IDENTIFICAÇÃO DE POLIMORFISMOS EM GENÓTIPOS DE *Coffea arabica* DE UMA COLEÇÃO DA ETIÓPIA[1]

Priscila Mary Yuyama[2]; David Pot[3]; Alexis Dereeper[4]; Stéphanie Pointet[5]; João Batista Gonçalves Dias da Silva[6]; Gustavo Hiroshi Sera[7]; Tumoru Sera[8]; Pierre Charmetant[9]; Douglas Silva Domingues[10]; Thierry Leroy[11]; Luiz Filipe Protasio Pereira[12]

1 Trabalho financiado pelo Consórcio Pesquisa Café, PHEGECO CAPES –Agropolis, FINEP Genocafé
2 Bolsista CAPES, MS, Londrina-PR, priscilayuyama@gmail.com
3. Pesquisador, PhD, CIRAD, Montpellier, França, david.pot@cirad.fr
4. Pesquisador, PhD, IRD, Montpellier, França, alexis.dereeper@ird.fr
5 Engenheira, CIRAD, Montpellier, França, stephanie.pointet@cirad.fr
6 Coordenador do Centro Tecnológico da COCARI, MS, COCARI, Mandaguari-PR, ctc@cocari.br
7 Pesquisador, DSc, IAPAR, gustavosera@iapar.br
8 Pesquisador, DSc, IAPAR, tsera@iapar.br
9 Pesquisador, PhD, CIRAD, Montpellier, França, pierre.charmetant@cirad.fr
10 Pesquisador, DSc, IAPAR, doug@iapar.br
11 Pesquisador, PhD, CIRAD, Montpellier, França, thierry.leroy@cirad.fr
12 Pesquisador, PhD, Embrapa Café, Brasília-DF, filipe.pereira@embrapa.br

RESUMO: Os marcadores moleculares são ferramentas importantes para acelerar os programas de melhoramento. Para o cafeeiro, uma espécie perene, o uso de marcadores é particularmente desejável devido ao tempo e recursos gastos para o lançamento de uma nova cultivar. Duas espécies do gênero *Coffea* são responsáveis por quase toda a produção de café: *Coffea arabica* e *C. canephora*. Contudo, para *C. arabica,* o número de marcadores polimórficos é relativamente baixo comparado a *C. canephora* e outras culturas, uma vez que a espécie apresenta baixa diversidade genética. Muitos estudos com marcadores genéticos foram feitos para analisar a diversidade da *C. arabica*, mas os resultados não foram eficientes para a discriminação genotípica detalhada e mapeamento genético. O Instituto Agronômico do Paraná (IAPAR) possui uma coleção de 132 acessos de *C. arabica* originários da Etiópia, que apresentam variabilidade fenotípica com potencial para serem utilizados para exploração da diversidade. Neste sentido, este estudo buscou analisar a diversidade nucleotídica pela identificação de polimorfismos, SNPs e INDELs, de uma população do centro de origem de *C. arabica,* associado com o sequenciamento de nova geração. O RNA-seq de dois tecidos, frutos e folhas, de quatro genótipos de *C. arabica* de uma população da Etiópia, *C. arabica* cv. Mundo Novo e de um dos seus ancestrais de *C. arabica – C. eugenioides*, foram sequenciados pela metodologia Illumina HiSeq2000. Os *reads* obtidos foram processados e posteriormente as sequências foram mapeadas em uma referência de *C. canephora* para identificação dos polimorfismos. Foram feitas duas estratégias: i) na primeira estratégia, foi utilizado uma ferramenta chamada SNiPloid com critérios de cobertura para o polimorfismo identificado e ii) uma segunda estratégia que considera os polimorfismos encontrados diretamente dos arquivos de detecção dos polimorfismos. Os resultados identificaram um número grande de polimorfismos. Na primeira estratégia, foram encontrados pelo menos 5.500 SNPs potenciais para a genotipagem e na segunda, 103.791 SNPs potenciais. Para essa última, ainda é necessário estabelecer critérios e filtros para escolher os polimorfismos que serão inicialmente genotipados. Os dados também mostraram a importância de utilizar um grupo mais diverso de genótipos associado com o sequenciamento de nova geração para detecção de SNPs. Este trabalho será importante para direcionar futuros trabalhos na caracterização da diversidade genética em *C. arabica*, além de estudos de mapeamento genético por associação.

