PRISCILA MARY YUYAMA

CARACTERIZAÇÃO DO TRANSCRIPTOMA DE FOLHAS E FRUTOS DE *Coffea eugenioides* E IDENTIFICAÇÃO DE POLIMORFISMOS DE ACESSOS DE *Coffea arabica* DA ETIÓPIA

Londrina
2014

Universidade Estadual de Londrina

Instituto Agronômico do Paraná

Brasileira de Pesquisa Agropecuária

PRISCILA MARY YUYAMA

CARACTERIZAÇÃO DO TRANSCRIPTOMA DE FOLHAS E FRUTOS DE *Coffea eugenioides* E IDENTIFICAÇÃO DE POLIMORFISMOS DE ACESSOS DE *Coffea arabica* DA ETIÓPIA

Londrina
2014

PRISCILA MARY YUYAMA

CARACTERIZAÇÃO DO TRANSCRIPTOMA DE FOLHAS E FRUTOS DE *Coffea eugenioides* E IDENTIFICAÇÃO DE POLIMORFISMOS DE ACESSOS DE *Coffea arabica* DA ETIÓPIA

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

Orientador: Prof. Dr. Luiz Filipe Protasio Pereira

Londrina
2014

Catalogação elaborada pela Divisão de Processos Técnicos da Biblioteca Central da Universidade Estadual de Londrina.

Dados Internacionais de Catalogação-na-Publicação (CIP)

Y95c Yuyama, Priscila Mary.
Caracterização do transcriptoma de folhas e frutos de *Coffea eugenioides* e identificação de polimorfismos de acessos de *Coffea arabica* da Etiópia / Priscila Mary Yuyama. – Londrina, 2014.
89 f. : il.

Orientador: Luiz Filipe Protasio Pereira.
Tese (Doutorado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2014.
Inclui bibliografia.

1. Café – Melhoramento genético – Teses. 2. Sequência de nucleotídeos – Teses. 3. Polimorfismo (Genética) – Teses. 4. Marcadores biológicos – Teses. 5. Biotecnologia vegetal – Teses. I. Pereira, Luiz Filipe Protasio. II. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. III. EMBRAPA. IV. Instituto Agronômico do Paraná. V. Título.

CDU 631.52:633.73

PRISCILA MARY YUYAMA

CARACTERIZAÇÃO DO TRANSCRIPTOMA DE FOLHAS E FRUTOS DE *Coffea eugenioides* E IDENTIFICAÇÃO DE POLIMORFISMOS DE ACESSOS DE*Coffea arabica* DA ETIÓPIA

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

BANCA EXAMINADORA

_____
Prof. Dr. Luiz Filipe Protasio Pereira
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Café

_____
Profa. Dra. Eveline Teixeira Caixeta
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Café

_____
Prof. Dr. Jorge Mauricio Costa Mondego
Instituto Agronômico de Campinas – IAC

_____
Prof. Dr. Luiz Gonzaga Esteves Vieira
Universidade do Oeste Paulista – UNOESTE

_____
Prof. Dr. Ricardo Vilela Abdelnoor
Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA Soja

Londrina, 21 de março de 2014.

À minha Família dedico.

AGRADECIMENTOS

A Universidade Estadual de Londrina - UEL, ao curso de Pós Graduação em Genética e Biologia Molecular e seus professores pela oportunidade e aprendizado durante esses quatro anos. Agradeço também a coordenadora do Programa, Dra. Ana Lúcia Dias e a secretária Sueli.

Ao Instituto Agronômico do Paraná e a Universidade Estadual de Londrina, pelo acolhimento e estrutura disponibilizada durante a minha tese.

Ao CIRAD por ter me proporcionado uma boa estrutura e organização na realização deste trabalho e pela experiência profissional proporcionada na minha vida acadêmica.

A CAPES e CAPES/Fundação Agrópolis pelas bolsas concedidas, no Brasil e na França, respectivamente.

A minha Família, pelo amor e carinho, pela confiança, paciência, compreensão e por toda dedicação e esforço investidos na minha formação. A minha admiração, respeito e agradecimento especial a essas pessoas tão importantes na minha vida. Amo vocês!

Ao meu orientador Dr. Luiz Filipe Protasio Pereira, pela orientação, parceria, pelas oportunidades oferecidas, pela confiança e pelo exemplo de profissional e pessoa.

Ao meu co-orientador, Dr. Thierry Leroy, pela orientação, confiança e por todo suporte oferecido durante o meu doutorado sanduíche.

À Pierre Charmetant e a Família Charmetant, por toda ajuda e carinho, em Londrina e em Montpellier.

Ao Dr. Douglas Silva Domingues, pelo incentivo e contribuições científicas para a realização deste trabalho.

Ao Dr. David Pot por toda ajuda proporcionada em Montpellier e na realização deste trabalho.

A Dra. Eveline Teixeira Caixeta, Dr. Jorge Mauricio Costa Mondego e Dr. Ricardo Vilela Abdelnoor por terem aceitado o convite para participar da banca examinadora e apreciação do trabalho.

Ao Dr. Luiz Gonzaga Esteves Vieira, por ter aceitado inicialmente a orientação no doutorado e por ter aceitado o convite na banca examinadora.

Ao Dr. Gonçalo Amarante Guimarães Pereira, Marcelo Carazzolle, Osvaldo e Jaime por terem ajudado na realização deste trabalho.

Ao João Batista Gonçalves Dias da Silva, Coordenador do Centro Tecnológico da COCARI, por ter cedido às plantas de *Coffea eugenioides*.

Ao Arthur e Fernando por terem ajudado nas coletas das plantas da coleção da Etiópia.

A todo o pessoal do LBI, pelo companheirismo, amizade, pelos momentos prazerosos e pelo aprendizado: Tiago, Renata, Karina (Kah), Dudu, Lucinéia, Cícera, Sueli, Giselly, Gislaine, Rafinha, João, Mari, Camilla, Julia, Carol, Kenia, Andrea, Rafael, Juliana, Bruna, Vivi, Yumi, Lívia, Sandra, Alessandra, Paloma, Leo, Dr. Juarez Pires Tomaz e Dr. Eduardo Fermino.

A Suzana (Suh), pela contribuição nos trabalhos e pela amizade.

À toda a equipe de bioinformática do CIRAD, pela ajuda no trabalho, por todo suporte, gentileza, paciência, diversão e carinho proporcionado durante o meu doutorado sanduíche. Parabenizo ao Manuel Ruiz e Stéphanie Sidibe-Bocs pela organização e competência da equipe. Agradeço ao Fred de Lamotte, Jean-François, Gautier, Bertrand e Dominique This. Em especial, agradeço aos meus amigos de “bureau” Felix, Guilhem e Marilyne. Agradeço pela amizade, ajuda, por me “suportarem” e por me entenderem muitas vezes só com os olhos.

Especialmente também agradeço ao Alexis Dereeper e Stéphanie Pointet pela ajuda no trabalho.

Aos meus queridos amigos brasileiros na França: Gabriel, Michelle, Maíra, Isabela, Brígida, Iara e Bia. Obrigada pela amizade e por ter aprendido “sobre a vida” com vocês.

Aos meus queridos amigos. Agradeço de coração por tudo que fizeram e fazem por mim.

Ao Dr. André Luís Laforga Vanzela, meu orientador da graduação e do mestrado.

Aos meus amigos e colegas do Laboratório de Citogenética Vegetal, na UEL.

Agradeço a tantas pessoas que me ajudaram nessa trajetória e que foram essenciais com uma ajuda, uma palavra de conforto, um gesto de compreensão ou que simplesmente torceram por mim.

Agradeço a Deus pela oportunidade de ter conhecido lugares incríveis e pessoas maravilhosas. Obrigada por ter me protegido, por ter me proporcionado tantos momentos de felicidade, de aventura, de aprendizado e pela oportunidade de sempre evoluir como profissional e como pessoa.

Muito obrigada!

Out here the nights are long, the days are lonely
I think of you and I'm working on a dream
I'm working on a dream

Now the cards I've drawn's a rough hand, darling
I straighten the back and I'm working on a dream
I'm working on a dream

I'm working on a dream
Though sometimes it feels so far away
I'm working on a dream
And I know it will be mine someday

Working on a dream - Bruce Springsteen

YUYAMA, Priscila Mary. **Caracterização do Transcriptoma de Folhas e Frutos de *Coffea eugenioides* e Identificação de Polimorfismos de Genótipos de *Coffea arabica* da Etiópia.** 2014. 89 f. Tese (Doutorado em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Londrina, 2014.

RESUMO

O café é uma das principais *commodities* agrícolas do mundo e o Brasil ocupa a posição de maior produtor e exportador de café mundial. *Coffea arabica* e *C. canephora* respondem pela maior parte dessa produção. As espécies do gênero são diplóides, exceto *C. arabica*, um alotetraplóide formado de uma recente hibridação de duas espécies diplóides, *C. canephora* e *C. eugenioides*. *C. arabica* apresenta uma estreita base genética, principalmente pelas suas características biológicas, recente história evolutiva e origem das espécies cultivadas. O RNA-Seq tem sido utilizado em trabalhos de anotação e identificação de polimorfismos de nucleotídeo único (SNP) em plantas, com obtenção de um grande volume de dados e resultados robustos. Neste trabalho, foram obtidos 98.635.514 *reads* em *Coffea* a partir da tecnologia de sequenciamento Illumina. As sequências foram obtidas de folhas e frutos a partir de genótipos selvagens de *C. arabica* originários da Etiópia, *C. arabica* cv. Mundo Novo e *C. eugenioides*. No primeiro trabalho, foi feita a caracterização do transcriptoma de *C. eugenioides* a partir de 36.935 contigs, obtidos de uma montagem *de novo*. As sequências foram anotadas baseadas em bancos de dados de proteínas não-redundantes (nr) do GenBank, Swiss-Prot, Gene Ontology (GO), InterproScan, PlantCyc e KEGG. Além disso, 10 contigs com maior expressão em órgãos de folha e fruto de *C. eugenioides* foram selecionados para validar a sua expressão por qPCR. O segundo trabalho desenvolveu análises de RNA-Seq de todos os genótipos sequenciados. Foi possível identificar 1.410 SNPs potenciais em cinco genótipos de *C. arabica* a partir de uma referência *de novo* de *C. canephora*. Um total de 311 SNPs foram validados em 128 genótipos selvagens de *C. arabica* e cinco cultivares de *C. arabica* através do Sequenom MassARRAY. A análise da estrutura da coleção com o programa *Strucutre* demonstrou quatro sub-populações. Assim, foi desenvolvido um atlas do transcriptoma de *C. eugenioides* como potencial referência para estudos futuros em *Coffea* e foram obtidos um grupo de SNPs validados. Esses resultados podem beneficiar o desenvolvimento de estudos de associação genética e auxiliar nos trabalhos de melhoramento de cafeeiros.

Palavras-chaves: Anotação. Café. Genotipagem. qPCR. SNPs.

YUYAMA, Priscila Mary. **Transcriptome Characterization of Leaves and Fruits of *Coffea eugenioides* and Polymorphisms Identification in Genotypes of *Coffea arabica* from Ethiopia.** 2014. 89 p. Tese (Doutorado em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Londrina, 2014.

ABSTRACT

Coffee is one of the most important agricultural commodities worldwide and Brazil stands out as the main coffee productor and exporter. *Coffea arabica and C. canephora* account with the most part of this production. The genus species are diploid, except *C. arabica* which is an allotetraploid from a recent hybridization of two diploid species or related species, *C. canephora* and *C. eugenioides*. *C. arabica* presents a narrow genetic diversity, mainly due its biological characteristics, recent evolutionary history and origin of cultivated genotypes. RNA-seq has been done in several works of annotation and SNPs identification in plants, with the production of large volume of data and robust results. In this work, we report the generation of a total of 98,635,514 reads in *Coffea* using Illumina sequencing. Sequences were obtained from leaves and fruits using wild *C. arabica* genotypes from Ethiopia, *C. arabica* cv. Mundo Novo and *C. eugenioides*. The *C. eugenioides* transcriptome was characterized from 36,935 contigs, obtained of a *de novo* assembled. Sequences were successfully annotated based on the Genbank non-redundant (Nr), Swiss-Prot, Gene Ontology (GO), InterproScan, PlantCyc and KEGG protein database. Furthermore, 10 highly expressed contigs from leaf and fruit were selected to confirm their expression by qPCR. Second work developed RNA-seq analysis of all genotypes sequenced and it was possible discovered 1,410 potential SNPs in five *C. arabica* genotypes using a *C. canephora* reference *de novo* assembled. They were validated 311 SNPs in 128 wild genotypes and five cultivars of *C. arabica* on the Sequenom MassARRAY system. Structure analysis of collection demonstrated four sub-populations. Thus, we present an overview of *C. eugenioides* transcriptome as a potential reference for future studies in *Coffea* and we obtained a set of SNPs to genotyping. These results may benefit the development of genetic association and support future studies in coffee breeding

Keywords: Annotation. Coffee. Genotyping. qPCR. SNPs.

SUMÁRIO

1 INTRODUÇÃO

O café é uma das principais *commodities* agrícolas do mundo. A produção mundial estimada no período de 2013/2014 é de 146,3 milhões de sacas (60 Kg) e o Brasil ocupa a posição de maior produtor e exportador mundial de café (CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO, 2013). Para assegurar a produtividade e potencializar o poder competitivo do país na produção e comercialização do café, os programas de pesquisa e melhoramento buscam soluções para problemas que possam prejudicar a cafeicultura, principalmente relacionados a estresses bióticos e abióticos, além de investir no desenvolvimento de cultivares com qualidade diferencial da bebida a fim de conquistar mercados mais exigentes (LEROY et al. 2006; DOS SANTOS et al. 2011; FERNANDEZ et al. 2012; MARRACCINI et al. 2012).

Duas espécies do gênero *Coffea* são responsáveis por quase toda a produção de café: *Coffea arabica* L. e *C. canephora* Pierre ex A. Froehner. *C. arabica* é uma espécie autógama e alotetraplóide ($2n = 4\times = 44$), ou seja, é uma espécie que combina dois conjuntos de genomas distintos (cromossomos homeólogos), formado a partir de um evento de hibridação de genomas divergentes e duplicação genômica. A espécie apresenta uma origem recente, de 100 a 500 mil anos (ANTHONY et al. 2010), embora exista uma estimativa mais recente de aproximadamente 10 a 50 mil anos (CENCI et al. 2012). Essa hibridação ocorreu entre duas espécies diplóides alógamas, *C. canephora* e *C. eugenioides* S. Moore (LASHERMES et al. 1999).

C. arabica responde por aproximadamente 65% da produção mundial, mas é uma espécie que apresenta baixa diversidade genética em consequência da sua origem, biologia reprodutiva e evolução. Essas características, associadas ao fato de ser uma espécie perene, são fatores que dificultam o desenvolvimento de novas cultivares (LOS SANTOS-BRIONES & HERNÁNDEZ-SOTOMAYOR, 2006). Nesse sentido, diversas ferramentas biotecnológicas podem ser empregadas com o objetivo de minimizar essas dificuldades, como a cultura de tecidos, a transformação genética e o uso de marcadores moleculares na seleção assistida (CAIXETA et al. 2008).

Os marcadores moleculares são uma importante ferramenta biotecnológica, porque podem possibilitar o acesso à variabilidade genética e a

escolha precoce e de forma mais precisa dos genótipos de interesse, o que permite um melhor gerenciamento de tempo e recursos para o melhoramento vegetal (FERREIRA & GRATTAPAGLIA, 1998). Recentemente, os estudos de associação genômica ampla (*Genome Wide Association Studies* – GWS) e seleção genômica ampla (*Genome Wide Association Studies* - GWAS) tornaram-se ferramentas importantes no melhoramento genético de plantas. Eles superam várias limitações do mapeamento tradicional de genes por fornecerem alta resolução e em populações bem estudadas podem prever e associar as variações genéticas com as variações fenotípicas (BRACHI et al. 2011).

A classe de marcadores com maior abundância de modificações nas sequências genômicas dos organismos são os SNPs (Polimorfismo de Nucleotídeo Único) (RAFALSKI 2002). Esses marcadores apresentam alto potencial informativo pela maior saturação do genoma. A procura desses polimorfismos no transcriptoma de *C. arabica* permitiu a identificação dos polimorfismos que diferenciavam os subgenomas de *C. arabica* – subgenoma *C. canephora* (CaCc) e subgenoma *C. eugenioides* (CaCe), mas não foram potenciais para identificar polimorfismos entre as cultivares de *C. arabica* (VIDAL et al. 2010).

Apesar da importância no entendimento da expressão de genes nos subgenomas dos ancestrais de *C. arabica*, existem poucos estudos desenvolvidos em *C. eugenioides*, uma vez que a maioria dos trabalhos foram realizados nas duas principais espécies de interesse comercial, *C. arabica* e *C. canephora*. Nesse sentido, a geração de um grande número de sequências a partir de dados de sequenciamento de nova geração (NGS) podem permitir estudos globais sobre o transcriptoma de *C. eugenioides* e compreender a expressão de determinados genes nessa espécie. Nos últimos anos houve um grande aumento nos estudos com essas novas tecnologias de sequenciamento com eficiência e custo efetivo (METZER et al. 2010) e que vem permitindo a anotação de dados genômicos e transcriptômicos de plantas assim como a identificação de polimorfismos (NOVAES et al. 2008; BLANCA et al. 2011).

Neste estudo, a partir de dados de RNA-Seq, foram desenvolvidos dois trabalhos. Primeiro, foi possível obter uma análise global do transcriptoma de *C. eugenioides* e analisar os genes mais expressos em órgãos de folha e fruto. Estes dados revelaram possíveis genes que poderiam contribuir para certas características de *C. arabica*. Alguns genes de *C. eugenioides* foram validados por qPCR e podem

ser futuros genes candidatos em estudos genéticos de *Coffea* bem como auxiliar na compreensão da expressão dos genes presentes nos subgenomas de *C. arabica*.

No segundo trabalho, a análise da diversidade nucleotídica a partir de dados de NGS em um painel mais diverso de genótipos permitiu a identificação de vários polimorfismos inter e intraespecífico e mostram a importância de utilizar um grupo mais diverso de genótipos associado com o número abundante de dados obtidos com essas novas tecnologias.

2 OBJETIVOS

- Caracterizar o transcriptoma de folhas e frutos de *C. eugenioides*;

- Identificar polimorfismos em *C. arabica* com potencial para genotipagem baseados em dados de transcriptoma de acessos selvagens de *C. arabica*, e dos seus ancestrais, *C. canephora* e *C. eugenioides*.

3 REVISÃO DE LITERATURA

3.1 IMPORTÂNCIA ECONÔMICA

O café é considerado uma das principais *commodities* agrícolas do mundo. A produção mundial (2013/2014) é estimada em 146,3 milhões de sacas (60 Kg) e o Brasil ocupa a posição de maior produtor e exportador de café mundial e responde por aproximadamente 35% do mercado (CONAB, 2013). A produção mundial de café apresentou nos últimos 12 anos um aumento médio de 1,02% ao ano. O consumo mundial registra um aumento médio de 1,65%, o que demonstra um descompasso no balanço da demanda de café para consumo com relação à produção (CONAB, 2013). *Coffea arabica* (Arabica) e *C. canephora* (Robusta ou Conilon) respondem pela maior parte da produção mundial, aproximadamente 65% e 35 % do mercado mundial, respectivamente (ICO - International Coffee Organization, 2014).

O café destaca-se no Brasil econômica e socialmente desde a chegada das primeiras mudas vindas da Guiana Francesa, em meados do século XVIII. Diante de sua rápida adaptação ao ambiente e ao clima, o produto adquiriu importância no mercado, e transformou-se em um dos principais itens de exportação, desde o Império até os dias atuais. A princípio restrita aos Estados do Pará e do Maranhão, a produção de café expandiu-se e atualmente são quinze Estados produtores, com destaque para Minas Gerais, Espírito Santo, São Paulo, Bahia, Paraná e Rondônia (MAPA - Ministério da Agricultura, Pecuária e Abastecimento, 2013). Apesar do aumento da sua produção nos últimos 15 anos, as áreas produtivas de café no Brasil apresentaram uma redução de 6,29% no período de 2007 a 2013 (CONAB, 2013).

3.2 CARACTERÍSTICAS DE C. ARABICA, C. CANEPHORA E C. EUGENIOIDES

C. arabica é uma espécie alotetraplóide e autógama ($2n = 4\times = 44$) de um recente cruzamento entre os ancestrais *C. canephora* e *C. eugenioides* a 100-500 mil anos (ANTHONY et al. 2010) embora exista uma estimativa mais recente, de aproximadamente 10-50 mil anos (CENCI et al. 2012). Essa hibridação ocorreu entre duas espécies diplóides, *C. canephora* ($2n = 2\times = 22$) e *C. eugenioides*

(2*n* = 2× = 22) (LASHERMES et al. 1999). As três espécies de *Coffea* apresentam variabilidade fenotípica e são melhores adaptadas em condições específicas.

C. arabica cresce em ambientes com variações de amplitude térmica de 20 a 24 °C e apresenta melhor qualidade da bebida em comparação a *C. canephora* (DAMATTA & RAMALHO, 2006; LEROY et al. 2006; PRIVAT et al. 2008), além de apresentar baixa diversidade genética comparada a outras espécies de *Coffea* (VIDAL et al. 2010).

