento de genes candidatos ("*Candidate Gene Approach*") relaciona polimorfismos em genes candidatos já conhecidos por afetar um fenótipo em particular. Os genes candidatos são selecionados a partir de análise mutacional, vias bioquímicas definidas em outras espécies modelo ou outros estudos de ligação (ZHU et al., 2008).

Estudos de associação têm avançado em café, como demonstrou um recente estudo de ANDRADE et al. (2017), no qual desenvolveram e validaram um Chip de genotipagem (26K Axiom SNP array) para *C. canephora* com boa cobertura das sequências. Nesse chip foram validados 20.982 SNPs visando posteriores estudos de GWAS. Recentemente nosso grupo de pesquisa publicou um estudo pioneiro de GWAS para qualidade da bebida de café utilizando acessos silvestres de café Arábica, no qual foram detectados 21 SNPs significativamente associados aos conteúdos de Lipídeos Totais, Cafestol, Caveol e Razão Caf/Cav e 9 putativos genes (SANT'ANA et al., 2018). No

entanto esse trabalho tinha como uma das limitações a utilização de apenas um genoma de referência (*C. canephora*) para a identificação dos SNPs.

Para o melhor entendimento dos mecanismos genéticos relacionados com a qualidade da bebida de café, neste trabalho utilizamos dados de GBS de *C. arabica* para realizar um novo *Pipeline* TASSEL para montagem, alinhamento e busca de SNPs nos dois parentais diploides da espécie (*C. canephora* e *C. eugenioides*). Foi realizada análise de diversidade e estrutura genética nos dois painéis gerados, além de estudo associativo para identificar novas variantes SNPs que expliquem características relacionadas com a qualidade da bebida, assim como genes candidatos.

22 MATERIAL E MÉTODOS

22.1 MATERIAL VEGETAL

No presente estudo foram utilizados acessos Etíopes cedidos pela FAO (MEYER, 1968), além de cultivares tradicionais e variedades de *C. arabica* nas análises de diversidade, estrutura e nos estudos de associação. A coleção da Etiópia é cultivada em no IAPAR em uma estação experimental localizada em Londrina, Brasil (23°23’00’’S e 51°11’30’’W). O solo é lato solo distrófico vermelho, e a média de chuva e temperatura são 1,500 mm/ano e 21°C, respectivamente. A coleção da FAO vem de sementes polinizadas da original coleção do CATIE (Costa Rica), introduzido no Brasil em 1976, e transferidas do Instituto Agronômico de Campinas (IAC) para o IAPAR.

22.2 FENOTIPAGEM

Foram utilizados 4 anos de coleta dos materiais de *C. arabica* da Etiópia (2011, 2012, 2015 e 2016) e média dos anos de maior produtividade (2011/2015). Assim, um total de 101 genótipos fenotipados incluindo acessos da Etiópia, além de cultivares tradicionais e variedades de *C. arabica* foram utilizados nos estudos de associação (Tabela Suplementar 1).

Os frutos em estágio cereja foram selecionados entre os meses de Maio e Junho, secos até obterem 12% de umidade, e beneficiados para remoção da casca e pergaminho. Para caracterizar o tamanho dos grãos de café verde beneficiados, os mesmos foram passados peneiras com tela 14 (0,56 mm), 15

(0,59 mm), 16 (0,63 mm), 17 (0,67 mm) e 18 (0,71 mm). Os grãos considerados ardidos (marrons), pretos e brocados foram removidos nessa etapa por não serem características próprias dos genótipos, e sim gerados pelo ambiente. Posteriormente os grãos de café foram mergulhados em nitrogênio líquido (-196°C), moídos em moinho de disco (PERTEN 3600, Kungens Kurva, Suécia), e passados em peneiras com malha 0,5 mm para controlar o tamanho das partículas. As amostras moídas foram mantidas congeladas a -18 C até posterior análise.

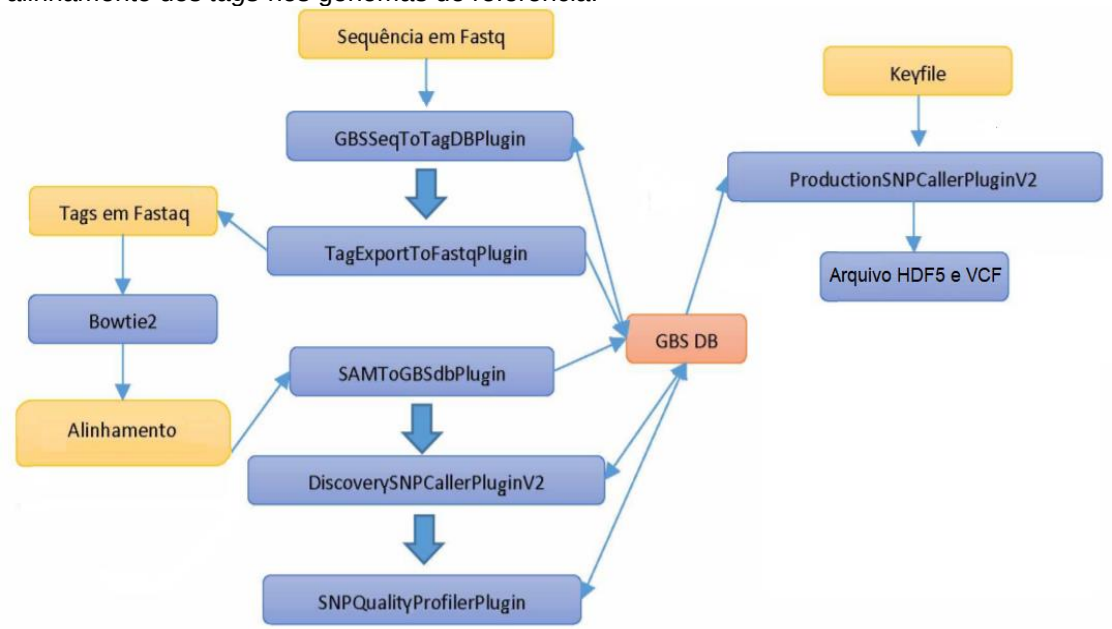
A composição química das amostras coletadas foi analisada utilizando a Espectroscopia no Infravermelho Próximo (NIRS), no espectrofotômetro NIRSystem 6500, com reflectância difusa e região espectral de 400 a 2500 nm. Modelos de predição desenvolvidos anteriormente (SCHOLZ et al 2014a, b) foram empregados para cada componente. Foram realizadas varreduras entre 1100 e 2300 nm em intervalos de 2 nm, utilizando-se uma cubeta retangular contendo 6 g de café moídos e em temperatura ambiente. Todos os cálculos foram feitos usando os *softwares* ISIscan e WinISI 4.5 (Infrasoft International, Port Matilda, Pennsylvania-USA).

22.3 Genotipagem por Sequenciamento, *Pipeline* TASSEL-GBS e Controle de Qualidade dos Dados

Os DNAs foram extraídos de folhas de plantas individuais usando o protocolo CTAB com modificações (DOYLE e DOYLE, 1990). A genotipagem pos sequenciamento (GBS) foi conduzida seguindo os protocolos descritos por Elshire et al. (2011) e realizada pela *Genomic Diversity Facility LIMS*, em *Cornell University* (Ithaca, NY – USA). Um total de 159 genótipos de *C. arabica*, incluindo genótipos repetidos; 2 controles negativos; acessos da etiópia; dois parentais de *C. arabica*; cultivares tradicionais e variedades foram distribuídos em duas placas de 96 poços e enviados para o sequenciamento *single-end* utilizando o sequenciador *Illumina HiSeq*2000 (Figura Suplementar 2). Para a preparação das bibliotecas GBS a enzima de restrição *PstI* (CTGCAG) e adaptadores com *barcodes* foram ligadas em cada amostra. As amostras foram agrupadas em uma única biblioteca e amplificadas por PCR. Cada biblioteca foi sequenciada em *reads* de 100 pb em duas linhas de sequenciamento.

Os *reads* resultantes do sequenciamento dos genótipos (formato Fastq) foram utilizados para montagem dos *tags*, alinhamento e identificação dos SNPs pelo *Pipeline* TASSEL - GBS versão 5.0 Standalone (GLAUBITZ et al., 2014) de acordo com a figura 1 e os passos descritos a seguir.

Figura 1. Fluxograma do *pipeline* TASSEL-GBS versão 5.0 para montagem das sequências e alinhamento dos *tags* nos genomas de referência.



DB – Banco de dados; GBS – Genotyping by Sequence; HDF5 e VCF – formatos de arquivo.

22.3.1 *GBSSeqToTagDBPlugin*

O *plugin* GBSSeqToTagDBPlugin utiliza os arquivos fastq como entrada, identifica os *tags* de várias linhas de sequenciamento e armazena esses dados em um banco de dados local (GLAUBITZ et al., 2014).

22.3.2 *TagExportToFastqPlugin*

O *plugin* TagExportToFastqPlugin recupera os tags distintos no banco de dados e os reformata em um arquivo fastq que pode ser lido pelos programas de alinhamento BWA ou Bowtie 2 (GLAUBITZ et al., 2014).

22.3.3 *Bowtie 2*

Com as sequências dos *tags* trimadas e limpas usamos o programa Bowtie 2 que é uma ferramenta eficiente para alinhar os *tags* obtidos nos

genomas de referência (LANGMEAD et al., 2009). Os genomas de referência utilizados para o mapeamento dos *tags* e identificação dos SNPs foram dos dois parentais diploides da espécie, *C. canephora,* publicado em 2014 (DENOEUD et al., 2014) e *C. eugenioides*, gentilmente cedido pelo Consórcio Genoma Café Arábica (ACGC). O arquivo de saída é um arquivo no formato .SAM que foi utilizado para “chamar” os SNPs em etapas posteriores do *Pipeline* (SCHMUTZ et al., 2010).

22.3.4 *SAMToGBSdbPlugin*

Esse plugin realiza a leitura dos arquivos .SAM para determinar as potenciais posições dos *tags* no genoma de referência. A ferramenta atualiza o banco de dados atual com informações sobre as posições dos *tags* (GLAUBITZ et al., 2014).

22.3.5 *DiscoverySNPCallerPluginV2*

DiscoverySNPCallerPlugin identifica os marcadores SNPs a partir dos *tags* alinhados no genoma de referência. A posição dos SNPs e os dados dos alelos são enviados para o banco de dados (GLAUBITZ et al., 2014).

22.3.6 *SNPQualityPofilerPlugin*

SNPQualityProfilerPlugin é utilizado para gerar informações sobre a qualidade dos putativos SNPs quanto à posição, cobertura do marcador e profundidade para um determinado conjunto de amostras (GLAUBITZ et al., 2014).

22.3.7 *ProductionSNPCallerPluginV2*

Essa ferramenta realiza a conversão do arquivo do arquivo fastq atualizado para arquivos nos formatos HDF5 e VCF (GLAUBITZ et al., 2014). Leituras de sequências de baixa qualidade que foram alinhadas no genoma de referência são removidas do banco de dados (LANGMEAD; SALZBERG, 2012).

Com o programa TASSEL versão 5.2.42 (GLAUBITZ et al., 2014) foi realizado um controle de qualidade (CQ) para a detecção de putativos SNPs, visando diminuir a frequência de falsos positivos (erro tipo I) e consequentemente futuras associações espúrias. Os critérios no CQ são específicos para cada estudo onde se busca a redução de SNPs falsos positivos sem comprometer o poder estatístico dos métodos de associação (CHAN; HAWKEN; REVERTER, 2009).

Para melhor confiabilidade da análise e visando minimizar possíveis viéis e erros nos resultados do GWAS foi realizado acompanhamento nos dois arquivos quanto a PCoA e à heterozigosidade observada e esperada com os filtros para *Call Rate* (*Call Rate* > 70, 80 e 90%) e para heterozigosidade (Ho < 0.5, 0.7 e 0.9). Foram retirados alelos raros, filtrando apenas os SNPs com uma frequência alélica mínima maior que 5% (limiar MAF ≥ 0.05). Os SNPs com mais de duas bases nitrogenadas em todos os genótipos (considerados multialelos) foram retirados. Após os filtros foi observado que os alelos obtidos em ambos subgenomas possuíam uma MAF ≥ 0.05 e ≤ 0.40 em todo o painel.

22.4 Análise de Diversidade Genética

De acordo com o conjunto de SNPs obtidos para ambos subgenomas, estimamos a média do número de alelos (Na), porcentagem de *loci* polimórficos (*P*), heterozigosidade observada (Ho) e esperada (He); índice de Shannon (*H'*), heterozigosidade esperada corrigida pelo tamanho amostral (*uHe*) e número de alelos privados utilizando o programa GenAlEx 6 (PEAKALL e SMOUSE, 2012).

22.5 Análise de estrutura popuplacional

Para analisar a estrutura genética da população foi realizada análise *bayesiana* implementada pelo *software* STRUCTURE versão 2.3.4 para calcular o número mais provável de subpopulações *K* com 10 independentes rodadas e modelo *admixture* variando de 2 a 11 pela simulação *MCMC* (*Markov Chain Monte Carlo*) com 10^5 interações *burn-in* e 10^5 ciclos de rodada. Um plot com a composição genômica de todos os genótipos para os SNPs de *C. arabica*

mapeados em ambos subgenomas foi gerado com base no critério *ΔK* do *software Structure Harvester* para estimar o melhor nível de estrutura.

Nós realizamos também análises de coordenadas principais (PCoA) para os subgenomas separados a partir de uma matriz de covariância usando o programa XLStat versão 2018.1 (ADDINSOFT, 2010) para visualizar as relações genéticas entre os genótipos. A PCoA foi obtida também a partir da união dos dois subgenomas pelo pscote do R "GAPIT" para fins de comparação.

22.6 ANÁLISE DE DESEQUILÍBIO DE LIGAÇÃO

As posições físicas dos SNPs de *C. arabica* mapeados em ambos subgenomas *C. canephora* e *C. eugenioides* foram utilizadas para calcular a extensão do decaimento do DL. O DL intracromossomal entre todos os SNPs aos pares foi estimado pelo programa HaploView 4.2 (BARRETT et al., 2005) e pelo pacote do R 'LDcorSV' versão 1.3.1 (MANGIN et al., 2012). Foram utilizados no programa HaploView 4.2 os seguintes parâmetros: somente pares de marcadores a uma distância de no máximo 20 Mpb, *cutoff* do valor de *P* do Equilíbrio de *Hardy-Weinberg* (1.10^{-6}), porcentagem mínima de genótipos (75%) e frequência alélica mínima (MAF ≥ 0.05).

Os blocos de haplótipos foram inferidos pelo programa HaploView 4.2 (BARRETT et al., 2005) por meio do algoritmo EM utililzando o método de partição/ligação descrito por QIN et al. (2002). Este método infere as fases dos haplótipos e gera estimativas de suas frequências populacionais com base em máxima verossimilhança.

No pacote do R 'LDcorSV' o decaimento do DL entre os marcadores mapeados em ambos subgenomas *C. canephora* e *C. eugenioides* foi também estimado pelos parâmetros r^2 clássico (sem correção) e r^2_{VS} (considera as matrizes *Q* e *K*). Somente pares de marcadores a uma distância de no máximo 20 Mb foram considerados e uma matriz de *Kinship* IBS (Identify-by-state) foi calculada pelo programa TASSEL versão 5.2.42 (GLAUBITZ et al., 2014). A matriz de estrutura populacional (matriz *Q*) para ambos subgenomas foi gerada pelo programa STRUCTURE versão 2.3.4 (*K* = 2) (EARL e VON HOLDT et al., 2012).

22.7 MAPEAMENTO POR ASSOCIAÇÃO GENÔMICA AMPLA PARA COMPOSTOS RELACIONADOS À QUALIDADE DA BEBIDA DE CAFÉ

A análise de associação foi realizada para identificar variantes SNPs associados com variação natural para o conteúdo de nove compostos relacionados com a qualidade da bebida de café: Proteínas Totais (PT), Lipídeos Totais (LT), Cafeína, Caveol, Cafestol, Sacarose, Açúcares Totais (AT), Açúcares Redutores (AR) e Ácidos Clorogênicos (ACGs), mais Diterpenos Totais, Média dos Diterpenos e Razão Caf/Cav. Foi utilizado no estudo 4 anos de fenotipagem, sendo o ano de 2011, 2012, 2015, 2016 mais a média para os dois anos de maior produtividade (2011/2015).

A análise de associação foi realizada por oito métodos associativos, sendo 4 métodos implementados pelo pacote do R 'mrMLM' (mrMLM, FASTmrEMMA, ISIS EM-BLASSO e pLARmEB) (WANG et al., 2016); 2 pelo pacote do R 'GAPIT' (cMLM – MLM compresso e MLM regular) (LIPKA et al., 2012); e 2 pelo programa TASSEL 5.2.20 (GLM – Modelo Linear Generalizado e MLM – Modelo Linear Misto) (BRADBURY et al., 2007). Para o estudo de associação foram unidos os dois arquivos hapmap obtidos dos SNPs de *C. arabica* mapeados em ambos subgenomas, criando um genoma artificial correspondente aos 22 cromossomos diploides básicos da espécie, o que incluiu 101 genótipos e 4.517 SNPs (Tabela Suplementar 1).

Em GWAS, a estrutura populacional e relação de parentesco entre os genótipos podem resultar em associações espúrias (YU et al., 2006). Para a condução do mapeamento por associação do genoma artificial, a matriz de PCoA (Matriz *Q*) foi calculada pelo método *VanRaden* no pacote do R 'GAPIT' (LIPKA et al., 2012) para corrigir a influência da estrutura da população nos modelos de associação.