PALAVRAS-CHAVE: *Coffea arabica*, Etiópia, polimorfismos e SNPs.

IDENTIFICATION OF POLYMORPHISMS IN *Coffea arabica* GENOYTPES OF AN ETHIOPIA COLLECTION

ABSTRACT: Molecular markers are important tools to speed up the crop breeding programs, and perennial species like coffee, the markers are particularly interesting due to the time and resources spent for introduce a new cultivar. Two species of the genus *Coffea* are responsible for almost all coffee production: *Coffea arabica* and *C. canephora*. However, the number of polymorphic markers currently available is very low in *C. arabica* compared to *C. canephora* and other crops, because of its narrow genetic basis. Many studies of genetic markers have been done to analyze the diversity of this specie, but the results were not efficient to discriminate genotypes efficiently and developing genetic maps. The Instituto Agronômico do Paraná (IAPAR) has a collection of 132 accessions of *C. arabica* from Ethiopia, which have phenotypic variability with the potential to be used for diversity exploration. Therefore, this study aimed to analyze the nucleotide diversity for detecting polymorphisms, SNPs and INDELs, from population of *C. arabica*, from the center of origin, in Ethiopia, associated with next-generation sequencing. RNA-seq of two tissues, fruit and leaves,

from four genotypes of a *C. arabica* Ethiopian population, *C. arabica* cv. Mundo Novo and one of the ancestors of *C. arabica* - *C. eugenioides*, were sequenced with Illumina HiSeq2000. The reads obtained were processed and the sequences were mapped into a scaffold of *C. canephora* to identify the polymorphisms. Two strategies were performed: in the first strategy, i) we used a tool called SNiPloid with several coverage criteria for identification of the polymorphism. In the second strategy, ii) we consider the polymorphisms found directly in files of polymorphisms detection. The results identified a large number of polymorphisms. In the first strategy, at least 5,500 potential SNPs were found for genotyping and in the second strategy, 103,791 potential SNPs. For this last one, it is still necessary to establish criteria and filters to choose which polymorphisms will initially be genotyped. The data also showed the importance of using a more diverse group of genotypes associated with next-generation sequencing to detect SNPs. In addition, these results will be important to direct future work on the characterization of genetic diversity in *C. arabica*, and studies of genetic and association mapping.

KEY WORDS: *Coffea arabica*, Ethiopia, polymorphism, SNPs.

INTRODUÇÃO

O café é considerado uma das mais importantes *commodities* agrícolas do mundo e duas espécies são responsáveis pela maior parte da produção de café: *Coffea arabica* L. e *C. canephora* Pierre. *Coffea arabica* é uma espécie alotetraplóide e autógama ($2n = 4\times = 44$) originária de uma recente hibridização de duas espécies diplóides ou espécies relacionadas, *C. canephora* ($2n = 2\times = 22$) e *C. eugenioides* Moore ($2n = 2\times = 22$) (Lashermes et al., 1999; Yu et al., 2011).

Como consequência da sua autogamia e história evolutiva, *C. arabica* apresenta uma estreita base genética e este problema é somado ao fato dos principais genótipos cultivados, como Mundo Novo, Caturra e Catuaí, terem sido selecionados de apenas duas bases populacionais, Typica e Bourbon (Anthony et al., 2002). Assim, mesmo com os avanços no melhoramento tradicional do cafeeiro nos últimos anos, a sua eficiência é limitada pela estreita base genética da mesma (Anthony et al., 2002; Maluf et al., 2005). A dificuldade de identificação de polimorfismos faz com que até o momento não existam mapas genéticos completos para *C. arabica* e os trabalhos de aplicação de marcadores moleculares em melhoramento sejam escassos para a espécie.

