# MONTAGEM DO TRANSCRIPTOMA DE *Coffea eugenioides* E IDENTIFICAÇÃO DE GENES DIFERENCIALMENTE EXPRESSOS EM FOLHAS E FRUTOS[1]

Danilo Ribeiro de Brito[2], Suzana Tiemi Ivamoto-Suzuki[3], Paulo Maurício Ruas[4], Claudete de Fátima Ruas[5], Luiz Filipe Protasio Pereira[6]

[1]Trabalho financiado pelo Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café – Consórcio Pesquisa Café
[2]Mestrando, UEL, Londrina, PR, Brasil, ribeiro1990brito@gmail.com,
[3]Pós-doutoranda, UNESP, Rio Claro, SP, Brasil, profsuzanatiemi@gmail.com,
[4]Professor, UEL, Londrina, PR, Brasil, pmruas@uel.br,
[5]Professora, UEL, Londrina, PR, Brasil, ruas@sercomtel.com.br,
[6]Pesquisador, PhD, Embrapa-café, Londrina, Brasil, filipe.pereira@embrapa.br.

**RESUMO**: O café é uma das principais *commodities* agrícolas do Brasil. Dentre as 124 espécies do gênero, *Coffea arabica* e *Coffea canephora* representam a maior parte da produção de café mundial. *C. arabica* é um alotetraplóide, originado a partir de uma hibridação natural entre *C. canephora* e *C. eugenioides*. O parental *C. eugenioides* ainda é muito pouco estudado, mas muitas características genéticas dos cafés Arábicas comerciais foram herdadas desta espécie. Desta forma, é importante aumentar o conhecimento sobre os genes funcionalmente ativos no transcriptoma de *C. eugenioides.* Em um trabalho anterior foi realizado a montagem *ab initio* do transcriptoma de *C. eugenioides*. O presente projeto teve como objetivo realizar a montagem, anotação funcional dos transcritos de folhas e frutos de *C. eugenioides* utilizando o seu genoma como referência para o alinhamento das sequências. Foram utilizadas duas bibliotecas de RNA-Seq (folhas e frutos) de *C. eugenioides* sequenciadas via plataforma Illumina Hiseq2000. Primeiramente foi realizado o alinhamento das bibliotecas contra o genoma de referência, disponibilizado pelo Consórcio ACGC, através do software HISAT2. Para montagem dos *reads* alinhados em transcritos foi utilizado o software StringTie e em seguida, o software Kallisto para a quantificação dos transcritos. O software DESeq2 foi utilizado para identificação de genes diferencialmente expressos em frutos e folhas. Para anotação utilizamos o BlastX contra o banco de dados de sequências não redundantes (NCBI-nr). Foi realizada uma análise comparativa com a ferramenta BlastN contra bancos de dados de transcriptomas de *C. arabica* e *C. canephora* pré-existentes. O resultado da montagem de transcritos identificou 16.743 *contigs* únicos para *C. eugenioides*, estas sequências apresentaram similaridade de 82,06% com ESTs de *C. arabica*, 91,38% em CDS de *C. canephora*, 94,87% com o RNA-Seq de *C. arabica*. Em relação ao trabalho de transcritoma *ab initio*, foi observado 98,07% de similaridade e foram identificados 322 novos transcritos, das quais 36 não foram descritos nas bases de dados da *Coffea* spp. Além disso, foram identificados 414 e 509 genes diferencialmente mais expressos em folhas e frutos, respectivamente.

**PALAVRAS-CHAVE:** café, RNA-seq, transcriptoma.

# TRANSCRIPTOME ASSEMBLY OF *Coffea eugenioides* AND IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES IN LEAVES AND FRUITS