Ambos modelos de único locus do GAPIT (MLM regular e cMLM – MLM compresso) foram utilizados visando observar diferenças nos resultados de significância. No modelo MLM regular cada genótipo foi considerado como um grupo. Já para reduzir o tempo computacional e aumentar o poder de detecção dos QTNs (Quantitative Trait Nucleotide) utilizamos o cMLM no qual todos os genótipos foram atribuídos a um só grupo como efeito aleatório, e com isso há redução da matriz de parentesco. As matrizes *Q e K* foram incorporadas nos

dois modelos como covariáveis de efeitos fixos e aleatórios. O limiar de significância para a associação SNP - característica fenotípica foi determinado pelo critério da razão de falsas descobertas (FDR, do inglês False Discovery Rate) (q) < 0.05.

Para os modelos testados no TASSEL foram utililzados SNPs com o limiar de frequência do alelo menor (MAF ≥ 0.05). Para corrigir a estrutura da população no modelo GLM, incorporou-se a matriz *Q* de PCoA como covariável e no modelo MLM as matrizes *Q* e *K*, o que fazem com que o segundo método tenha maior poder estatístico, além do algoritmo EMMA, que visa reduzir o tempo computacional. O limiar $\alpha = 0.05/p$ (YANG et al., 2014) para a correção de *Bonferroni* foi utilizado com um valor de corte $-\log_{10}(P) \geq 4.955$ e significante *P*-value de $1x10^{-5}$ (correção baseada no número de SNPs testados $P < 0.05/4.517$) para considerar se um SNP está significativamente associado à característica fenotípica analisada.

Para o programa mrMLM nós incorporamos a matriz *Q* de PCoA como uma covariável (*Q* = 2), como também a matriz *K* (*Kinship*) como uma segunda covariável, no qual foi gerada pelo próprio mrMLM (matriz *K*). Os parâmetros utilizados para o método de único locus no mrMLM foram, crítico *P*-value em rMLM = 0.01; procurar genes candidatos a um raio de 20 Kpb e crítico *LOD score* para mrMLM = 3. O limiar de significância para a associação SNP-característica fenotípica foi determinado utilizando a estatística *Wald.*

A maioria das características quantitativas são controladas por alguns genes com grandes efeitos e numerosos poligenes de efeitos menores. No entanto, as abordagens de varredura do genoma unidimensional para GWAS não correspondem ao modelo genético verdadeiro para esses tipos de características (YI, 2008). Para resolver este problema, metodologias multi-locus foram utilizadas; por exemplo, FASTmrEMMA, ISIS EM-BLASSO e pLARMEB. Para o método ISIS-EM-BLASSO foi utilizado o crítico *P-value* = 0.01. Já para o método pLARmEB os parâmetros utilizados foram, crítico *LOD score* = 2 e número de variáveis potencialmente associadas selecionadas por LARS = 50. Para o método FASTmrEMMA, todos putativos QTNs com *P-value* ≤ 0.005 e *LOD score* = 3 foram considerados significativamente associados. Os 4 modelos de associação foram realizados usando o pacote do R “mrMLM”.

Nós consideramos como putativos SNPs para posteriores análises de busca dos genes candidatos, aqueles presentes em no mínimo em dois métodos associativos, como também SNPs encontrados em mais de dois anos de fenotipagem. As sequências completas do genoma de referência *C. canephora* estão disponíveis no banco de dados disponível na Web do *Coffee Genome Hub* (http://coffee-genome.org/). Esse banco de dados foi utilizado para identificar candidatos genes de *C. canephora* localizados em um intervalo de 100 Kpb dos SNPs significativamente associados,

Para comparar a distribuição dos valores fenotípicos dos indivíduos com base aos diferentes genótipos para os SNPs significativos (homozigotos para os diferentes alelos e os heterozigotos), foram construídos gráficos box plots com auxílio do programa XLStat versão 2018.1 (ADDINSOFT, 2010).

23 RESULTADOS E DISCUSSÃO

23.1 GBS E IDENTIFICAÇÃO DE SNPS EM AMBOS SUBGENOMAS

Devido à ausência do genoma de *C. arabica,* foram utilizados os subgenomas de seus ancestrais diploides (*C. canephora* e *C. eugenioides*) como referência para montagem, mapeamento dos *tags* GBS de *C. arabica* e busca dos SNPs. O alto grau de conservação entre *C. arabica* e seus os parentais diploides é bem conhecido (LASHERMES et al., 1999; CENCI et al., 2012) e nos permitiu o mapeamento de tags de dados de GBS para a identificação dos SNPs.

A biblioteca GBS rendeu aproximadamente 48 milhões de *reads single-end* para 159 genótipos de *C. arabica,* que resultaram na montagem de um total de 1.661.838 *tags*. Desses, 13.42% dos *tags* se alinharam em posições únicas no genoma de *C. canephora*, sendo identificados um total de 126.617 SNPs. Após os filtros de controle de qualidade para os SNPs de *C. arabica* mapeados em *C. canephora* foram obtidos um total de 1.719 SNPs bialélicos com uma cobertura média de 68X. Para o genoma *C. eugenioides,* dos 1.661.838 *tags* gerados, 17.87% se alinharam em posições únicas, sendo identificados um total de 130.730 SNPs. Após os filtros de controle de qualidade para os SNPs de *C. arabica* mapeados em *C. eugenioides*, foram obtidos um total de 2.949 SNPs bialélicos com uma cobertura média de 47X.

Sant'Ana et al., 2018 utilizou o Pipeline TASSEL-GBS versão 3 para montagem e alinhamento dos tags no genoma de *C. canephora*, no qual obteve 6.210.920 tags e 20% alinharam em posições únicas. Desses, 6.696 SNPs foram identificados com uma cobertura média de 39X. Após os filtros de controle de qualidade (MAF>0.05, Call Rate 80% e Heterozigosidade <0.9), 2.587 SNPs foram obtidos.

Os marcadores SNPs de *C. arabica* mapeados em ambos subgenomas foram utilizados nas posteriores análises de diversidade, estrutura populacional e análise de associação para compostos relacionados com a qualidade da bebida de café.

23.2 Análise da Diversidade Genética

Como já relatado, apesar da ampla distribuição do cultivo de café Arábica, o número de cultivares utilizadas é muito pequeno (LABOUISSE et al., 2008) e a estreita base e diversidade genética das cultivares resultou em uma cultura com comportamento agronômico homogêneo (LASHERMES et al., 1999). Entre os grupos genéticos analisados para ambos subgenomas (Leste, Oeste e o das Cultivares), o grupo Oeste obteve maior diversidade intragrupo com maiores valores de *Na, P* %, *H'* e Ho em relação ao grupo Leste e o grupo das Cultivares (Tabela 1). Com os valores de *uHe,* que medem a diversidade genética ponderada pelo tamanho amostral de cada grupo, o grupo Oeste da Etiópia também apresentou o maior valor, como esperado. Assim como nos resultados de diversidade obtidos no capítulo 1 com os marcadores SSRs, espera-se também que o grupo do lado Leste apresente valores de variabilidade intermediários entre o grupo Oeste e o das Cultivares. Porém no grupo das Cultivares estão nove genótipos que possuem introgressão de *C. canephora* (IAPAR59, IPR99, IPR100, IPR101, IPR102, IPR103, IPR104, IPR105 e IPR107), o parental diploide e de natureza alógama de *C. arabica*.

Os alelos privados para ambos subgenomas foram observados apenas no grupo do lado Oeste (124 e 147 alelos, respectivamente), reforçando a maior variabilidade dos materiais localizados nessa região. A maior diversidade genética dos acessos da coleção da Etiópia em relação às cultivares tradicionais/ variedades corrobora com os resultados do capítulo 1, e é

consistente com a maior variabilidade genética e fenotípica observada nos capítulos 1 e 2, assim como também em prévios estudos (SILVESTRINI et al., 2007; SCHOLZ et al., 2016, IVAMOTO et al., 2017; TRAN et al., 2017; SANT'ANA et al., 2018).

Nossa coleção de acessos do lado Oeste do Vale do Rift tem mostrado uma fonte valiosa de alelos favoráveis que poderão ser explorados pelos programas de melhoramento genético.

Tabela 1. Tamanho amostral (N), média do número alelos (NA), porcentagem de *loci* polimórficos (*P*%), índice de *Shannon* (*H'*), heterozigosidade observada (Ho), heterozigosidade corrigida pelo tamanho amostral (*uHe*) e número de alelos privados para cada grupo genético identificado pela análise do STRUCTURE (K = 3) para os 1.719 SNPs de *C. arabica* mapeados em *C. canephora* **(A)** e para os 2.949 SNPs de *C. arabica* mapeados em *C. eugenioides* **(B)**.

A)

Grupo Genético	N	Na	*P* (%)	*H'*	Ho	*uHe*	Alelos Privados
Leste	9	1.79	79.00	0.34	0.22	0.22	0
Oeste	105	2.00	100.00	0.39	0.24	0.24	124
Cultivares	16	1.85	85.40	0.36	0.28	0.23	0
Total	130	1.88	88.13	0.36	0.24	0.23	124

B)

Grupo Genético	N	Na	*P* (%)	*H'*	Ho	*uHe*	Alelos Privados
Leste	9	1.79	79.59	0.32	0.21	0.19	0
Oeste	104	2.00	100.00	0.36	0.23	0.22	147
Cultivares	16	1.86	86.84	0.32	0.21	0.20	0
Total	129	1.88	88.81	0.33	0.21	0.21	147

23.3 DESEQUILÍBRIO DE LIGAÇÃO E ESTRUTURA POPULACIONAL

Foram utilizados os 1.719 e 2.949 SNPs de *C. arabica* mapeados em *C. canephora* e *C. eugenioides* respectivamente, para estimar o padrão de decaimento do DL em todo o genoma e a estrutura nos dois painéis.

Os parâmetros r^2 e r^2_{vs} foram estimados como uma função da distância física entre os marcadores SNPs. Para o parâmetro r^2_{vs} (corrigido para o viés de estrutura e parentesco) nós observamos um valor do DL abaixo de $r^2_{vs} = 0.2$ a uma distância a 66 Kpb e 22 Kpb para ambos subgenomas. Considerando os valores de r^2 (sem correção) nós obtivemos um DL decaindo abaixo de $r^2 = 0.2$ a 89 Kbp e 40 Kbp para *C. canephora* e *C. eugenioides,* respectivamente (Figura 2A e 2B). Visando aumentar a confiabilidade do tamanho da janela física que flanquea os polimorfismos para a identificação de genes causais, o padrão de decaimento do DL ao longo do genoma também foi calculado pelo programa

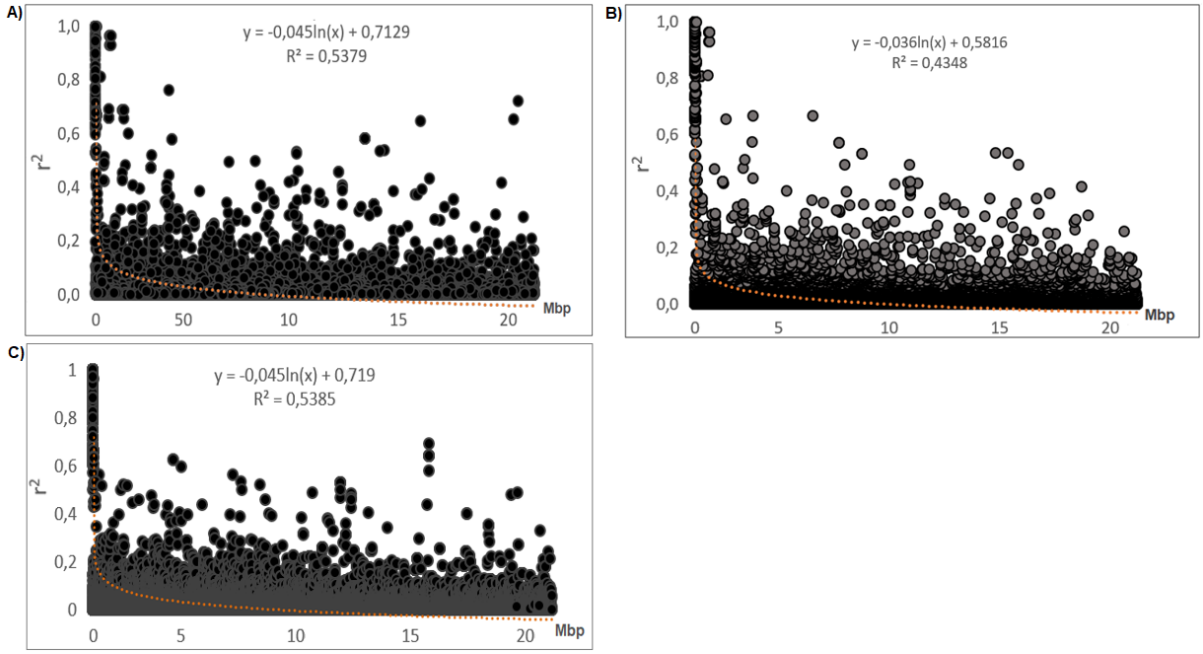
HaploView 4.2 (BARRETT et al., 2005), onde obtivemos um DL decaindo abaixo de $r^2 = 0.2$ a 102 Kbp (Figura 2C).

Assim, com a medida de r^2vs (correção para estrutura e parentesco), houve um decaimento do DL mais rápido a uma curta distância para ambos subgenomas. Baseado nesses resultados e no DL obtido por Sant'Ana et al. (2018), a busca dos genes candidatos no banco de dados de *C. canephora* foi realizada a uma janela física de 102 Kbp.

Sant'Ana et al. (2018) utilizando ambos parâmetros de DL para 107 genótipos e 2.587 SNPs (filtros para Ho < 0.9, *Call Rate* 80% e MAF > 0.05) mapeados no genoma de *C. canephora*, obtiveram uma maior queda do DL utilizando o parâmetro r^2vs (185 Kpb a 0.2), com um declínio exponencial do DL em relação à distância, demonstrando a eficiência dessa medida na correção do viés na sua análise (diferença de 113 Kpb).

Análises anteriores do nosso laboratório (dados não publicados), demonstram que *C. eugenioides* possui mais alelos em heterozigose do que *C. canephora*, e isso justifica um decaimento do DL maior a curta distância dos SNPs de *C. arabica* mapeados nesse subgenoma. É importante ressaltar que o DL foi calculado nos dois genomas ancestrais como referência e que esse resultado possa ser diferente quando os SNPs de *C. arabica* forem mapeados no próprio subgenoma da espécie.

Figura 2. **(A)** Decaimento de DL em função da distância genética entre 1.719 marcadores SNPs mapeados em *C. canephora* feito pelo pacote do R 'LDcorSV'; **(B)** Decaimento de DL em função da distância genética entre 2.949 marcadores SNPs mapeados em *C. eugenioides* feito pelo pacote do R 'LDcorSV' e **(C)** Decaimento de DL em função da distância genética entre 1.719 marcadores SNPs mapeados em *C. canephora* feito pelo programa HaploView. Ambos utilizando a função Y= pr1*ln (x) + pr2.



Sabe-se que a estreita base genética das cultivares de *C. arabica* tem resultado em uma cultura com comportamento agronômico homogêneo, incluindo plantas com alta susceptibilidade a estresses bióticos e abióticos (LASHERMES et al., 2009). A estrutura populacional baseada em estimativa *bayesiana* (STRUCTURE) (valor de ΔK) para os SNPs mapeados em ambos subgenomas indicou que o número mais provável de grupos em que a população pode ser dividida foi $K = 2$, embora a estruturação com 3 grupos ($K = 3$) também apresentou alto valor para ambos subgenomas (Figura 3 e 4).

Na análise *bayesiana* com $K = 2$ para ambos subgenomas, todas as cultivares tradicionais, variedades e acessos do lado Leste do Grande Vale do Rift foram agrupadas conjuntamente (grupo Q1). O grupo Q2 foi exclusivamente composto pelos acessos selvagens do lado Oeste do Vale do Rift. Trabalhos anteriores de caracterização desta coleção da Etiópia utilizando marcadores SSRs, assim como o nosso discutido no capítulo 1, resultaram na subdivisão dos genótipos também em dois grupos (K = 2), os acessos de ambos lados do Vale do Rift e as cultivares (ANTHONY et al., 2001; SILVESTRINI et al., 2007) (Figuras 3 e 4).

Por outro lado, na estruturação com *K* = 3, todas as cultivares tradicionais, variedades e acessos do lado Leste do Grande Vale do Rift foram agrupados no mesmo grupo (Q1) junto com 3 acessos do lado Oeste (E261, E302, E333). Além disso, os acessos do lado Oeste se subdividiram em dois grandes grupos, no qual o grupo Q2 foi formado por acessos do lado Oeste do Vale do Rift, e o grupo Q3 foi formado preferencialmente por acessos coletados em regiões florestais e de parques de reservas (20 acessos do tipo silvestre). Esses resultados corroboram com os resultados de Santana et al. (2018), no qual de acordo com os seus resultados de *ΔK,* o painel que inclui acessos da Etiópia se subdividiu em *K* = 2 e *K* = 3 grupos.

Figura 3. Estrutura populacional entre 130 acessos de *C. arabica.* **(A)** Plot de *ΔK* vs. *K* com a evolução de valores *ΔK* (eixo y) de acordo com o número de grupos genéticos (eixo x), **(B)** e **(C)** Plot de barras para o coeficiente de estimação de membros de grupo (Q) para 130 acessos de *C. arabica* baseado em 1.719 SNPs mapeados em *C. canephora* para *K* = 2 e *K* = 3.