As novas tecnologias de sequenciamento de nova geração, como 454 e Illumina/Solexa, permitiram avanços na geração de um grande volume de dados, com custo efetivo e resultados robustos. Muitos trabalhos foram feitos para identificação de um alto número de SNPs (polimorfismo de nucleotídeo único) para fins de construção de mapas genéticos (Metzker et al., 2010; Blanca et al., 2011). Por exemplo, essas tecnologias foram usadas para identificar SNPs em dados de transcriptomas de algodão, alfafa e melão (Blanca et al., 2011; Byers et al., 2011; Yang et al., 2011). A maioria dos trabalhos de identificação de SNPs foram desenvolvidos para organismos diplóides (Garber et al., 2011; Langmead e Salzberg, 2012), mas esses trabalhos são insuficientes para mapeamento de *reads* de organismos poliplóides por dois motivos. Primeiro, o mapeamento de *reads* de um poliplóide para um genoma diplóide resulta em mapeamento diferentes porque um subgenoma pode apresentar maior mapeamento do que o outro. Segundo, as ferramentas existentes não podem designar resultados quantitativos de mapeamento de um subgenoma e de outro. Outro ponto é a variação do número de cópias, o que causa problemas diferentes de interpretação dos resultados (Kitzman et al., 2012; Page et al., 2013). Porém, as análises de identificação de SNPs em poliplóides são facilitadas quando existe a presença dos parentais diplóides ou espécies relacionadas (Wu e Nacu 2010).

Vidal et al. (2010) identificaram polimorfismos a partir de dados de transcriptoma de *Coffea* spp. a partir de sequências de ESTs do banco de dados do Projeto Brasileiro do Genoma Café. Neste estudo, foram identificados 23.062 SNPs e 2.165 INDELS (Inserções/Deleções) em ESTs de *C. arabica* e *C. canephora*. Apesar da grande quantidade de SNPs identificados, não foi possível detectar variabilidade dentro dos subgenomas de *C. arabica*, o que evidencia a dificuldade de encontrar marcadores com potencial para genotipagem e mapeamento da espécie. Provavelmente um dos fatores que contribuíram para essa falta de informação foi o fato do estudo envolver apenas duas cultivares de *C. arabica* muitos próximas geneticamente (Vidal et al., 2010).

Posteriormente, Yanagui (2012) analisou genes envolvidos na biossíntese de açúcares, diterpenos, gene cloroplastídico e álcool desidrogenase em genótipos de *C. arabica* do centro de origem, na Etiópia, e genótipos comerciais, além de *C. canephora* e *C. eugenioides*. A utilização de germoplasma proveniente do centro de origem apresenta uma maior diversidade genética e fenotípica do que cultivares comerciais. Foram analisados cerca de 9 kb e foram encontrados 360 polimorfismos entre *C. arabica* e seus ancestrais, *C. canephora* e *C. eugenioides*. Porém, 12 polimorfismos mostraram variabilidade entre os genótipos, ou seja, a existência de variabilidade intraespecífica, ainda que em frequência reduzida, que pode ser aplicada para genotipagem. Esses resultados mostraram a importância de utilizar um grupo mais diverso de genótipos para detecção de SNPs.

Neste sentido, em função da necessidade de identificação de marcadores intraespecíficos com potencial para o melhoramento de *C. arabica* e a baixa variabilidade observada em estudos recentes, esse trabalho buscou avaliar a diversidade nucleotídica de *C. arabica* numa coleção de genótipos do centro de origem da espécie a partir de dados de sequenciamento de nova geração. Os resultados deste trabalho também deverão fornecer um grande número de marcadores SNPs para serem utilizados para a genotipagem de uma população para fins de mapeamento genético de *C. arabica*.