C. canephora cresce melhor em regiões de baixas altitudes, com temperatura média de 22 a 26 °C e é caracterizada pela alta produtividade, tolerância a pragas e doenças, estresse a seca e conteúdo alto de cafeína. Porém, a sua bebida é considerada de qualidade inferior quando comparado a *C. arabica*. Assim, ela é mais utilizada em *blends* com *C. arabica* e na indústria de cafés solúveis (DAMATTA et al. 2006).

Outro ancestral de *C. arabica*, *C. eugenioides* cresce em regiões de altitudes mais elevadas, temperaturas entre 18 a 23 °C e é adaptada em regiões próximas as bordas de florestas, regiões secas. Ela não é produzida em escala comercial em função da baixa produtividade e tamanho pequeno dos grãos. No melhoramento, a espécie é utilizada para reduzir os níveis de cafeína e para melhorar a qualidade da bebida (MAZZAFERA & CARVALHO, 1991).

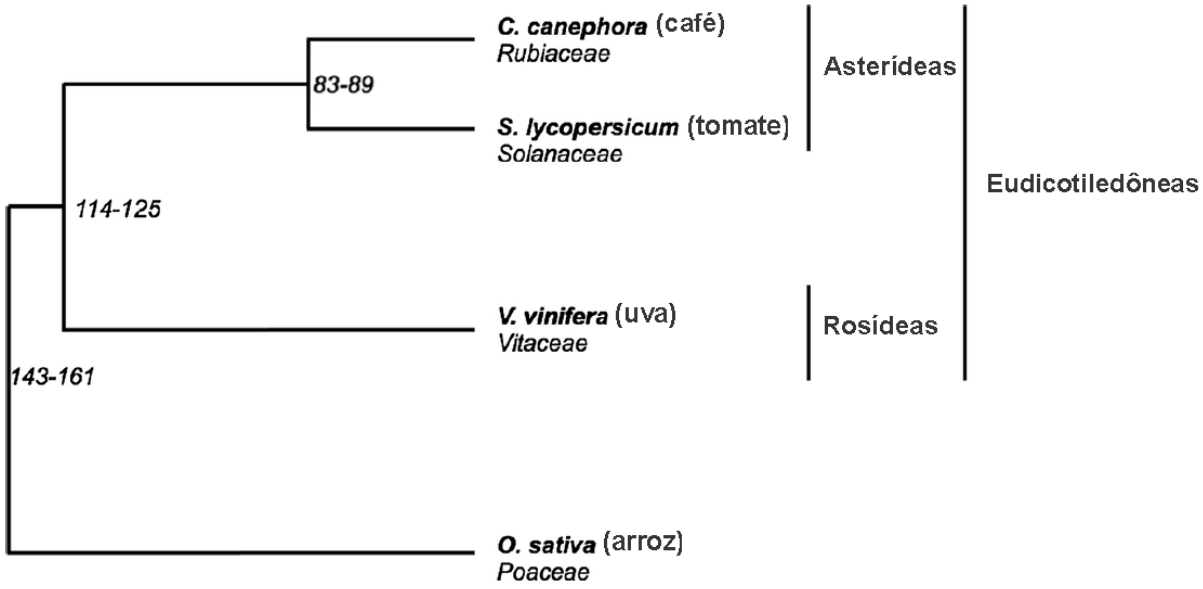
3.3 Aspectos Taxonômicos e Centro de Origem de *Coffea*

O cafeeiro pertence à divisão das Fanerógamas, classe Angiosperma, subclasse Eudicotiledônea, ordem Rubiales, família Rubiaceae, tribo Coffeeae, subtribo Coffeinae, gêneros *Coffea* e *Psilanthus* (GUERREIRO FILHO et al. 2008). O gênero *Coffea* apresenta 124 espécies, o qual inclui espécies do gênero *Psilanthus* sp. (DAVIS et al. 2011). Atualmente, as análises morfológicas e genéticas e a baixa diversidade genética suportam a inclusão de *Coffea* e *Psilanthus* em um único gênero (LOMBELLO & PINTO-MAGLIO, 2003, 2004; MAURIN et al. 2007; DAVIS et al. 2011). Todas as espécies de café compartilham a morfologia típica de grão de café, ou seja, um sulco no lado plano do grão (DAVIS et al. 2005).

A família com filogenia mais próxima do café (Rubiaceae) pertence à família Solanaceae. Esta família inclui espécies importantes economicamente, como o tomate, a batata, o pimentão, o tabaco, a petúnia e a berinjela. Ambas as famílias

pertencem ao clado Asteridae I das plantas dicotiledôneas e divergiram do seu ancestral comum a aproximadamente 83-89 milhões de anos (Figura 1). A família Solanaceae compartilha com o café uma grande similaridade em tamanho do genoma (NOIROT et al. 2003), número cromossômico básico (11 e 12 para café e tomate, respectivamente) e arquitetura cromossômica (PINTO-MAGLIO & DA CRUZ, 1998), ausência de poliploidização (WU et al. 2006) e expressão de genes no fruto e semente (LIN et al. 2005). Outra espécie utilizada para estudos comparativos com *Coffea* é *Vitis vinifera* L. membro da família Vitaceae no clado das Rosídeas (GUYOT et al. 2012). *V. vinifera* e *Coffea* divergiram a aproximadamente 114-125 milhões de anos (Figura 1), mas apresentam um alto nível de conservação genética (WIKSTROM et al. 2001; GUYOT et al. 2012).

Figura 1 – Relacionamento taxonômico entre *S. lycopersicum* L. (tomate, Solanaceae), *C. canephora* (café, Rubiaceae) e *V. vinifera* (uva, Vitaceae). A escala de tempo da divergência das famílias das angiospermas é indicada em milhões de anos (adaptado de Guyot et al. 2012).



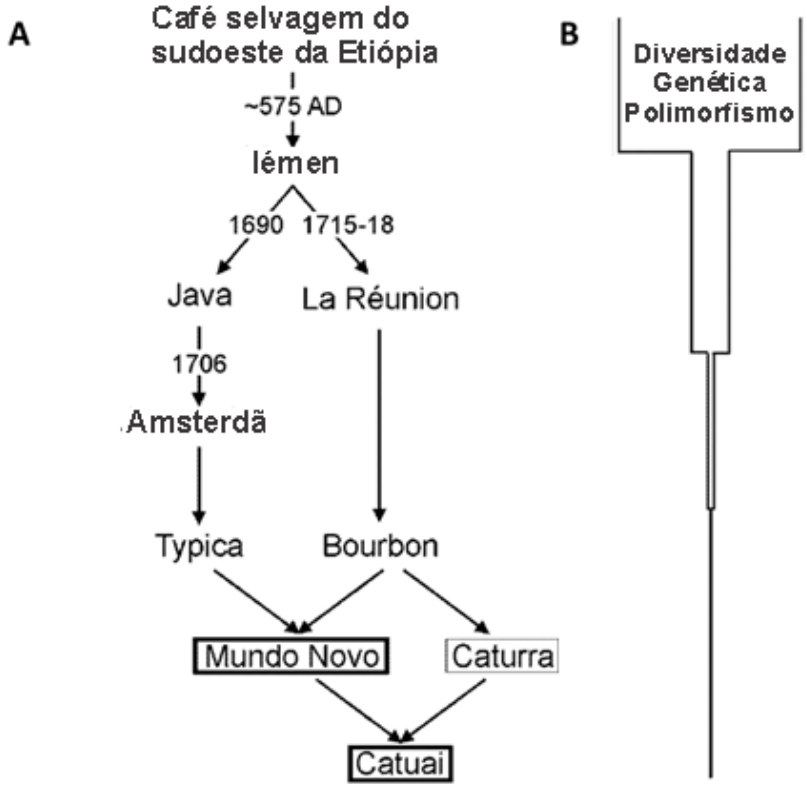
As espécies de *Coffea* têm como centro de origem as florestas intertropicais da África e das ilhas de Madagascar e Mascarenhas (LASHERMES et al. 1997; ANTHONY et al. 2010). A restrita diversidade genética e as análises filogenéticas sugerem um rápido modo de especiação de *Coffea* (ANTHONY et al. 2010). Atualmente, as espécies nativas de *Coffea* estão localizadas na África

tropical, Madagascar, Comores e Ilhas Mascarenhas (CHEVALIER 1947; BRIDSON & VERDCOURT, 1988) e com a inclusão de *Psilanthus* (DAVIS et al. 2011), ocorre também no sudoeste da Ásia (subcontinente indiano), no sul da Ásia tropical (Camboja, Myanmar, Tailândia, Vietnã), sudoeste da Ásia (Java, Pequenas Ilhas da Sonda, Filipinas, Papua-Nova Guiné) e Australasia (Australia: Queensland) (DAVIS et al. 2011).

3.4 *COFFEA ARABICA* E COLEÇÃO DA ETIÓPIA

No Brasil, a maior parte das cultivares de *C. arabica* derivaram basicamente de duas variedades botânicas: Typica e Bourbon (MENDES et al. 2008). Deste modo, a diversidade genética dos genótipos cultivados é pequena, o que gera restrições tanto para os programas de melhoramento, quanto para os trabalhos de busca de marcadores polimórficos para auxiliarem esses programas (ANTHONY et al. 2002; VIDAL et al. 2010) (Figura 2).

Figura 2 – A. Origem das cultivares modernas de *C. arabica*. B. Representação da diminuição da diversidade da espécie com a dispersão e desenvolvimento de cultivares. Fonte: Adaptado de Anthony et al. (2002) e Vidal et al. (2010)



No Brasil, a introdução de Typica se deu em 1727 e foi a única cultivar explorada comercialmente até meados do século XIX (MENDES et al. 2008). As primeiras lavouras de café tiveram origem das sementes de mudas de uma única planta Typica existente no Jardim Botânico de Amsterdã, na Holanda. Em 1859, a cultivar Bourbon foi trazida da Ilha de Reunião e introduzida no Brasil por considerarem de alta produtividade. Os principais genótipos cultivados de café, por exemplo, Mundo Novo, Catuai e Caturra, foram selecionados principalmente dessas duas bases populacionais, Typica e Bourbon (ANTHONY et al. 2002), sendo uma dos motivos da baixa diversidade genética de *C. arabica*.

Uma variabilidade genética maior é encontrada nas coleções de acessos provenientes do centro de origem, a Etiópia, e que podem ser utilizadas como fonte de genes de interesse para o melhoramento do cafeeiro (HEIN & GATZWEILER, 2006; AERTS et al. 2012). Algumas características alvo foram variáveis entre os genótipos da Etiópia como, por exemplo, níveis de cafeína (SILVAROLLA et al. 2004) e diterpenos (PAGIATTO, 2013); resistência a doenças e pragas: como *Colletotrichum kahawae* (VAN DER VOSSEN & WALYARO, 1980); *Hemileia vastatrix* (KUSHALAPPA & ESKES, 1989); *Meloidogyne incognita* (ANZUETO et al. 2001) e produtividade (BERTRAND et al. 2005). Devido ao seu valor econômico, pois constituem uma grande fonte de alelos para o melhoramento genético, os recursos genéticos de *C. arabica* da Etiópia, vêm sendo caracterizados, por meio de estudos fenotípicos (SILVAROLLA et al. 2004) e moleculares (CHAPARRO et al. 2004; AGA et al. 2005; SILVESTRINI et al. 2007). Esses estudos mostraram que existe uma diferenciação tanto fenotípica como molecular, quanto a origem do material na Etiópia, entre as regiões Leste e Oeste do Vale do Rift.

O Instituto Agronômico do Paraná (IAPAR) possui uma coleção de 132 acessos de *C. arabica* originários da Etiópia (centro de origem) e a avaliação desses acessos com marcadores moleculares constitui uma ferramenta importante para a caracterização da diversidade genética e para o mapeamento de características de interesse. Esta caracterização genotípica de uma coleção de cafeeiros provenientes da Etiópia será importante para o mapeamento de genes e busca de marcas de interesse agronômico.

3.5 MARCADORES MOLECULARES EM CAFEEIRO

O melhoramento tradicional de *C. arabica* é um processo demorado em função de vários fatores. O café é uma planta perene e o seu ciclo de reprodução leva em torno de quatro anos, sendo a seleção realizada em plantas com seis a oito anos. Vários ciclos de cruzamento e seleções são requeridos até a obtenção de genótipos superiores. Para garantir a fidelidade da propagação das sementes, são necessários seis ciclos de homozigosidade, de modo que um programa de melhoramento requer aproximadamente 28 anos (LOS SANTOS-BRIONES & HERNÁNDEZ-SOTOMAYOR, 2006).

Esses problemas tornam o melhoramento tradicional do café uma tarefa difícil e de custo elevado. Dessa forma, no melhoramento genético do cafeeiro, é relevante o uso de ferramentas biotecnológicas como o uso de marcadores de DNA e não somente as avaliações fenotípicas. Marcadores de DNA aumentam o sucesso do melhoramento porque apresentam herança simples enquanto características morfológicas são complexas e podem envolver heranças qualitativas ou quantitativas (CAIXETA et al. 2008).

Os estudos de marcadores microssatélites (SSR) em café têm sido realizados com o objetivo de suprir a necessidade do desenvolvimento de programas de melhoramento para esta cultura. Dentre eles, Moncada & McCouch (2004) analisaram 34 locus SSRs em espécies diplóides e tetraplóides em *Coffea.* Este trabalho indicou que alguns híbridos de espécies diplóides e tetraplóides poderiam ser utilizados nos programas de melhoramento. Silvestrini et al. (2007) utilizaram 16 locus SSRs em 115 acessos de *Coffea*, dentre eles acessos do centro de origem (Etiópia), cultivares comerciais, material originado do Iémen e outras espécies. Outro exemplo de emprego destes marcadores pode ser observado no trabalho de Hendre et al.(2008) no qual se obteve nove novos locus que poderiam ser utilizados em mapas de ligação de *C. canephora*.

Cristancho & Gaitán (2008) construíram bibliotecas de SSRs em café. Dos 12 locus SSRs desenvolvidos, nove foram polimórficos em genótipos diplóides, enquanto cinco foram polimórficos em genótipos tetraplóides, o que confirma a grande diversidade genética em espécies diplóides a partir destes marcadores. Missio et al. (2009) construíram duas bibliotecas genômicas enriquecidas, com o objetivo de desenvolver novos locus SSR. Um total de 96

primers de SSRs foram testados em dois genótipos de *C. arabica* e 90 novos locus SSRs foram validados para futuros estudos genéticos na espécie.

C. canephora apresenta maior diversidade do que *C. arabica* como foi observado em trabalhos com o uso de marcadores RFLP, SSR e SNPs. Esses marcadores mostraram eficiência em caracterizar os grupos ou populações e discriminar os diferentes genótipos (GOMEZ et al. 2009; MUSOLI et al. 2009). Além disso, existem mapas de ligação (RAPD, RFLP e SSR) (COULIBALY et al. 2003) e mapas de sintenia (LEFEBVRE-PAUTIGNY et al. 2010) para esta espécie.

Uma das principais finalidades dos marcadores moleculares é o desenvolvimento de mapas genéticos e a caracterização da base molecular de características quantitativas (QTL – *Quantitative Trait Loci*) em plantas. Em recente trabalho com *C. canephora*, foram identificados QTLs relacionados a uma série de características agronômicas, demonstrando o potencial de aplicação do mapeamento para trabalhos de melhoramento no cafeeiro (LEROY et al. 2011). Entretanto, no caso de *C. arabica*, poucas informações de mapeamento estão disponíveis. Até o momento, alguns mapas parciais, pouco saturados, foram publicados para essa espécie usando marcadores RAPD (TEIXEIRA-CABRAL et al. 2004; OLIVEIRA et al. 2007) e AFLP (PEARL et al. 2004), mas nenhum mapa genético completo da espécie foi disponibilizado. Entretanto, alguns trabalhos pontuais de identificação de QTLs para resistência a ferrugem foram realizados com sucesso (LASHERMES et al. 2001; PRAKASH et al. 2004; MAHÉ et al. 2008; DIOLA et al. 2011).

3.6 Polimorfismos de Nucleotídeo Único (SNPs)

Os SNPs são modificações de um único nucleotídeo e os INDELs são inserções/deleções de poucos pares de base, que caracterizam o conjunto de alelos de um locus específico (BROOKES, 1999). Os SNPs e INDELs podem ser encontrados em regiões codantes e não codantes dos genes (RAFALSKI, 2002). Quando localizados em regiões não codantes são denominados anônimos, quando são localizados em regiões codantes podem ser sinônimos (não alteram o aminoácido de tradução) ou não-sinônimos (alteram o aminoácido e consequentemente a proteína) (BROOKES, 1999).

As substituições nos polimorfismos não sinônimos podem ser conservativas ou não, dependente da característica dos aminoácidos trocados. Assim, podem gerar alterações estruturais e funcionais na proteína. Os SNPs sinônimos não modificam a sequência protéica, mas podem gerar ou suprimir códons de terminação ou poliadenilação na molécula de RNAm (altera a sua estabilidade), podem promover o *splicing* alternativo, modificar códons de iniciação de tradução e alterar o padrão de expressão gênica (KWOK & GU, 1999).

Os SNPs são causados por modificações de ponto, e podem ser classificados como marcadores de transversão, quando a substituição é de uma purina por uma pirimidina e vice-versa, ou de transição, quando ocorrem trocas entre duas purinas ou duas pirimidinas (BROOKES, 1999). Esses constituem as formas mais abundantes de variações do genoma, fornecendo uma fonte para análise de diversidade e saturação de mapas genéticos, auxilia em estudos de associação e mapas de ligação, seleção de BACs e construção de mapas físicos e seleção assistida por marcadores (SAM) (BROOKES, 1999; KWOK & GU,1999; RAFALSKI, 2002; GUPTA & RUSTGI, 2004).

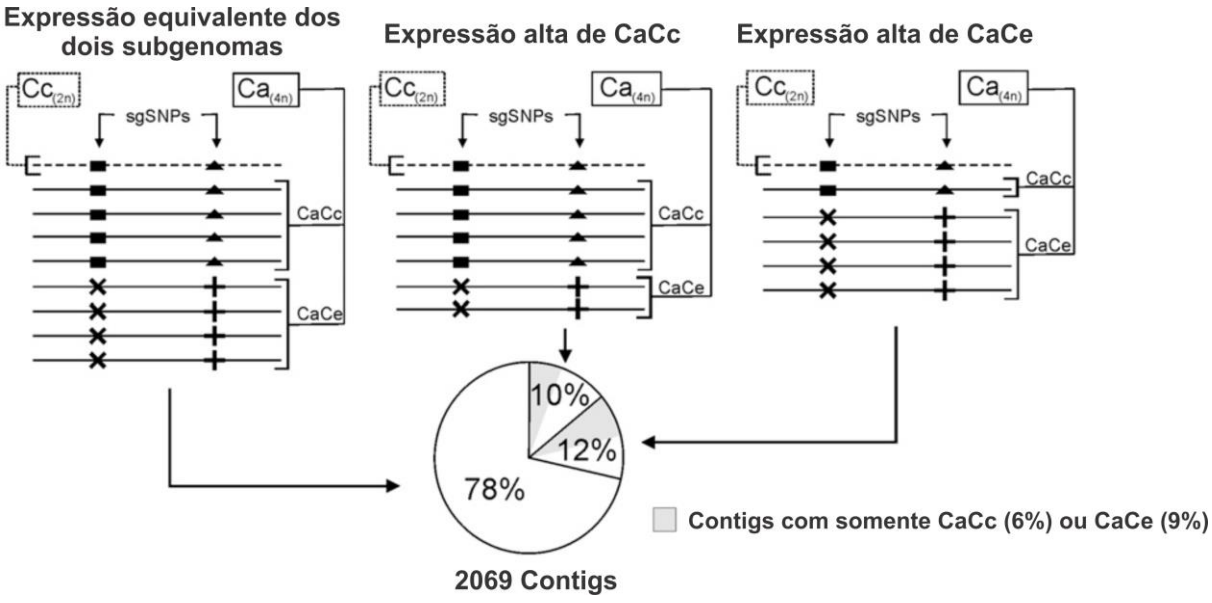
Algumas espécies, como o milho, apresentam alta diversidade nucleotídica; outras como *Arabidopsis*, sorgo e tomate possuem diversidade nucleotídica menor (RAFALSKI, 2002). Existem inúmeras estratégias para prospecção e genotipagem de SNPs e a escolha da metodologia mais adequada é baseada no organismo a ser estudado, no número de polimorfismos genotipados, no tipo de informação obtido, nos materiais, informações disponíveis e custo (TABASSUM & LAKHANPAUL, 2006). Porém, o avanço do NGS vem tornando viável a identificação de SNPs em larga escala para a maioria das espécies.

A identificação de um grande número de marcadores através do NGS tem permitido a aplicação de novos conceitos no melhoramento genético, como *Genome Wide Selection* (GWS) e *Genome Wide Associaton Studies* (GWAS) baseado nas associações de marcadores individuais com traços herdados quantitativamente (GANAL et al. 2009). A alta densidade de polimorfismos de nucleotídeo único (SNP) permite que blocos pequenos de haplótipos sejam correlacionados com variações das características quantitativas. Estas pesquisas tem permitido estudos em doenças humanas (HINDORFF et al. 2009) e em plantas tem obtido sucesso na identificação de locus que possam explicar as variações fenotípicas. Os resultados de GWAS em plantas explicaram grande proporção das

variações fenotípicas encontradas em *Arabidopsis thaliana* L., arroz e milho (BRACHI et al. 2011).