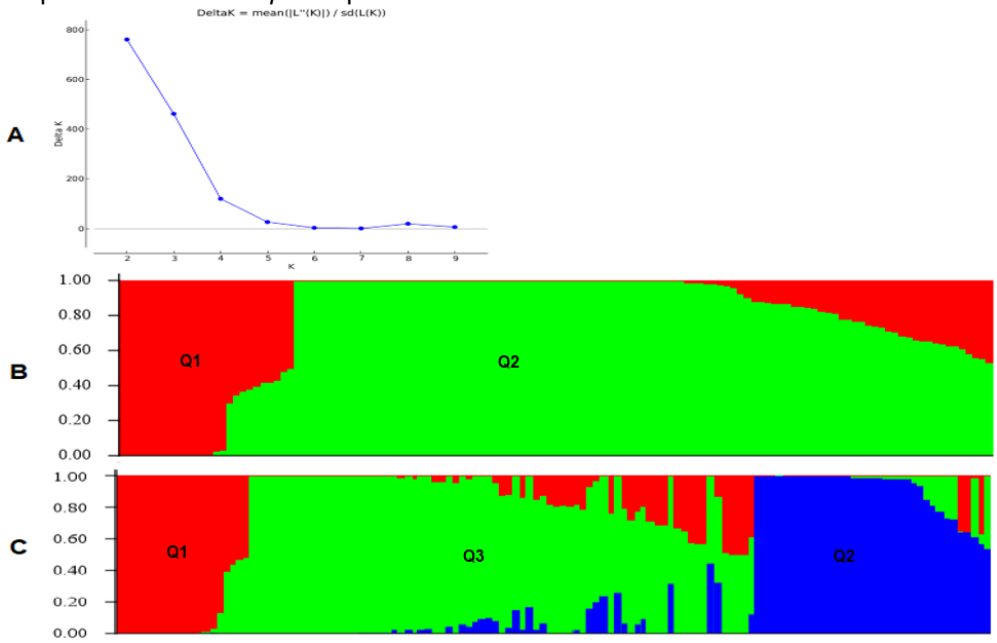
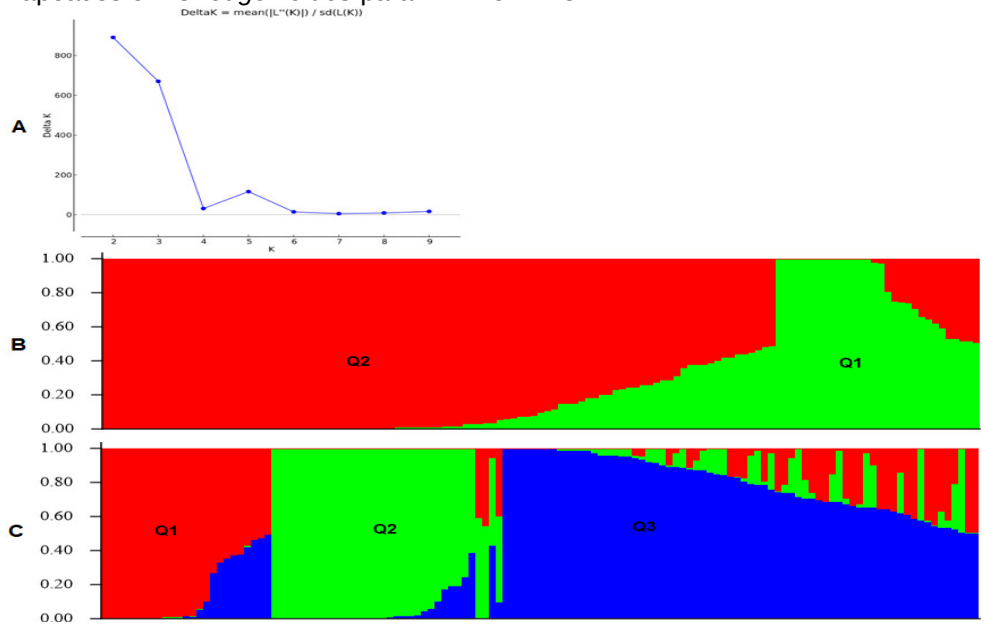


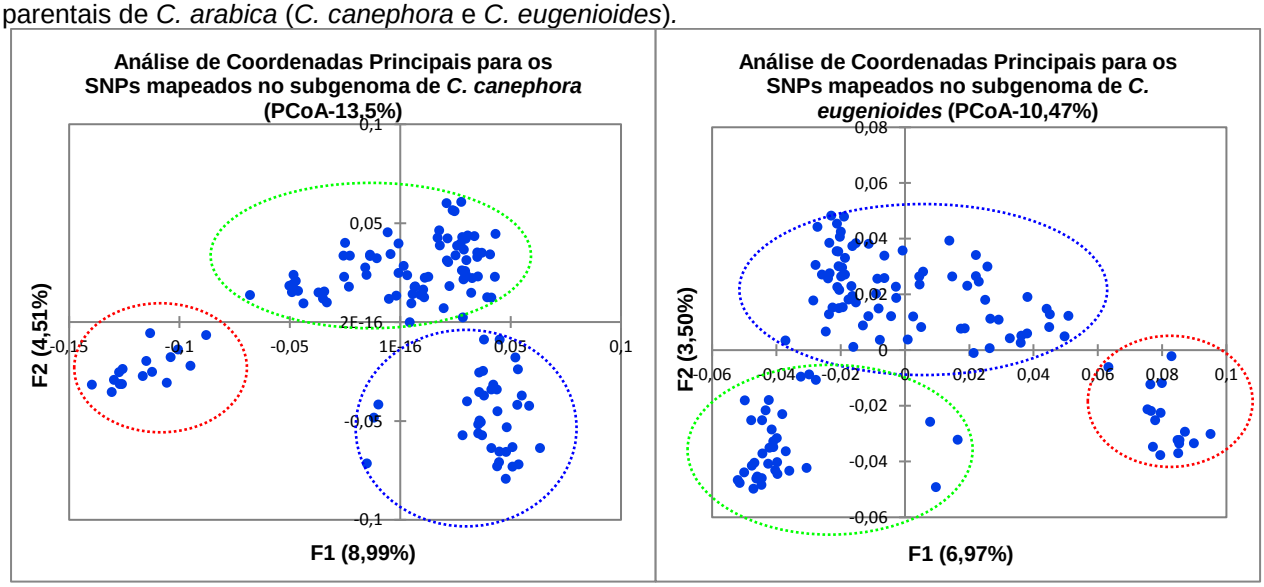
Figura 4. Estrutura populacional entre 129 acessos de *C. arabica.* **(A)** Plot de *ΔK* vs. *K* com a evolução de valores *ΔK* (eixo y) de acordo com o número de grupos genéticos (eixo x), **(B)** e **(C)** Plot de barras para o coeficiente de estimação de membros de grupo (Q) para 129 acessos de *C. arabica* baseado em 2.949 SNPs mapeados em *C. eugenioides* para K = 2 e *K* = 3.



Na análise de coordenadas principais (PCoA) para os dois subgenomas, as duas primeiras coordenadas explicaram 13,5% da variação genética total (Figura 5). Similar aos resultados do STRUCTURE, as cultivares tradicionais e variedades foram geneticamente próximas aos acessos do lado Leste do Grande Vale do Rift do que os acessos lado Oeste, que se subdividiram em dois grupos.

Para fins de comparação, a estrutura genética com o genoma artificial gerado pela união dos dois subgenomas (101 acessos da coleção da Etiópia e 4.517 SNPs) também foi explorada por PCoA. O gráfico de dispersão tridimensional (CP1, CP2 e CP3) envolvendo todos os genótipos também agrupou os genótipos em 2 e 3 subpopulações (Tabela Suplementar 3).

Figura 5. Análise de Coordenadas Principais (PCoA) para SNPs mapeados nos subgenomas dos parentais de *C. arabica* (*C. canephora* e *C. eugenioides*).



Sabe-se que a variação genética de materiais de *C. arabica* espontâneos e subspontâneos de florestas da Etiópia é maior que em materiais que passaram por seleção (MEYER, 1965). Além disso, parte dos acessos do lado Oeste do Vale do Rift são de florestas e parques de reservas florestais e possuem características geneticas e fenotipicas contrastantes para o melhoramento da espécie (acessos silvestres). Porém, esse *pool* de genes mais selvagens tem sido potencialmente ameaçado pela fragmentação e degradação de florestas e pela hibridização introgressiva com variedades de café localmente melhoradas (AERTS et al., 2013).

Este trabalho reforça a importância de se preservar o germoplasma de *C. arabica* do centro de Origem, principalmente acessos do lado Oeste do Vale do Rift, uma vez que a fragmentação florestal pode ter um impacto negativo na diversidade genética das espécies florestais, resultando em aumento da endogamia (YOUNG et al., 1996). No geral, esses resultados podem nos ajudar a definir quais materiais do lado Oeste do Vale do Rift que estão no grupo Q3 são importantes para se preservar, visando o melhoramento de *C. arabica*

23.4 Estudo de Associação Genômica Ampla (GWAS)

Foram identificados um total de 33 SNPs associados às características fenotípicas analisadas em 4 anos fenotipagem (2011, 2012, 2015, 2016) e média (ano 2011/2015) (Tabela 1).

Tabela 1. Marcadores SNPs de *C. arabica* mapeados em ambos subgenomas (*C. canephora* e *C. eugenioides*) associados aos compostos relacionados com a qualidade da bebida de café detectados por 4 métodos associativos e entre 4 anos e média (anos 2011/2015).

Característica	SNP	Subgenoma	Ano de Fenotipagem	GLM	mrMLM (*P-value*)	ISIS-EM-BLASSO (*LOD-value*)	pLARmEB (*P-value*)
Cafeína	S1_31.043.806	*C. canephora*	2015	__	2.62 e-05	4.06	__
Cafeína	S6_18.302.851	*C. canephora*	Média (2011/2015)	__	3.62 e-05	2.49	__
Proteínas Totais	S0_61.349.521	*C. canephora*	2011	__	__	3.68	3 e-04
Proteínas Totais	S1_31.043.837	*C. canephora*	2012	__	4.28 e-07	2.05	__
Proteínas Totais	S1_31.043.806	*C. canephora*	2012/2015/2016	__	__	4.46	2.00 e-04
Proteínas Totais	S2_6.527.554	*C. canephora*	2015/2016	__	__	3.97	8.17 e-08
Proteínas Totais	S10_26.252.138	*C. canephora*	2011/ 2015/ 2016 e Média (2011/2015)	__	1.02 e-07	3.68	__
Lipídeos Totais	CH12_58.140.623	*C. eugenioides*	2016	__	1.52 e-06	5.01	__
Açúcares Redutores	S4_5.942.227	*C. canephora*	2011	__	4.61 e-05	5.29	__
Caveol	S0_91.208.415	*C. canephora*	2011	__	__	3.37	4.06 e-05
Caveol	S2_14.156.413/.429/.434/.457	*C. canephora*	2011	__	2.29 e-07	3.41	__
Caveol	CH14_15.094.727/.750/.755/.771	*C. eugenioides*	2011	__	1.47 e-08	3.78	__
Caveol	CH15_3.781.665	*C. eugenioides*	Média (2011/2015)	__	1.16 e-06	3.99	__
Cafestol	S6_7.853.747/.804/.861/.871	*C. canephora*	2016	4.88 e-06	3.28 e-07	__	__
Cafestol	CH18_8.093.457/.514/.571/.581	*C. eugenioides*	2016	3.69 e-06	1.09 e-07	__	__
Razão Caf/Cav	S2_14.156.413/.429/.434/.457	*C. canephora*	2011	4.67 e-05	__	__	__
Razão Caf/Cav	CH13_544.661	*C. eugenioides*	2016	4.10 e-07	8.54 e-05	__	__
Razão Caf/Cav	S8_23.722.236	*C. canephora*	2011/2016	1.07 e-05	__	4.78	__

Dos oito modelos de associação utilizados, 4 modelos identificaram associações com valores significativos (GLM, mrMLM, ISIS-EM-BLASSO e pLARmEB) dentre os 4 anos e média para Cafeína, PT, LT, AR, Caveol, Cafestol, Razão Caf/Cav. O método ISIS-EM-BLASSO e mrMLM foram os métodos que identificaram o maior número de SNPs associados, 14 e 12, respectivamente. Para os SNPs associados às características fenotípicas e posteriormente mineração de genes candidatos, foram considerados SNPs

associados em no mínimo dois métodos associativos e/ ou SNPs encontrados em mais de dois anos de fenotipagem.

Dos 33 SNPs significativamente associados, 22 foram encontrados em *C. canephora* e 11 em *C. eugenioides*. Dos SNPs mapeados em *C. canephora*, 2 SNPs foram associados ao conteúdo de cafeína (2015 e média 2011/2015); 4 SNPs foram associados ao conteúdo de PT (2011, 2012, 2015, 2016 e média 2011/2015); 1 SNP foi associado a LT (ano 2016); 1 SNP foi associado a AR (2011); 5 SNPs foram associados ao conteúdo de Caveol (2011 e média 2011/2015); 4 SNPs foram associados ao conteúdo de Cafestol (ano 2016) e 5 SNPs foram associados à Razão Caf/Cav (ano 2011 e 2016). Pelo fato de não existir anotação para *C. eugenioides*, a busca por genes causais nas proximidades dos SNPs foi realizada apenas nos SNPs mapeados em *C. canephora*.

Uma característica comum dos métodos associativos baseados em MLM é uma varredura genômica unidimensional (único locus), testando um marcador por vez que passa pela correção de múltiplos testes para o limiar do teste de significância. No entanto, esse modelo não facilita boa estimativa dos efeitos dos marcadores porque o modelo nunca corrige se uma característica for realmente controlada por múltiplos *loci*, que é o caso das características complexas (WANG et al., 2016). Embora seja um método conservador e não corrige quando uma determinada característica é afetada por múltiplos *loci*, é um método eficiente para controle de falsos positivos, pois incorpora as matrizes Q e K simultaneamente como efeito fixo e aleatório através das amostras (WANG et al., 2016).

Neste trabalho o método mrMLM (modelo linear misto com efeito do SNP multi-locus aleatório) com 12 SNPs associados foi o segundo modelo com maior poder de detecção dos QTNs. O método mrMLM trata os efeitos dos SNPs como aleatório, e devido à natureza multi-locus não é necessário a correção de *Bonferroni* para calcular o limiar do *P-value*.

No método GLM a matriz de estrutura populacional foi ajustada como efeito fixo, porém não utiliza a matriz de parentesco (matriz *kinship*) como uma potencial causa da associação genótipo-fenótipo. Uma notável característica é que somente as associações relacionadas com Cafestol e Razão Caf/Cav foram

detectadas por meio deste método. Os diterpenos foram identificados por HPLC (*High performance liquid chromatography*), uma técnica que detecta especificamente cada composto. Já nos métodos baseados em MLM, a estrutura populacional é ajustada como um efeito fixo, enquanto o parentesco entre os indivíduos é incorporado como variância-covariância de efeitos genéticos aleatórios para os indivíduos, reduzindo o erro tipo I (falso-positivos) pela incorporação dos dois fatores simultaneamente (YU et al., 2006).

O método ISIS-EM-BLASSO foi o que detectou o maior número de SNPs associados (14 SNPs). Esse modelo usa uma abordagem iterativa de rastreamento de independência (ISIS) na redução do número de SNPs para um tamanho moderado. A expectativa maximização bayesiana (EM) é usada para estimar todos os efeitos dos SNPs selecionados para a detecção dos verdadeiros QTNs no modelo reduzido. A simulação de Monte Carlo que valida esse método possui maior poder empírico na detecção de QTNs e maior precisão na estimação do efeito do QTN. É o mais rápido comparado com o eficiente modelo de associação misto (EMMA) e mrMLM (TAMBA et al., 2017).

No método FASTmrEMMA, uma nova matriz é construída para obter um novo modelo genético que inclui apenas variação de QTNs e erro residual normal. Todos os putativos QTNs que passaram pelo estringente limiar (*P-value* ≤ 0.005) no primeiro passo são incluídos em um modelo multi-locus para detecção de verdadeiros QTNs, estimados pela Bayes Empírica de Expectativa e Maximização (EMEB). Devido à característica multi-locus, a correção de *Bonferroni* é substituída por um critério de seleção menos rigoroso (WEN et al., 2016).

Relata-se que em modelos multi-locus se o número de marcadores for muito maior que o tamanho amostral, o que é o nosso caso, o efeito dos marcadores pode ser incluído em um único modelo e estimados de maneira imparcial ou até mesmo essas abordagens de redução podem falhar (WEN et al., 2016; TAMBA et al., 2017). Esse deve ser o provável motivo pelo qual obtivemos poucos SNPs e nenhum significativamente associados pelos modelos multi-locus pLARmEB e FASTmrEMMA. WHEN et al. (2016) afirmam que devemos considerar em como reduzir o número de efeitos dos marcadores em

um modelo multi-locus levando em conta o tempo computacional, que pode ser um fato limitante para esse tipo de abordagem.