MATERIAL E MÉTODOS

O RNA total de dois tecidos (frutos inteiros e folhas maduras) de quatro genótipos de *C. arabica* da população da Etiópia (E-007/CAF087, E-123A/CAF231, E-238/CAF022 e E-516/CAF069), *C. arabica* cv. Mundo Novo e um dos ancestrais de *C. arabica*, *C. eugenioides*, foram extraídos a partir do protocolo de Chang et al. (1993) e a concentração foi determinada com o NanoDrop 1000 Spectrophotometer (NanoDrop, Wilmington, DE). Para verificar a qualidade da extração do RNA, as amostras foram submetidas à eletrofose com tampão TAE 1× em gel de agarose 1% (p/v) e coloração com brometo de etídio (0,5 μg/ml) sob luz UV.
A partir do RNA extraído, foram construídas as bibliotecas de cDNA e posterior sequenciamento com a tecnologia Illumina HiSeq2000, na Universidade da Carolina do Norte, Estados Unidos. Além disso, foram cedidos dados de RNA-seq de amostras de folha do outro ancestral de *C. arabica*, *C. canephora*, pelo CIRAD, para a montagem de uma referência. Os *reads* brutos foram filtrados para eliminação dos adaptadores, regiões de qualidade baixa (*score phred* min. 30) e *reads* com tamanho mínimo de 35 pares de bases (pb). Após processamento, as sequências de todos os genótipos de *C. arabica*, *C. canephora* e *C. eugenioides* foram mapeadas na referência de *C. canephora* a partir de uma montagem *de novo* com as sequências da espécie. Previamente a identificação dos polimorfismos, as sequências que mapearam em muitas regiões na referência (multi-mapeadas) foram eliminadas. Para detectar os SNPs e INDELs a partir dos dados de mapeamento, foi utilizado o *software* GATK (McKenna et al., 2010) e três ferramentas do GATK, em sequência, para análise dos polimorfismos: i) *UnifiedGenotyper* para a identificação dos polimorfismos; ii) *VariantFiltration* para estabelecer critérios de qualidade para os polimorfismos identificados, ou seja, os polimorfismos caracterizados em função da sua qualidade (PASS, Snp Cluster, LowQual, Hard to Validate) (Tabela 1); e iii) *ReadBackedPhasing* para a identificação dos haplótipos. Para isso, ele considera todos os *reads* dentro de um quadro "Bayesiano" e tenta identificar os haplótipos com a mais alta probabilidade.

Tabela 1. Descrição dos filtros utilizados para os polimorfismos identificados.

Filtros	Características
Hard to Validate	Os polimorfismos não podem ser classificados como PASS porque a informação é insuficiente
Low Qual	*Score Phred* muito baixo (<40)
Snp Cluster	Três polimorfismos ou mais em 10 pb
QD Filter	QD filter <1,5, qualidade por profundidade abaixo de 1,5
PASS	Score Phred > 40

A partir dos resultados obtidos, foram feitas duas estratégias para análise dos polimorfismos. Na primeira estratégia, foi feito a análise e seleção dos polimorfismos a partir de uma ferramenta do SNiPlay (Derepeer et al., 2011), chamada SNiPloid. Esta ferramenta compara os SNPs entre um tetraplóide e seus genomas parentais e permite classificar os SNPs encontrados. No nosso trabalho, foram comparados os SNPs obtidos para cada um dos genótipos de *C. arabica* (tetraplóide) com os SNPs obtidos para *C. eugenioides* (genoma parental). Para todos os genótipos de *C. arabica* e *C. eugenioides*, foram utilizados o mesmo genótipo como referência, *C. canephora*. Para análise neste programa, foram necessários arquivos de cobertura de cada base da sequência na referência (obtido pela ferramenta do GATK chamada *Depth of Coverage*) e os arquivos com as informações dos polimorfismos de cada genótipo, o VCF - *Variant Call Format*. Assim, foram necessários quatro arquivos para cada genótipo de *C. arabica* no SNiPloid: i) arquivo de cobertura de *C. eugenioides*; ii) arquivo VCF com as informações dos polimorfismos de *C. eugenioides*; iii) arquivo de cobertura de *C. arabica*; iv) arquivo VCF com as informações dos polimorfismos de *C. arabica*. Adicionalmente, foi estabelecido um valor de 20 sequências para uma cobertura mínima na referência, tanto para *C. arabica* quanto para *C. eugenioides* e somente os SNPs anotados com qualidade «PASS» foram analisados. Na segunda estratégia, os dados foram organizados e analisados no próprio arquivo VCF e os polimorfismos foram filtrados a partir de tabelas do programa Excel. Somente os polimorfismos anotados como «PASS» fizeram parte da seleção.