As análises de polimorfismos em café mostraram que *C. arabica* apresenta uma frequência maior de polimorfismos (0,393 SNPs por 100 pb) do que *C. canephora* (0,169 SNPs por 100 pb), mas grande parte dos polimorfismos encontrados são diferenças entre os dois subgenomas de *C. arabica* (VIDAL et al. 2010). A divergência entre os subgenomas podem indicar que existe um mecanismo para prevenir a homogeneização do subgenoma CaCc e CaCe. Além disso, o mesmo trabalho demonstrou que podem existir contribuições diferentes em *C. arabica* derivados de ancestrais específicos, ou seja, uma expressão diferencial dos homeólogos (Figura 3). Apesar da grande quantidade de SNPs detectados, nenhum deles apresentou potencial para genotipagem. Porém, apenas duas cultivares muitos próximas de *C. arabica* foram utilizadas, o que explica em parte esses resultados (VIDAL et al. 2010).

Figura 3 – Variabilidade da frequência dos genes homeólogos nos *contigs*. A esquerda, o painel mostra que em 78% dos *contigs* a frequência dos ESTs de CaCc e CaCe foram equivalentes. Os painéis do meio e da direita mostram que em 10% dos *contigs* a frequência de CaCc foi mais alta que CaCe, enquanto em 12% dos *contigs*, a frequência de CaCe foi mais alta do que CaCc, o que indica que *C. arabica* mostra expressão com particionamento dos genes homeólogos. Adaptado de Vidal et al. (2010).



A fim de aumentar o número de marcadores polimórficos para genotipagem, foi realizado um trabalho de análise inicial da diversidade nucleotídica de genes de interesse dentro de *C. arabica*. Para tanto, foram selecionados 12 genótipos de *C. arabica*, sendo oito acessos da Etiópia e quatro cultivares comerciais. A análise da diversidade genotípica de 18 fragmentos de genes de interesse demonstrou uma freqüência maior de polimorfismos no genoma de *C. arabica*, comparado ao trabalho de Vidal et al. (2010). Foi possível verificar uma freqüência de SNPs representativos de *C. arabica*, com uma freqüência de 1/530 nt, que apesar de baixa, evidenciava o potencial de utilização dos recursos genéticos desta população (YANAGUI et al. 2010; PEREIRA et al. 2011). Quando analisada as regiões de éxon, a freqüência diminui para 1/1285 nt, consequentemente diminuindo o potencial de utilização de SNPs e INDELs para identificação de polimorfismos associados a genes de interesse.

A identificação e uso de marcadores co-dominantes e bialélicos, como os SNPs, pode auxiliar os estudos em *C. arabica*, cuja variabilidade é limitada. Técnicas modernas de sequenciamento de nova geração, como Illumina/Solexa, associadas a estudos no transcriptoma podem ser ferramentas importantes para a identificação de polimorfismos e o futuro desenvolvimento de mapas genéticos e identificação de QTLs para a espécie.

O RNA-Seq permite que sejam feitas análises do perfil transcricional do organismo em estudo associado ao NGS. Para plantas, foram desenvolvidos vários trabalhos de identificação de polimorfismos nas regiões transcritas e que podem gerar informações úteis no melhoramento genético da espécie, por exemplo, no centeio, que permitiu a identificação de 5.234 SNPs a partir de dados de NGS (HASENEYER et al. 2011). Apesar da dificuldade na busca de polimorfismos em organismos alopoliplóides, principalmente pelas limitações das ferramentas de bioinformática (YANG et al. 2011; BYERS et al. 2012; PAGE et al. 2013), recentemente estas dificuldades estão sendo superadas. Por exemplo, para *Brassica napus* L., foram encontrados 41.593 polimorfismos entre duas culivares de *Brassica* (TRICK et al. 2009) e em alfafa, foram identificados 10.826 SNPs entre dois genótipos analisados (YANG et al. 2011). Portanto, estratégias de NGS aplicados a estudos de transcriptoma podem ser uma alternativa viável para descoberta de SNPs em espécies poliplóides.

A fim de se obter uma compreensão maior do transcriptoma de *Coffea* sp. assim como a busca de polimorfismos de nucleotídeos em *C. arabica*, o presente estudo desenvolveu dois trabalhos: um trabalho de análise do transcriptoma de *C. eugenioides* e um segundo trabalho de análise da variabilidade nucleotídica em genótipos de *C. arabica* do centro de origem da espécie, na Etiópia.

4 REFERÊNCIAS

AGA, E.; BEKELE, E.; BRYNGELSSON, T. Inter-simple sequence repeat (ISSR) variation in forest coffee trees (*Coffea arabica* L.) populations from Ethiopia. **Genetica**, v. 124, n. 2-3, p. 213–221, jul. 2005.

AERTS, R. et al. Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian montane rainforests. **Evolutionary Applications**, v. 6, n. 2, p. 243-252, fev. 2013.

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

ANTHONY, F. et al. Adaptive radiation in *Coffea* subgenus *Coffea* L. (Rubiaceae) in Africa and Madagascar. **Plant Systematics and Evolution**, v. 285, n. 1-2, p. 51–64, mar. 2010.

ANZUETO, F. et al. Resistance to *Meloidogyne incognita* in Ethiopian *Coffea arabica* accessions. **Euphytica**, v. 118, n. 1, p. 1-8, mar. 2001.

BERTRAND, B. et al. *Coffea arabica* hybrid performance for yield, fertility and bean weight. **Euphytica**, v. 141, n. 3, p. 255–262, jan. 2005.

BLANCA, J. et al. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). **BMC Genomics**, v. 12, p. 104, fev. 2011.

BRACHI, B.; MORRIS, G.P.; BOREVITZ, J.O. Genome-wide association studies in plants: the missing heritability is in the field. **Genome Biology**, v. 12, n. 10, p. 232, out. 2011.

BRIDSON, D.M.; VERDCOURT, B. Flora of Tropical East Africa. In: POLHILL, R.M. (Ed.). **Rubiaceae**. London: 1988. p. 727.

BROOKES, A.J. The essence of SNPs. **Gene**, v. 234, n. 2, p.177-186, jul. 1999.

BYERS, R.L. et al. Development and mapping of SNP assays in allotetraploid cotton. **Theoretical and Applied Genetics**, v. 124, n. 7, p. 1201-1214, maio 2012.

CAIXETA, E.V., et al. Biotecnologia aplicada ao melhoramento genético do cafeeiro. In: CARVALHO, C.H.S. (Org.). **Cultivares de Café: origem, características e recomendações**. Brasília: Embrapa Café, 2008. p. 103-128.

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 Molecular Biology**, v. 78, n. 1-2, p.135–145, jan. 2012.

CHAPARRO, A.P. et al. Genetic variability of *Coffea arabica* L. accessions from Ethiopia evaluated with RAPDs. **Genetic Resources and Crop Evolution**, v. 51, n. 3, p. 291–297, maio 2004.

CHEVALIER, A. Les caféiers du globe. **Systématique des caféiers et faux caféiers, maladies et insectes nuisibles**. Fascicle III**.** Paris: Encyclopédie Biologique, 1947. 357 p.

CONAB – Companhia Nacional de Abastecimento, 2013 [http://www.conab.gov.br/OlalaCMS/uploads/arquivos/13_09_12_17_54_48_12_cafe.pdf]

COULIBALY, I. et al. 2003. AFLP and SSR polymorphism in a *Coffea* interspecific backcross progeny [(*C. heterocalyx* x *C. canephora*) x *C. canephora*]. **Theoretical and Applied Genetics**, v. 107, n. 6. p. 1148–1155, out. 2003.

CRISTANCHO, M.A.; GAITÁN A.L. Isolation, characterization and amplification of simple sequence repeat loci in coffee. **Crop Breeding and Applied Biotechnology**, v. 8, n. 4, p. 321-329, dez. 2008.

DAMATTA, F.M.; RAMALHO, J.D.C. 2006. Impacts of drought and temperature stress on coffee physiology and production: a review. **Brazilian Journal of Plant Physiology**, v. 18, n. 1, p.55-81, jan./mar. 2006.

DAVIS, A.P.; BRIDSON, D.; RAKOTONASOLO, F. A reexamination of *Coffea* subgenus Baracoffea and comments on the morphology and classification of *Coffea* and *Psilanthus* (Rubiaceae-Coffeeae). In: KEATING, R.C.; HOLLOWELL, V.C.; CROAT, T. (Ed.). **Festschrift for William G. D'Arcy: the legacy of a taxonomist (Monograph in Syst Bot 104).** St. Louis: MBG Press, 2005. p. 398–420.

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

DIOLA, V. et al. High-density genetic mapping for coffee leaf rust resistance. **Tree Genetics & Genomes**, v. 7, n. 6, p. 1199-1208, dez. 2011.

DOS SANTOS, T.B. et al. Expression of three galactinol synthase isoforms in *Coffea arabica* L. and accumulation of raffinose and stachyose in response to abiotic stresses. **Plant Physiology and Biochemistry**, v. 49, n. 4, p. 441-448, abr. 2011.

FERNANDEZ, D. et al. 2012. 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. **Molecular Plant Pathology**, v. 13, n. 1, p. 17-37, jan. 2012.

FERREIRA, M.E.; GRATTAPAGLIA, D. Aplicações de Marcadores Moleculares na Genética e Melhoramento de Plantas. In: FERREIRA M.E., GRATTAPAGLIA D. **Introdução ao uso de marcadores moleculares em análise genética.** 3. ed. Brasília: EMBRAPA-CERNAGEN, 1998. p. 69-116.

GANAL, M.W.; ALTMANN, T.; RÖDER, M.S. 2009. SNP identification in crop plants. **Current Opinion in Plant Biology,** v. 12, n. 2, p. 211-217, abr. 2009.

GOMEZ, C. et al. Current genetic differentiation of *Coffea canephora* Pierre ex A. Froehn in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and recent human activities. **BMC Evolutionary Biology**, v. 16, n. 9, p.167, jul. 2009.

GUERREIRO FILHO, O. et al. Origem e Classificação Botânica do Cafeeiro. In: Carvalho C.H.S. (Org.). **Cultivares de café: origem, características e recomendações**. Brasília: Embrapa Café, 2008. p. 21-35.

GUPTA, P.K.; RUSTGI, S. Molecular markers from the transcribed/expressed region of the genome in higher plants. **Functional & Integrative Genomics**, v. 4, n. 3, p. 139–162, jul. 2004.

GUYOT, R. et al. Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum* Sp.) and rosid (*Vitis vinifera*) clades. **BMC Genomics**, v. 13, p. 103, mar. 2012.

HASENEYER, G. et al. From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). **BMC Plant Biology**, v. 38, n. 11, p. 131, set. 2011.

HEIN, L., GATZWEILER, F. The economic value of coffee (*Coffea arabica*) genetic resources. **Ecological Economics**, v. 60, n. 1, p. 176-185, nov. 2006.

HENDRE, P.S. et al. 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 Biology**, v. 8, p. 51, abr. 2008.

HINDORFF, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. **Proceedings of the National Academy of Sciences**, v. 106, n. 23, p. 9362-9367, jun. 2009.

ICO - International Coffee Organization, 2014 [http://www.ico.org]

KUSHALAPPA, A.C.; ESKES, A.B. Advances in Coffee Rust Research. **Annual Review of Phytopathology,** v. 27, n. 1, p. 503-531, set. 1989.

KWOK, P.Y.; GU, Z. Single nucleotide polymorphism libraries: why and how are we building them? **Molecular Medine Today**, v. 5, n. 12, p. 538-543, dez. 1999.

LASHERMES, P. et al. Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. **Theoretical and Applied Genetics**, v. 94, n. 6-7, p. 947–955, jun. 1997.

LASHERMES, P. et al. Molecular characterisation and origin of the *Coffea arabica* L. genome. **Molecular Genetics and Genomics**, v. 261, n. 1, p. 259–266, mar. 1999.

LASHERMES, P. et al. Genetic linkage map of *Coffea canephora*: effect of segregation distortion and analysis of recombination rate in male and female meioses. **Genome**, v. 44, n. 4, p. 589–596, 2001.

LEFEBVRE-PAUTIGNY, F. et al. High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. **Tree Genetics & Genomes**, v. 6, n. 4, p. 565–577, jul. 2010.

LEROY, T. et al. Genetics of coffee quality. **Brazilian Journal of Plant Physiology**, v. 18, n. 1, p. 229–242, jan./mar. 2006.

LEROY, T. et al. Improving the quality of African robustas: QTLs for yield- and quality-related traits in *Coffea canephora*. **Tree Genetics & Genomes**, v. 7, n. 4, p. 781-798, ago. 2011.

LIN, C. et al. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. **Theoretical and Applied Genetics**, v. 112, n. 1, p. 114-130, dez. 2005.

LOMBELLO, R.A.; PINTO-MAGLIO, C.A.F. Cytogenetic studies in *Psilanthus ebracteolatus* Hiern., a wild diploid coffee species. **Cytologia**, v. 68, n. 4, p. 425–429, 2003.

LOMBELLO, R.A.; PINTO-MAGLIO, C.A.F. Cytogenetic studies in *Coffea* L. and *Psilanthus* Hook.f. using CMA/DAPI and FISH. **Cytologia**, v. 69, n. 1, p. 85–91, 2004.

LOS SANTOS-BRIONES, C.; HERNÁNDEZ-SOTOMAYOR, S.M.T. Coffee biotechnology. **Brazilian Journal of Plant Physiology**, v. 18, n. 1, p. 217-227. 2006.

MARRACCINI, P. et al. Differentially expressed genes and proteins upon drought acclimation in tolerant and sensitive genotypes of *Coffea canephora*. **Journal of Experimental Botany**, v. 63, n. 11, p. 4191–4212, abr. 2007.

MAHÉ, L. et al. Development of sequence characterized DNA markers linked to leaf rust (*Hemileia vastatrix*) resistance in coffee (*Coffea arabica* L.). **Molecular Breeding**, v. 21, n. 1, p. 105–113, jan. 2007.

MAURIN, O. et al. Towards a phylogeny for *Coffea* (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. **Annals of Botany**, v. 100, p. 1565–1583, dez. 2007.

MAPA - Ministério da Agricultura, Pecuária e Abastecimento, 2013.

MAZZAFERA, P.; CARVALHO, A. Breeding for low seed caffeine content of coffee (*Coffea* L.) by interspecific hybridization. **Euphytica**, v. 59, n. 1, p. 55–60, nov. 1991.

MENDES, A.N.G. et al. História das primeiras cultivares de cafés plantadas no Brasil. In: CARVALHO, C.H.S. (Org.). **Cultivares de café: origem, características e recomendações**. Brasília: Embrapa Café, 2008. p. 69-78.

METZKER, M.L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31-46, jan. 2010.

MISSIO, R.F. et al. Development and validation of SSR markers for *Coffea arabica* L. **Crop Breeding and Applied Biotechnology**, v. 9, p. 361-371. 2009.

MONCADA, P.; MCCOUCH, S. Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. **Genome**, v. 47, n. 3, p. 501-509, jun. 2004.

MUSOLI, P. et al. Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda. **Genome**, v. 52, n. 7, p. 634-646, jul. 2009.

NOIROT, M., et al. Genome size variations in diploid african *Coffea* species. **Annals of Botanny**, v. 92, n. 5, p. 709-714, nov. 2003.

NOVAES, E. et al. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. **BMC Genomics**, v. 9, p. 312, jun. 2008.

OLIVEIRA, C.B. et al. Partial map of *Coffea arabica* L. and recovery of the recurrent parent in backcross progenies. **Crop Breeding and Applied Biotechnology**, v. 7, n. 2, p. 196-203, mar. 2007.

PAGE, J.T.; GINGLE, A.R.; UDALL, J.A. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. **G3 (Bethesda)**, v. 3, n. 3, p. 517-25, mar. 2013.

PAGIATTO, N.F. **Análise de diterpenos e cafeína em uma coleção da etiópia de *Coffea arabica***. 2013. 77 fls. Dissertação de Mestrado (Pós Graduação em Biotecnologia) – Universidade Estadual de Londrina, Londrina, 2013.

PEARL, H.M. et al. Construction of a genetic map for Arabica coffee. **Theoretical and Applied Genetics**, v. 108, n. 5, p. 829-835, mar. 2004.

PEREIRA, L.F.P. et al. Analysis of nucleotide diversity in *Coffea* spp. In: XIX Plant Animal Genomes Conference, 2011, San Diego. **Anais**...San Diego: PAG-XIX Abstracts, 2011. p. W-153.

PINTO-MAGLIO, C.A.F.; CRUZ, N.D. Pachytene chromosome morphology in *Coffea* L. II. *C. arabica* L. complement. **Caryologia**, v. 51, n. 1, p. 19-35, 1998.

PRAKASH, N.S. et al. AFLP analysis of introgression in coffee cultivars (*Coffea arabica* L.) derived from a natural interspecific hybrid. **Euphytica**, v. 124, n. 3, p: 265–271, abr. 2002.

PRIVAT, I. 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**, v. 178, n. 4, p. 781-797, mar. 2008.

RAFALSKI, A. Applications of single nucleotide polymorphisms in crop genetics. **Current Opinion in Plant Biology**, v. 5, n. 2, p. 94-100, abr. 2002.

SILVAROLLA, M.B.; MAZZAFERA, P.; FAZUOLI, L.C. A naturally decaffeinated arabica coffee. **Nature**, v. 429, p. 826-826, jun. 2004.

SILVESTRINI, M. et al. Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. **Genetic Resources and Crop Evolution**, v. 54, n. 6, p.1367-1379, set. 2007.

TABASSUM, J.; LAKHANPAUL, S. Single nucleotide polymorfism (SNP) – methods and applications in plant genetics: a review. **Indian Journal of Biotechnology**, v. 5:453-459, out. 2006.

TEIXEIRA-CABRAL, T.A. et al. 2004. Single-locus inheritance and partial linkage map of *Coffea arabica* L. **Crop Breeding and Applied Biotechnology**, v. 4, n. 4, p. 416-421, dez. 2004.

TRICK, M. et al. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. **Plant Biotechnology Journal**, v. 7, n. 4, p. 334-346, maio 2009.

VAN DER VOSSEN, H.A.M.; WALYARO, D.J. Breeding for resistance to coffee berry disease in *Coffea arabica* L. II. Inheritance of the resistance. **Euphytica**, v. 29, n. 3, p. 777–791, nov. 1980.

VIDAL, R.O. et al. 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**, v. 154, n. 3, p. 1053-1066, nov. 2010.

WIKSTROM, N.; SAVOLAINEN, V.; CHASE, M.W. Evolution of the angiosperms: calibrating the family tree. **Proceedings of the Royal Society B: Biological Sciences**, v. 268, n. 1482, p. 2211-2220, nov. 2001.

WU, F. et al. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. **Genetics**, v. 174, v. 3, p.1407-1420, nov. 2006.

YANAGUI, K. **Diversidade nucleotídica de oito genes relacionados à qualidade da bebida de *Coffea arabica***. 2012. 53 fls. Dissertação de Mestrado (Pós Graduação em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Londrina, 2012.

YANG, S.S. et al. Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. **BMC Genomics**, v. 12, p. 199, abr. 2011.

5 ARTIGO 1

TRANSCRIPTOME SEQUENCING AND ANALYSIS OF DIFFERENTIAL GENE EXPRESSION IN *Coffea eugenioides*

Abstract

Background*:*
Coffea eugenioides is considered the ancestor of *C. arabica* allopolyploid, together with *C. canephora*. However, there are few molecular studies in this specie. This work delivers a global overview of transcriptionally genes in this specie using next-generation sequencing. RNA-Seq generates a large volume data and it has been used in several studies in the transcriptome level. A panel with coding genes in *C. eugenioides* open possibilities to discover potential candidate genes related with some important agronomic traits for coffee plants. Here, we present an effort for a large scale gene identification of *C. eugenioides* from RNA-Seq of leaves and fruits.

Results*:*
We obtained a total of 8435413 Illumina reads: 3688364 reads from leaves and 4747049 reads from fruits. A *de novo* assembled with all organs sequences was performed by Trinity and it generated 36935 contigs which were annotated against non-redundant (NCBI-nr) protein database (63.1% hits), Swiss-Prot (45.8% hits), Gene Ontology (48.9% hits), InterproScan (34.7% hits), PlantCyc (20.8% hits) and KEGG (2.2% hits). In addition, we examined an overview of differentially expressed genes between leaves and fruits using DESeq. Interestingly, several genes exclusively expressed in fruits did not find similarity in any database. We selected contigs more expressed in each organ to confirm transcriptional profile by qPCR.

Conclusion*:*
Our study provides a first gene catalog to *C. eugenioides*. These informations can increase knowledge about the mechanisms involved in the *C. arabica* homeologs expression. A general repertoire of differential gene expression in *C. eugenioides* in leaves and fruit opens new perspectives to studies in this specie and presents potential value for genetic improvement of coffee.

Keywords: Annotation, *Coffea eugenioides*, *De novo* assembly, Differential expression, qPCR, RNA-seq.