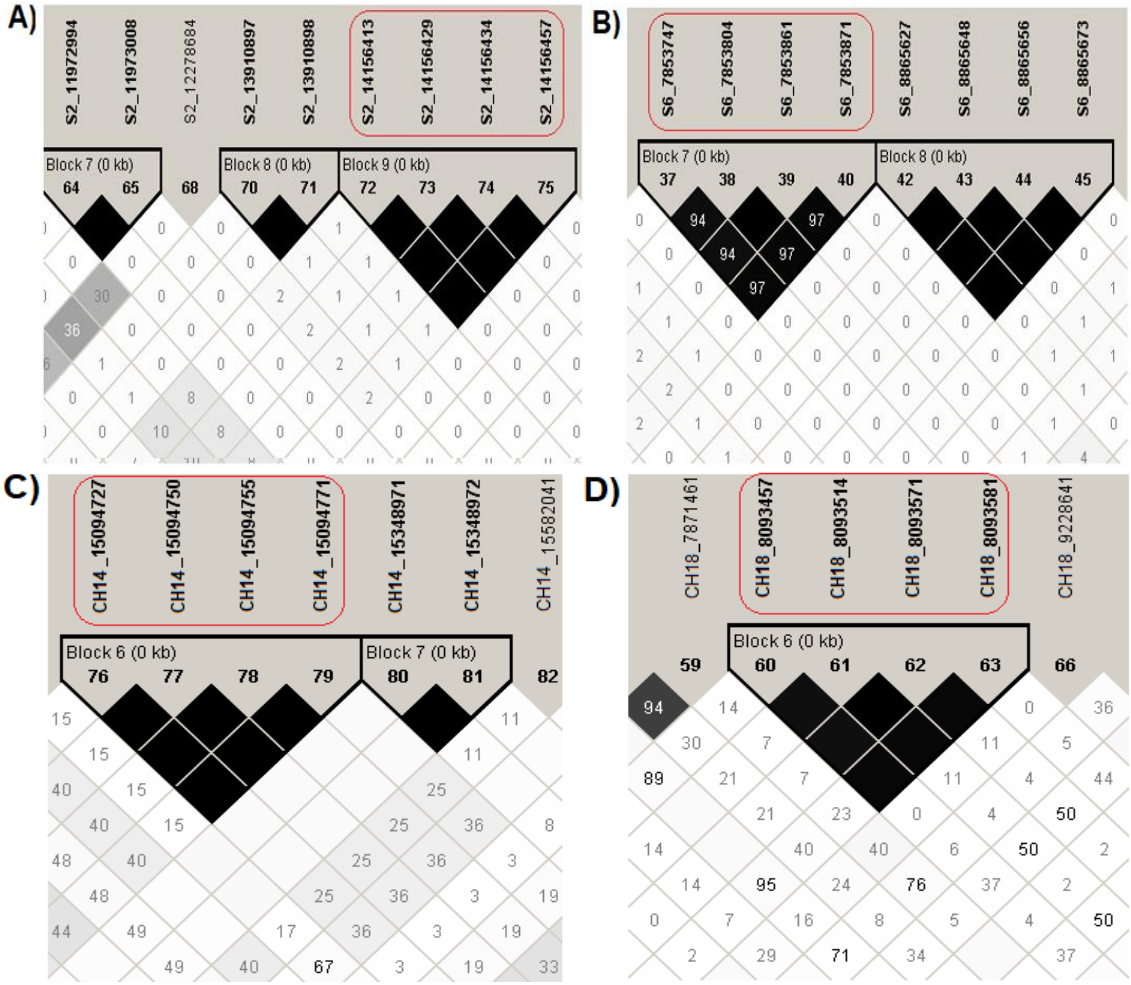
23.5 Blocos de haplótipos de *C. arabica* mapeados em ambos subgenomas *C. canephora* e *C. eugenioides* significativamente associados a Caveol, Cafestol e Razão Caf/Caveol

Os blocos de haplótipos foram gerados e analisados separadamente para cada cromossomo em ambos subgenomas de *C. canephora* e *C. eugenioides*. O DL obtido entre cada par de SNP pode ser calculado pelo HaploView 4.2 por meio de três medidas que se baseiam nas frequências alélicas da população em estudo: o LOD (*log of the likelihood odds ratio*) *score,* coeficiente de desvio padronizado (*D'*) e a medida do coeficiente de correlação ao quadrado (r^2). Em nosso estudo utlizamos a medida de r^2.

O HaploView permite representar o DL por meio de diferentes esquemas de cores. De acordo com GABRIEL et al. (2002), um par de SNPs está em forte DL se $r^2 = 1$. Sabe-se que blocos de haplótipos tendem a ser conservados, principalmente entre indivíduos que compartilham ancestralidade recente (SCHERER e CHRISTENSEN, 2016). Nos resultados das Figura 7 e 8 se destacam dois blocos de haplótipos de *C. arabica* mapeados nos cromossomos 2 e 6 de ambos subgenomas, *C. canephora* e *C. eugenioides* significativamente associados ao conteúdo de Caveol, Cafestol e Razão Caf/Cav e que explicam 17% e 38% da variação fenotípica. Os dois blocos de haplótipos de *C. arabica* correspondem a regiãos similares em ambos subgenomas a uma distância menos de 1 Kpb, demonstrando alta conservatividade dos alelos introgredidos de seus dois parentais diploides (*C. canephora* e *C. eugenioides*).

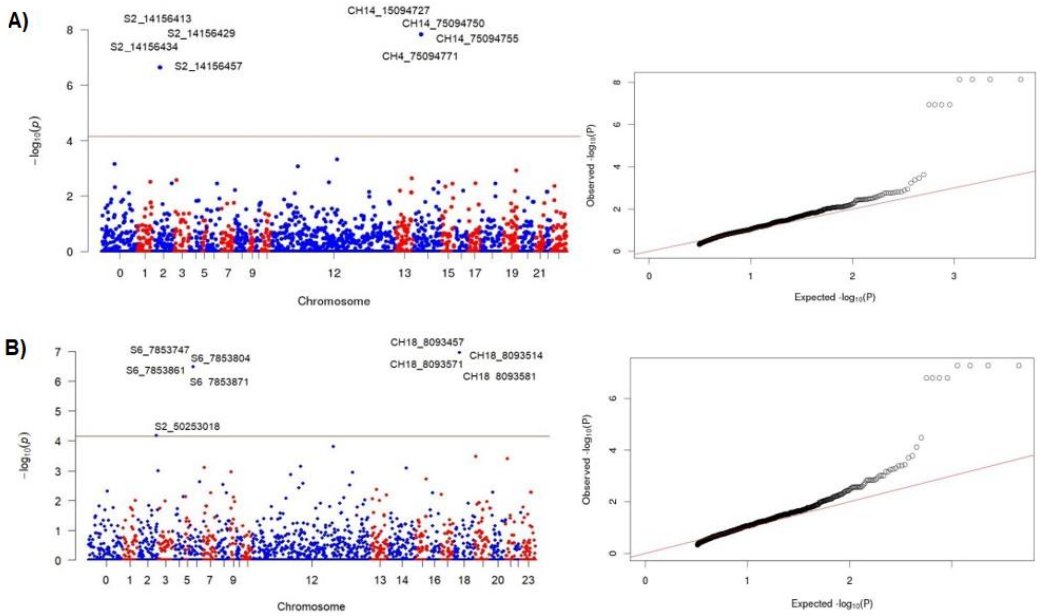
Assim, pelo ponto de vista do melhoramento do café Arábica, os genótipos de *C. arabica* que possuem o bloco de haplótipo S6_7.853.747/804/861/871 ou CH18_8.093.457/514/571/581 em homozigose possuem baixos conteúdos de Cafestol. Como também os genótipos que possuem o bloco de haplótipo S2_14.156.413/429/434/457 ou CH14_15.094.727/750/755/771 em heterozigose possuem altos conteúdos de Caveol (Figuras 13 e 14).

Figura 7. DL plot greado pelo programa HaploView para representar os dois blocos de haplótipos de *C. arabica* em forte DL encontrados nos cromossomos 2 e 6 dos subgenomas *C. canephora* (A e B) e *C. eugenioides* (C e D) em função dos valores de r^2 e associados aos teores de Cafestol, Caveol e Razão Caf/Cav.



O Branco indica $r^2 = 0$; tons de cinza, $0 < r^2 < 1$; preto, $r^2 = 1$, e os SNPs estão a poucos pares de bases uns dos outros.

Figura 8. Manhattan plots e Q-Qplots obtidos pelo modelo mrMLM para os blocos de haplótipos significativamente associados ao Caveol **(A)** Cafestol **(B)**, co-localizados aos genes candidatos da tabela 2 envolvidos nas vias metabólicas dos mesmos.



Linha vermelha no *Manhattan plot* indica o limiar dos valores de *P*; linha vermelha do Q-Q plot representam a distribuição nula esperada dos valores de *P* e os pontos representam a distribuição observada dos valores de *P.*

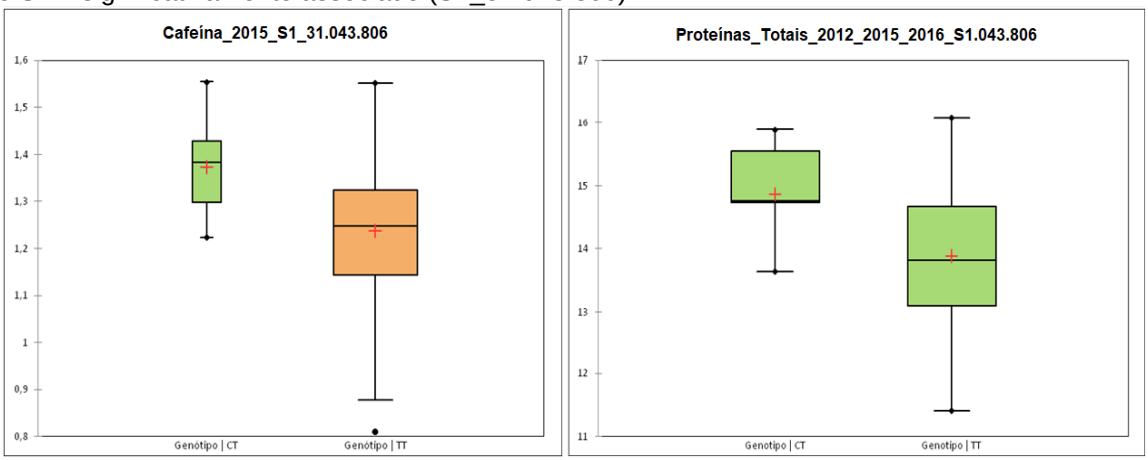
É interessante ressaltar que o bloco de haplótipo de *C. arabica* mapeado em *C. canephora* e significativamente associado ao Cafestol (Cromossomo 6) possui o mesmo padrão genotípico em todo o painel de estudo que o bloco de haplótipo mapeado em *C. eugenioides* (Figura 13). Já o bloco de haplótipo de *C. arabica* mapeado em *C. canephora* e significativamente associado ao Caveol e Razão Caf/Cav (Cromossomo 2) possui padrões genotípicos diferentes em todo o painel em ambos subgenomas (Figura 14). Esse resultado demonstra que o bloco de haplótipo associado ao Caveol e Razão Caf/Cav introgredido nos materiais de *C. arabica* é específico de cada parental diploide de *C. arabica* e está influenciando da mesma forma na concentração desse composto nesses materiais.

23.6 DISTRIBUIÇÃO DOS VALORES FENOTÍPICOS PARA OS SNPS CO-LOCALIZADOS AOS GENES CANDIDATOS

Comparando o conteúdo de Cafeína e Proteínas Totais entre diferentes genótipos para o SNP S1_31.043.806 significativamente associado nos anos 2012, 2015 e 2016, observamos que os indivíduos com o genótipo homozigoto TT para o alelo alternativo apresentaram baixas concentrações de Cafeína e os

indivíduos com o genótipo heterozigoto CT, altas concentrações de Proteínas Totais (Figura 9). Como foi observado na análise fenotípica para o ano de 2016 (Capítulo 2), assim como em estudos anteriores com esses materiais (SCHOLZ et al., 2016), o teor de cafeína nos grãos de café verde é correlacionado ao teor de proteínas (r = 0,65). Esse resultado demonstra a influência de uma variante SNP na composição química de diferentes compostos relacionados com a qualidade da bebida de café.

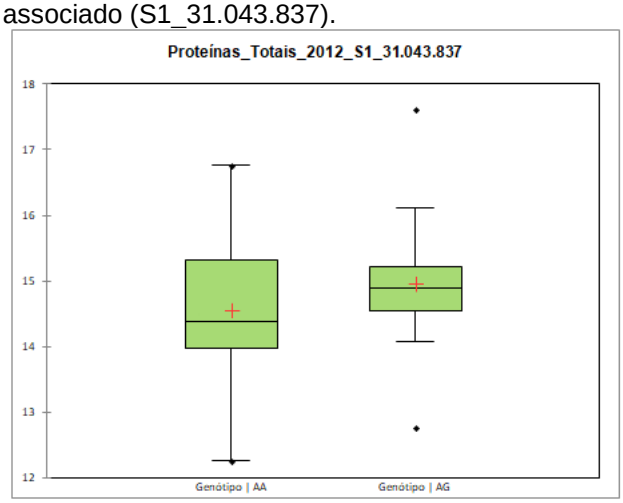
Figura 9. Distribuição do teor de Cafeína e PT (ano 2012, 2015 e 2016) segundo genótipos para o SNP significativamente associado (S1_31.043.806).



Os genótipos heterozigotos e homozigotos para o alelo alternativo estão representados como CT e TT e indicam alta e baixa concentração dos compostos associado a esses alelos. Os Box plots representam os quantis superiores e inferiores, com os valores medianos mostrados em uma linha no meio do box. Genes a uma janela de 100 Kpb a esses SNPs significativamente associados foram considerados candidados. Valores expressos em g.100 g^{-1}.

Comparando os valores de Proteínas Totais entre diferentes genótipos para o SNP S1_31.043.837 significativamente associado no ano 2012 (Figura 10), observamos que os indivíduos com o genótipo homozigoto AA para o alelo referência apresentaram maiores concentrações de PT em relação aos indivíduos com o genótipo heterozigoto AG.

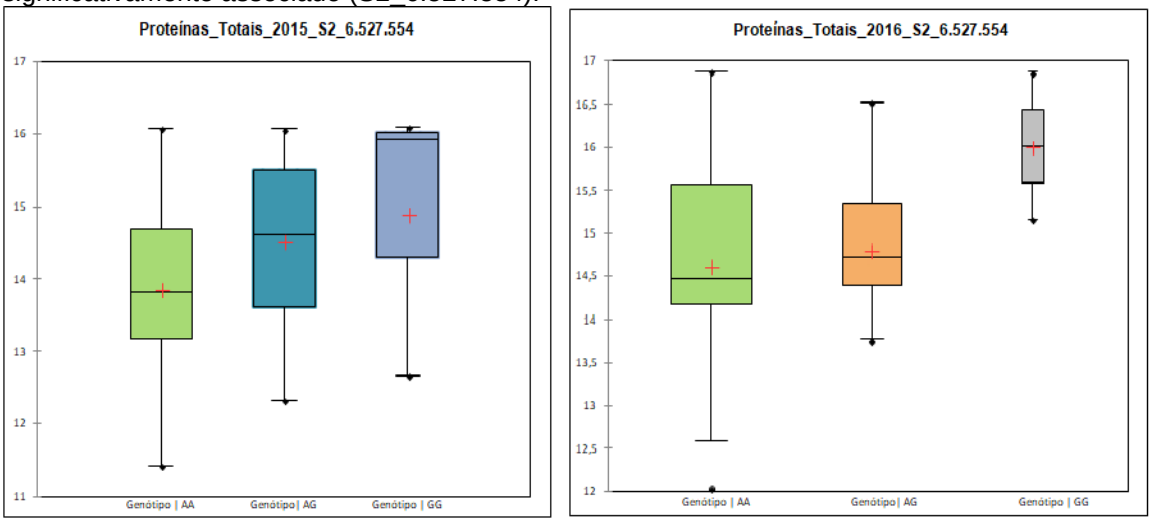
Figura 10. Distribuição do teor de PT (ano 2012) segundo genótipos para o SNP significativamente associado (S1_31.043.837).



O genótipo AA é homozigoto para o alelo referência, AG é o heterozigoto. Valores expressos em g.100 g^{-1}.

Comparando os valores de Proteínas Totais entre diferentes genótipos para o SNP S2_6.527.554 significativamente associado nos anos 2015 e 2016 (Figura 11), observamos que os indivíduos com o genótipo homozigoto GG para o alelo alternativo apresentaram maiores concentrações de PT em relação aos indivíduos com o genótipo heterozigoto AG e AA (homozigoto para alelo referência). O box plot mostra que o genótipo GG não alterou a concentração de PT nos diferentes anos, demonstrando que os genes co-localizados próximos a ele possam estar sob forte controle genético. De acordo com nossos resultados de busca dos genes candidatos, esse SNP obtido por GWAS está co-localizado ao gene Cc02_g08220 (*bHLH113*), um importante Fator de transcrição basal.

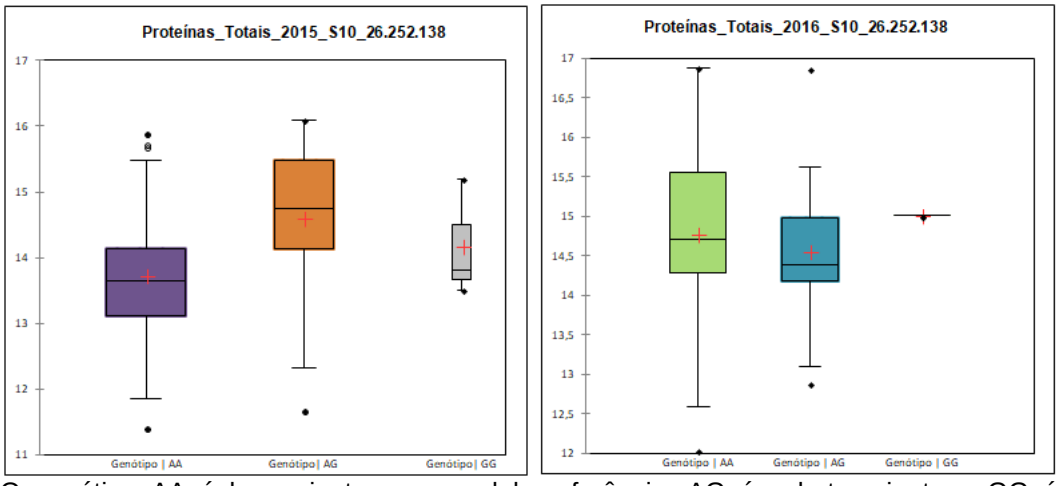
Figura 11. Distribuição do teor de PT (ano 2015 e 2016) segundo genótipos para o SNP significativamente associado (S2_6.527.554).



O genótipo AA é homozigoto para o alelo referência, AG é o heterozigoto e GG é homozigoto para o alelo alternativo. Valores expressos em g.100 g^{-1}.