RESULTADOS E DISCUSSÃO

Na primeira estratégia, os SNPs encontrados em *C. arabica* e *C. eugenioides* a partir da mesma referência de *C. canephora* foram comparados e classificados de acordo com o SNiPloid (Tabela 2).
Cada classe obtida com os resultados do SNiPloid representam um tipo de polimorfismo. A classe 1 representa a variabilidade intraespecífica em *C. eugenioides* (Ce), a classe 2 a variabilidade intraespecífica em *C. canephora* (Cc), a classe 3 ou 4 são os polimorfismos encontrados dentro de *C. arabica* e que não representam as diferenças polimórficas dos ancestrais, Cc e Ce, e podem ser potenciais para a genotipagem. Porém, o SNiPloid não conseguiu separar os polimorfismos representativos de CaCc (subgenoma *C. canephora* dentro de *C. arabica*) e CaCe (subgenoma *C. eugenioides* dentro de *C. arabica*). A classe 3 separa os SNPs identificados no subgenoma CaCe e a classe 4 separa os SNPs identificados no subgenoma CaCc. Assim, estas classes são importantes para selecionar os SNPs potenciais para a

genotipagem. A classe 5 representa a variabilidade interespecífica entre Cc e Ce, ou seja, os polimorfismos entre os ancestrais de *C. arabica*. A classe "Outros" são os SNPs que não puderam ser classificados no SNiPloid e a classe "SNP Heterozigosidade Genoma 1" representam os SNPs heterozigotos encontrados somente em *C. eugenioides*. Assim, a partir dos resultados das classes 3, 4 e classe 3 ou 4, foram identificados 23.068 SNPs entre os cinco genótipos analisados, dos quais pelo menos cerca de 5.500 SNPs tem potencial para a genotipagem.

Tabela 2. Classificação dos polimorfismos encontrados no SNiPloid.

Genótipos de *C. arabica*	1 (Intra Ce)[1]	2 (Intra Cc)[1]	3 ou 4 (Dentro de *C. arabica*)	3 (Dentro de CaCe)	4 (Dentro de CaCc)	5 (Inter Cc e Ce)[2]	Outros	SNP Heterozigosidade Genoma 1
E-007 CAF087	518	1326	1264	757	1096	5731	4168	4125
E-123A CAF231	863	2247	2319	1384	1813	9494	6248	6171
E-238 CAF022	924	2201	2262	1414	1840	9283	6113	6026
E-516 CAF069	864	2194	2227	1320	1812	9565	6195	6116
Mundo Novo	722	1982	2122	581	857	8628	5715	5666

[1]Variabilidade intraespecífica
[2]Variabilidade interespecífica

Na segunda estratégia, para a análise dos polimorfismos encontrados, os dados foram filtrados por meio de uma planilha de dados referentes às informações do arquivo VCF. A análise dos genótipos do centro de origem e da cultivar de *C. arabica*, permitiram a identificação de 334.273 polimorfismos com qualidade <<PASS>>. Destes, 329.127 foram SNPs e 5.146 INDELs, sendo que 331.191 apresentaram um alelo polimórfico, 3.043 com dois alelos polimórficos e 39 com três alelos polimórficos.

Foram encontrados 237.436 polimorfismos em *C. arabica*. Deste valor, 103.791 (44%) foram polimorfismos que correspondem a diferenças entre Cc e Ce e que não seriam importantes para o mapeamento: 56.929 polimorfismos (24%) foram polimorfismos que correspondem a diferenças entre os ancestrais Cc e Ce e 15.302 (6%) foram polimorfismos fixados entre os dois genomas no qual nenhuma divergência foi detectada. Por fim, 133.645 polimorfismos (56%) foram obtidos em *C. arabica* dentro dos subgenomas e que seriam potenciais marcadores para a genotipagem (Tabela 3).

Tabela 3. Polimorfismos encontrados em *C. arabica*.