**ABSTRACT**: Coffee is one of the main agricultural commodities in Brazil. Among 124 coffee species, *Coffea arabica* and *Coffea canephora* are the most economically important and the first two represent the major part of coffee production. *C. arabica* is an allotetraploid, originated from a natural hybridization between *C. canephora* and *C. eugenioides*. The parental *C. eugenioides* is still poorly studied, but many genetic characteristics of commercial Arabicas coffee are inherited from this species. Thus, it is important to increase our knowledge about the functionally active genes in the transcriptome of *C. eugenioides*. In a previous work, a *C. eugenioides* transcriptome *ab initio* assembly was accomplished. In this work, we perform a new assembling and functional annotation of the transcriptome of leaves and fruits of *C. eugenioides*, using its genome as a reference for sequences alignment. Two RNA-Seq (leaf and fruit) libraries of *C. eugenioides* were obtained via the Illumina Hiseq2000 platform. The alignment of libraries against the reference genome, provided by ACGC Consortium, was executed using the HISAT2 software. For assembly of the aligned transcripts, StringTie software was used. Transcript quantification was achieved with Kallisto software. DESeq2 was used to identify genes differentially expressed in fruits and leaves. For annotation, we used BlastX against the non-redundant sequence database (NCBI-nr). A comparative analysis was performed with BlastN against the preexisting *C. arabica* and *C. canephora* transcriptome data. The result of the transcripts assembly identified 16,743 unique contigs that presented 82.06% similarity with ESTs of *C. arabica*, 91.38% with CDS of *C. canephora*, 94.87% with RNA-Seq of *C. arabica*. In relation to the original *ab initio* assembly work, we observed 98.07% similarity, with 322 new transcripts identified, 36 of them not described in the *Coffea* spp databases. In addition, 414 and 509 genes were differentially and more expressed in leaves and fruits, respectively.

**KEY-WORDS:** coffee, RNA-seq, transcriptome.

## INTRODUCÃO

A cafeicultura brasileira apresenta números expressivos que traduzem a grande importância econômica e social que a atividade representa para o país (ABIC, 2019). O Brasil é o maior produtor e exportador de café do mundo e o segundo maior consumidor (CECAFE, 2018). O aumento do consumo da bebida do café nos últimos anos tem sido associado, dentre outros fatores, sua maior qualidade aliada a bons preços de mercado, assim como às suas propriedades funcionais e nutracêuticas (ABIC, 2019).

Em função de sua importância, é preciso aumentar o conhecimento sobre as bases genéticas que influenciam a composição bioquímica dos cafés. A maioria dos estudos genéticos de cafeeiros possuem foco na espécie de maior importância econômica, a alotetraplóide *Coffea arabica* (CARDOSO *et al.,* 2014; MOFFATO *et al.,* 2016; IVAMOTO *et al.,* 2017a), alguns no seu parental diplóide *Coffea canephora* (DENOUED *et al.,* 2014; DEREEPER *et al.,* 2014) e apenas um trabalho com dados de RNA-Seq para seu outro parental diplóide *Coffea eugenioides* (YUYAMA *et al.,* 2015).

Desta forma, é importante aumentar o conhecimento sobre a espécie *C. eugenioides*, como por exemplo, a identificação e caracterização de genes diferencialmente expressos através de estudos de transcriptômica. Uma das formas de se alcançar este objetivo, é realizar uma análise do transcriptoma em larga escala através do sequenciamento de RNAs mensageiros (RNAm), também conhecido como RNA-Seq. Essa técnica tem proporcionado resultados relevantes com relação à identificação, anotação e caracterização de genes de espécies de plantas (WANG *et al.,* 2009).

Este trabalho teve como objetivo, fazer uma comparação dos resultados de uma montagem baseada no genoma de referência com a montagem *ab initio* disponível até o momento (YUYAMA *et al.,* 2015) e aumentar o conhecimento básico sobre a espécie *C. eugenioides* a partir da identificação de genes diferencialmente expressos em folhas e, especialmente em frutos. Esta informação será utilizada para identificar e selecionar genes candidatos para as vias metabólicas de interesse relacionadas à composição bioquímica de grãos de café e servirá como base para novos projetos de melhoramento genético de cafeeiros visando o aumento da qualidade da bebida.

## MATERIAL E METODOS

Para a montagem do transcriptoma foram utilizadas duas bibliotecas (folhas e frutos) de RNA-seq de *C. eugenioides* disponíveis no NCBI (PRJNA273321). As análises foram realizadas no servidor do Laboratório Multiusuário de Bioinformática da Embrapa Informatica Agropecuária - CNPTIA. Foi utilizado o genoma de referência de *C. eugenioides* (versão 1.9) gentilmente cedido pelo ACGC (*Arabica Coffee Genome Consortium*) para o alinhamento dessas bibliotecas de RNA-Seq. A qualidade das sequências foi realizada com o software FastQC v.0.11.5 (ANDREWS, 2010) e a trimagem para a remoção dos adaptadores com o software Trimmomatic v.36 (BOLGER; LOHSE; USADEL, 2014).