Ainda comparando os valores de Proteínas Totais entre diferentes genótipos para o SNP S10_26.252.138 significativamente associado nos anos 2015 e 2016, observou-se que para o ano de 2015 o dobro de genótipos representam a variabiliade fenotípica de PT em relação ao ano de 2016, distorcendo assim a influência real do alelo entre os dois anos (Figura 12).

Figura 12. Distribuição do teor de PT (ano 2015 e 2016) segundo genótipos para o SNP significativamente associado (S10_26.252.138).

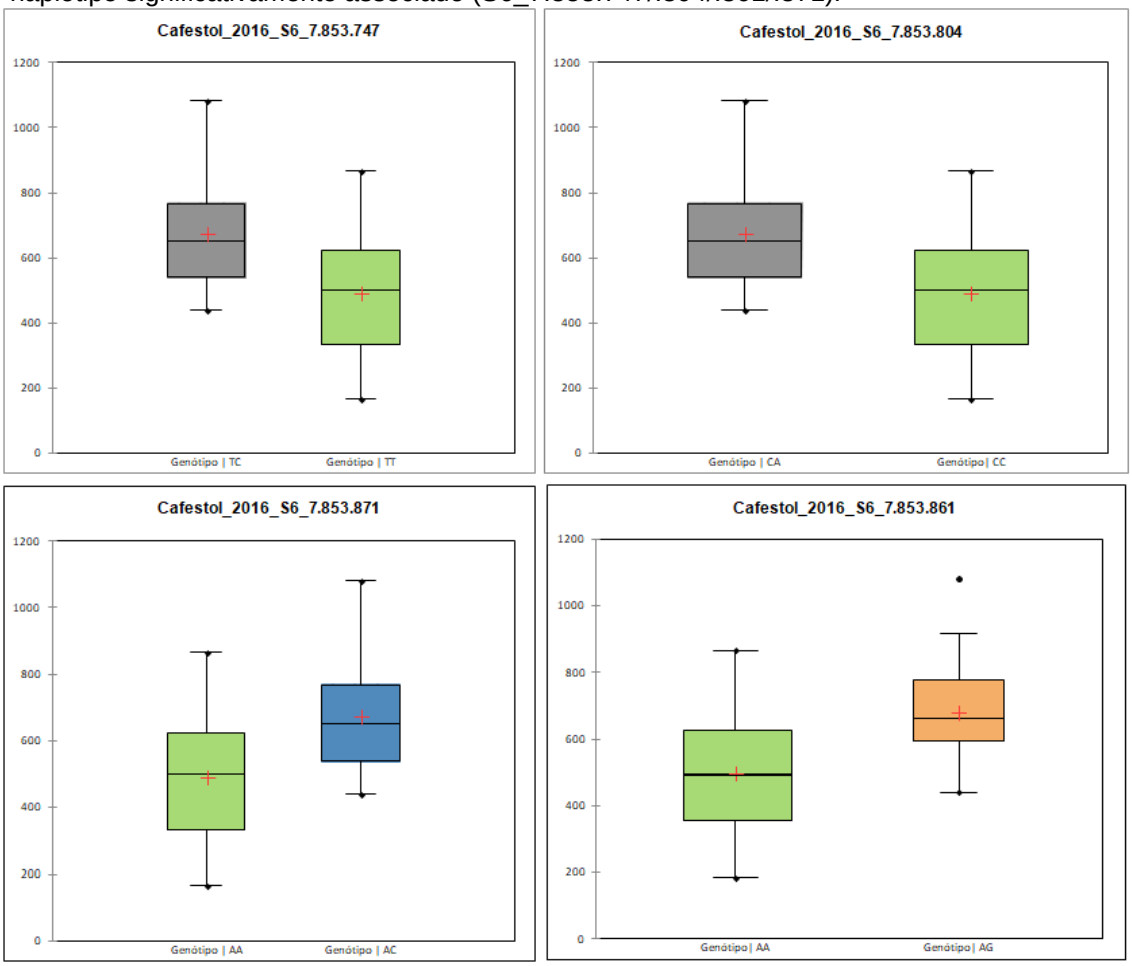


O genótipo AA é homozigoto para o alelo referência, AG é o heterozigoto e GG é homozigoto para o alelo alternativo. Valores expressos em g.100 g^{-1}.

Quanto aos valores de Cafestol entre diferentes genótipos para o bloco de haplótipo (S6_7.853.747, S6_7.853.804, S6_7.853.861, e S6_7.853.871) significativamente associado no ano 2016, observamos que os indivíduos com o

genótipo heterozigoto TC, CA, AC e AG apresentaram maiores concentrações desse composto em relação aos indivíduos com os genótipos em homozigose (Figura 13). É interessante ressaltar que esse bloco de haplótipo de *C. arabica* foi mapeado em ambos subgenomas, *C. canephora* e *C. eugenioides* e possui o mesmo padrão genotípico em todos os indivíduos analisados. Esse bloco de haplótipo foi discutido no ítem 23.4.

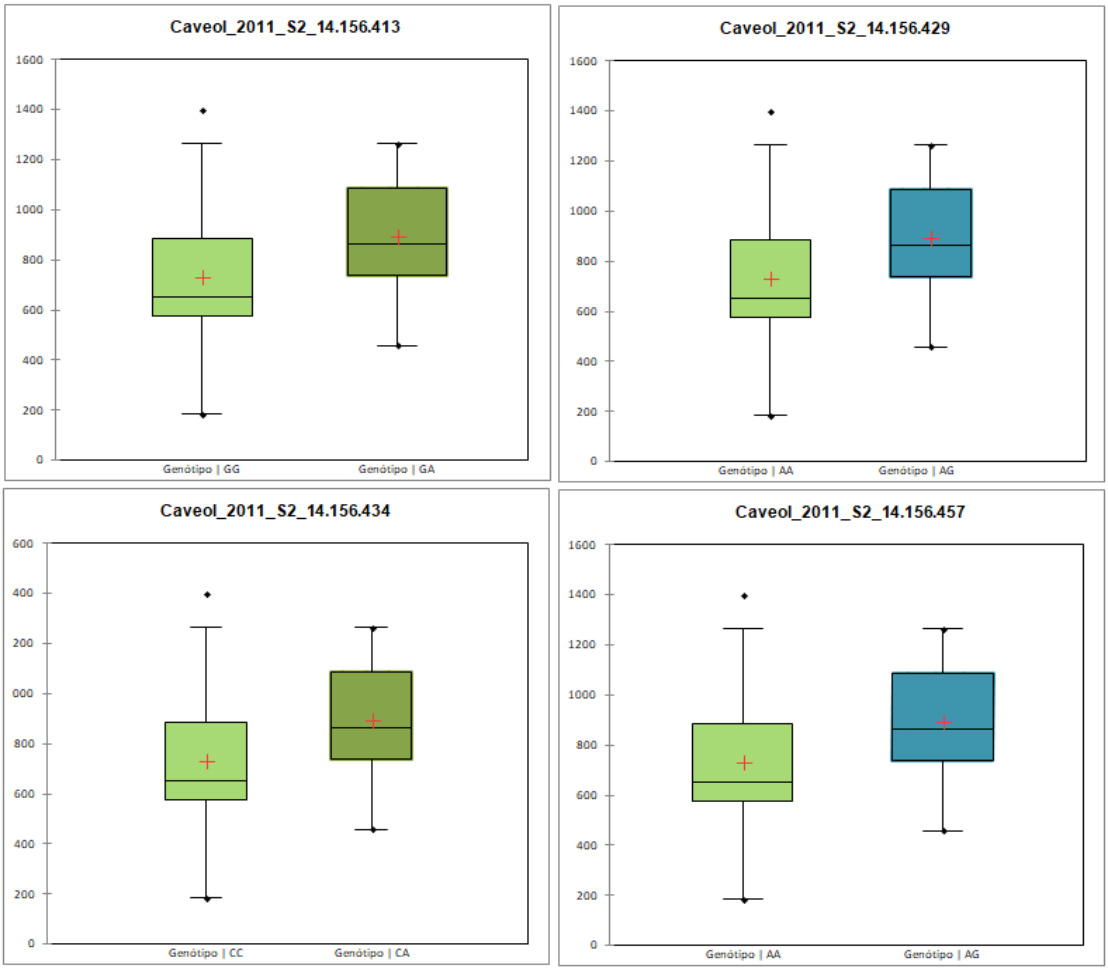
Figura 13. Distribuição do teor de Cafestol (ano 2016) segundo genótipos para o bloco de haplótipo significativamente associado (S6_7.853.747/.804/.861/.871).



O genótipo TT, CC e AA são homozigotos para o alelo referência e TC, CA, AC e AG são heterozigotos. Valores expressos em mg.100 g^{-1}.

Para os valores de Caveol e Razão Caf/Cav entre diferentes genótipos para o bloco de haplótipo (S2_14.156.413, S2_14.156.429, S2_14.156.434, S2_14.156.457) significativamente associado no ano 2011, observamos que os indivíduos com o genótipo heterozigoto GA, AG e CA apresentaram maiores concentrações de Caveol em relação aos indivíduos com os genótipos em homozigose (Figura 14).

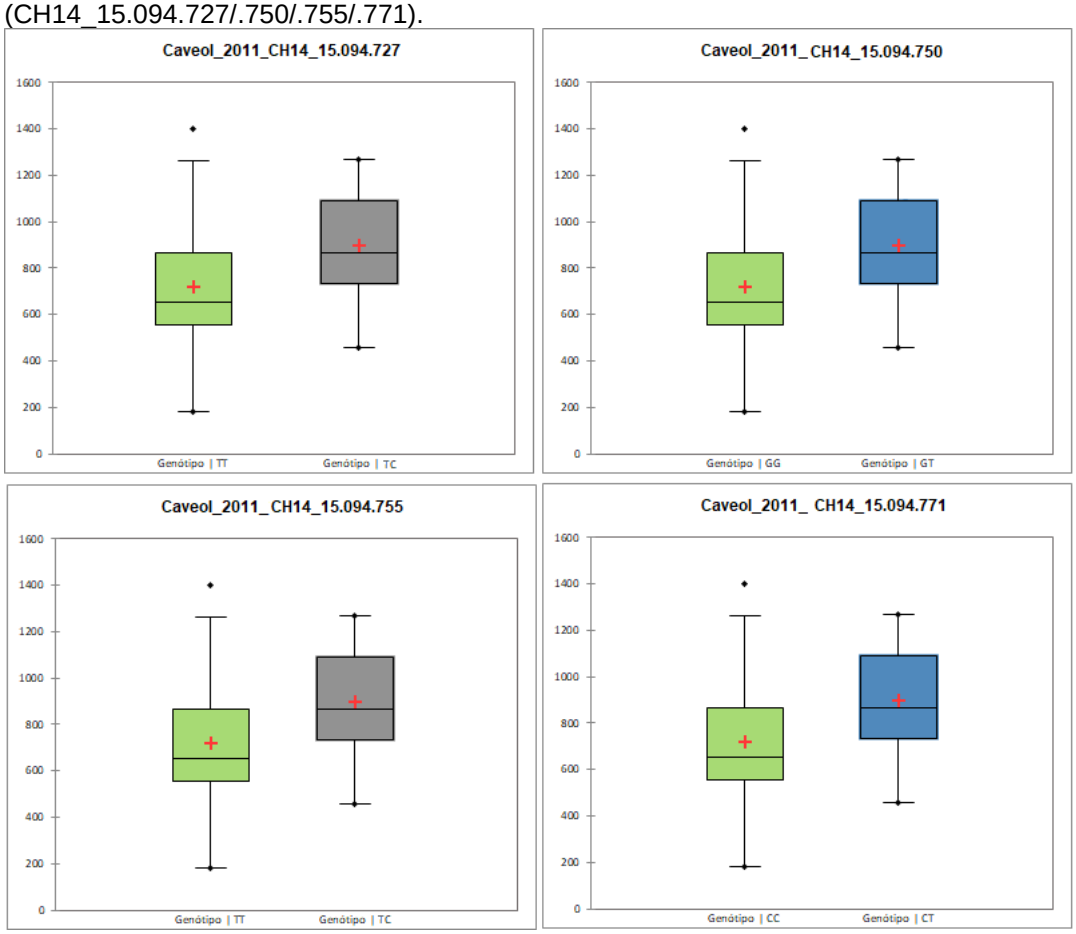
Figura 14. Distribuição do teor de Caveol (ano 2011) segundo genótipos para o bloco de haplótipo de *C. arabica* mapeado em *C. canephora* e significativamente associado (S2_14.156.413/.429/.434/.457).



O genótipo GG, AA e CC são homozigotos para o alelo referência e GA, AG e CA são heterozigotos. Valores expressos em mg.100 g^{-1}.

Esse mesmo bloco de haplótipo significativamente associado ao Caveol/ Razão Caf/Caveol foi mapeado em *C. eugenioides* e possui padrões genotípicos diferentes em todos os indivíduos analisados. Os indivíduos que possuem o genótipo heterozigoto (TC, GT e CT) apresentaram maiores concentrações de Caveol em relação aos indivíduos com os genótipos em homozigose (Figura 15). Esse bloco de haplótipo foi discutido no ítem 23.4.

Figura 15. Distribuição do teor de Caveol (ano 2011) segundo genótipos para o bloco de haplótipo de *C. arabica* mapeado em *C. eugenioides* e significativamente associado (CH14_15.094.727/.750/.755/.771).



O genótipo TT, GG e CC são homozigotos para o alelo referência e TC, GT e CT são heterozigotos. Valores expressos em mg.100 g^{-1}.

23.7 GENES CANDIDATOS CO-LOCALIZADOS AOS SNPS ASSOCIADOS

De acordo com os resultados da análise de decaimento do DL por todo o genoma, uma região total de 100 Kb foi identificada como candidata para a busca de genes co-localizados aos SNPs no subgenoma de *C. canephora.* Dos 22 loci SNPs significativamente associados às características fenotípicas, 17 SNPs de *C. arabica* mapeados em C. *canephora* são co-localizados a 13 genes candidatos envolvidos nas vias metabólicas da Cafeína, PT, Caveol, Cafestol e Razão Caf/Cav (Tabela 2 e Suplementar 4).

Tabela 2. Genes candidatos co-localizados próximos aos SNPs que apresentaram associação significativa para Cafeína, Proteínas Totais, Cafestol, Caveol e Razão Caf/Caveol detectados por no mínimo dois métodos associativos e/ou mais de dois anos analisados.

Característica	Ano	SNP	Gene Candidato	Distância (Kbp)	Anotação Funcional
Cafeína	2015	S1_31.043.806	Cc01_g12390	6.78	Subunidade 2 do subcomplexo α1 da NADH Desidrogenase (At5g47890)
			Cc01_g12480	42.77	Proteína (CNX3) de biossíntese do cofator Molibdopterina (Moco)
Proteínas Totais	2012/2015/2016	S1_31.043.806	Cc01_g12390	6.78	Subunidade 2 do subcomplexo α1 da NADH Desidrogenase (At5g47890)
			Cc01_g12480	42.77	Proteína (CNX3) de biossíntese do cofator Molibdopterina (Moco)
			Cc01_g12540	69.31	Hipotética enzima que realiza fosforilação proteica (GSCOCT00028192001)
Proteínas Totais	2015/2016	S10_26.252.138	Cc10_g15280	25	Putativa D-Carboxipeptidase (*CPD*)
	2015/2016	S2_6.527.554	Cc02_g08220	23	Putativo fator de transcrição (*bHLH113*)
Proteínas Totais	2012	S1_31.043.837	Cc01_g12390	6.75	Subunidade 2 do subcomplexo α1 da NADH Desidrogenase (At5g47890)
			Cc01_g12480	42.74	Proteína (CNX3) de biossíntese do cofator Molibdopterina (Moco)
			Cc01_g12540	69.28	Hipotética enzima que realiza fosforilação proteica (GSCOCT00028192001)
Cafestol	2016	S6_7.853.747/.804/.861/.871	Cc06_g09560	100	Aldeído desydrogenase família 2 membro C4 (*ALDH2C4*)
			Cc06_g09670	13	Flavina monooxigenase (*FCM*)
Caveol e Razão Caf/Cav	2011	S2_14.156.413/.429/.434/.457	Cc02_g15760	35	Citocromo P450 81D1 (*CYP81D1*)
			Cc02_g15870	210	Citocromo P450 81D1 (*CYP81D1*)

Caveol e Cafestol estão entre os 6 diterpenos exclusivos do gênero *Coffea* e possuem como diferença uma dupla ligação no hidrocarboneto aromático composto por 20 carbonos (PEREIRA E IVAMOTO, 2015). A via comum dos diterpenos em plantas inclui uma maquinaria de enzimas que consistem em diterpeno sintases (TPSs), monooxigenases da família *P450* (*CYPs*) e vários tipos de transferases e oxidorredutases que estão diretamente

envolvidas na biossíntese desses metabólitos secundários (IVAMOTO et al., 2015).

Sabe-se que as monooxigenases podem reconhecer esqueletos cauranos e realizar modificações de oxirredução levando a produção dos terpenoides e foram descritas como estando diretamente envolvidas na biossíntese desses compostos em plantas (SIRÉN et al., 2016). Estudos relatam as *P450s* como potenciais enzimas envolvidas nos estágios finais das vias biossintéticas de Cafestol e Caveol, porém a identificação das enzimas da família *P450* é difícil devido ao seu grande tamanho e diversidade (WRIESSNEGGER et al., 2014; IVAMOTO et al., 2017).

Até o momento alguns estudos de identificação e caracterização de putativas enzimas envolvidas na bioossíntese dos isoprenoides para o café foram relatados (TISKI et al., 2011; WANG et al., 2012; DEL TERRA et al., 2013; IVAMOTO et al., 2016; 2017; SANT'ANA et al., 2018).