Classificação dos polimorfismos em *C. arabica*	Número de SNPs em *C. arabica*
Polimorfismos entre os dois genomas Cc e Ce	103.791 (44%)
Polimorfismos dentro dos subgenomas CaCc e CaCe	133.645 (56%)
Total	237.436

A partir desses resultados, serão iniciados trabalhos de genotipagem para fins de construção de um mapa genético para *C. arabica*. Porém, será necessário estabelecer critérios para selecionar os polimorfismos mais relevantes, como cobertura dos polimorfismos e a representação do menor alelo (alelo polimórfico). Para algodão, foram estabelecidos critérios de cobertura ≥8 sequências e o menor alelo deveria ser representado pelo menos em 20% dos alelos observados (Byers et al., 2012). A partir desses critérios, foram encontrados um total de 11.834 e 1.679 SNPs não-gênicos, nos acessos de *Gossypium hirsutum* L. e *G. barbadense* L. em montagens de bibliotecas genômicas reduzidas, respectivamente. Para *Brassica napus* L., foram encontrados 41.593 polimorfismos entre duas cultivares de *Brassica*, Tapidor e Ningyou 7 (Trick et al., 2009). No trabalho, foram estabelecidos no mínimo quatro sequências para identificação dos polimorfismos.

Em *Coffea*, foram desenvolvidos alguns trabalhos anteriores para identificação de SNPs, mas foram encontrados poucos SNPs representativos para fins de melhoramento. Vidal et al. (2010) observaram que *C. arabica* apresentava uma frequência maior de polimorfismos (0,393 xSNPs por 100 pb) do que *C. canephora* (0,169 SNPs por 100 pb), mas grande parte dos polimorfismos encontrados foram diferenças entre os dois subgenomas de *C. arabica*. No nosso trabalho, foram encontrados um valor alto de polimorfismos potenciais para o melhoramento da espécie. A partir dessas informações, serão selecionados os SNPs para a genotipagem e que poderão ser utilizados na construção de mapas genéticos e estudos de associação para a espécie, com a possibilidade de serem úteis nos programas de melhoramento genético.

CONCLUSÕES

Os dados obtidos mostram a importância de utilização de um grupo mais diverso de genótipos associado com o sequenciamento de nova geração para detecção de polimorfismos. Desta forma, esses resultados auxiliarão nos avanço em estudos genéticos de cafeeiro, sendo o ponto de partida para a genotipagem e futuramente auxiliar na construção de mapas genéticos mais saturados para *C. arabica*.

REFERÊNCIAS BIBLIOGRÁFICAS

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

BLANCA, J. M.; CAÑIZARES, J.; ZIARSOLO, P.; ESTERAS, C.; MIR, G.; NUEZ, F.; GARCIA-MAS, J.; PICÓ M. B. (2011). Melon Transcriptome Characterization: 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.

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. *Theoretical and Applied Genetics* 124:1201-1214.

CHANG, S.; PURYEAR J.; CAIRNEY, J. A simple and efficient method for isolating RNA from pine trees. (1993). *Plant Molecular Biology* 11:113-116.

DEREEPER, A.; NICOLAS, S.; CUNFF, L. L.; BACILIERI, R.; DOLIGEZ, A.; PEROS, J. P.; RUIZ, M.; THIS, P. (2011). SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics* 12:134.

GARBER, M.; GRABHERR, M. G.; GUTTMAN, M.; TRAPNELL, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* 8:469–477.

KITZMAN, J. O.; SNYDERM M. W.; VENTURA, M.; LEWIS, A. P.; QIU, R.; SIMMONS, L. E.; GAMMILL, H. S.; RUBENS, C. E.; SANTILLAN, D. A.; MURRAY, J. C.; TABOR, H. K.; BAMSHAD, M. J.; EICHLER, E. E.; SHENDURE, J. (2012). Noninvasive whole-genome sequencing of a human fetus. *Science Translational Medicine* 4:137-176.

LANGMEAD, B.; SALZBERG, S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359.

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. *Molecular and General Genetics* 261:259-266.

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

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 Research* 20:1297–1303.

METZKER, M. L: (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* 11:31-46.

PAGE, J. T.; GINGLE, A. R.; UDALL, J. A. (2013). PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. *G3 (Bethesda)* 3:517-525.

TRICK, M.; LONG, Y.; MENG, J.; BANCROFT, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploidy *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnology Journal* 7:334–346.

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

WU, T. D.; NACU, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873–881.

YANAGUI, K. Diversidade nucleotídica de oito genes relacionados a qualidade da bebida de *Coffea arabica*. (2012). Dissertação (Mestrado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Brasil.

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.

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