5.1 Background

Coffee is one of the most important agricultural commodity worldwide. The genus has 124 species (Davis et al. 2011) and *Coffea arabica* L. and *C. canephora* Pierre ex A. Froehner account for approximately 65% and 35% of world production, respectively (http://www.ico.org). All species in the genus are diploid, except *C. arabica* which is an allotetraploid (2n = 4× = 44) probably derived from a recent hybridization event of two diploid species, *C. canephora* (2n = 2× = 22) and *C. eugenioides* S. Moore (2n = 2× = 22) (Lashermes et al. 1999). This hybridization event, probably followed by genome duplication (Combes et al. 2012), occurred around 100-500 thousand years ago (Anthony et al. 2010, Yu et al. 2011), or even earlier, between 10-50 thousand years ago (Cenci et al. 2012).

The species considered ancestors of *C. arabica* present different agroecological adaptation and characteristics. *C. canephora* grows better in lowlands and is characterized to higher productivity, tolerance to pests, drought stress and higher caffeine content, however, its beverage is considered of lower quality when compared with *C. arabica*. Therefore, it is used mostly by the instant coffee industry and/or in blends with *C. arabica* (DaMatta et al. 2006). *C. eugenioides* grows in highlands and near forest edges in Central-East Africa and it is not produced on a commercial scale due its low fruit production. In breeding, *C. eugenioides* was included to reduce caffeine levels and to improve cup quality, since *C. eugenioides* presents small fruits with low caffeine content, comparing with both *C. arabica* and *C canephora* (Mazzafera & Carvalho 1991). Meanwhile, *C. arabica* can be grown in regions with marked variations in thermal amplitude and it has a better cup quality, in comparison with *C. canephora* (DaMatta & Ramalho 2006; Leroy et al. 2006; Privat et al. 2008).

Advances in new generation sequencing technologies (NGS) such as Illumina/Solexa allowed progresses in the analysis of the transcriptome of various species generating a large volume of data with cost effective (Metzker et al. 2010). In addition, it has the advantage of higher sensitivity and greater dynamic range of gene expression than array techniques (Marioni et al. 2008). This methodology has since been applied to several organisms to answer questions regarding gene annotation, diversity and gene expression (Blanca et al. 2011; Toledo-Silva et al. 2013; Cardoso et al. 2014).

In coffee, most of transcriptome sequencing data relies on the two major cultivated species. Sanger EST sequencing projects were developed for *C. canephora* (Lin et al. 2005) and *C. arabica* (Vidal et al. 2010; Mondego et al. 2011) and more recently, transcriptome analysis using NGS was done in studies involved in biotic and abiotic interactions (Fernandez et al. 2012; Combes et al. 2013). From these works, it was observed that the two subgenomes contained in allotetraploid genome of *C. arabica*, subgenome *C. canephora* – CaCc and subgenome *C. eugenioides* – CaCe, do not contribute equally to the transcriptome. Given the regulation mechanisms between homeologous genes, those studies showed the complexity of the regulation in allopolyploids and indicated that genes useful for *C. arabica* in breeding programs could be present in its genome but are inactive due to partitioned expression (Vidal et al. 2010; Marraccini et al. 2011; Cotta et al. 2014). However, one of the drawbacks of those studies was the lack of *C. eugenioides* data in order to increase and improve the comparison of gene expression of those subgenomes.

Despite the strategic importance of understanding gene expression of the *C. arabica* coffee ancestors, there are few studies focusing on *C. eugenioides.* Up to the moment, there are only 58 sequences of *C. eugeniodes* deposited at NCBI database. Therefore, applying RNA-Seq to analyze leaves and fruits organs of *C. eugenioides*, we describe the generation of 36935 contigs *de novo* assembled from 8435413 reads obtained using Illumina. We present an overview of this transcriptome as a potential model for future studies in *Coffea* that could explain the mechanisms involved in the expression of homeologous genes in *C. arabica* and it also allows a comparative frame among *Coffea* gene expression in different species.

5.2 Results

5.2.1 Sequencing *C. eugenioides* transcriptome

We produced a total of 8435413 reads (3688364 reads from leaves and 4747049 reads from fruits). Due the absence of reference genomic sequences, a *de novo* RNA-Seq assembly was performed resulting a total of 36935 contigs of size length >200 bp and quality ≥20. The average length of these contigs was 701.52 base pairs (bp) and more than 8000 contigs had size higher than 1000 bp. The distribution of contigs according to their size is showed in Figure 1.

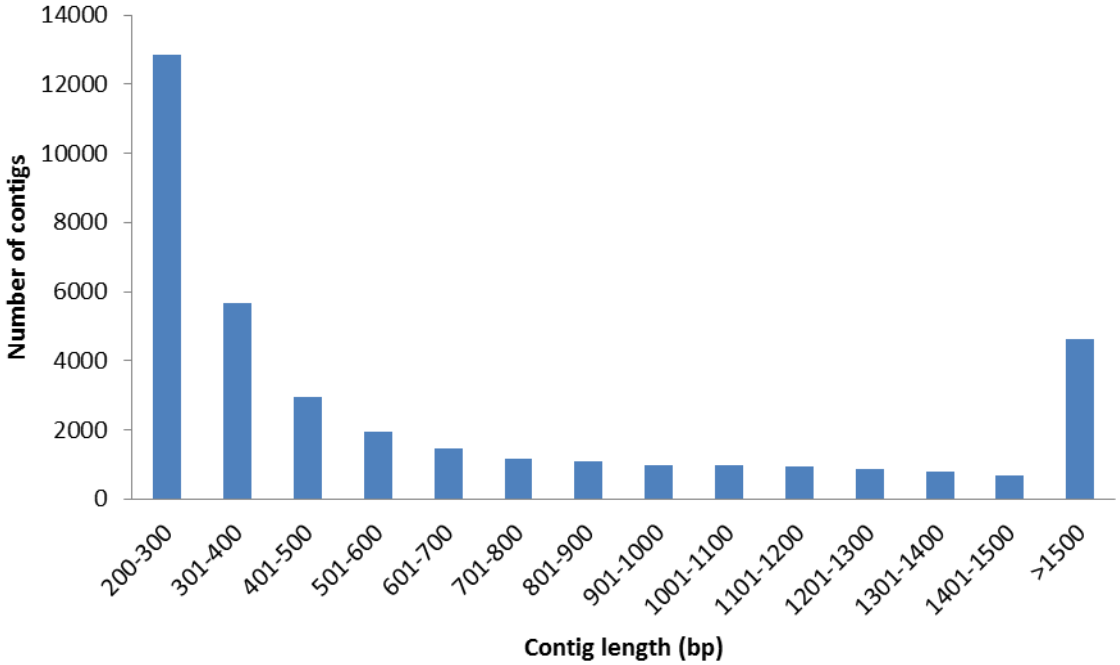


Figure 1. Length distribution of the *C. eugenioides* contigs with a *de novo* assembled from Illumina RNA-Seq.

5.2.2 Functional characterization

To estimate the accuracy of *de novo* assembly performed in the work, they were done annotations by similarity analysis against seven databases. A summary of these annotations is described in Table 1. The assembled sequences of *C. eugenioides* transcriptome were analyzed by similarity with BLASTX against available non redundant sequences from the NCBI database (NCBI-nr). A total of 23297 contigs (63.1% hits) have a hit above 1e-5 (Table 1). Among the NCBI-nr BLASTX top hits, 10706 contigs had their first hit in *Vitis vinifera* proteins (29.4%), followed by *Ricinus communis* (9.7%), *Populus trichocarpa* (8.7%), *Nicotiana tabacum* (1.2%) and *Glycine max* (1%). The 100 top-hit species distribution is presented in Additional File 1.

To compare *C. eugenioides* contigs with published data of coffee, a BLASTN was done between our dataset (36935 contigs) and 35113 contigs of *C. arabica* database developed by Mondego et al. (2011). Nearly half of *C. eugenioides* contigs (18567 contigs - 50.2%) had a homolog in *C. arabica* with >90% identity. Furthermore, to identify *C. eugenioides* contigs potentially encoding proteins with known function, a BLASTX analysis (1e-5) was performed using Swiss-Prot protein

databases (http://www.uniprot.org/downloads) and 16902 contigs were annotated (45.8% hits) (Table 1).

Table 1. Annotation summary of assembled *C*. *eugenioides* contigs.

Database	BLAST	Annotations[1]	Annotations Percentage[2]
NCBI-nr proteins	BLASTX	23297	63.1%
Coffea EST Project	BLASTN	18567	50.2%
Swiss-Prot	BLASTX	16902	45.8%
Gene Ontology(Go Slim)	BLASTX	18058	48.9%
InterProscan	BLASTX	12834	34.7%
PlantCyc	BLASTX	7669	20.8%
KEGG	BLASTX	802	2.2%

[1]Number of contigs annotated.
[2]Percentage of annotated contigs from *C. eugenioides* related to the total 36,935 contigs.

Functional characterization of contigs was performed assigning Gene Ontology annotations (GO), with the BLAST2GO software (Conesa et al. 2005). To provide a general representation of the annotation, the classification of GO Slim was obtained. A total of 18058 contigs (48.9% hits) could be assigned to one or more ontologies, i.e., one contig may be assigned to more than one ontology (Table 1). The number of GO terms per contig varied from 1 to 49. In total, 87640 GO terms were retrieved, 41.8% in the biological process, 32.4% in the molecular function and 25.8% in the cellular component category, respectively. A summary with the contigs annotated in each GO Slim term in biological process and molecular function categories is shown in Figure 2.

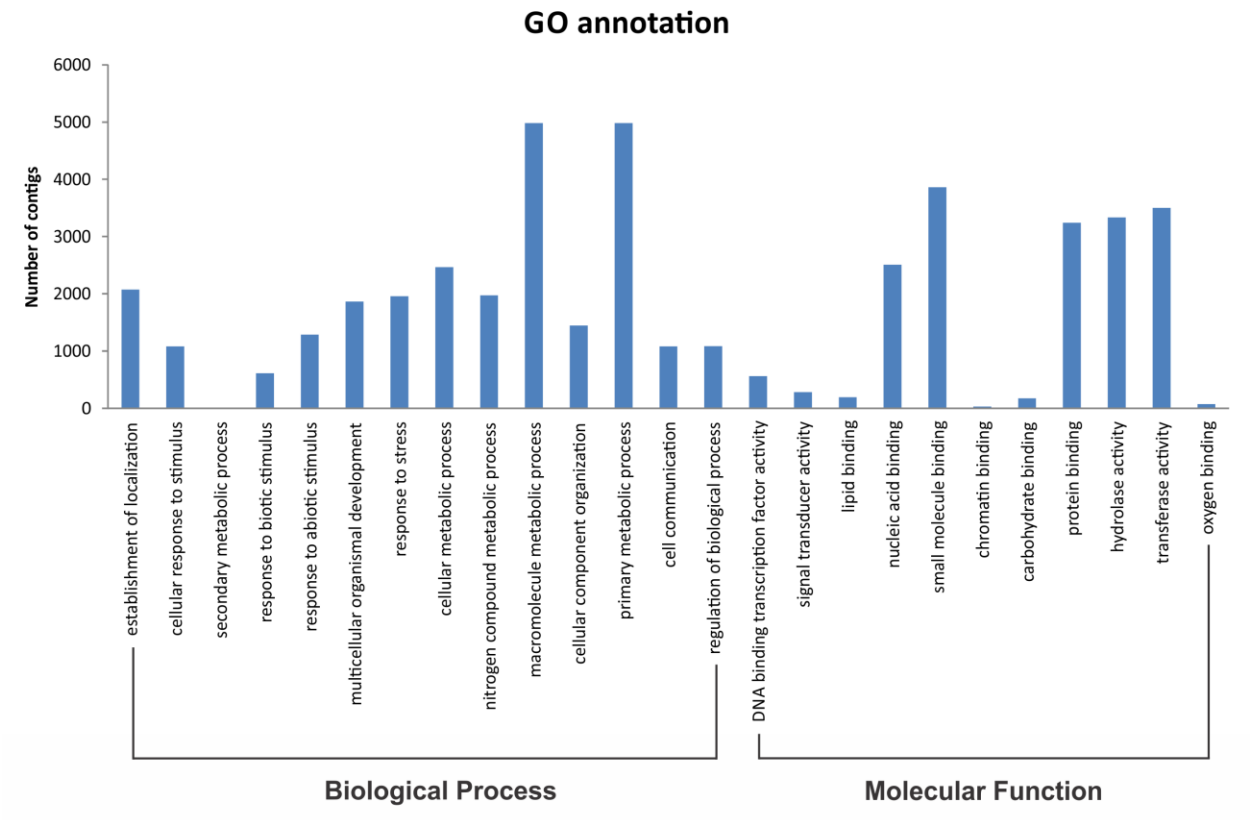


Figure 2. Number of *C. eugenioides* contigs in each functional category. *C. eugenioides* contigs were classified into different functional groups based on a set of GO slims in the Biological Process and Molecular Function categories.

We used the GO annotations to assign each contig to a set of GO Slims of the biological process and molecular functions categories. Metabolic process (GO:0044238; GO:0043170; GO:0044237; GO:0006807), response to stress (GO:0006950) and multicellular organismal development (GO:0007275) were the most highly represented groups under the Biological Process category. Under the molecular function category, assignments were mainly to the small molecule binding (GO:0036094), transferase activity (GO:0016740), hydrolase activity (GO:0016787), protein binding (GO:0005515) and nucleic acid binding (GO:0003676).

Conserved domains in *C. eugenioides* contigs were identified against the InterProscan databases. A total of 12834 contigs presented annotations, with 4961 different domains/families (Figure 3). InterPro domains/families were ranked according to the number of contigs contained in each InterPro domain, and the 20 most abundant InterPro domains/families are represented in Figure 3.

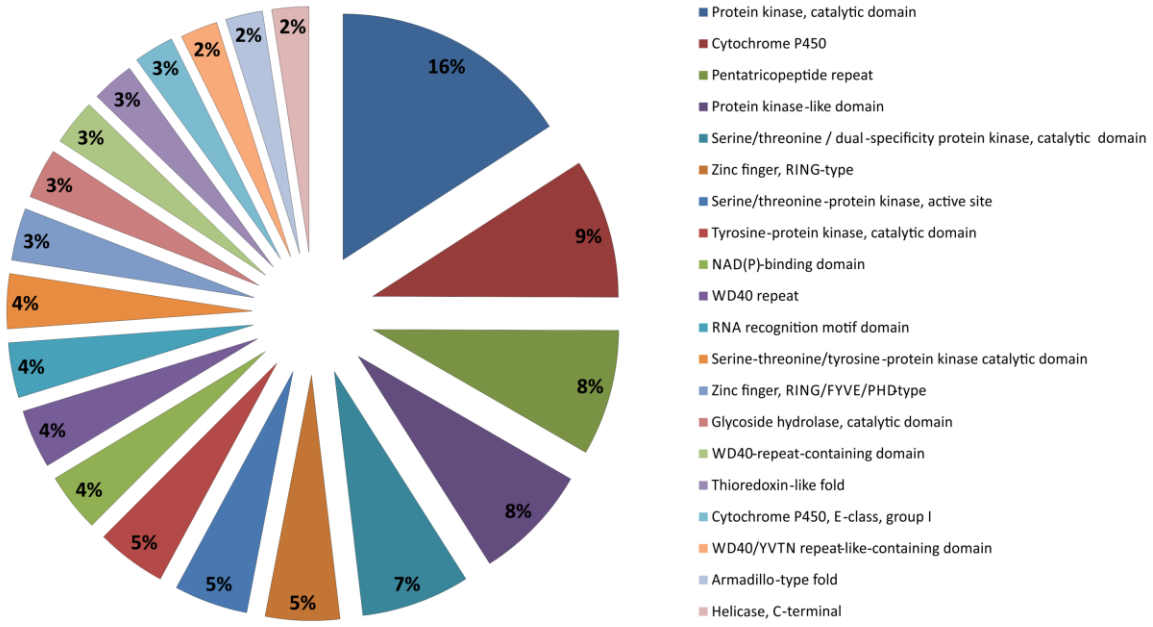


Figure 3. The 20 most abundant InterProscan categories present in the *C. eugenioides* contigs catalogue.

The 20 most frequent areas were protein kinase (IPR000719) with 2286 contigs (16%), cytochrome P450 (IPR001128) with 1321 contigs (9%), pentatricopeptide repeat (IPR002885) with 1192 contigs (8%), protein kinase-like domain (IPR011009) with 1104 contigs (8%) e serine/threonine- / dual specificity protein kinase (IPR002290) with 1028 contigs (7%).

The distribution of contigs into various metabolic pathways was verified using the PlantCyc (http://www.plantcyc.org) and KEGG database (Kanehisa et al. 2000). BLASTX analyses against PlantCyc database resulted in the annotation of 7669 contigs (20.8%). From KEGG, 802 contigs (2.2%) were assigned to 142 pathways and 374 enzymes (Figure 4). The starch and sucrose metabolism were the most abundant categories with 450 members/contigs (Additional File 2), followed by purine metabolism (393 members), methane metabolism (204 members) and glycolysis/glucogenesis (201 members) (Figure 4).

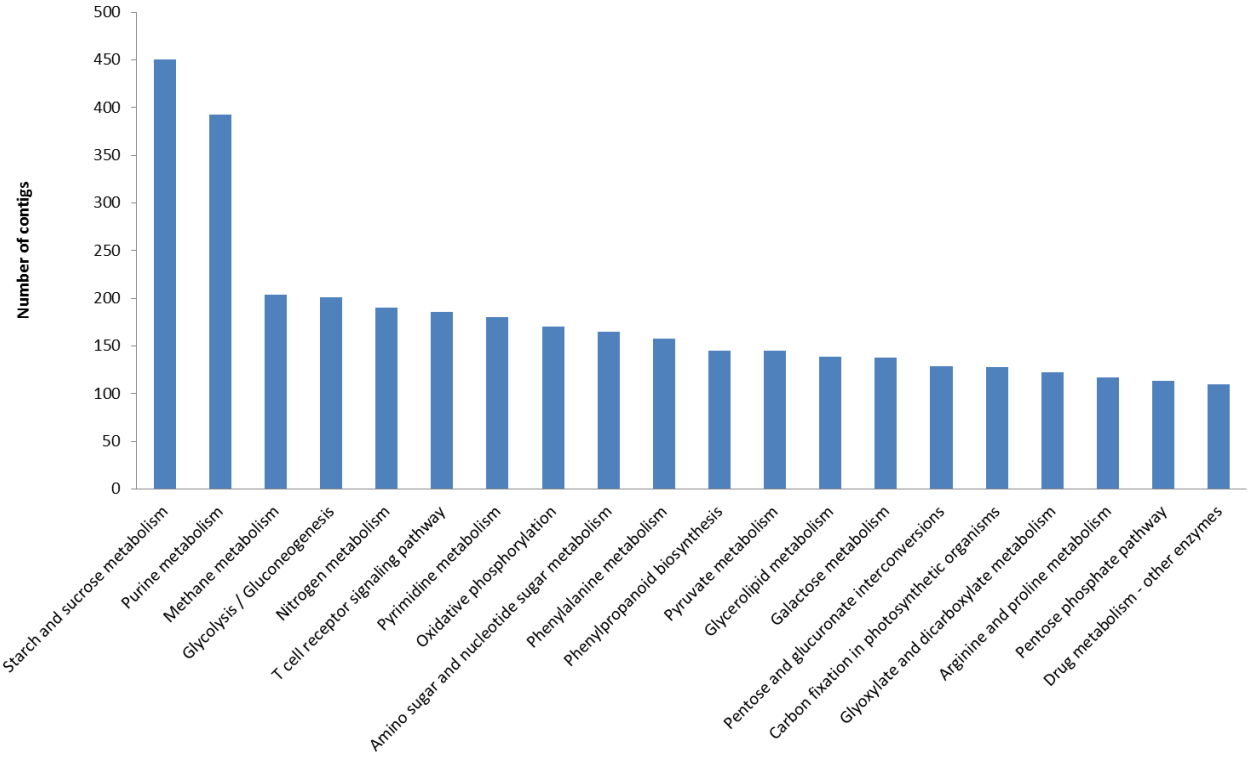


Figure 4. Top 20 pathways obtained in *C. eugenioides* transcriptome among a total of 802 predicted pathways available on KEGG database using BLAST2GO tool.

5.2.3 Analysis of differential gene expression

The assembled transcripts were mapped using Bowtie software and their respective read abundances (RPKM) were estimated by DESeq. Normalized reads from each organ were mapped against the *de novo* transcriptome as reference. Based on the RPKM list developed to each organ, 2050 contigs were considered as specific to leaves, 3299 contigs were exclusively found in fruit, and 31586 contigs were described for both organs (Figure 5).