Nesse trabalho nós identificamos um bloco de haplótipo (S2_14.156.413, S2_14.156.429, S2_14.156.434, S2_14.156.457 significativamente associado a Caveol e Razão Caf/Cav (ano 2011) e co-localizado a dois genes candidatos (Cc02_g15760 e Cc02_g15870) que codificam duas citocromos *P450* (*CYP81D1*). Porém essa monooxigenase não foi identificada em nenhuma via metabólica até o momento (WIESNER, SCHREINER e ZRENNER, 2014). SANT'ANA et al. (2018) identificaram um SNP (S11_29.778.697) em *C. canephora* associado a Cafestol que codifica a monooxigenase *CYP704.* Na análise do padrão de expressão, a *CYP704* demonstrou um padrão transcricional similar ao acúmulo de Caveol durante o desenvolvimento do fruto de café.

SANT'ANA et al. (2018) também encontraram SNPs localizados nos cromossomos 2 e 6 de *C. canephora* significativamente associados aos diterpenos (S2_45.775.221, S2_48.526.210, S6_12.529.278 e S6_7.853.861). Esses SNPs são co-localizados a putativos genes envolvidos da via dos diterpenos sugerindo que os genes da biossíntese desses compostos em café possam estar agrupados nos cromossomos 2 e 6. Para o arroz algumas enzimas da família *P450* foram encontradas no cromossomo 2 (OTOMO et al., 2004; SHIMURA et al., 2007)

Foi identificado outro bloco de haplótipo associado ao Cafestol (ano 2016) (S6_7.853.747, S6_7.853.804, S6_7.853.861 e S6_7.853.871) e co-localizado aos genes Cc06_g09560 e Cc06_g09670, que codificam uma Aldeído Desidrogenase (*ALDH2C4*) e uma Flavina Monooxigenase (*FCM*). Em comparação aos SNPs significativamente associados aos diterpenos no estudo de SANT'ANA et al. (2018), somente o SNP S6_7.853.861 co-localizado à *FCM* foi detectado em nosso estudo, no qual compõe bloco de haplótipo significativamente associado ao Cafestol. Esse recente trabalho também demonstrou em análise do padrão de expressão que a *FCM* é fortemente expressa nos estágios finais de maturação dos frutos de café, e pode ter um potencial papel na composição final de lipídeos e seus derivados em grãos de café verde. Sabe-se que nesse estudo foi utilizado um filtro mais estringente (Ho < 0.5), onde os SNPs significativamente associados aos diterpenos do estudo anterior foram retirados pelos filtros de controle de qualidade. Porém o número de marcas utilizadas nesse estudo foi maior (4.517 SNPs) além de 4 anos de fenotipagem e média para as análises, aumentando a chance das variantes SNPs encontradas na população estarem associadas aos compostos analisados.

TAKEUCHI et al. (1980) relataram que uma *ALDH* citosólica desempenha papel na biossíntese dos terpenoides em batata doce infectada pelo fungo *C. fimbriata*, e o estudo propôs que a via dos terpenoides consiste de uma piruvato descarboxilase, aldeído desidrogenase (*ALDH*) e acetil-CoA sintetase. PADDON et al. (2013) também demonstraram que a *ALDH1* catalisa a formação de ácido dihidroartemisínico na via da Artemisina, um sesquiterpenoide com função terapêutica para a malária.

Em *Arabdopsis thaliana* uma monooxigenase (*CYP74A1)* faz parte das *CYPs* envolvidas na formação de ácido jasmônico, um hormônio de defesa em plantas, e com isso aldeídos de 6 C são gerados (LAUDERT et al., 1996; BATE et al., 1998). As Aldeído Desidrogenases (*ALDH*) oxidam os aldeídos gerando ácidos carboxílicos. Pelos aldeídos serem deletérios aos sistemas biológicos (WEI et al., 2009), faz da *ALDH2C4* uma potencial enzima que realiza detoxificação dos aldeídos na via dos diterpenos.

Em um estudo sobre o panorama evolutivo de cafeína em diferentes espécies de plantas foi identificada a evolução convergente de 3 distintas N-methyltransferases (NMTs) envolvidos na biossíntese da cafeína localizados no cromossomo 1, no qual expandiram por meio de uma duplicação sequencial idependentemente dos genes de cacao e chá, sugerindo que a cafeína em dicotiledônias possui origem polifilética (DENOEUD et al., 2014). Grãos de café Arábica contém até 1.8% de cafeína e desde os anos 70 a demanda por café descafeinado tem aumentado rapidamente, devido aos potenciais efeitos adversos à saúde associados ao seu consumo excessivo, como já mencionado (ASHIHARA e SUZUKI, 2004).

O SNP (S1_31.043.806) localizado no cromossomo 1 foi significativamente associado ao conteúdo de Cafeína (ano 2015) e co-localizado ao gene Cc01_g12390. O gene Cc01_g12390 é a subunidade 2 do subcomplexo α1 da NADH desidrogenase (ubiquinona). As ubiquinonas são um grupo de benzoquinonas lipossolúveis envolvidas no transporte de elétrons e portando são oxirredutases (SOOLE e MENZ, 1995). Sabe-se que em humanos, microorganismos e outros animais as *P450* são responsáveis por quase todos processos metabólicos da cafeína realizando desmetilação, e dentre eles intermediam a formação de paraxantina e também de ácido 1,7-dimetilúrico a partir da desmetilação da cafeína (KALOW e TANG, 1991; FUHR, 1992; NEHLIG, 2017).

Em plantas, até onde sabemos poucas investigações relataram qualquer tipo de atividade desmetilase, principalmente em café (MAZZAFERA, 2004). Assim, as oxirredutases parecem participar de sequenciais reações do catabolismo da cafeína em café (ASHIHARA e CROZIER, 2001).

O mesmo SNP (S1_31.043.806) significativamente associado ao teor de Cafeína (ano 2015) e PT (ano 2012, 2015 e 2016) está co-localizado ao gene Cc01_g12480, responsável por codificar a proteína *CNX3* envolvida na biossíntese de cofator redox Molibdênio-molibdopterina (Moco). Em plantas, os genes *CNXs* estão envolvidos na síntese de Moco e existem diversas apoproteínas dependente de Moco (Mo-enzimas) que catalizam reações redox, sendo uma delas a Xantina Desidrogenase (*XDH*), uma enzima envolvida no catabolismo das purinas (VITÓRIA e MAZZAFERA, 1999; MENDEL e BITTNER,

2006; HILLE et al., 2011). O gene *CNX3* foi identificado na matriz mitocondrial em *Arabidopsis thaliana* e foi relatado como potencialmente envolvido na produção de ácido 1,7-Dimetilúrico a partir de paraxantina (TESCHNER et al., 2010). Assim como em um estudo sobre a degradação de purinas em folhas e frutos de *C. arabica* e *C. dewevrei,* a Moco foi essencial na atividade da enzima *XDH* (VITÓRIA e MAZZAFERA, 1999).

O SNP S1_31.043.806 foi significativamente associado ao teor de PT (ano 2012, 2015 e 2016), como também está a 31Kpb do SNP S1_31.043.837 (ano 2012), co-localizado à subunidade 2 da NADH desidrogenase (ubiquinona) (Cc01_g12390), à *CNX3* (Cc01_g12480) e a uma hipotética enzima com provável função de fosforilação proteica (Cc01_g12540). Nossos resultados demonstram o papel constitutivo desses genes na via das PT e da Cafeína.

Entre os SNPs associados com PT (ano 2015 e 2016) um (S2_6.527.554) está co-localizado ao gene Cc02_g08220, no qual sintetiza um putativo fator de transcrição (*bHLH113*), considerado hélice-alfa-hélice básico (do inglês – basic helix-loop-helix). Os fatores de transcrição (FTs) *bHLH* são uma das maiores famílias de FTs em eucariotos, e a segunda maior classe de fatores de transcrição em plantas (MURRE et al., 1994; ZHANG et al., 2015). Esses FTs possuem um domínio altamente conservado entre várias espécies de plantas e estão envolvidos no controle transcricional de genes que participam de diversas vias metabólicas, atuando como ativadores ou repressores transcricionais (TOLEDO-ORTIZ et al*.,* 2003; PIRES e DOLAN, 2010).

Vários FTs *bHLHs* foram identificados em importantes plantas modelo e não-moldelos, e com base em análises de GWAS, 167, 190 e pelo menos 191 FTs *bHLHs* foram relatados nos genomas *Arabidopsis thaliana L*, *Oriza sativa, Nicotiana tabacum L*. e *Vitis vinifera L*, respectivamente (LI et al., 2006; JAILLON et al., 2007; RUSHTON et al., 2008; CARRETERO-PAULET et al., 2010). Seus papéis incluem regulação da deiscência dos frutos, carpelos, anteras e desenvolvimento de células epidérmicas, sinalização fitoquímica, biossíntese de flavonoides, terpenoides, alcaloides, monooxigenases, sinalização de hormônios e resposta a estresses bióticos e abióticos (QUATTROCCHIO et al., 1998; DOMBRECHT et al., 2007; SHINOZAKI & YAMAGUCHI-SHINOZAKI, 2007; STEYN et al., 2009; HICHRI, et al., 2010; FELLER, MACHEMER e

GROTEWOLD, 2011; SCHWEIZER et al., 2013; LI et al., 2014; 2016; WANG et al., 2015; Qi et al., 2015; YAMAMURA et al., 2015.

Zhang et al., 2015 em um estudo da expressão de genes da família *bHLH* em vários tecidos da *Salvia miltiorrhiza Bungue*, uma planta chinesa com propriedades medicinais, identificaram genes *bHLH* com altas expressões e desses, sete foram potencialmente envolvidos na biossíntese das tanshinonas, um importante grupo de constituintes farmacologicamente ativos dessa planta. OLIVAS et al. (2016) buscando entender a arquitetura genética da resposta de *Arabdopsis thaliana* frente a múltiplos estresses encontraram *bHLHs* envolvidos em resposta à seca.

Entre os SNPs associados às PT (ano 2015 e 2016) um (S10_26.252.138) é co-localizado ao gene Cc10_g15280, que sintetiza uma putativa *D-Carboxipeptidase* (*CPD*). As carboxipeptidases realizam proteólise de polipeptídeos e pertencem a família das serina carboxipeptidases (enzimas que usam um resíduo de serina como sítio ativo). Elas possuem uma conservada ‘tríade catalítica’ composta por Ser-Asp-His (serina, ácido aspártico e um resíduo de histidina) e catalizam a hidrólise de peptídeos, ésteres e amidas, liberando aminoácidos, alcoóis e grupos amônia a partir dos terminais C dos peptídeos, ésteres e amidas (BREDDAM, 1986).

Dadas suas características enzimáticas na hidrólise de proteínas, as serina carboxipeptidases têm sido amplamente utilizadas como ferramentas de biologia molecular para determinação, troca e síntese da sequência C-terminal de peptídeos, sendo importantes na modificação pós-traducional durante a maturação da proteína funcional (FENG e XUE, 2006). A carboxipeptidase D cliva preferencialmente resíduos Arg e Lis no terminal C dos polipeptídeos.

24 CONCLUSÃO

A tecnologia GBS junto ao *Pipeline* TASSEL utilizando os genomas de referência dos ancestrais diploides do alopoliploide *C. arabica* forneceu 1.719 e 2.949 SNPs de *C. arabica* para análise de diversidade e estrutura que agrupou os genótipos em 2 e 3 principais grupos. Dentre os agrupamentos, um grupo adicional foi formado principalmente de acessos silvestres do lado Oeste do Vale do Rift demonstrando maior diversidade genética do que os outros grupos.

No estudo de associação com 4.517 SNPs e 5 anos de fenotipagem, esse trabalho identificou 33 SNPs associados a compostos relacionados à qualidade da bebida de café, sendo 11 SNPs mapeados em *C. eugenioides* e 22 SNPs mapeados em *C. canephora*. Desses 22 SNPs significativamente associados aos compostos relacionados à qualidade da bebida 17 são co-localizados a 13 genes candidatos anotados em *C. canephora.*

Este trabalho demonstrou a riqueza alélica e a importância de se conservar os acessos com maior variabilidade do lado Oeste do Vale do Rift do banco de germoplasma do IAPAR, como também identificou novos SNPs informativos e putativos genes causais para a engenharia genética, sendo o ponto inicial para obter plantas com desejáveis conteúdos de compostos associados com a qualidade da bebida de café e/ou visando a saúde.

25 REFERÊNCIAS

ADDINSOFT. XLSTAT: your data analysis solution. Paris: Addinsoft, 2010.

AERTS, R. et al. Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian mountain rainforests. **Evol. Appl**. 6, 243–252, 2013.

ANDRADE, A.C.; DE A. CARNEIRO, F.; DA SILVA JR, O.B.; MARRACCINI, P.; GRATTAPAGLIA, D**. Towards GWAS and genomic prediction in coffee: development and validation of a 26K SNP chip for Coffea canephora**. In: Proceedings Plant and Animal Genome XXV Conference. San Diego, 2017.

ASHIHARA H.; CROZIER A. Caffeine: a well known but little mentioned compound in plant science. **Trends. Plant. Sci**, 6: 407-413, 2001.

ASHIHARA H.; SUZUKI T. Distribution and biosynthesis of caffeine in plants. **Front. Biosci**, 9: 1864-1876, 2004.

ANTHONY, F.; BERTRAND, B.; QUIROS, O.; WILCHES, A.; LASHERMES, P.; BERTHAUD, J.; CHARRIER, A. Genetic diversity of wild coffee (*Coffea arabica L.*) using molecular markers. **Euphytica**, 118: 53-65, 2001.

ATAWONG, A.; HASEGAWA, M AND KODAMA, O. Biosynthesis of Rice Phytoalexin: Enzymatic Conversion of 3β-Hydroxy-9β-pimara-7,15-dien-19,6β-olide to Momilactone A. **Biosci. Biotechnol. Biochem**, 66, 566–570, 2002.

A. TAKEUCHI.; K. J. SCOTT.; K. OBA AND I. URITANI. **Plant & Cell Physiol**, 21, 917, 1980.

BARRE, P.; AKAFFOU, S.; LOUARN, J.; CHARRIER, A.; HAMON, S.; NOIROT, M. Inheritance of caffeine and heteroside contents in an interspecific cross between a cultivated coffee species Coffea liberica var dewevrei and a wild species caffeine free C. pseudozanguebariae. **Theor. Appl Genet**, 96:306–311, 1998.

BARRETT, J.C.; B. Fry J. Maller M. J. Daly. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics***,* Volume 21, Issue 2, 15 January 2005, p. 263–265, 2005.

BATE, N.J.; SIVASANKAR, S.; MOXON, C.; RILEY, J.; THOMPSON, J.E.; ROTHSTEIN, S.J. Molecular characterization of an Arabidopsis gene encoding hydroperoxide lyase, a cytochrome P-450 that is wound inducible. **Plant Physiology**, 117(4):1393, 1998.

BRADBURY, P. J et al. TASSEL: Software for association mapping of complex traits in diverse samples. **Bioinformatics***,* 23, 2633–263, 2007.

BREDDAM K. Serine carboxypeptidases. A review. Carlsberg Res Commun. 51: 83 – 128, 1986.

CAI, C.; YE, W.; ZHANG, T.; GUO, W. Association analysis of fiber quality traits and exploration of elite alleles in upland cotton cultivars/accessions (Gossypium hirsutum L.). **J. Integr. Plant. Biol**, 56:51–62, 2014.

CARRETERO-PAULET, L. et al. Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. **Plant. Physiol**, 153, 1398–1412, 2010.

CENCI, A.; COMBES, M. C. & LASHERMES, P. Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. **Plant Mol. Biol**, l78, 135–45, 2012.

CHAN, E, K, F.; HAWKEN, R.; REVERTER, A. The combined effect os SNP-marker and phenotype atributes in genome-wide association studies. **Animal. Genetics**, v.40, n.2, p.149-56, abr, 2009.

CONAB – Companhia Nacional de Abastecimento. Acompanhamento da safra brasileira - safra 2018 - N. 8. Levantamento março, 2018.

DAVIS, A.P.; TOSH, J.; RUCH, N.; FAY, M.F. 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**, 167(4): 357-377, 2011

DE CASTRO, R. D.; MARRACCINI, P. Cytology, biochemistry and molecular changes during coffee fruit development. **Brazilian Journal of Plant Physiology**, v.18, n.1, p.175-199, 2006.

DEL TERRA, L.; LONZARICH, V.; ASQUINI, E.; NAVARINI, L.; GRAZIOSI, G.; LIVERANI, F.S et al. Functional characterization of three Coffea arabica L. monoterpene synthases: insights into the enzymatic machinery of coffee aroma. **Phytochemistry**. 89:6–14, 2013.

DENOEUD, F et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, n. 345, p.1181-1184, 2014.

DOMBRECHT, B.; XUE, G.P.; SPRAGUE, S.J.; KIRKEGAARD, J.A.; ROSS, J.J.; REID, J.B.; FITT, G.P.; SEWELAM, N.; SCHENK, P.M.; MANNERS J.M et al. MYC2 differentially modulates diverse jasmonate-dependent functions in *Arabidopsis*. **Plant Cell**, 19: 2225–2245, 2007.

DOYLE, J.J.; DOYLE, J.L. **Isolation of plant DNA from fresh tissue.** Focus 12: 13 15, 1990.

EARL, D. A & VON HOLDT, B. M. Structure harvester: A website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conserv. Genet. Resour**, 4, 359–361, 2012.

FENG, Y.; XUE Q. The serine carboxypeptidase like gene family of rice (*Oryza sativa* L. ssp. japonica). **Funct Integr Genomics**, 6: 14–24, 2006.

FELLER, A.; MACHEMER, K.; BRAUN, E.L.; GROTEWOLD, E. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors**. Plant J**. 66:94–116, 2011.

FERRÃO, M.F.; FURTADO, J.C.; NEUMANN, L.G.; KONZEN, P.H.A.; MORGANO, M.A.; BRAGAGNOLO, N.; FERREIRA, M.M.C. **Técnica não destrutiva de análise de tanino em café empregando espectroscopia no infravermelho e algoritmo genético**. Tecno-lóg. Santa Cruz do Sul, v.7, n.1, p. 9-26, 2003.