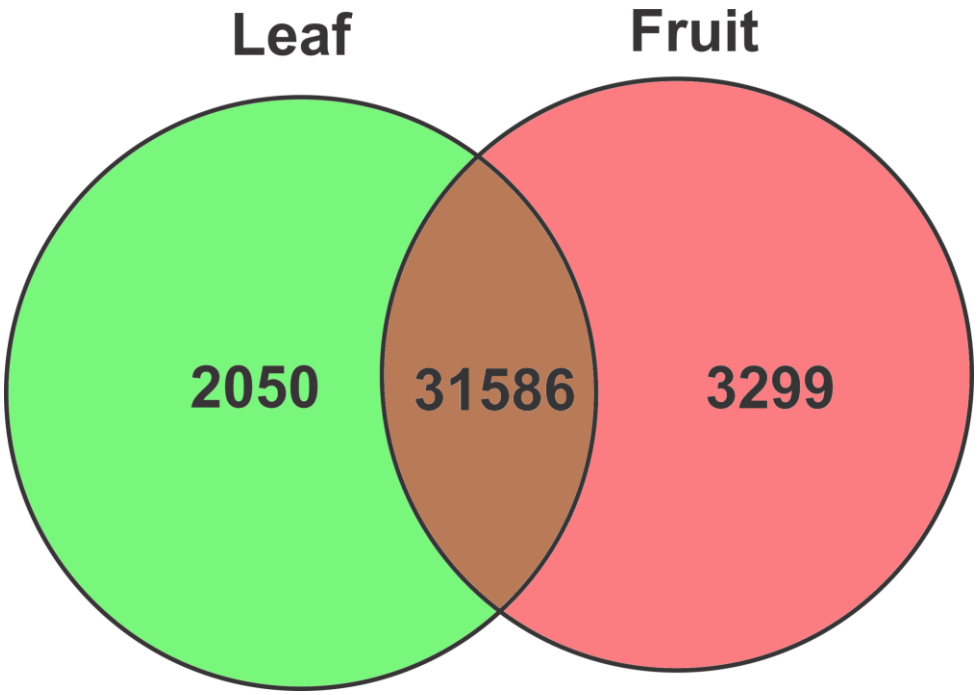


Figure 5. Venn Diagram indicated contigs of each organ that mapped in the reference transcriptome. 2050 contigs were specifically founded in leaves and 3299 contigs were specifically observed in fruit. 31586 contigs were described for both organs.

BLAST2GO was used for a GO functional enrichment analysis of exclusive genes in leaves and fruits (Figure 6). In leaves, biological process category provided phosphorylation (GO:0016310), protein phosphorylation (GO:0006468), carbohydrate catabolic process (GO:0016052), hexose metabolic process (GO:0019318) and single-organism carbohydrate catabolic process (GO:0044724) (Figure 6A). Molecular function category found catalytic activity (GO:0003824), small molecule binding (GO:0036094), nucleotide binding (GO:0000166), nucleoside phosphate binding (GO:1901265) and transferase activity (GO:0016740) (Figure 6A).

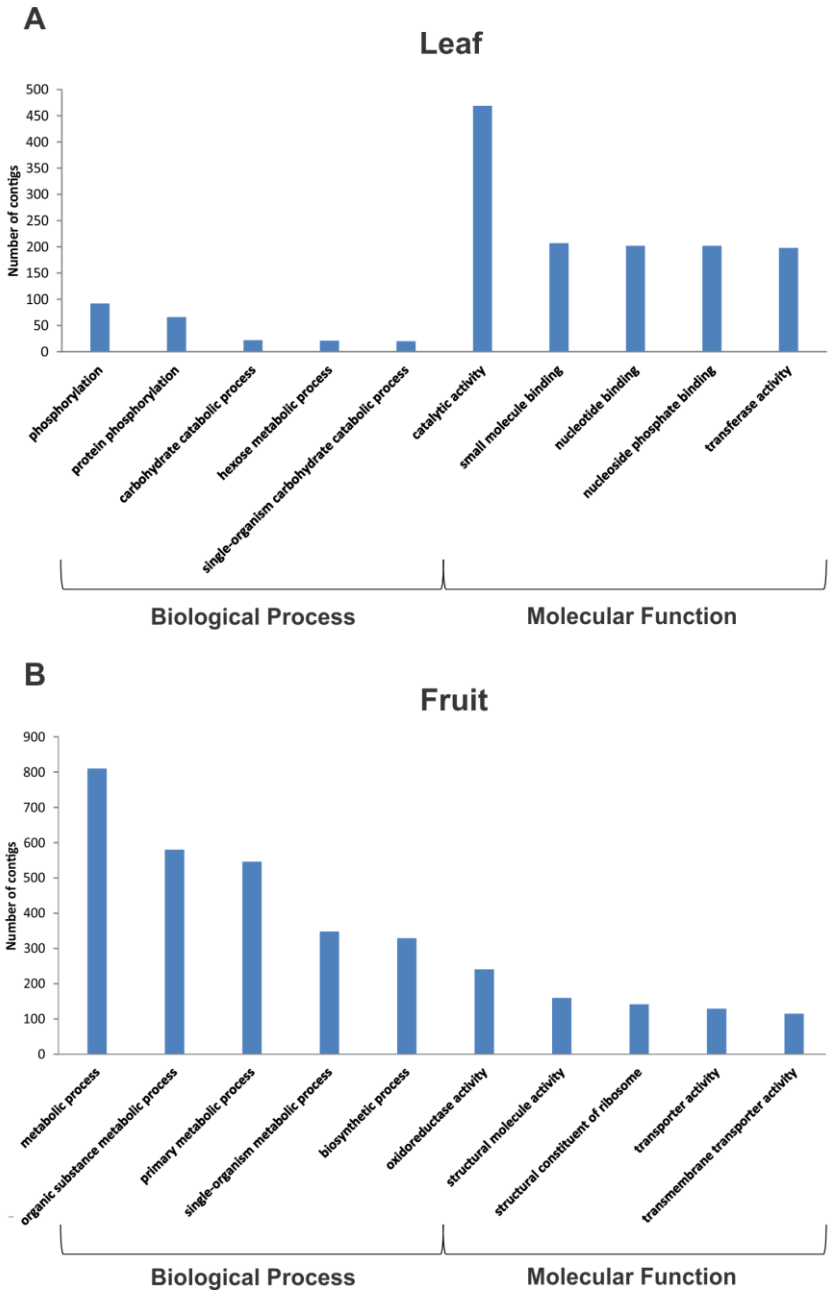


Figure 6. GO term distribution of contigs differentially expressed in *C. eugenioides*. A. Top-hits GO distribution of expressed contigs in leaves. B. Top-hits GO distribution of expressed contigs in fruit.

In fruits, biological process category provided assignments mainly to metabolic process (GO:0008152), organic substance metabolic process (GO:0071704), primary metabolic process (GO:0044238), single-organism metabolic process (GO:0044710) and biosynthetic process (GO:0009058) (Figure 6B). In molecular function category, oxidoreductase activity (GO:0016491), structural molecule activity (GO:0005198), structural constituent of ribosome (GO:0003735), transporter activity (GO:0005215) and transmembrane transporter activity (GO:0022857) (Figure 6B).

5.2.4 qPCR

We used qPCR technique to validate transcriptional pattern of 10 selected contigs with high expression levels in leaves and fruits and all genes displayed a transcriptional pattern in agreement with DESeq analysis. *Ce14433*, *Ce15205*, *Ce2770*, *Ce14847* and *Ce10671* unigenes were mostly expressed in leaves, and *Ce14834*, *Ce13100*, *Ce13451, Ce9246* and *Ce13525* unigenes were higher expressed in fruits (Figure 7). An interesting result was obtained for *Ce14834* and *Ce13451* unigenes, in function that they were amplified only in fruit. These contigs were manually annotated against *Arabidopsis* proteins – TAIR database (Table 2) (http://www.arabidopsis.org).

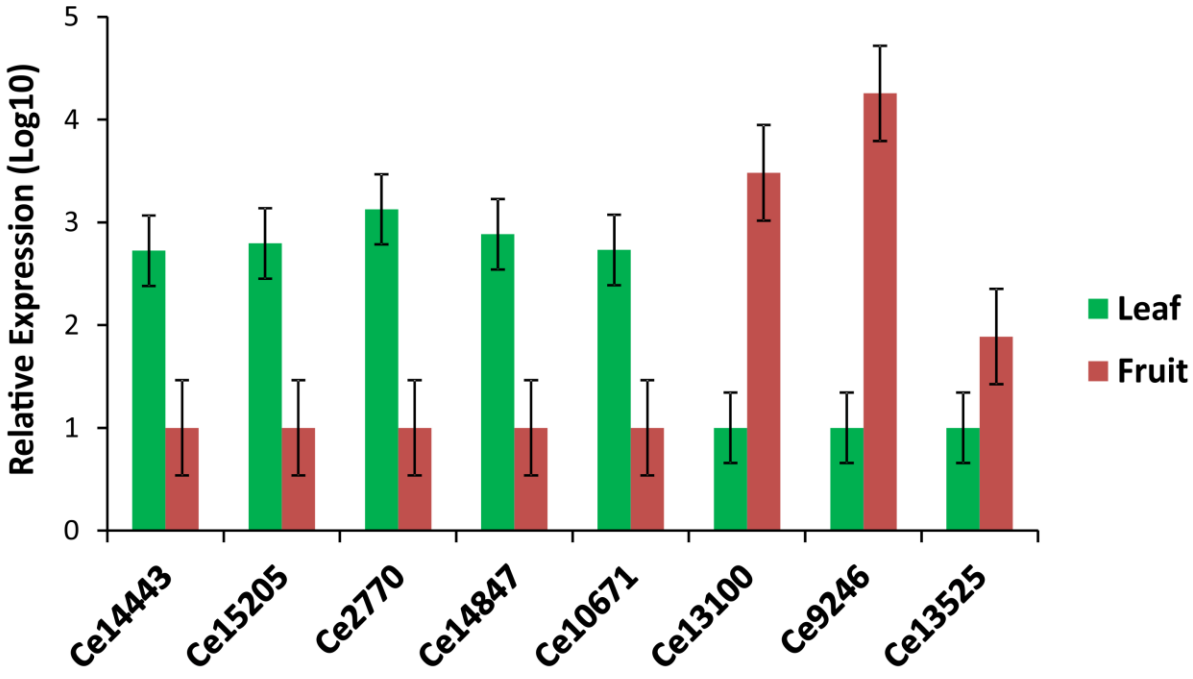


Figure 7. Relative expression values of unigenes up-regulated in leaves (green) and fruits (red) by qPCR (log10 scale). Bars represent the standard error values.

Annotation process according BLASTX on TAIR database showed that high expressed genes in leaves encoded UDP-glycosyltransferase (*Ce14433*), germin 3 (*Ce15205*), glucose-6-phosphate/phosphate translocator (*Ce2770*), chitinase (*Ce14847*) and BURP domain-containing protein (*Ce10671*), while the high expressed genes in fruits were related to BURP domain-containing protein (*Ce14834*), serine carboxypeptidase *(Ce13100)*, cytochrome P450 (*Ce13451*),

transcription factor (*Ce9246*) and oxidoreductase, zinc-binding dehydrogenase family protein (*Ce13525*).

5.3 Discussion

5.3.1 Assembly and annotation of *C. eugenioides* transcriptome

This report represents the first overview of *C*. *eugenioides* transcriptome using NGS. We reported a comprehensive annotation of it transcriptome in several databases and determined genes which are differentially expressed in leaves and fruits. Contig length ranged from 200 bp to 10 kb, with a mean length of approximately 700 bp, which is similar to recent analyses in *Panicum maximum* Jacq (758 bp) (Toledo-Silva et al. 2013) and *Cocos nucifera* L. (752 bp) (Fan et al. 2013) using Illumina sequencing, but it even greater than those for *Camellia sinensis* (L.) Kuntze (355 bp) (Shi et al. 2011) and *Cymbidium sinense* (G. Jackson ex H. C. Andrews) Willdenow (612 bp) (Zhang et al. 2013).

63.1% of *C. eugenioides* contigs were similar to proteins described in the NCBI-nr database. The specie with higher similarity was *V. vinifera* with 10706 hits and similar results were also reported analyzing *Coffea* sp. contigs from Sanger ESTs (Mondego et al. 2011). *C. arabica* and *C. canephora* have few sequences deposited in the Genbank database compared to other species with draft genome sequence available, so it is expected that *Coffea* sequences share more similarity with plants from the Asteridae clade (e.g. Solanaceae species) and with *V. vinifera*, in function of high synteny, possibly derived of the same ancestral genome (Mondego et al. 2011; Guyot et al. 2012). Interestingly, we observed that 9.7% contigs had *R. communis* as top hit in our data.

C. eugenioides contigs were analyzed for sequence similarity against the 35133 *C. arabica* contigs database (BLASTN, identity 90%) (Mondego et al. 2011) which approximately 50% sequences had matched. These results reinforce the idea that our study probably contains new potential genes in *Coffea*.

5.3.2 Functional annotation

Several databases of functional annotation were used in order to predict potential genes and their biological functions. Nearly half of *C. eugenioides* contigs found annotations onto nr, Swiss-Prot and GO and these numbers may indicate not-described genes in database. We restricted informations to GO in TAIR database, that represents robust informations associated with plants and notwithstanding it was found considerable information.

GO annotations in *C. eugenioides* transcriptome found terms associated with macromolecule metabolic process and primary metabolic process in the category biological process. These terms could reflect expression associated with basal process, reforcing informations observed by Vidal et al. (2010) that identified contributions of subgenome CaCe in *C. arabica* of proteins associated with citric acid cycle, pentose-phosphate shunt and related with photosynthesis.

Small molecule binding, transferase activity, hydrolase activity, protein binding and nucleic acid binding were annotated in the molecular function category. Mondego et al. (2011) focused their study in this category in *C. arabica* and *C. canephora* and observed similar categories, demonstrated the high similarity among *C. eugenioides* and other species of coffee. However, top domains like small molecule binding and transferase activity may be indicated of process associated with content of sugars and transport. Furthermore, these terms could indicate proteins related sugar metabolism, especially sucrose, as was seen for pathways in KEGG, which identified abundant pathway starch and sucrose metabolism.

Sucrose has an important role in determine the coffee cup quality. It is one of the major resources of the free reducing sugars participating in the Maillard reaction that occurs during the roast of the coffee grain and generate a significant number of products like caramel, sweet and dark colors (Holscher & Steinhart 1995). *C. canephora* accumulates less sucrose than *C. arabica* because *C. canephora* present proteins activities that participate mainly in two stages, early in grain development that prevent the accumulation of sucrose and in the final stage of grain, with less capacity for sucrose re-synthesis (Privat et al. 2008). In the other hand, *C. arabica* produces and accumulates more sucrose during the development of grain comparing with *C. canephora*, and CaCe subgenome could be contribute for this characteristic in the allopolyploid *C. arabica*.

In InterProscan proteins domains, similar results were found between the *C.*

eugenioides transcriptome and sequences of ESTs from *C. arabica* and *C. canephora* (Lin et al. 2005; Mondego et al. 2011). They showed several families in common such as protein kinases, serine threonine kinases, tyrosine kinases, cytochrome P450 monooxygenases and pentatricopeptide repeat. Moreover, most families of predicted proteins found in *Coffea* were serine/threonine kinases, pentatricopeptide repeat and cytochrome P450. Although zinc-finger is among the most abundant proteins in eukaryotic genomes (Laity et al. 2001), it is the first time that this domain is one of the most represented in a *Coffea* transcriptome.

5.3.3 RNA-Seq differential expression and validation by qPCR

A comparative expression analysis of up-regulated contigs in leaves and fruits were done from the choice of gene expression performed in silico. Ten genes were chosen to develop qPCR studies, all contigs exhibited distinct expression patterns between leaves and fruits and the relative transcript abundances in each RNA organ were quantified in relation to the expression of a constitutive transcribed gene (*GAPDH*). We analyzed the DESeq data and it confirmed the previously reported expression profiles in all cases.

In leaves, UDP-Glycosyltransferase (*Ce14433*) was found with greater relative expression comparing to fruits. Vidal et al. (2010) annotated UDP-glucose 4-epimerase with greater abundance in the subgenome CaCe in *C. arabica*. This enzyme carries out the reversible epimerization of UDP-glucose to UDP-galactose, the cognate substrate for galactosyltransferases and could influence the activity of glycosyltransferases (Lee et al. 1999). *C. arabica* domains proteins had a greater percentage of UDP-glucuronosyl transferases than *C. canephora*, proteins related to sugar metabolism and transport, annotations related to better cup quality.

Other gene observed with expression in leaves was *Ce15205,* a germin-like protein. Germins and germin-like proteins are a large gene family with a wide distribution among plants, presents critical roles in plant development and in plant defense response. Peanut plants treated with a serial of biotic and abiotic stresses significantly increase the expression of germin gene in leaves (Wang et al. 2013).

Ce2770 gene was identified as a glucose 6-phosphate/phosphate translocator and it was related with transport of glucose 6-phosphate into plastids for use as a precursor for starch (and fatty acid) biosynthesis and/or as a substrate for the oxidative pentose phosphate pathway (Bowsher et al. 1992). Glucose-6-

phosphate/phosphate translocator was found constitutively present in defined cells of the leaves, such as the bundle sheath cells and stomatal guard cells in *Arabidopsis thaliana* (Kunz et al. 2010).

A chitinase was also annotated with higher expression in leaves (*Ce14847*). A high constitutive level of chitinase activity was detected in the intercellular fluid of healthy leaves in *C. arabica* indicate the participation of these PR proteins in plant defense and during somatic embryogenesis (Fernandez et al. 2004; Guerra-Guimarães et al. 2009).

Leaves (*Ce10671*) and fruits (*Ce14834*) presented high expression of BURP domain-containing protein, subunit of polygalacturonase. BURP domain proteins comprise a broadly distributed, but it is a protein family that in plants are functionally poorly understood. Lin et al. (2005) identified high expression in libraries of *C. canephora* as well as Vidal et al. (2010) in the subgenome CaCe. BURP genes were described with relative expression in the leaf sheath and lamina in rice plant induced by cold and salt stresses (Ding et al. 2009). Additionally, during the fruit ripening, polygalacturonases are involved with increase softness of the fruit (Sitrit et al. 1996) and the enhanced biomass production in cotton (Xu et al. 2012).

Serine carboxypeptidase (*Ce13100*) was identified in fruit and was related to a large family of protein hydrolyzing enzymes that play roles in multiple cellular processes. The physiological role of serine carboxypeptidase remains unknown in plants, but serine carboxypeptidases was involved in brassinosteroid signal transduction (Li et al. 2001) and in *Prunus mume* a serine carboxypeptidase was found with expression during fruit ripening (Mita et al. 2006). Other gene described in this organ, cytochrome P450, family 79, subfamily B, polypeptide 2 (*Ce13451*) are also important in the biosynthesis of lipids, steroids, and other secondary metabolites (Tsukamoto et al. 2004) and it was reported as an important domain protein in InterproScan annotations in this work.

A transcription factor was found highly expressed in fruits (*Ce9246*). This transcription factor was related to *Arabidopsis* locus AT4G18650 that encodes a *DOG1-LIKE 4* (*DOGL4*) transcription factor. *DOGL4* is a member of a small gene family that includes *DOG1*, which functions in control of seed dormancy (Heisel et al. 2013). However, the function of *DOGL4* is currently unknown (Bentsink et al. 2006). The last gene described in fruit (*Ce13525*) belongs to the oxidoreductase, zinc-binding dehydrogenase family protein. It was found with high expression in

subgenome CaCe (Vidal et al. 2010) and also related with light and dark reactions of photosynthesis, which found oxidoreductases more prevalent in *C. arabica* than *C. canephora*.

5.4 Conclusion

To our knowledge, this is the first transcriptome profile in *C. eugenioides* using RNA-Seq and the first large report on the analysis of leaf and fruit genes in this specie. Our data reveal genes in coffee that could explain certain characteristics present in subgenome CaCe of *C. arabica*. We identified genes with expression patterns directly related with basal processes and sugar metabolism, which corroborate with previous informations about differential homeologous expression of the subgenome CaCe in *C. arabica*. Moreover, this report demonstrates several genes without any previous description in literature. This suite of experiments and results present new genes that provide the basis for further gene expression studies in *Coffea* genus and it will also be a potential important tool to coffee breeding programs.

5.5 Methods

5.5.1 Plant materials

Mature fruits and young leaves were harvested from 10-year-old individuals of *C. eugenioides* maintained at the Technological Center of Cooperativa Agropecuária e Industrial (COCARI), Mandaguari, PR, Brazil (latitude (S): 23°30'52''; longitude (W): 51°42' 86"). The region has 650 m height and mean annual temperature of 22-23 °C. All samples were collected in July, 6th, 2011 (between 9 am and 11 am). After collection, the samples were immediately frozen in liquid nitrogen and stored at -80 °C until RNA extraction.

5.5.2 RNA extraction

Leaves and fruits of *C. eugenioides* were ground to a fine powder with liquid nitrogen in using cooled mortar and pestle. Total RNA was isolated following the protocol of Chang et al. (1993). The integrity of RNA samples was examined by 1% agarose gel electrophoresis and the samples were treated with DNase (RNase-free). The quality and the concentration of extracted RNA samples were determined using the NanoDrop® ND-1000 spectrophotometer (NanoDrop, Wilmington, DE) and the absence of contamination with genomic DNA was confirmed by PCR using glyceraldehyde *3*-phosphate dehydrogenase (GAPDH) primers in 100 ng of RNA (data not shown).

5.5.3 RNA sequencing

The mRNA sequencing was performed at the High Throughput Sequencing Facility at the Carolina Center for Genome Sciences (University of North Carolina, USA). For each sample, 10 µg total RNA were used to prepare the mRNAseq library according to the protocol provided by Illumina. The gel extraction step was modified by dissolving excised gel slices at room temperature to avoid the underrepresentation of AT-rich sequences (Quail et al. 2008). Library quality control and quantification were performed using a Bioanalyzer Chip DNA 1000 series II (Agilent). All libraries were tagged and multiplexed in Illumina HiSeq™ 2000, in order to generate 100 base-pair (bp) single-end sequences.