FLINT-GARCIA, S.; THORNSBERRY, J AND BUCKLER, E. Structure of linkage disequilibrium in plants. **Annu. Rev. Plant Biol**, 54:357–374, 2003.

FUHR, U: Biotransformation of caffeine and theophylline in mammalian cell lines genetically engineered for expression of single cytochrome P450 isoforms. **Biochem. Pharm**, 43, 2, 225-235, 1992.

GABRIEL, S.B.; SCHAFFNER, S.F.; NGUYEN, H.; MOORE, J.M.; ROY, J.; BLUMENSTIEL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M et al. The structure of haplotype blocks in the human genome. **Science**, 296, 2225–2229, 2002.

GLAUBITZ, J. C. et al. TASSEL-GBS: A high capacity Genotyping-by-Sequencing analysis pipeline. **PLoS One**, 9, e90346, 2014.

HAO, D.Y AND M.M. Yeoman: Evidence in favour of an oxidative N-demethylation of nicotine to nornicotine in tobacco cell cultures. J Plant Physiol 152, 420-426, 1998.

HICHRI, I. et al. The basic helix-loop-helix transcription factor MYC1 is involved in the regulation of the flavonoid biosynthesis pathway in grapevine. **Mol. Plant**, 3, 509–523, 2010.

HILLE R., NISHINO T., BITTNER F. Molybdenum enzymes in higher organisms**. Coord. Chem. Ver**, 255, 1179–1205, 2011.

HUANG, X.H et al. Genome-wide association studies of 14 agronomic traits in rice landraces. **Nat Genet**, 42:961–976, 2010.

HUBER, M AND T.W. BAUMANN. The first step of caffeine degradation in coffee - still a mistery. in Symposium Future Trends in Phytochemistry. Rolduc, The Netherlands, The Phytochemical Society of Europe, 1998.

IVAMOTO, S. T.; DOMINGUES, D. S.; VIEIRA, L. G. E & PEREIRA, L. F. P. Identification of the transcriptionally active cytochrome P450 repertoire in *Coffea arabica*. **Gen. Mol. Res**, 14, 2399–2412, 2015.

IVAMOTO, S. T.; SAKURAY, L. M.; FERREIRA, L. P.; KITZBERGER, C. S. G.; SCHOLZ, M. B. S.; POT, D.; LEROY, T.; VIEIRA, L. G. E.; DOMINGUES, D. S.; PEREIRA, L. F. P. Diterpenes biochemical profile and transcriptional analysis of

cytochrome P450s genes in leaves, roots, flowers, and during Coffea arabica L. fruit development. **Plant Physiology and Biochemistry** (Paris), v. 111, p. 340-347, 2016.

IVAMOTO, S. T. et al. Diterpenes biochemical profile and transcriptional analysis of cytochrome P450s genes in leaves, roots, flowers, and during *Coffea arabica* L. fruit development. **Plant Physiol. Biochem**,111, 340–347, 2017.

KALOW, W. AND B.K. Tang: Use of caffeine metabolite ratios to explore CYP1A2 and xanthine oxidase activities. **Clin Pharmacol Ther**, 50, 508-519, 1991.

LANGMEAD, B.; TRAPNELL, C.; POP, M.; SALZBERG, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Bio**, 110:25, 2009.

LASHERMES, P.; COMBES, M.C.; ROBERT, J.; TROUSLOT, P.; D'HONT, A.; ANTHONY F, et al. Molecular characterisation and origin of the *Coffea arabica L*. genome. **Mol. Gen. Genet**, Mar;261(2):259–66, 1999.

LI, R.; WELDEGERGIS, B.T.; LI, J.; JUNG, C.; QU, J.; SUN, Y.W.; QIAN, H.M.; TEE, C.; VAN LOON J.J.A.; DICKE, M et al. Virulence factors of Geminivirus interact with MYC2 to subvert plant resistance and promote vector performance. **Plant Cell**, 26: 4991–5008, 2014.

LIPKA, A.E.; TIAN, S.; WANG, Q.S.; PEIFFER, J.; LI, M.; BRADBURY, P.J.; GORE, M.A.; BUCKLER, E.S.; ZHANG, Z.M. **GAPIT: Genome association and prediction integrated tool**. Bioinformatics, 28 (18), pp. 2397-239, 2012.

LABOUISSE, J. P.; BELLACHEW, B.; KOTECHA, S & BERTRAND, B. Current status of coffee (*Coffea arabica L*.) genetic resources in Ethiopia: implications for conservation. *Genet. Resour*. **Crop. Evol***,* 55, 1079–1093, 2008.

LANGMEAD, B.; SALZBERG, S.L. Fast gapped-read alignment with Bowtie 2. **Nat. Methods**, 9: 357–35, 2012.

LAUDERT, D.; PFANNSCHMIDT, U.; LOTTSPEICH, F.; HOLLANDER-CZYTKO H.; WEILER, E.W. Cloning, molecular and functional characterization of Arabidopsis thaliana allene oxide synthase (CYP 74), the first enzyme of the octadecanoid pathway to jasmonates. **Plant. Mol. Biol**, 31(2):323-335, 1996.

LEROY, T.; DE BELLIS, F.; LEGNATÉ, H.; KANANURA, E.; GONZALES, G.; PEREIRA, L.F.P.; ANDRADE, A.C.; CHARMETANT, P.; MONTAGNON, C.; CUBRY, P.; MARRACCINI, P.; POT, D.; DE KOCHKO A**.** Improving the quality of African robustas: QTLs for yield-and quality-related traits in *Coffea canephora*. **Tree Genet. Genom**, 7, 781–798, 2011.

LI, X.L.; ZHANG, H.M.; AI, Q.; LIANG, G.; YU D.Q. Two bHLH Transcription Factors, bHLH34 and bHLH104, Regulate Iron Homeostasis in *Arabidopsis thaliana*. **Plant Physiol**, 170: 2478–2493, 2016.

MAZZAFERA, P. Catabolism of caffeine in plants and microorganisms. **Front. Biosci**, 9: 1348-1359, 2004.

MANGIN, B.; SIBERCHICOT, A.; NICOLAS, S.; DOLIGEZ, A.; THIS, P.; CIERCO-AYROLLES, C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. **Heredity**, 108: 285–291, 2012.

MARCHINI, J.; CARDON, L.R.; PHILLIPS, M.S.; DONNELLY, P. The effects of human population structure on large genetic association studies**. Nat Genet**, 36(5):512–517, 2004.

MENDEL, R.; BITTNER F. Cell biology of molybdenum. Biochim**. Biophys. Acta**, 1763: 621–635, 2006.

MÉROT-L'ANTHOËNE, V.; MANGIN, B.; LEFEBVRE-PAUTIGNY, F. et al. Comparison of three QTL detection models on biochemical, sensory, and yield characters in *Coffea canephora*. **Tree Genet. Genom**,10, 1541–1553, 2014.

MEYER, G. F. Notes on wild *Coffea arabica* from Southwestern Ethiopia, with some historical considerations. **Econ. Bot**, 19, 136–151, 1965.

MEYER, F. G. et al. FAO. **Coffee Mission to Ethiopia**, 1964–65. FAO, Rome, Italy, 1968.

MONCADA, M. D.; TOVAR, E.; MONTOYA, J. C.; ET AL. A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. **Tree Genet. Genom**, 12, 5, 2016.

MOSER. G.; LEE, S.H.; HAYESC, B.J, et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. **PLoS. Genet**, 11: e1004969, 2015.

MURRE, C.; BAIN, G.; VAN DIJK, M. A.; ENGEL, I.; FURNARI, B. A.; MASSARI, M. E.; MATTHEWS, J. R.; QUONG, M. W.; RIVERA, R. R.; STUIVER, M. H. Structure and function of helix-loop-helix proteins. **Biochim. Biophys. Acta**, v. 1218, n. 2, p. 129-135, June 1994.

NEHLIG, A. **Interindividual differences in caffeine metabolism and factors driving caffeine consumption**. Pharmacol Rev, 70(2)384–411, 2017.

OTOMO, K.; KANNO, Y.; MOTEGI, A.; KENMOKU, H.; YAMANE, H.; MITSUHASHI, W.; OIKAWA, H.; TOSHIMA, H.; ITOH, H.; MATSUOKA, M.; SASSA, T AND TOYOMASU, T. Diterpene cyclases responsible for the biosynthesis of phytoalexins, momilactones A, B, and oryzalexins A-F in rice. **Biosci. Biotechnol. Biochem**, 68, 2001–2006, 2004.

PADDON, C.J.; WESTFALL ,P.J.; PITERA, D.J et al. High-level semisynthetic production of the potent antimalarial artemisinin. **Nature**, 496:528–532, 2013.

PEAKALL, R. & SMOUSE, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research an update. **Bioinformatics**, 28, 2537–2539, 2012.

PEREIRA, L. F. P & IVAMOTO, S. T. Chapter 6: C**haracterization of coffee genes involved in isoprenoid and diterpene metabolic pathways**. In: *Coffee in Health and Disease Prevention* (Preedy, R. V. Ed.). London: Academic Press, 45-51, 2015.

PESTANA, K. N.; CAPUCHO, A. S.; CAIXETA, E. T.; DE ALMEIDA, D. P.; ZAMBOLIM, E. M.; CRUZ, C. D.; & SAKIYAMA, N. S. Inheritance study and linkage mapping of resistance loci to Hemileia vastatrix in Híbrido de Timor UFV. **Tree Genetics & Genomes**, v. 11, n.4, p.1-13. 2015.

PIRES, N AND DOLAN, L. Origin and diversification of basic-helix-loophelix proteins in plants. **Mol. Biol. Evol**, 27, 862–874, 2010.

PRADHAN, S.K et al. Population structure, genetic diversity and molecular marker-trait association analysis for high temperature stress tolerance in rice. **Plos One**, 11, 2016.

PRAKASH, N. S.; COMBES, M. C.; SOMANNA, N.; LASHERMES, P. AFLP analysis of introgression in coffee cultivars (*Coffea arabica L*.) derived from a natural interspecific hybrid. **Euphytica**, v. 124, p. 265-271, 2002

QI, T.C.; HUANG, H.; SONG, S.S.; XIE, DX. Regulation of jasmonate-mediated stamen development and seed production by a bHLH-MYB complex in *Arabidopsis*. **Plant Cell**, 27: 1620–1633, 2015.

QIN, Z.S. et al. Partition-Ligation EM algorithm for haplotype inference with single nucleotide polymorphisms. **Am. J. Hum. Genet**, 71, 1242–1247, 2002.

QUATTROCCHIO, F.; WING, J. F.; VAN DER WOUDE, K.; MOL, J. N & KOES, R. Analysis of bHLH and MYB domain proteins: Species specific regulatory differences are caused by divergent evolution of target anthocyanin genes. **The Plant. J**, 13, 475–488, 1998.

RUSHTON, P. J. et al. Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. **Plant. Physiol**, 147, 280–295, 2008.

SANT'ANA, GUSTAVO C.; PEREIRA, LUIZ F. P.; POT, DAVID.; IVAMOTO, SUZANA T.; DOMINGUES, DOUGLAS S.; FERREIRA, RAFAELLE V.; PAGIATTO, NATALIA F.; DA SILVA, BRUNA S. R.; NOGUEIRA, LÍVIA M.; KITZBERGER, CINTIA S. G.; SCHOLZ, MARIA B. S.; DE OLIVEIRA, FERNANDA F.; SERA, GUSTAVO H.; PADILHA, LILIAN.; LABOUISSE, JEAN-PIERRE; GUYOT, ROMAIN; CHARMETANT, PIERRE; LEROY, THIERRY.

Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica L*. **Sci. Reports**, v. 8, p. 465, 2018.

SCHERER, A.; CHRISTENSEN, G, BRYCE. Concepts and relevance of genome-wide association studies. **Science Progress**, 99(1), 59 – 67(9), 2016.

SCHMUTZ, J., S.B. CANNON, J. SCHLUETER, J. MA, T. MITROS, W. NELSON, D.L. HYTEN, Q. SONG, J.J. THELEN, J. CHENG, D. XU, U. HELLSTEN, G.D. MAY, Y. YU, T. SAKURAI, T. UMEZAWA, M.K. BHATTACHARYYA, D. SANDHU, B. VALLIYODAN, E. LINDQUIST, M. PETO, D. GRANT, S. SHU, D. GOODSTEIN, K. BARRY, M. FUTRELLGRIGGS, B. ABERNATHY, J. DU, Z. TIAN, L ZHU, N. GILL, T. JOSHI, M. LIBAULT, A. SETHURAMAN, X.-C. ZHANG, K. SHINOZAKI, H.T. NGUYEN, R.A. WING, P. CREGAN, J. SPECHT, J. GRIMWOOD, D. ROKHSAR, G. STACEY, R.C. SHOEMAKER, AND S.A. Jackson. Genome sequence of the paleopolyploid soybean. **Nature**, 463:178–183, 2010.

SCHOLZ, M.B.S; PAGIATTO, N.F; KITZBERGER, C.S.G; PEREIRA, L.F.P; DAVRIEUX, F; CHARMETANT, P; LEROY, T. **Validation of nearinfrared spectroscopy for the quantification of cafestol and kahweol in green coffee**. Food Res Int, 61:176–182, 2014a.

SCHOLZ, M.B.S.; KITZBERGER, C.S.G.; PEREIRA, L.F.P.; DAVRIEUX, F.; POT, D.; CHARMETANT, P; LEROY, T. Application of near infrared spectroscopy for green coffee biochemical phenotyping. **J. Near. Infrared. Spectrosc**, 22:411–421, 2014b.

SHEHZAD, T.; I. HIROYOSHI AND O. KAZUTOSHI. Genome-wide association mapping of quantitative traits in sorghum (*Sorghum bicolor* (L.) Moench) by using multiple models. **Breeding. Sci**, 59: 217-227, 2009.

SHINOZAKI K and YAMAGUCHI-SHINOZAKI K. Gene networks involved in drought stress response and tolerance. **J. Exp. Bot**, 58:221-227, 2007.

SCHOLZ, M.B.S.; PAGIATTO, N.F.; KITZBERGER, C.S.G.; PEREIRA, L.F.P.; DAVRIEUX, F.; CHARMETANT, P.; LEROY, T. Validation of nearinfrared spectroscopy for the quantification of cafestol and kahweol in green coffee. **Food. Res. Int**, 61:176–182, 2014a.

SCHOLZ, M.B.S.; KITZBERGER, C.S.G.; PEREIRA, L.F.P.; DAVRIEUX, F.; POT, D.; CHARMETANT, P.; LEROY, T. Application of near infrared spectroscopy for green coffee biochemical phenotyping. **J. Near. Infrared. Spectrosc**, 22:411–421, 2014b.

SCHWEIZER, F.; FERNÁNDEZ-CALVO, P.; ZANDER, M.; DIEZ-DIAZ, M.; FONSECA, S.; GLAUSER, G.; LEWSEY, M.G.; ECKER, J.R.; SOLANO, R AND REYMOND, P. Arabidopsis Basic Helix-Loop-Helix Transcription Factors MYC2, MYC3, and MYC4 Regulate Glucosinolate Biosynthesis, Insect Performance, and Feeding Behavior. **Plant Cell**, 25, 3117-3132, 2013.

SHIMURA, K.; OKADA, A.; OKADA, K.; JIKUMARU, Y.; KO, K-W.; TOYOMASU, T.; SASSA, T.; HASEGAWA, M.; KODAMA, O.; SHIBUYA, N.; KOGA, J.; NOJIRI, H.; YAMANE, H. Identification of a biosynthetic gene cluster in rice for momilactones. **J. Biol. Chem**, 282:34013–34018, 2007.

SILVESTRINI S.; JUNQUEIRA M.G.; FAVARIN A.C.; GUERREIRO-FILHO O.; MALUF M.P.; SILVAROLLA M.B.; COLOMBO C.A. Genetic diversity and structure of Ethiopian, Yemen and Brazilian Coffea arabica L. accessions using microsatellites markers. **Genetic Resources Crop Evolution**, 54, 6: 1367-1379, 2007.

SOOLE, K.L.; MENZ, R.I. Functional molecular aspects of the NADH dehydrogenases of plant mitochondria. **Journal of Bioenergetics and Biomembranes**, p. 27, n. 4, p. 397–406, 1995.

STEYN, W. J.; WAND, S. J.; JACOBS, G.; ROSECRANCE, R. C & ROBERTS, S. C. Evidence for a photoprotective function of low-temperature induced anthocyanin accumulation in apple and pear peel. **Physiol. Plant**. 136, 461–472, (2009).

SYRÉN, P. O.; HENCHE, S.; EICHLER, A.; NESTL, B. M & HAUER, B. Squalene-hopene cyclases-evolution, dynamics and catalytic scope. **Curr. Opin. Struct. Biol**, 41, 73–82, 2016.

TAMBA, C. L.; NI, Y. L & ZHANG, Y. M. Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. **PLoS. Comput. Biol**, 13, e1005357, 2017.

TISKI, I.; MARRACCINI, P.; POT, D.; VIEIRA, L.G.E.; PEREIRA, L.F.P. Characterization and expression of two cDNA encoding 3-hydroxy-3-methylglutaryl coenzyme A reductase isoforms in coffee (*Coffea arabica* L.). **OMICS J. Integr. Biol**, 15(10):719–27, 2011.

TRAN, H. T. M. et al. Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.). **Tree. Genet. Genom**, 13, 54, 2017.

TOLEDO-ORTIZ, G.; HUQ, E AND QUAIL, P.H. The Arabidopsis basic/helixloop-helix transcription factor family. **Plant Cell**, 15, 1749–1770, 2003.

TURNER, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. **bioRxiv**, 005165, 2014.