5.5.4 RNA-seq data processing

To obtain high-quality clean read data for *de novo* assembly, the raw reads from mRNA-seq were filtered by discarding the reads with adaptor contamination and regions of low quality (quality <20).The processed reads of both organs were merged and assembled with Trinity assembler (Grabherr et al. 2011), using an optimized k-mer length of 25 for *de novo* assembly. Only contigs >200 bp were used for further annotation.

5.5.5 Annotation and classification of contigs

All contigs were compared using BLASTX against the NCBI non-redundant sequence database (nr), with e-value cutoff of 1e-5 and it was done a BLASTN against the EST database developed by Mondego et al. (2011) in *C. arabica* to verify the representation of our data with available data. The same way, BLASTX with e-value cutoff of 1e-5 was done in Swiss-Prot. Functional annotation describing biological processes, molecular function and cellular component was performed using BLAST2GO version v.2.7.0 (Conesa et al. 2005), considering GO Slim annotations provided by TAIR. Furthermore, InterProscan (Quevillon, et al. 2005) and KEGG database were used to observe protein domains and metabolic pathways, respectively.

5.5.6 Differential expression

Bowtie software (Langmead et al. 2009) was used in default parameters to map the processed reads against the reference *de novo* assembled transcriptome, allowing the maximum of 3 mismatches. The normalized expression level for each contig of the reference transcriptome was expressed as reads per kilobase of transcript sequence per million reads (RPKM) (Mortazavi et al. 2008). After this step, mapped sequence counts were processed using the ‘DESeq’ package (Anders & Huber 2010) to estimate the transcript level. For each contig, the fold change of the expression between the samples was analyzed and the statistical significance estimated using an adjusted p-value (Benjamin-Hochberg method). Differentially expressed genes identified by DESeq were required to have a 2-fold-change and $p \leq 0.05$. Additionally, BLAST2GO was also used for a GO functional enrichment analysis of exclusive contigs of leaves and fruits, by performing Fisher’s exact test with a robust FDR correction.

5.5.7 qPCR and data analysis

Based on transcriptional activity pattern showed by DESeq results, 10 contigs were selected for expression analysis by qPCR to validate the results (Table 2). They were chosen following the criteria: genes with the most fold change normalized (log2foldchange) and a minimum of one read for either leaves and fruits.

Complementaries DNAs (cDNAs) of leaves and fruit organs of *C. eugenioides* were synthesized using SuperScript III Reverse Transcriptase (Invitrogen) following the manufacturer's instructions in a final volume of 20µl using 5µg of total RNA.

The transcript abundance for differentially expressed genes was analyzed by qPCR, in a 7500 Fast Real-Time PCR System (Applied Biosystems) using the SYBR Green PCR Master Mix (Applied Biosystems). Basic procedures of qPCR followed previous publications in coffee plants (Marraccini et al. 2012). The reaction mixture contained 12.5µl of SYBR Green PCR Master Mix, 0,5µl of each primer (5 µM), 1µl of cDNA and 10,5µl of Milli-Q water. qPCR conditions were 95°C for 5 min, followed by 40 cycles of 94°C for 30s, 62°C for 60s and 72°C for 30s and last step of 72°C for 10 min. Melting curves were analyzed to verify the presence of a single product including a negative control. All reactions were performed with three technical replicates and we have followed the minimum information for publication of qPCR experiments (MIQE) according to Bustin et al. (2009).

Data were analyzed to determine cycle threshold (Ct) values. The specificity of the PCR products generated for each set of primers was verified by analyzing the Tm (dissociation) of amplified products. PCR efficiency (E) was determined using LinReg (Ramakers et al. 2003), using only qPCR reactions with efficiency >94%. Expression levels were calculated by applying the formula $(1 + E)^{-\Delta\Delta Ct}$ where $\Delta Ct_{\text{target}} = Ct_{\text{target gene}} - Ct_{\textit{CaGAPDH}}$ and $\Delta\Delta Ct = \Delta Ct_{\text{target}} - \Delta Ct_{\text{reference sample}}$, the calibrator organ was chosen always being the minimum relative valueof samples. Gene expression levels were normalized with the expression of *GAPDH* gene as endogenous control (Barsalobres-Cavallari et al. 2009).

Table 2. Primers designed for candidate genes validation by qPCR analysis

Contig	Transcript AGI	Annotation	E-value	Primer sequence
Ce14433	AT1G22400.1	UDP-Glycosyltransferase superfamily protein	e-165	F: 5' GCCAAGCTCCTCCACCAAA 3' R: 5' GCATCAGGACCGCTGGAT 3'
Ce15205	AT5G20630.1	germin 3	2e-65	F: 5' CTCCAGGGTGCGTGTGAAA 3' R: 5' CGTTCCCTGGTGTGAATGG 3'
Ce2770	AT1G61800.1	glucose-6-phosphate/phosphate translocator 2	e-149	F: 5' GCATTGAGGACCTTCTTGTTGTAG 3' R: 5' TGCAGCGCAGAAGCTTAAGAT 3'
Ce14847	AT5G24090.1	chitinase A	2e-86	F: 5' GGCCAAACACCGGAACTG 3' R: 5' CAGGCTCTGGCAAACCTCTATC 3'
Ce10671	AT1G23760.1	BURP domain-containing protein	2e-36	F: 5' ACGCGTCCAACCATCAATT 3' R: 5' TTCAAAAACTGCCATAGGTGACA 3'
Ce13100	AT4G30810.1	serine carboxypeptidase-like 29	e-156	F: 5' GAGGGCTTGTTTAGGCTTGTGT 3' R: 5' GAGGATGGACTCAGCAGTATGAAG 3'
Ce13451	AT4G39950.1	cytochrome P450, family 79, subfamily B, polypeptide 2	9e-66	F: 5' TGTGCCGGAAAATGAAGGA 3' R: 5' ACGTGGGCTGGCATGTG 3'
Ce9246	AT4G18650.1	transcription factor	5e-66	F: 5' GAAAGGAGTGTTGGATGTGTTGAA 3' R: 5' CTTTTCTTCCCCCATTTTCTCA 3'
Ce13525	AT4G21580.1	oxidoreductase, zinc-binding dehydrogenase family protein	2e-67	F: 5' TGGAGTGAACTTTTTGAACCAGAAT 3' R: 5' TTTACGAATCCCCATGGATCTT 3'
Ce14834	AT1G23760.1	BURP domain-containing protein	9e-27	F: 5' CCCACTAAAACTCTCCGCTAAAAT 3' R: 5' TTTTCTCAACATCGCCTTTTGA 3'
GAPDH	AT1G13440	glyceraldehyde-3-phosphate dehydrogenase		F: 5’ AGGCTGTTGGGAAAGTTCTTC 3’ R: 5’ ACTGTTGGAACTCGGAATGC 3’

5.6 References

Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**:R106.

Anthony F, Diniz LEC, Combes MC, Lashermes P: **Adaptive radiation in *Coffea* subgenus *Coffea* L. (Rubiaceae) in Africa and Madagascar**. *Plant Syst Evol* 2010, **285**:51–64.

Barsalobres-Cavallari CF, Severino FE, Maluf MP, Maia IG: **Identification of suitable internal control genes for expression studies in *Coffea arabica* under different experimental conditions.** *BMC Mol Bio* 2009, **10**:1.

Bentsink L, Jowett J, Hanhart CJ, Koornneef M: **Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*.** *Proc Natl Acad Sci U S A* 2006, **103(45)**:17042-17047.

Blanca J, Cañizares J, Roig C, Ziarsolo P, Nuez F, Picó B: **Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae).** *BMC Genomics* 2011, **12**:104.

Bowsher CG, Boulton EL, Rose J, Nayagam S, Emes MJ: **Reductant for glutamate synthase is generated by the oxidative pentose-phosphate pathway in nonphotosynthetic root plastids.** *Plant J* 1992*,* **2:** 893–898.

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT: **The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.** *Plant Physiol* 2010, **154**:1053-1066.

Cardoso DC, Martinati JC, Giachetto PF, Vidal RO, Carazzolle MF, Padilha L, Guerreiro-Filho O, Maluf MP: **Large-scale analysis of differential gene expression in coffee genotypes resistant and susceptible to leaf miner-toward the**

identification of candidate genes for marker assisted-selection. *BMC Genomics* 2014, **15(1)**:66.

Cenci A, Combes MC, Lashermes P: **Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments.** *Plant Mol Biol* 2012, **78**:135–145.

Chang S, Puryear J, Cairney J: **A simple and efficient method for isolating RNA from pine trees.** *Plant Mol Biol Rep* 1993, **11**:113–116.

Combes MC, Cenci A, Baraille H, Bertrand B, Lashermes P: **Homeologous gene expression in response to growing temperature in a recent Allopolyploid (*Coffea arabica* L.)**. *J Hered* 2012, **103**:36-46.

Combes MC, Dereeper A, Severac D, Bertrand B, Lashermes P: **Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures**. *New Phytologist* 2013, **200**:251-260.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **BLAST2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**: 3674–3676.

Cotta MG, Barros LMG, Almeida JD, De Lamotte F, Barbosa EA, Vieira NG, Alves GSC, Vinecky F, Andrade AC, Marraccini P: **Lipid transfer proteins in coffee: isolation of *Coffea* orthologs, *Coffea arabica* homeologs, expression during coffee fruit development and promoter analysis in transgenic tobacco plants.** *Plant Mol Biol* 2014:1-21.

DaMatta FM, Ramalho JDC: **Impacts of drought and temperature stress on coffee physiology and production: a review.** *Braz J Plant Physiol* 2006, **18**:55-81.

Davis AP, Tosh J, Ruch N, Fay MF: **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* 2011, **167**:357–377.

Ding X, Hou X, Xie K, Xiong L: **Genome-wide identification of BURP domain-containing genes in rice reveals a gene family with diverse structures and responses to abiotic stresses.** *Planta* 2009, **230**:149–163.

Fan H, Xiao Y, Yang Y, Xia W, Mason AS, Xia Z, Mason AS, Xia Z, Qiao F, Zhao S, Tang H: **RNA-Seq Analysis of *Cocos nucifera*: Transcriptome Sequencing and *De Novo* Assembly for Subsequent Functional Genomics Approaches.** *PloS One* 2013, **8**:e59997.

Fernandez D, Santos P, Agostini C, Bon MC, Petitot AS, Silva MC, Guerra-Guimarães L, Ribeiro A, Argout X, Nicole M: **Coffee (*Coffea arabica*L.) genes early expressed during infection by the rust fungus (*Hemileia vastatrix*)**. *Mol Plant Pathol* 2004, **5**:527-536.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F,Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptomeassembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29(7):**644-52.

Guerra-Guimarães L, Silva MC, Struck C, Loureiro A, Nicole M, Rodrigues Jr CJ, Ricardo CPP: **Chitinases of *Coffea arabica* genotypes resistant to orange rust *Hemileia vastatrix.*** *Biol Plantarum* 2009, **53(4)**:702-706.

Guyot R, Lefebvre-Pautigny F, Tranchant-Dubreuil C, Rigoreau M, Hamon P, Leroy T, Hamon S, Poncet V, Crouzillat D, Kochko A: **Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum* sp.) and rosid (*Vitis vinifera*) clades**. *BMC Genomics* 2012, **13**:103.

Heisel TJ, Li CY, Grey KM, Gibson SI: **Mutations in HISTONE ACETYLTRANSFERASE1 affect sugar response and gene expression in *Arabidopsis***. *Front Plant Sci* 2013, **17(4)**:245

Holscher W, Steinhart H: **Aroma compounds in green coffee.** In: *Food Flavors: generation, analysis and process influence*. Edited by Charalambous. Amsterdam: Elsevier; 1995:785-803.

International Coffee Organization [http://www.ico.org]

Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucl Acids Res* 2000, **28**:27–30.

Kunz HH, Häusler RE, Fettke J, Herbst K, Niewiadomski P, Gierth M, Bell K, Steup M, Flügge UI, Schneider A: **The role of plastidial glucose-6-phosphate/phosphate translocators in vegetative issues of *Arabidopsis thaliana* mutants impaired in starch biosynthesis**. *Plant Biology* 2010, **12(s1)**: 115-128.

Laity JH, Lee BM, Wright PE: **Zinc finger proteins: new insights into structural and functional diversity.** *Curr Opin Struc Biol* 2001, **11**:39-46.

Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**:R25.

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A: **Molecular characterisation and origin of the *Coffea arabica* L. genome.** *Mol Gen Genet* 1999, **261:**259–266.

Lee FK, Gibson BW, Melaugh W, Zaleski A, Apicella MA: **Relationship between UDP-Glucose 4-Epimerase Activity and Oligoglucose Glycoforms in Two Strains of *Neisseria meningitidis*.** *Infect Immun* 1999, **67(3):**1405-1414.

Leroy T, Ribeyre F, Bertrand B, Charmetant P, Dufour M, Montagnon C, Marraccini P, Pot D: **Genetics of coffee quality**. *Braz J Plant Physiol* 2006, **18**:229–242.

Li J, Lease KA, Tax FE, Walker JC: **BRS1, a serine carboxypeptidase, regulates BRI1 signaling in *Arabidopsis thaliana*.** *Proc Natl Acad Sci USA* 2001, **98(10)**:5916-5921.

Lin C, Mueller LA, Mc Carthy J, Crouzillat D, Petiard V, Tanksley SD: **Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts**. *Theor Appl Genet* 2005, **112**:114-130.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.

Marraccini P, Freire, LP, Alves, GS, Vieira NG, Vinecky F, Elbelt S, Ramos HJO, Montagnon C, Vieira LGE, Leroy T, Pot D, Silva VA, Rodrigues GC, Andrade AC: ***RBCS1* expression in coffee: *Coffea* orthologs, *Coffea arabica* homeologs, and expression variability between genotypes and under drought stress.** *BMC Plant Biol* 2011, **11**:85.

Mazzafera P, Carvalho A: **Breeding for low seed caffeine content of coffee (*Coffea* L.) by interspecific hybridization**. *Euphytica* 1991, **59**:55–60.

Metzker ML: **Sequencing technologies-the next generation**. *Nat Rev Genet* 2010, **11**:31-46.

Mita S, Nagai Y, Asai T: **Isolation of cDNA clones corresponding to genes differentially expressed in pericarp of mume (*Prunus mume*) in response to ripening, ethylene and wounding signals.** *Physiol Plant* 2006, **128(3):**531–545.

Mondego JMC, Vidal RO, Carazzolle MF, Tokuda EK, Parizzi LP, Costa GGL, Pereira LFP, Andrade AC, Colombo CA, Vieira LGE, Pereira GAG, Brazilian Coffee Genome Project Consortium: **An EST-based analysis identifies new genes and**

reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. *BMC Plant Biol* 2011, **11**:30.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**:621–628.

Plant Metabolic Network (PMN), [http://www.plantcyc.org].

Privat I, Foucrier S, Prins A, Epalle T, Eychenne M, Kandalaft L, Caillet V, Lin C, Tanksley S, Foyer C, McCarthy J: **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 Phytol* 2008, **178:**781-797.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: **A large genome center's improvements to the Illumina sequencing system.** *Nat Methods* 2008, **5**:1005–1010.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic Acids Res* 2005, **33**:W116-W120.

Ramakers C, Ruijter JM, Deprez RH, Moorman AF: **Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data**. *Neurosci Lett* 2003, **339(1)**:62-66.

Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T, Wan XC: **Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds.** *BMC Genomics* 2011, **12**:131.

Sitrit Y, Hadfield KA, Bennett AB, Bradford KJ, Downie AB: **Expression of a polygalacturonase associated with tomato seed germination.** *Plant Physiol* 1999, **121(2)**:419-28.

The Swiss-Prot Database [http://www.uniprot.org/downloads]

The TAIR Database: The Arabidopsis Information Resource [http://www.arabidopsis.org]

Toledo-Silva G, Cardoso-Silva CB, Jank L, Souza AP: ***De novo* transcriptome assembly for the tropical grass *Panicum maximum* Jacq**. *PloS One* 2013, **8**:e70781.

Tsukamoto S, Tomise K, Aburatani M, Onuki H, Hirorta H, Ishiharajima E, Ohta T: **Isolation of cytochrome P450 inhibitors from strawberry fruit, *Fragaria ananassa*.** *J Nat Prod* 2004, **67:**1839-1841.

Vidal RO, Mondego JM, Pot D, Ambrósio AB, Andrade AC, Pereira LF, Colombo CA, Vieira LG, Carazzolle MF, Pereira GA: **A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*.** *Plant Physiol* 2010, **154**:1053-1066.

Wang T, Chen X, Zhu F, Li H, Li L, Yang Q, Chi X, Yu S, Liang X: **Characterization of Peanut Germin-Like Proteins, *AhGLPs* in Plant Development and Defense.** *PLoS One* 2013 **8**:e61722.

Xu B, Gou JY, Li FG, Shangguan XX, Zhao B, Yang CQ, Wang LJ, Yuan S, Liu CJ, Chen XY: **A cotton BURP domain protein interacts with α-expansin and their co-expression promotes plant growth and fruit production.** *Mol Plant* 2013, **6(3):**945-58.

Yu Q, Guyot R, de Kochko A, Byers A, Navajas-Pérez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, 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* 2011, **67(2)**:305-317.

Zhang J, Wu K, Zeng S, Silva JAT, Zhao X, Tian CE, Xia H, Duan J: **Transcriptome analysis of *Cymbidium sinense* and its application to the identification of genes associated with floral development**. *BMC Genomics* 2013, **14**:279.

ADDITIONAL FILES I

Additional File 1. Species distribution of 100 BLASTX top hits against available non redundant sequences from the NCBI database (NCBI-nr).

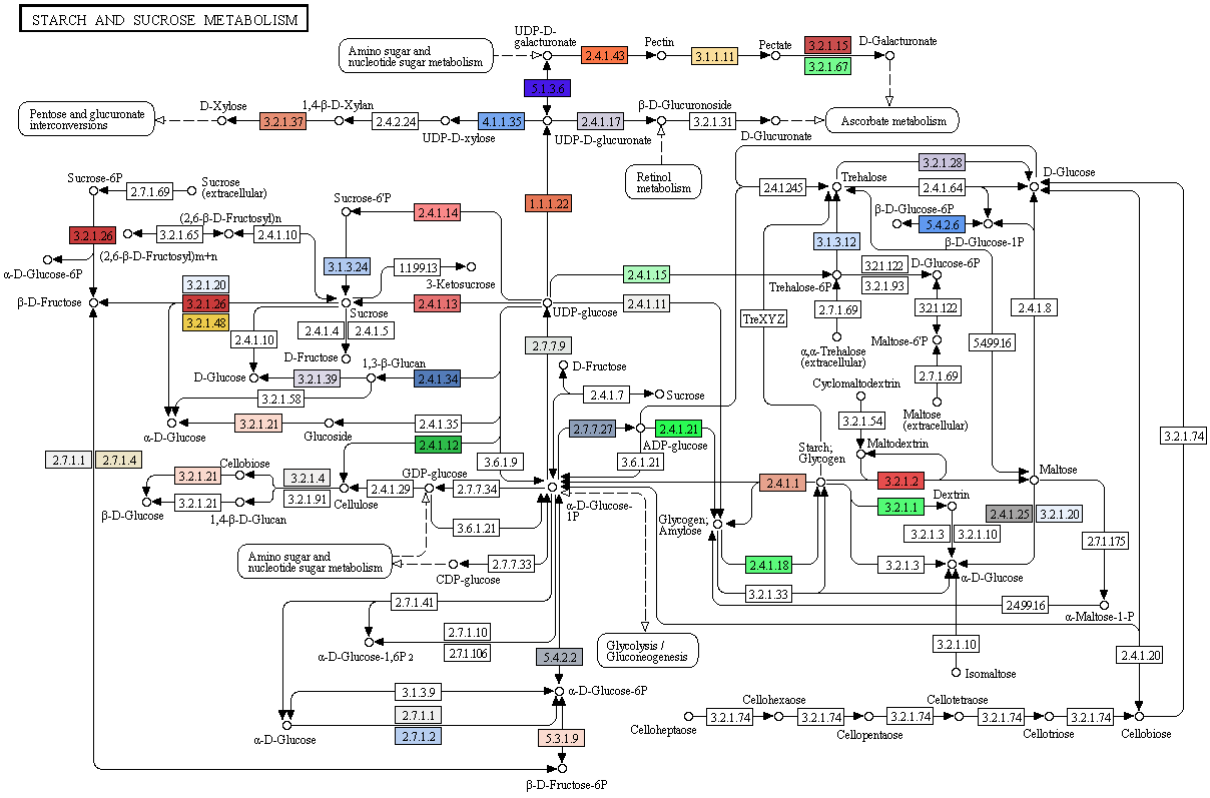
Species	Hits number	% Hits
Vitis vinifera	10706	29,4%
Ricinus communis	3546	9,7%
Populus trichocarpa	3173	8,7%
Nicotiana tabacum	452	1,2%
Glycine max	369	1,0%
Solanum lycopersicum	327	0,9%
Arabidopsis lyrata	309	0,8%
Arabidopsis thaliana	232	0,6%
Solanum tuberosum	223	0,6%
Pyrenophora teres	217	0,6%
Leptosphaeria maculans	160	0,4%
Phaeosphaeria nodorum	156	0,4%
Pyrenophora tritici-repentis	147	0,4%
Oryza sativa	144	0,4%
Coffea arabica	143	0,4%
Mycosphaerella graminicola	139	0,4%
Glomerella graminicola	121	0,3%
Catharanthus roseus	100	0,3%
Medicago truncatula	92	0,3%
Sorghum bicolor	79	0,2%
Solanum demissum	73	0,2%
Nicotiana benthamiana	72	0,2%
Coffea canephora	58	0,2%
Petunia x hybrida	55	0,2%
Capsicum annuum	52	0,1%
Glossina morsitans	51	0,1%
Gossypium hirsutum	50	0,1%
Cucumis melo	49	0,1%
Zea mays	44	0,1%
Hordeum vulgare	43	0,1%

Jatropha curcas	43	0,1%
Malus domestica	40	0,1%
Camellia sinensis	40	0,1%
Drosophila virilis	34	0,1%
Beta vulgaris	33	0,1%
Ipomoea nil	33	0,1%
Drosophila willistoni	27	0,1%
Drosophila ananassae	26	0,1%
Drosophila grimshawi	26	0,1%
Drosophila mojavensis	24	0,1%
Drosophila melanogaster	24	0,1%
Nicotiana sylvestris	23	0,1%
Antirrhinum majus	22	0,1%
Ceratitis capitata	20	0,1%
Nicotiana plumbaginifolia	20	0,1%
Picea sitchensis	19	0,1%
Olea europaea	19	0,1%
Sclerotinia sclerotiorum	18	0,0%
Actinidia deliciosa	17	0,0%
Nicotiana attenuata	16	0,0%
Botryotinia fuckeliana	16	0,0%
Solanum chacoense	16	0,0%
Daucus carota	16	0,0%
Drosophila pseudoobscura	16	0,0%
Capsicum chinense	15	0,0%
Cucumis sativus	14	0,0%
Ipomoea batatas	14	0,0%
Selaginella moellendorffii	14	0,0%
Saccharomyces cerevisiae	14	0,0%
Nectria haematococca	14	0,0%
Sesamum indicum	14	0,0%
Coffea spp*. mixed genomic*	14	0,0%
Solanum pennellii	14	0,0%

Panax notoginseng	13	0,0%
Physcomitrella patens	13	0,0%
Hevea brasiliensis	12	0,0%
Davidiella tassiana	12	0,0%
Elaeis guineensis	11	0,0%
Trichoderma reesei	11	0,0%
Melampsora larici-populina	10	0,0%
Salvia miltiorrhiza	10	0,0%
Drosophila simulans	10	0,0%
Drosophila erecta	10	0,0%
Verticillium albo-atrum	10	0,0%
Panax ginseng	10	0,0%
Ipomoea trifida	10	0,0%
Gibberella zeae	10	0,0%
Lotus japonicus	9	0,0%
Solanum peruvianum	9	0,0%
Penicillium marneffei	9	0,0%
Thellungiella halophila	9	0,0%
Rauvolfia serpentina	8	0,0%
Carica papaya	8	0,0%
Solanum bulbocastanum	8	0,0%
Triticum aestivum	8	0,0%
Helianthus annuus	8	0,0%
Gardenia jasminoides	8	0,0%
Fragaria x ananassa	7	0,0%
Drosophila persimilis	7	0,0%
Aspergillus terreus	7	0,0%
Prunus persica	7	0,0%
Solanum nigrum	7	0,0%
Lycium barbarum	7	0,0%
Ajellomyces capsulatus	7	0,0%
Talaromyces stipitatus	7	0,0%
Populus tomentosa	7	0,0%

Gentiana triflora	7	0,0%
Neosartorya fischeri	7	0,0%
Pyrus communis	7	0,0%
Camellia oleifera	6	0,0%

Additional File 2. *Coffea eugenioides* contigs involved in the starch and sucrose pathway. All colored E.C. numbers (square) are a contig/contig with similarity (>1e-5) to their respective protein described for this metabolic pathway (http://www.plantcyc.org).



6. ARTIGO 2

SNP DETECTION AND GENOTYPING IN WILD TYPE *Coffea arabica* COLLECTION

Abstract

Coffea arabica is an allotetraploid specie of coffee originated from a recent hybridization of two diploid species, *C. canephora* and *C. eugenioides*. *C. arabica* has lower genetic diversity as consequence of its autogamy and recent evolutionary history. RNA-sequencing (RNA-Seq) produces a large volume of data and allows rapid identification of a large number of genetic markers, mainly single nucleotide polymorphisms (SNPs). In this study, RNA-Seq using Illumina sequencing of *C. arabica* genotypes from wild Ethiopian genotypes of *C. arabica* and it ancestors, *C. canephora* and *C. eugenioides*, generated 111,664,604 reads. Computational analysis allowed discovering 382,474 polymorphisms. According filters established in this study, considering quality and depth, as well as to separate SNPs related to the allopolyploid subgenomes, 1,410 SNPs were selected with potential to be used in genotyping. From those, 469 SNPs were validated in 135 genotypes of wild *C. arabica* from Ethiopia and cultivars. Using Sequenom MassARRAY system, 311 SNPs (66,4 %) were able to detect polymorphism between the genotypes used. Further diversity analysis through Structure software revealed no significant associations between the groups according with their geographical origin. Furthermore, the cultivars were separated in a specific group. The identification of SNPs in RNA-Seq data associated with analysis of a population with variability phenotypic was effective in the identification of SNPs. Additionally, this work can orientate future genetic studies of Genome Wide Selection Studies using genotypes of *C. arabica* from the Ethiopia wild type collection.

Keywords: *Coffea arabica*, genotyping, single nucleotide polymorphism, RNA-seq, wild genotypes.

6.1 Introduction

Coffee is one of the most important agricultural commodities worldwide. The genus *Coffea* has 124 species (Davis et al. 2011) and it is relatively young, approximately 500 thousand years ago (Anthony et al. 2010). *Coffea* species are diploid except *Coffea arabica* L., which is an allotetraploid (2*n* = 4 × = 44) and derived from a recent interspecific hybridization between two diploid species, *C. canephora* Pierre ex A. Froehner and *C. eugenioides* S. Moore (2*n* = 2 × = 22) (Lashermes et al. 1999). This hybridization occurred probably in the plateaus of Ethiopia (Lashermes et al. 1999) between 100-500 thousand years ago (Anthony et al. 2010) although there was a recent estimation of the origin, approximately 10-50 thousand years ago (Cenci et al. 2012).

C. arabica presents a narrow genetic diversity, mainly due its biological characteristics, such as autogamy, perennial plant and its recent evolutionary history. Furthermore, the main cultivated genotypes, including Catuai, Caturra and Mundo Novo, were selected from only two base populations, Typica and Bourbon (Anthony et al. 2002; Labouisse et al. 2008). The narrow genetic base of those cultivars (Anthony et al. 2002) has resulted in a crop with homogenous agronomic behavior (Lashermes et al. 2009), but also with a high susceptibility to biotic and climatic hazards, and a low adaptability in response to environmental changes or changing of market demands (Labouisse et al. 2008; Jaramillo et al. 2011).

The limited diversity in cultivars of coffee hinders the identification of genes/alleles that provide resistance to biotic/abiotic stress and cup quality, making necessary the exploitation of new sources of genetic diversity in *C. arabica*. Wild *C. arabica* population are a valuable source to genetic analysis and breeding of cultivars, including varieties with differences in the content of coffee biochemistry compounds such as caffeine, sucrose, chlorogenic acids, diterpenes as well as resistance to pests and pathogens such as root nematodes, coffee leaf rust and coffee berry disease (Hein and Gatzweiler 2006; Aerts et al. 2013). The Instituto Agronômico do Paraná (IAPAR), Londrina, Brazil, has a collection of 132 accessions of *C. arabica* plants originally collected during an expedition to Ethiopia in 1964-1965, organized by the Food and Agriculture Organization of the United Nations (FAO 1968). Seed samples were divided up and grown in several countries including Brazil.

The low diversity of the *C. arabica* has been demonstrated in studies with different molecular markers (Lashermes et al. 1999; Anthony et al. 2001; Silvestrini et al. 2007). To overcome these problems, genetic markers as single nucleotide polymorphisms (SNPs) constitute a powerful tool for mapping and marker-assisted breeding (Blanca et al. 2012). SNPs are the most abundant variations in genomes and they have been replacing the microsatellites in many model and non-model plants for building and saturating genetic maps (Deleu et al. 2009; Rafalski et al. 2002). In coffee, Vieira et al (2006) and Lin et al. (2005) generated above 267,000 ESTs from *C. arabica* and *C. canephora* and nearly 47,000 ESTs from *C. canephora*, respectively, using the Sanger sequencing. These data were assembled and used to discovery SNPs between the cultivars Mundo Novo and Catuai, as well as between *C. canephora* and *C arabica* (Vidal et al. 2010). A total of 25,133 SNPs were identified within *Coffea* EST databases, but the analysis in two *C. arabica* cultivars did not allow the detection of polymorphisms between them. Furthermore, polymorphisms within subgenomes (589 in subgenome *C. canephora* - CaCc and 371 in subgenome *C. eugenioides* - CaCe) were not specific to one of cultivars that suggest the maintenance of a residual subgenome heterozigosity.

Next-generation sequencing (NGS) have been enabled the production of large volume of data, cost-effective and robust results. Several works have been done to identify a high number of SNPs in transcriptome NGS data with purposes of genetic maps (Metzker et al. 2010; Blanca et al 2011). Those technologies were used in order to identify thousands of SNPs in transcriptome data of cotton, alfalfa and melon (Blanca et al. 2011; Byers et al. 2011; Yang et al. 2011).

In this work, we employed Illumina RNA sequencing (RNA-Seq) of leaves and fruits samples of *C. arabica* genotypes from an Ethiopia collection to identify SNPs. Our objectives were identified large-set SNPs markers in this allopolyploid and validate SNPs using Sequenom MassARRAY system. This work reports the discovery and use of a large number of SNPs in *C arabica*.

6.2 Materials and methods

6.2.1 Plant materials

C. arabica genotypes were collected in Ethiopia by FAO between the years 1964-1965 (FAO 1968) and samples were sent to six internationals institutes (India, Tanzania, Ethiopia, Costa Rica, Peru and Portugal). Seeds of these accessions were transferred to the Instituto Agronômico de Campinas (IAC), Campinas, Brazil, which provided 132 genotypes to IAPAR.

Previous phenotyping and microsatellite analysis of genotypes permited to select four diverse genotypes of *C. arabica* from the Ethiopia collection (data not shown): E-007/087, E-123A/231, E-238/022 and E-516/069. Samples of young leaves and mature fruits were harvested of these four genotypes and one genotype of *C. arabica* cv. Mundo Novo at Instituto Agronômico do Paraná (IAPAR), Londrina, PR, Brazil (latitude (S): 23°21', longitude (W): 51°9'). These samples were collected in June, 5th, 2011 (between 9 am and 12 am). Three genotypes of *C. eugenioides* were collected at the Technological Center of Cooperativa Agropecuária e Industrial (COCARI), Mandaguari, PR, Brazil (latitude (S): 23°30'52''; longitude (W): 51°42' 86"). These samples were collected in July, 6th, 2011 (between 9 am and 11 am). After collection, the samples were immediately frozen in liquid nitrogen and stored at -80 °C until RNA extraction.

6.2.2 RNA extraction

Leaves and fruits of *Coffea* genotypes were ground to a fine powder with liquid nitrogen using cooled mortar and pestle. Total RNA was isolated following the protocol of Chang et al. (1993). The integrity of RNA samples was examined by 1 % agarose gel electrophoresis and the samples were treated with DNase (RNase-free). The quality and the concentration of extracted RNA samples were determined using the NanoDrop® ND-1000 spectrophotometer (NanoDrop, Wilmington, DE) and the absence of contamination with genomic DNA was confirmed by PCR using glyceraldehyde *3*-phosphate dehydrogenase (GAPDH) primers in 100 ng of RNA (data not shown).

6.2.3 RNA sequencing

The mRNA sequencing was performed at the High Throughput Sequencing Facility at the Carolina Center for Genome Sciences (University of North Carolina, USA). For each sample, 10 µg of total RNA were used to prepare the mRNAseq library according to the protocol provided by Illumina HiSeq^{TM}2000. The gel extraction step was modified by dissolving excised gel slices at room temperature to avoid the underrepresentation of AT-rich sequences (Quail et al. 2008). Library quality control and quantification were performed using a Bioanalyzer Chip DNA 1000 series II (Agilent). Twelve libraries were barcoded and could be multiplexed on a lane in Illumina HiSeq™2000, in order to generate 100 base-pair (bp) single-end sequences. *C. canephora* library was provided ARCAD group, at CIRAD - La recherche agronomique pour le développement (unpublished data).

6.2.4 RNA-Seq data processing and detection of single nucleotide polymorphisms

The *C. canephora* reads were preprocessed to remove adapters and reads with low quality scores. After, reads were assembled using ABySS v 1.2.1 (Simpson et al. 2009) followed by one step of CAP3 v.1 program (Huang and Madan 1999). This *C. canephora* assemble was used as a reference for the SNPs discovery. For other samples, reads were processed to remove adapters, reads with low quality and after, processed sequences were aligned in the *C. canephora* reference. For SNP discovery, it was used the GATK toolkit (McKenna et al. 2010) using the Unified Genotyper module to obtain SNPs list and allelic data.

VariantFiltration module of GATK was used to characterize polymorphisms in terms of quality (PASS, SnpCluster, LowQual, HARD_TO_VALIDATE) (Table 1). Only polymorphisms annotated "PASS" done part of the pre-selection.

Table 1. Terms definition of VariantFiltration module of GATK.

Filters definition	Definition
HARD_TO_VALIDATE	Polymorphisms cannot be classified as PASS because information was insufficient
LowQual	Score Phred very low (<40)
SnpCluster	Three polymorphisms or more in 10 bp
QD_filter	QD filter <1,5, quality by depth below 1,5
PASS	Score Phred > 40

The follow filter criteria were established to select true SNPs: i) SNPs with ancestors informations, *C. canephora* and *C. eugenioides*; ii) Q >1000 (quality); iii) only SNPs represented with a minimum depth of 100× in both five *C. arabica* genotypes (sum of the depth of all five *C. arabica* genotypes); iv) minimum depth 10× in each *C. arabica* genotype; v) only the sites for which at least four *C. arabica* genotypes were available; vi) when only the information of four *C. arabica* genotypes were available, have a minimum of two-fold detected.

6.2.5 Genotyping of SNPs

DNA was extracted from leaves from 135 plants: 128 of *C. arabica* from Ethiopia, 5 commercial genotypes of *C. arabica* (Bourbon, Catuai Vermelho, Mundo Novo, IAPAR 59 and Typica) and two parents of a mapping population, (E-335/219 and Catuai) and Tupi (Additional File 1). Samples were ground in liquid nitrogen according Doyle and Doyle (1990). DNA concentrations were quantified with a NanoDrop ND-1000 spectrophotometer (Thermo Scientific) and diluted to 20-25 ng/µl.

Primers were designed using the Sequenom MassARRAY Assay Design software. A total of 469 SNPs were genotyped in the 135 plants described above using the Sequenom MassARRAY system (Sequenom, San Diego, CA, USA) at the Iowa State University Genomic Technologies Facility, USA. The SNP assay is based on the single-base extension of locus-specific primers followed by mass spectrometry to detect polymorphisms, yielding allele-specific information (Gabriel et al. 2009). The mass intensities should be proportional to the abundance of each allele.

6.2.6 Statistical analysis

The model-based clustering method Structure 2.3.4 (Pritchard et al. 2000) was used for detecting the population structure of the different genotypes and the multilocus data. We tested 135 genotypes and 40 SNPs to show an overview data. Structure was executed using statistic method ΔK (Evanno et al. 2005), 20 runs to each value *K*, application of 10,000 cycles and 10,000 interactions of Markov Chain Monte Carlo (MCMC) to analysis. Structure was run to test the hypothesis of two and five (*K* = 2-5) sub-populations. STRUCTURE HARVESTER was used to assess and visualize likelihood values across multiple values of *K* and detection of the number of genetic groups that best fit the data (Earl et al. 2012).

6.3 Results and discussion

Using Illumina sequencing, we generated a total of 98,635,514 reads of 100 bp, single-end protocol from 5 genotypes of *C. arabica* (cv. Mundo Novo and four wild *C. arabica* genotypes) and *C. eugenioides*, in addition to 13,029,090 reads of *C. canephora*, which were used to construct a reference (Table 2).

Table 2. Description of *Coffea* sequences obtained with Illumina.

Genotypes	Organ	Reads	Total reads
C. arabica (E-007/087)	Leaf	3,244,131	4,695,631
	Fruit	1,451,500	
C. arabica (E-123A/231)	Leaf	14,057,294	22,878,125
	Fruit	8,820,831	
C. arabica (E-238/022)	Leaf	19,576,566	38,097,601
	Fruit	18,521,035	
C. arabica (E-516/069)	Leaf	3,326,834	14,536,621
	Fruit	11,209,787	
C. arabica cv. Mundo Novo	Leaf	3,835,373	9,992,123
	Fruit	6,156,750	
C. eugenioides	Leaf	3,688,364	8,435,413
	Fruit	4,747,049	
*C. canephora**	Leaf	13,029,090	13,029,090

*Reads provided by ARCAD project/CIRAD

After processing, alignment of pooled Illumina sequences to the *C. canephora* map resulted in 52,318,179 sequences aligned. The identification of polymorphic sites was done with GATK toolkit which revealed 382,474 polymorphisms. From this, 334,273 polymorphisms were classified as PASS by GATK, once the Variant Call Format (VCFs) carry filters to each polymorphism and the polymorphisms with good quality have PASS in their FILTER field. A summary of analysis is represented through a pipeline (Figure 1).

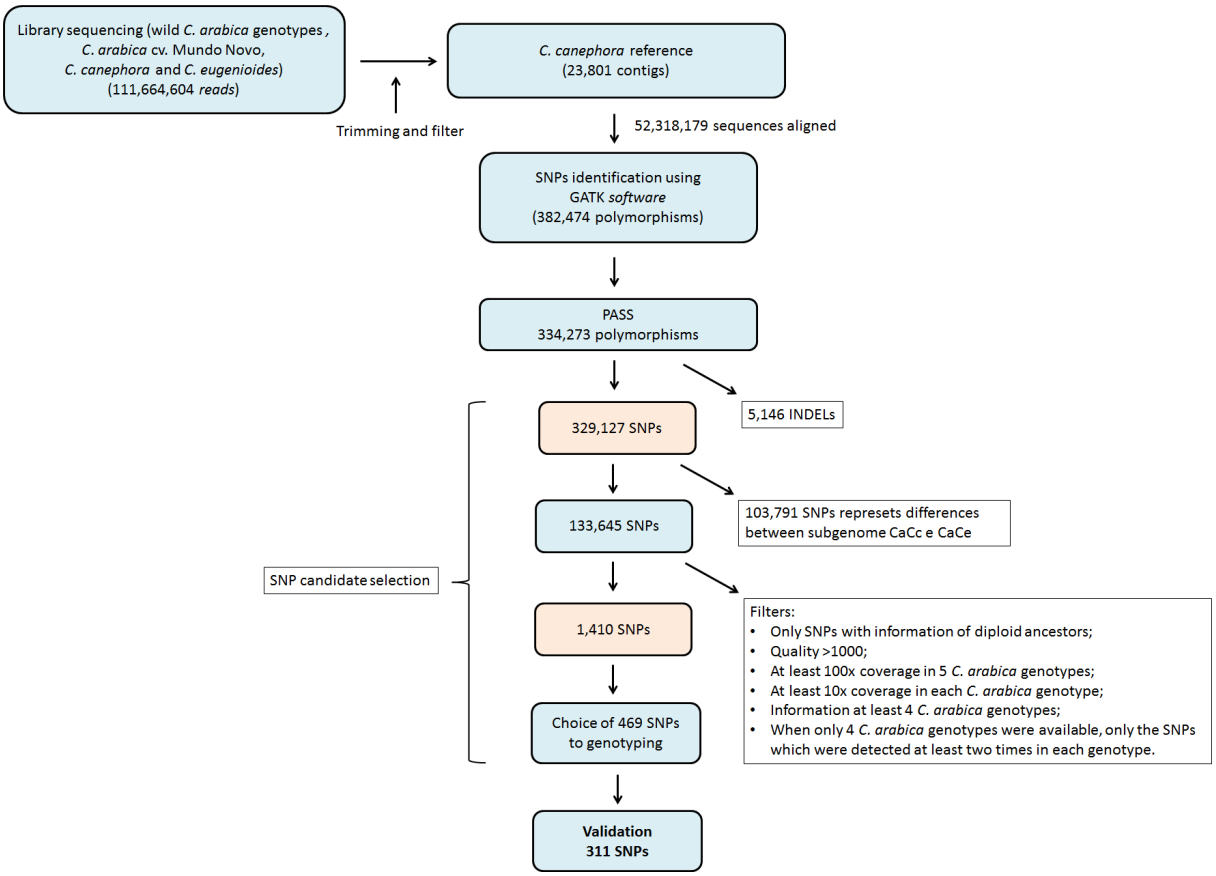


Figure 1. Pipeline description for data cleaning, assembly, and filters used to SNPs selection to genotyping.