VITÓRIA, A.P. AND P. MAZZAFERA. Xanthine degradation and related enzymes activities in leaves and fruits of two Coffea species differing in caffeine catabolism. **J. Agric. Food. Chem**, 47, 5, 1851-1855 (1999).

WANG, Q.; HILLWIG, M.L.; WU, Y.; PETERS, R.J. CYP701A8: a rice entkaurene oxidase paralog diverted to more specialized diterpenoid metabolism. **Plant. Physiol**, 158:1418–25, 2012.

WANG, J.Y.; HU, Z.Z.; ZHAO, T.M.; YANG, Y.W.; CHEN, T.Z.; YANG, M.L et al. Genome-wide analysis of bHLH transcription factor and involvement in the infection by yellow leaf curl virus in tomato (*Solanum lycopersicum*). **BMC. Genomics**, 16: 39, 2015.

WANG, S. B. et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. **Sci. Rep**, 6, 19444, 2016.

WEI, Y.; LIN, M.; OLIVER, D.J.; SCHNABLE, O.S. The roles of aldehyde dehydrogenases (ALDHs) in the PDH bypass of Arabidopsis. **BMC. Biochem**, Mar 25; 10:7, 2009.

WEN, W.; LIU, H.; ZHOU, Y.; JIN, M.; YANG, N.; LI, D.; LUO, J.; XIAO, Y.; PAN, Q.; TOHGE, T., et al. Combining quantitative genetics approaches with regulatory network analysis to dissect the complex metabolism of the maize kernel. **Plant Physiol**, 170:136–146, 2016.

WIESNER, M.; SCHREINER, M & ZRENNER, R. Functional identification of genes responsible for the biosynthesis of 1-methoxy-indol-3-ylmethyl-glucosinolate in *Brassica rapa* ssp. chinensis. **BMC Plant Biol**, 14, 124, 2014.

WRIESSNEGGER.; AUGUSTIN.; ENGLEDER.; LEITNER.; MULLER.;KALUZNA. ; SCHURMANN.; MINK.; ZELLNIG.; SCHWAB.; PICHLER. **Production of the sesquiterpenoid (+)- nootkatone by metabolic engineering of *Pichia pastoris***. Metab. Eng., 24C, p. 18-29, 2014.

YAMAMURA, C.; MIZUTANI, E.; OKADA, K et al. Diterpenoid phytoalexin factor, a bHLH transcription factor, plays a central role in the biosynthesis of diterpenoid phytoalexins in rice. **Plant J**, 84 (6), 2015.

YANG, N.; LU, Y.L.; YANG, X.H.; HUANG, J.; ZHOU, Y.; ALI, FH.; WEN, W.W.; LIU, J.; LI, J.S AND YAN JB. Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. **PLoS Genet,** 10, 2014.

YI, NXU, S. Bayesian LASSO for quantitative trait loci mapping. **Genetics**. 179(2): 1045–55, 2008.

YU, J.M.; PRESSOIR, G.; BRIGGS, W.H.; BI, I.V.; YAMASAKI, M, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, 38: 203–208, 2006.

VIDAL, R.O.; MONDEGO, J.M.C.; POT, D.; AMBRÓSIO, A.B.; ANDRADE, A.C.; PEREIRA, L.F.P.; COLOMBO, C.A.; VIEIRA, L.G.E.; CARAZZOLLE, M.F.; PEREIRA, G.A.G. 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 Physiology**, 154(3), 1053-1066, 2010.

VITÓRIA, A.P AND P. MAZZAFERA. Xanthine degradation and related enzymes activities in leaves and fruits of two Coffea species differing in caffeine catabolism. **J. Agric. Food. Chem**, 47, 5, 1851-1855, 1999.

YOUNG, A.; BOYLE, T & BROWN, T. The population genetic consequences of habitat fragmentation for plants. **Trends Ecol. Evol**, 11, 413–418,1996.

ZHANG J.; LIU B.; LI M.; FENG D.; JIN H.; WANG P.; LIU J.; XIONG, F.; WANG J.; WANG, H.B. The *bHLH* transcription factor bHLH104 interacts with IAA-LEUCINE RESISTANT3 and modulates iron homeostasis in *Arabidopsis*. **Plant Cell**, 27: 787–805, 2015.

ZHU, C.; GORE, M.; BUCKLER, E.S.; YU, J. Status and Prospects of Association Mapping in Plants. **The Plant Genome**, 1:5-20, 2008.

26 MATERIAL SUPLEMENTAR

Suplementar 1. Lista dos genótipos do gênero *Coffea* enviados para a Genotipagem por Sequenciamento (GBS).

Genótipos	Nº na coleção	Genótipos	Nº na coleção	Genótipos	Nº na coleção
1	BA10_057	54	E209_031	107	E456_062
2	E007_087	55	E213_211	108	E457_477
3	E012_136	56	E218_581	109	E458_097
4	E016_298	57	E220_127	110	E464_417
5	E017_419	58	E221_214	111	E466_125
6	E018_494	59	E233_015	112	E467_045
7	E021_011	60	E237_071	113	E478_408
8	E022_163	61	E238_022	114	E481_238
9	E025_308	62	E254_284	115	E486_189
10	E030_075	63	E261_052	116	E490_516
11	E037_676	64	E265_101	117	E494_173
12	E038_043	65	E267_090	118	E505_140
13	E039_434	66	E268_067	119	E511_157
14	E041_079	67	E270_044	120	E514_129
15	E044_122	68	E272_143	121	E516_069
16	E046_021	69	E279_618	122	E534_036
17	E047_267	70	E283_096	123	E546_118
18	E055_005	71	E287_029	124	E552_323
19	E057_497	72	E298_382	125	E565_010
20	E061_126	73	E301_111	126	E571_072
21	E068_014	74	E302_083	127	E621_139
22	E071_258	75	E308_049	128	M7846_67
23	E080_584	76	E315_081	129	SEL_106
24	E081_041	77	E320_145	130	Catuai_V26
25	E085_085	78	E324_093	131	M_Novo38
26	E087_194	79	E325_522	132	Typica
27	E089_391	80	E326_124	133	Bourbon
28	E114_447	81	E327_032	134	IAPAR59
29	E116_061	82	E331_280	135	L1C1
30	E118_213	83	E332_023	136	L3C3
31	E123a_231	84	E333_104	137	*C. canephora*
32	E123b_121	85	E335_219	138	*C. eugenioides*
33	E124_245	86	E338_218	139	Java1
34	E126_359	87	E340_179	140	Java6
35	E130_169	88	E344_008	141	Java15
36	E131_018	89	E351_248	142	Java26
37	E146_012	90	E363_735	143	Et34_1
38	E148_254	91	E364_059	144	Et34_2
39	E152_017	92	E368_600	145	Et34_5
40	E159_180	93	E370_196	146	Et34_6
41	E164_417	94	E383_142	147	Et34_8
42	E169_180	95	E386_131	148	Et34_12
43	E174_164	96	E389_133	149	Et34_14
44	E179_650	97	E401_643	150	Et34_17

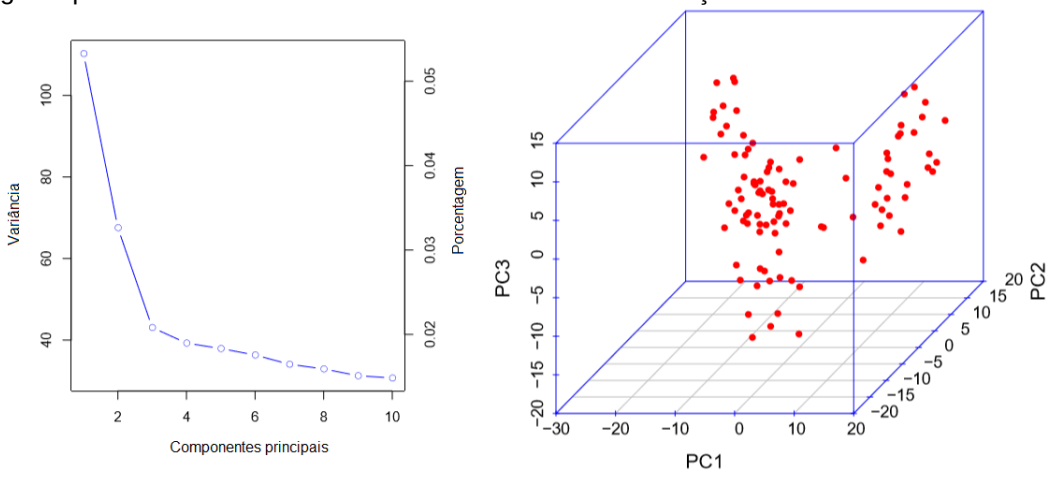
45	E180_070	98	E404_135	151	Et19_14
46	E181_358	99	E408_001	152	IPR99
47	E183_138	100	E409_114	153	IPR100
48	E189_119	101	E419_098	154	IPR101
49	E190_013	102	E428_109	155	IPR102
50	E196_117	103	E439_094	156	IPR103
51	E199_011	104	E442_279	157	IPR104
52	E201_134	105	E450_235	158	IPR105
53	E208_193	106	E454_107	159	IPR107

Suplementar 2. Amostragem dos 101 genótipos da coleção de *C. arabica* do IAPAR utilizados nos estudos de associação (GWAS).

N° de genótipos	Genótipos	Origem dos Acessos	N° de genótipos	Genótipos	Origem dos Acessos
1	Mundo Novo	Cultivar	52	E261	O
2	Typica	Cultivar	53	E265	O
3	Bourbon	Cultivar	54	E267	O
4	Catuaí Vermelho	Cultivar	55	E268	O
5	M7846	O	56	E270	O
6	E007	L	57	E272	O
7	E012	L	58	E279	O
8	E017	L	59	E283	O
9	E018	L	60	E301	O
10	E021	L	61	E302	O
11	E022	L	62	E315	O
12	E025	O	63	E320	O
13	E037	L	64	E325	O
14	E038	O	65	E326	O
15	E041	O	66	E327	O
16	E044	O	67	E331	O
17	E047	O	68	E332	O
18	E055	O	69	E333	O
19	E057	O	70	E338	O
20	E061	O	71	E340	O
21	E071	O	72	E344	O
22	E080	O	73	E364	O
23	E081	O	74	E368	O
24	E085	O	75	E386	O
25	E089	O	76	E401	O
26	E114	O	77	E408	O
27	E116	O	78	E409	O
28	E123a	O	79	E428	O
29	E123b	O	80	E439	O
30	E124	O	81	E442	O
31	E131	O	82	E450	O
32	E146	O	83	E454	O
33	E148	O	84	E456	O

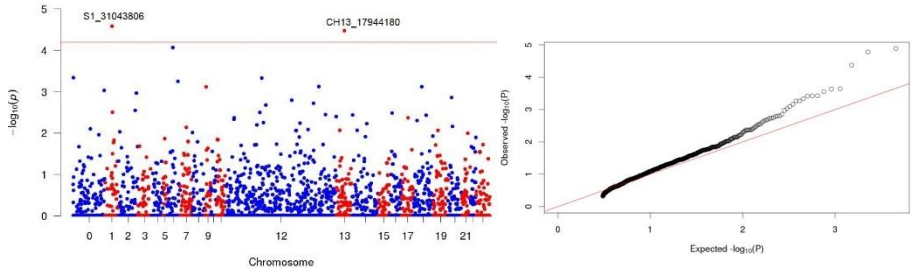
34	E152	O	85	E457	O
35	E159	O	86	E458	O
36	E174	O	87	E464	O
37	E179	O	88	E466	O
38	E183	O	89	E467	O
39	E189	O	90	E478	O
40	E190	O	91	E481	O
41	E199	O	92	E486	O
42	E201	O	93	E490	O
43	E208	O	94	E494	O
44	E209	O	95	E505	O
45	E213	O	96	E511	O
46	E218	O	97	E514	O
47	E220	O	98	E552	O
48	E221	O	99	E565	O
49	E233	O	100	E571	O
50	E237	L	101	E621	O
51	E254	O			

Suplementar 3. Gráfico de dispersão tridimensional (CP1, CP2 e CP3) dos 101 genótipos e 4.517 SNPs utilizados nas análises de associação.

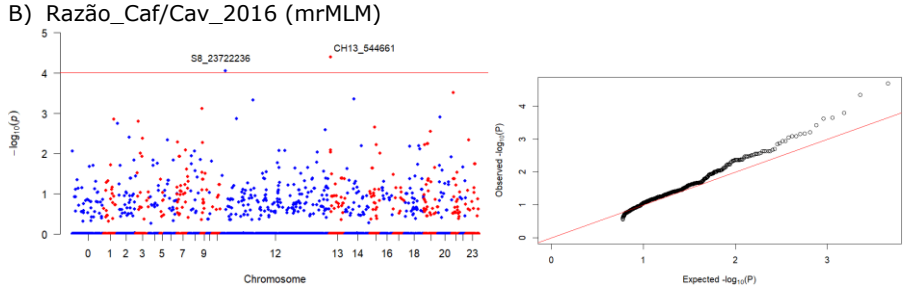


Suplementar 4. Manhattan plots, Q-Qplots e LOD *score* obtidos pelos métodos mrMLM e ISIS-EM-BLASSO para todos os SNPs significativamente associados aos compostos relacionados à qualidade da bebida de café e co-localizados aos genes da tabela 2. Linha vermelha no *Manhattan plot* indica o limiar dos valores de *P*; linha vermelha do Q-Q plot representa a distribuição nula esperada dos valores de *P* e os pontos representam a distribuição observada dos valores de *P.*

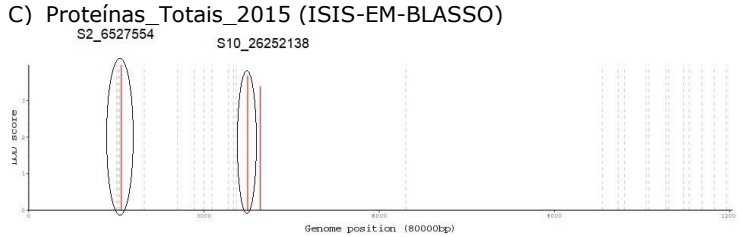
A) Cafeína_2015 (mrMLM)



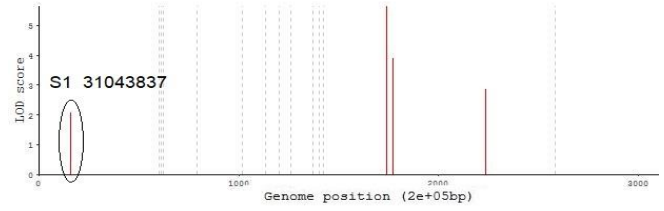
B) Razão_Caf/Cav_2016 (mrMLM)



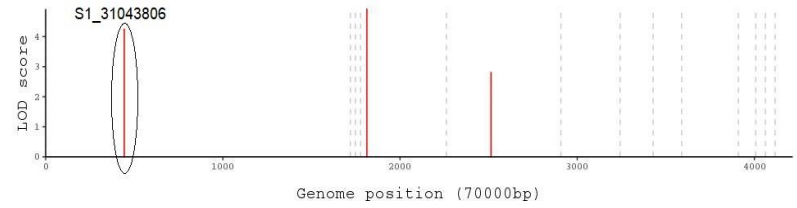
C) Proteínas_Totais_2015 (ISIS-EM-BLASSO)



D) Proteínas_Totais_2012 (ISIS-EM-BLASSO)



E) Proteínas_Totais_2016 (ISIS-EM-BLASSO)



27 CONCLUSÕES GERAIS

Na primeira parte desse trabalho, por meio dos SSRs foi possível observar a estruturação do painel em 2 grupos. Esse resultado foi similar aos resultados obtidos do capitulo 3, onde também foi possivel diferenciar as cultivares e variedades dos acessos do Grande Vale do Rift. No entanto, o estudo de estrutura com maior número de SNPs e de genótipos permitiu uma melhor separação (K=3) dos materiais, ressaltanto um grupo adicional composto por genótipos de maior variabilidade, no qual a maioria são de parques de reservas florestais e de florestas.

Nossos resultados baseados tanto nos SSRs como nos SNPs podem nos ajudar a definir quais acessos do lado Oeste do Vale do Rift são mais importantes para se preservar com boa representação genética da coleção da FAO, visando estudos de seleção de genitores contrastantes.

As análises de estrutura realizadas posteriormente demonstraram que a espécie *C. eugenioides* possui uma posição mais próxima dos materiais cultivados e dos genótipos da região Leste do Vale do Rift do que *C. canephora*. Essas análises corroboram com os resultados obtidos no capítulo 1 e reforçam a estreita prossimidade desse parental diploide de *C. arabica* com os materiais melhorados.

As análises com dados fenotípicos para a safra de 2016 do capítulo 2 corroboram com os resultados genotípicos sobre a variabilidade existente nos materiais selvagens da Etiópia. Com esses resultados é possível selecionar potenciais genótipos para produção de uma bebida com qualidade diferenciada e com benefícios à saúde. Nesse trabalho se destaca os grupos 1 e 3 obtidos pela AAH onde se encontram genótipos com altas concentrações de Caveol, Lipídeos Totais, Proteínas Totais, Sacarose e Açúcares Totais, bem como baixos de Cafeína, Cafestol e ACGs.

Nossos resultados demonstram que utilizando um filtro mais restritivo que o de SANT'ANA et al., 2018 após a realização de um novo *Pipeline* TASSEL (He < 0.5), foi possível detectar SNPs significativamente associados e genes candidatos às vias metabólicas dos compostos químicos analisados. O filtro mais restritivo foi selecionado após acompanhamento por heterozigosidade e análise de estrutura e visou evitar o máximo possível da inflação de falsos

positivos (erro tipo I). Entretanto essa abordagem eliminou SNPs associados anteriormente (SANT'ANA et al., 2018). Por isso pode-se propor valores de heterozigosidade intermediários visando identificar o mesmo nível de associação dos dois trabalhos.