Initially, 329,127 SNPs and 5,146 INDELs (insertion/deletion) were discovered. From those, 133,645 polymorphisms did not show differential expression between homeologs *C. arabica* genotypes. From the VCF files, we detected SNPs and established criteria to choose reliable SNP sites to the genotype following the steps described in the material and methods section. This higher confidence alignment resulted in the identification of 1,410 SNPs potential SNPs in five *C. arabica* genotypes within 911 contigs. The discovery of SNPs in polyploid species, such as the allotetraploid *C. arabica*, is more difficult than in diploid species because a gene may be represented in a determinate number in different alleles. For alleles present in a single dose (or low dose), direct re-sequencing of PCR products using conventional Sanger sequencing methods could fail in detect the corresponding SNP and INDELs or also create uninterpretable sequences. Thus, cloned products sequencing are necessary to detect these SNPs, which enhance costs in function of using conventional dideoxy methods (Bundock et al. 2009). By contrast, NGS is an effectible tool to discover polymorphisms in polyploidy organism due the production

of a relatively large number of sequences and the possibility to achieve a sequence depth to identified true SNPs, including SNPs with low dosage, i.e., those present on one or a few homologues, and it can be discriminated from sequencing errors. Another drawback for SNPs discovery in *C. arabica* is the lack of a true genome reference. The strategy to map the *C arabica* and *C. eugenioides* reads in the *C. canephora* reference, allow us to remove part of the SNPs which correspond to the difference between the two subgenomes in *C. arabica*, and identify SNPs which could be used for genotyping.

Our work identified an effective number of SNPs, 1,410 SNPs, using reliable parameters that did not correspond the differences between homeologs - CaCc and CaCe. Others studies developed in *C. arabica* cultivars found a large number of SNPs that represent homeologs SNPs between CaCc and CaCe (Vidal et al. 2010, Combes et al. 2013). Using EST Sanger database, Vidal et al. (2010) found several SNPs within the CaCc that were related to *C. canephora* polymorphisms, but none of them were specific to Catuai and Mundo Novo. The low diversity of *C. arabica* data used as well as the lack of depth in sequencing could have contributed for those results.

In our work, several strategies contributed substantially to the increase the identification of SNPs number in *C. arabica*. Informations of two ancestors of *C. arabica*, *C. canephora* and *C. eugenioides*, were used to identify SNPs in wild *C. arabica* collection. Thus, it was possible identify SNPs that represent differences between the subgenomes CaCc and CaCe and SNPs within subgenomes. Large dataset of the transcriptome reference of *C. canephora* allowed the *C. arabica* and *C. eugenioides* sequences be aligned and increase the data representation and consequently the number of SNPs. Finally, wild *C. arabica* genotypes of this population showed the potential resources in search of genetic markers.

Although there are difficult to find SNPs in organism polyploid, studies with NGS have been reported for other species. In sugarcane, an approach to search of polymorphism in target genes obtained 1,013 and 1,632 SNPs between two varieties from sugarcane. In alfalfa (*Medicago sativa* (L.) *sativa*), a total of 10,826 SNPs were identified between the two genotypes using Illumina GA-II platform (Yang et al. 2011). In cotton, a total of 11,834 SNPs in genic regions and 1,679 SNPs in intergenic regions were identified between accessions of *Gossypium hirsutum* L. and

Gossypium barbadense L., respectively, in genome reduction assemblies (Byers et al. 2012).

Once we defined a set of SNPs with high confidence, we chose arbitrarily 469 SNPs to genotype. Sequenom MassARRAY iPLEX assays were designed for genotyping and these assays were used to analysis the genomic DNA of 135 *C. arabica* individuals, including Ethiopian colletion of IAPAR, commercials genotypes and two parental of a crossing population. A set of 311 SNPs of the evaluated SNPs were validated.

The nature of the organism in study, number of polymorphisms, the type of information to be obtained, customs and previous studies that it have been developed, i.e., the available database, are factors that define the strategies to SNPs identification and genotyping (Tabassum and Lakhanpaul 2006). Sequenom was used in other allopolyploids species such as alfalfa (Yang et al. 2011) and *Tragopogon miscellus* Ownbey (Buggs et al. 2012). In alfafa, out of 55 SNPs randomly selected for experimental for validation by Sequenom, 47 (85%) were polymorphic between the two genotypes (Yang et al. 2011). In other polyploids, for example, *Jatropha curcas* L., a total of 78 putative SNPs were validated and 60 SNPs provided results consistents among 148 global collections of *J. curcas* lines (Gupta et al. 2011).

Structure analysis was used to determine the underlying groups (*K*) in the collection. The *K* value determines the number of sub-populations and the best results were to 4 sub-populations (*K* = 4), according statistical analysis of STRUCTURE HARVESTER (Figure 2).

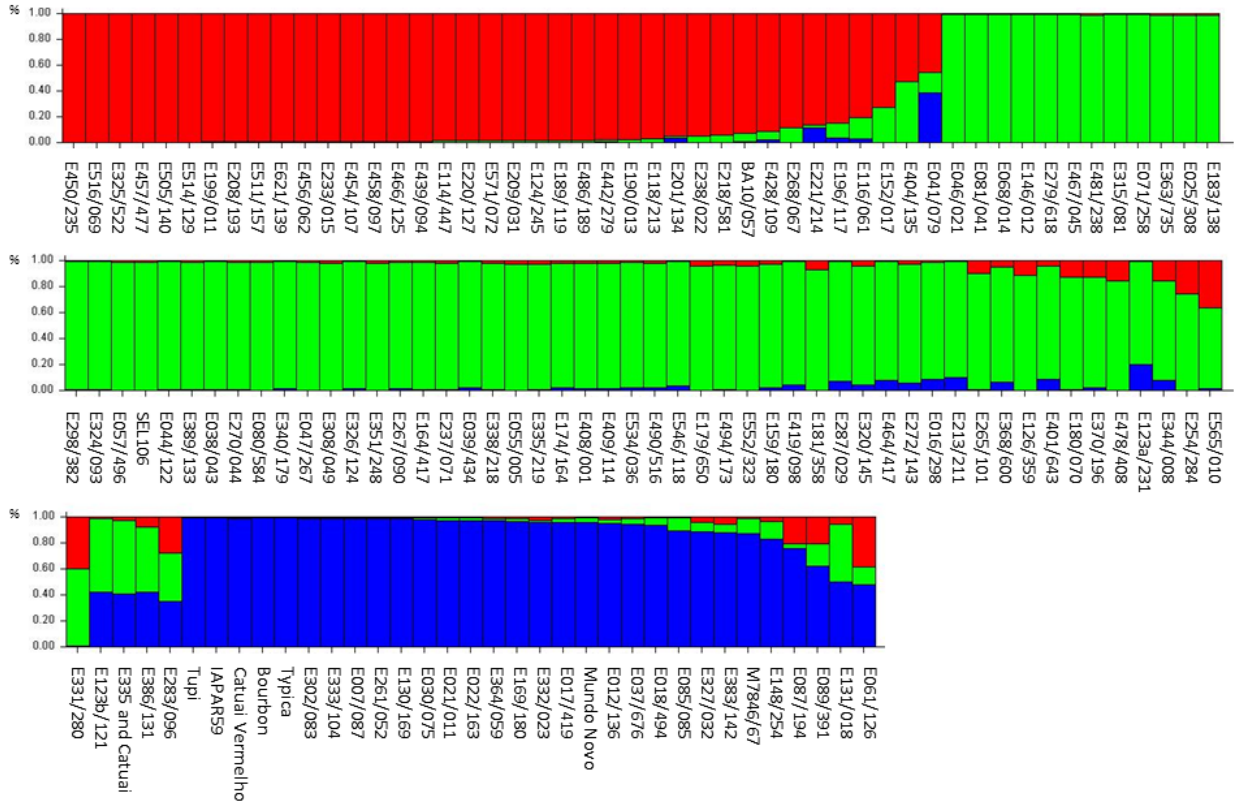


Figure 2. Population structure of *C. arabica* from Ethiopia, 5 commercial genotypes of *C. arabica* (Bourbon, Catuai Vermelho, Mundo Novo, IAPAR 59 and Typica) and two parents of a mapping population (E-335/219 and Catuai) and Tupi. Value of *K* (number of clusters) = 3.

There were no significant separation of groups according with their geographical origin (Montagnon and Bouharmont 1996, Silvestrini et al. 2007), which observed two groups separated among the accessions of *C. arabica*, from Ethiopia, one group of east and the other group from west of Rift Valley. However, our results showed the discrimination between genotypes of *C. arabica* and commercial genotypes and demonstrate the potential use of analyzed population in future studies of association of markers with phenotypic characteristics.

Population structure is an important association factor in association mapping studies (Breseghello and Sorrells 2006). Association mapping has been used as the method of choice for identifying loci involved in the inheritance of complex traits (Risch and Merikangas 1996). This method involves the identification of markers with significant allele-frequency differences among individuals associated with the phenotype of interest and a set of individuals. A statistical association between genotypes at a marker locus and the phenotype is usually considered to be evidence

of close physical linkage between the marker and a locus. Individuals used in association mapping studies can be grouped by the level of population structure and within-group relatedness. The complex evolutionary and breeding history in maize (Flint-Garcia et al. 2005), *Arabidopsis thaliana* (L.) Heynh. (Nordborg et al. 2005) and rice (Garris et al. 2005) has undoubtedly created both population structure and complex familial relationships. To reduce this risk of errors, estimates of population structure must be included in association analysis.

The work demonstrated that RNA-seq is an effective technology to identify a large number of valid SNPs with great efficiency and accuracy. The SNPs identified were abundant in different species and they could assist in the generation of dense genetic maps in *C. arabica*. This study also demonstrated the importance and potential of the Ethiopian coffee collection as a resource to future studies in coffee to genetic improvement of *C. arabica* using Marker Assisted Selection (MAS), Genome Wide Selection (GWS) or even Genome-Wide Association Studies (GWAS)

6.4 References

Aerts R., Berecha G., Gijbels P., Hundera K., Glabeke S., Vandepitte K., Muys B., Roldán-Ruiz I., Honnay O. 2013. Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian montane rainforests. Evol Appl 6:243-252.

Anthony F.; Bertrand B.; Quiros O.; Wilches A.; Lashermes P.; Berthaud J.; Charrier A. 2001. Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. Euphytica 118:53-65.

Anthony F., Combes M.C., Astorga C., Bertrand B., Graziosi G., Lashermes P. 2002. The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. Theor Appl Genet 104:894-900.

Anthony F., Diniz L.E.C., Combes M.C., Lashermes P. 2010. Adaptive radiation in *Coffea* subgenus *Coffea* L. (Rubiaceae) in Africa and Madagascar. Plant Syst Evol 285:51–64.

Blanca J.M., Cañizares J., Ziarsolo P., Esteras C., Mir G., Nuez F., Garcia-Mas J., Picó M.B. 2011. Melon transcriptomec: simple sequence repeats and single nucleotide polymorphisms discovery for high throughput genotyping across the species. The Plant Genome 4:118–131.

Blanca J., Cañizares J., Roig C., Ziarsolo P., Nuez F., Picó B. 2011. Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). BMC Genomics 12:104.

Breseghello F., Sorrells M.E. 2006. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177.

Buggs R.J., Renny-Byfield S., Chester M., Jordon-Thaden I.E., Viccini L.F., Chamala S., Leitch A.R., Schnable P.S., Barbazuk W.B., Soltis P.S., Soltis D.E. 2012. Next-generation sequencing and genome evolution in allopolyploids. Am J Bot 99:372-82.

Bundock P.C., Eliot, F.G., Ablett G., Benson A.D., Casu R.E., Aitken K.S., Henry R.J. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J, 7:347-354.

Byers R.L., Harker D.B., Yourstone S.M., Maughan P.J., Udall J.A. 2012. Development and mapping of SNP assays in allotetraploid cotton. Theor Appl Genet 124:1201-14.

Cenci A., Combes M.C., Lashermes P. 2012. Genome evolution in diploid and tetraploid *Coffea* species. Plant Mol Biol. 78:135–145.

Chang S., Puryear J., Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep.11:113–116.

Combes M.C., Dereeper A., Severac D., Bertrand B., Lashermes P. 2013. Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. New Phytol 200:251-260.

Davis A.P., Tosh J., Ruch N., Fay M.F. 2011. Growing coffee: *Psilanthus* (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of *Coffea*. Bot J Linn Soc 167:357–377.

Deleu W., Esteras C., Roig C., González-To M., Fernández-Silva I., Blanca J., Aranda M.A., Arús P., Nuez F., Monforte A.J., Picó M.B., Garcia-Mas J. 2009. A set of EST-SNPs for map saturation and cultivar identification in melon. BMC Plant Biol 9:90.

Doyle J.J., Doyle J.L 1990. Isolation of plant DNA from fresh tissue. Focus 12:13-15.

Earl D.A., vonHoldt B.M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359-361.

Evanno G., Regnaut S., Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-20.

FAO. 1968. FAO Coffee Mission to Ethiopia, 1964–65. Rome, Italy, 200 pp.

Flint-Garcia S.A., Thuillet A., Yu J., Pressoir G., Romero S.M., Mitchell S.E., Doebley J.F., Kresovich S., Goodman M.M., Buckler E.S. 2005. Maize association population: A high resolution platform for QTL dissection. Plant J 44:1054-1064.

Gabriel S., Ziaugra L., Tabbaa D. 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet 60:2–12.

Garris A.J., Tai T.H., Coburn J., Kresovich S., McCouch S. 2005. Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631-1638.

Gupta P., Idris A., Mantri S., Asif M.H., Yadav H.K., Roy J.K., Tuli R., Mohanty C.S., Sawant S.V. 2012. Discovery and use of single nucleotide polymorphic (SNP) markers in *Jatropha curcas* L. Mol Breeding 30:1325-1335.

Hein L., Gatzweiler F. 2006. The economic value of coffee (*Coffea arabica*) genetic resources. Ecol Econ 60:176-185.

Huang X., Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res 9:868-877.

Jaramillo J., Muchugu E., Vega F.E., Davis A., Borgemeister C., Chabiolaye A. (2011). Some like it hot: the influence and implications of climate change on coffee berry borer (*Hypothenemus hampei*) and coffee production in East Africa. PLoS ONE 6:1–14.

Labouisse J.P., Bellachew B., Kotecha S., Bertrand B. 2008. Current status of coffee (*Coffea arabica* L.) genetic resources in Ethiopia: implications for conservation. Genet Resour Crop Ev 55:1079–1093.

Lashermes P., Combes M. C., Robert J., Trouslot P., D'Hont A., Anthony F., Charrier A. 1999. Molecular characterisation and origin of the *Coffea arabica* L. genome. Mol Gen Genet 261:259-266.

Lashermes P., Benoît B., Hervé. E. Breeding coffee (*Coffea arabica*) for sustainable production. In: Jain M., Priyadarshan P.M., editors. Breeding Plantation Tree Crops: Tropical Species. New York: Springer; 2009. p. 525–544.

Lin C., Mueller L.A., Mc Carthy J., Crouzillat D., Petiard V., Tanksley S.D. 2005. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. Theor Appl Genet 112:114-130.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303.

Metzker M.L. 2010. Sequencing technologies-the next generation. Nat Rev Genet. 11:31-46.

Montagnon C., Bouharmont P. 1996. Multivariate analysis of phenotypic diversity of *Coffea arabica*. Genet Res Crop Evol 43:221–227.

Nordborg M., Hu T.T., Ishino Y., Jhaveri J., Toomajian C., Zheng H., Bakker E., Calabrese P., Gladstone J., Goyal R., Jakobsson M., Kim S., Morozov Y., Padhukasahasram B., Plagnol V., Rosenberg N.A., Shah C., Wall J.D., Wang J, Zhao K., Kalbfleisch T., Schulz V., Kreitman M., Bergelson J. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Quail M.A., Kozarewa I., Smith F., Scally A., Stephens P.J., Durbin R., Swerdlow H, Turner D.J. 2008. A large genome center's improvements to the Illumina sequencing system. Nat Methods 5:1005–1010.

Rafalski J.A. 2002. Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162:329-333.

Risch N., Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516-7.

Silvestrini S., Junqueira M.G., Favarin A.C., Guerreiro-Filho O., Maluf M.P., Silvarolla M.B., Colombo C.A. 2007. Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. Genet Resour Crop Evol 54:1367-1379.

Simpson J.T., Wong K., Jackman S.D., Schein J.E., Jones S.J., Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117-1123.

Tabassum J., Lakhanpaul, S. 2006. Single polymorfism (SNP) – Methods and applications in plant genetics: A review. Indian Journal of Biotechnology, 5:453-459.

Vidal R.O., Mondego J.M., Pot D., Ambrósio A.B., Andrade A.C., Pereira L.F., Colombo C.A., Vieira L.G., Carazzolle M.F., Pereira G.A. 2010. A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. Plant Physiol 154:1053-1066.

Vieira et al. 2006. Brazilian coffee genome project: an EST-based genomic resource. Braz J Plant Physiol 18:95-108.

Yang S.S., Tu Z.J., Cheung F., Xu W.W., Lamb J.F.S., Jung H.J.G.; Vance C.P.; Gronwald J.W. 2011. Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. BMC Genomics 12:199.

ADDITIONAL FILE II

Additional File 1. Genotyping from the panel of 135 genotypes of *C. arabica*.

BA10/057	E114/447	E213/211	E332/023	E464/417
E007/087	E116/061	E218/581	E333/104	E466/125
E012/136	E118/213	E220/127	E335/219	E467/045
E016/298	E123a/231	E221/214	E338/218	E478/408
E017/419	E123b/121	E233/015	E340/179	E481/238
E018/494	E124/245	E237/071	E344/008	E486/189
E021/011	E126/359	E238/022	E351/248	E490/516
E022/163	E130/169	E254/284	E363/735	E494/173
E025/308	E131/018	E261/052	E364/059	E505/140
E030/075	E146/012	E265/101	E368/600	E511/157
E037/676	E148/254	E267/090	E370/196	E514/129
E038/043	E152/017	E268/067	E383/142	E516/069
E039/434	E159/180	E270/044	E386/131	E534/036
E041/079	E164/417	E272/143	E389/133	E546/118
E044/122	E169/180	E279/618	E401/643	E552/323
E046/021	E174/164	E283/096	E404/135	E565/010
E047/267	E179/650	E287/029	E408/001	E571/072
E055/005	E180/070	E298/382	E409/114	E621/139
E057/496	E181/358	E302/083	E419/098	M7846/67
E061/126	E183/138	E308/049	E428/109	SEL106
E068/014	E189/119	E315/081	E439/094	Catuai Vermelho
E071/258	E190/013	E320/145	E442/279	Mundo Novo

E080/584	E196/117	E324/093	E450/235	Typica
E081/041	E199/011	E325/522	E454/107	Bourbon
E085/085	E201/134	E326/124	E456/062	IAPAR 59
E087/194	E208/193	E327/032	E457/477	E-335/219 × Catuai
E089/391	E209/031	E331/280	E458/097	Tupi

7 CONSIDERAÇÕES FINAIS

A partir de dados de RNA-seq, foi possível desenvolver dois trabalhos. Primeiro, foi possível desenvolver uma análise global do transcriptoma de *C. eugenioides* e dos genes mais expressos em órgãos de folha e fruto dessa espécie. Os resultados revelaram possíveis correspondências entre os genes de *C. eugenioides* e o subgenoma CaCe de *C. arabica*. Alguns genes de *C. eugenioides* foram validados por qPCR e podem ser futuros genes candidatos nos estudos genéticos em *Coffea*.

Tanto os dados de anotação quanto a identificação e validação de alguns genes mais expressos em folha e fruto de *C. eugenioides* podem auxiliar no entendimento da evolução de *C. arabica*, principalmente mecanismos relacionados com a expressão diferencial de homeólogos. Por exemplo, genes de interesse econômico em *C. eugenioides* podem estar presentes em *C. arabica*, mas estão inativos, ou seja, não são expressos. Assim, o entendimento da base transcriptômica de *C. eugenioides* pode inicialmente ajudar no direcionamento de cruzamentos em café bem como ajudar no entendimento dos mecanismos evolutivos envolvidos na formação do alopoliplóide.

Os dados de transcriptoma de *C. eugenioides* também foram importantes para servir de referência para a busca de SNPs em *C. arabica*. Neste segundo trabalho, a análise da diversidade nucleotídica a partir de dados de NGS em um painel mais diverso de genótipos permitiu a identificação de 1410 SNPs com potencial para a genotipagem, um número expressivo de SNPs em *C. arabica* não descrito na literatura. Foram escolhidos 469 SNPs para a genotipagem em 129 indivíduos de uma coleção de *C. arabica* da Etiópia e validação de 311 SNPs a partir da metodologia de genotipagem MassARRAY. A partir do programa *Structure*, os resultados revelaram quatro sub-populações nessa coleção.

Mesmo com a baixa diversidade genética observada em trabalhos anteriores em *C. arabica*, esse trabalho mostra o potencial de fonte de recursos genéticos da coleção da Etiópia para futuros estudos de melhoramento genético. Assim, os SNPs validados inicialmente podem auxiliar nos trabalhos de associação uma vez que a coleção da Etiópia vem sendo caracterizada fenotipicamente com perspectivas do uso dessa coleção também em estudos de GWS e GWAS.