Para alinhamento dos *reads* contra o genoma de referência do *C. eugenioides* foi utilizado o alinhador HISAT2 versão 2.1.0 com parâmetros *default* (KIM; LANGMEAD; SALZBERG, 2015). O software SamTools versão 1.6 (LI *et al.,* 2009) converteu os arquivos de saída do alinhamento (SAM) para o formato aceito pelo StringTie (BAM). Para a identificação dos transcritos, foi utilizado o software StringTie versão 1.3.4 (PERTEA *et al.,* 2015), juntamente com a sua função *merge*, todos com os (parâmetros *default*). Apenas os transcritos com no mínimo 200bp de tamanho seguiram para as próximas análises.

A quantificação a abundância de transcritos a partir de dados de RNA-Seq foi realizada com o software Kallisto versão 0.44 (BRAY *et al.,* 2016). Todos os transcritos foram anotados através da análise comparativa realizada pela ferramenta BlastX contra o banco de dados de sequências não redundantes (NCBI-nr) disponíveis na plataforma do *National Center for Biotechnology Information* (NCBI). O parâmetro de corte utilizado foi de *e-value* menor ou igual a $1e^{-5}$. Além do NCBI-nr, foi realizada uma segunda análise comparativa dos transcritos de *C. eugenioides* contra o banco de dados curados de proteínas curado (Swiss-Prot - *The UniProt Consortium*, 2014) através da ferramenta BlastX.

A anotação funcional dos *contigs* montados e de seus respectivos processos biológicos, funções moleculares e componentes celulares foi realizada utilizando o software Blast2GO Pro 5 (CONESA *et al.,* 2005). O banco de dados do *Kyoto Encyclopedia of Genes and Genomes* (KEGG) foi usado para identificar as vias metabólicas dos genes putativamente encontrados no transcriptoma (KENEHISA *et al.,* 2018).

A análise comparativa entre o transcriptoma obtido por este trabalho e os previamente publicados em cafés (MONDEGO *et al.,* 2011; DENOEUD *et al.,* 2014; YUYAMA, *et al.,* 2015; IVAMOTO *et al.,* 2017) foi realizada utilizando a ferramenta Blast.

A análise de expressão gênica diferencial, entre os transcritos das bibliotecas de folha e fruto, foi realizada pelo software DESeq2 (LOVE; HUBER; ANDERS, 2014), cuja finalidade foi quantificar o número total de sequências alinhadas em cada transcrito gerado pelo software Kallisto. Os parâmetros utilizados foram p-value < 0.05 em combinação com Log2FC >= 1,58 para transcritos positivamente (*up*) regulados e <= -1,58 para transcritos negativamente (*down*) regulados. Após a identificação, foram verificadas as anotações nos bancos NCBI-nr e Swiss-Prot e também a análise GO (*Gene Ontology*) via Blast2GO dos genes identificados como diferencialmente expressos. O software TransDecoder (https://transdecoder.github.io/) foi utilizado para identificar os unigenes que possuem sequências completas para a codificação de proteínas (*full-lenghts*).

## RESULTADOS E DISCUSSÃO

Do total de 3.618.185 de *reads* para a biblioteca de folhas e 4.657.475 para a biblioteca de frutos, foram alinhados com o software HISAT2 contra o genoma de referência 82,98% (3.002.213) e 79,42% (3.769.938) das sequencias respectivamente (Tabela 1). Esses resultados para os alinhamentos estão de acordo com os resultados apresentados pelo protocolo de referência para este tipo de análise (PERTEA *et al.*, 2016)

**Tabela 1 –** Resultado do alinhamento das bibliotecas contra o genoma de referência.

| Informações do Alinhamento | Folhas | Frutos |
|---|---|---|
| *Reads* de alta qualidade | 3.618.185 | 4.657.475 |
| Porcentagem de *reads* alinhados 1 vez | 82,98% (3.002.213) | 81,14% (3.779.132) |
| Porcentagem de *reads* não alinhados | 17.02% (615.972) | 18,86% (878.343) |
| Porcentagem de reads alinhados mais de 1 vez | 0% (0) | 0% (0) |

A identificação de possíveis transcritos resultou em um total de 16.743 contigs (13.668 de folhas e 14.543 frutos) foram identificados com comprimento médio de 1.248 pares de base (pb). Nossos resultados corroboram com um estudo anterior realizado com o transcriptoma de *C. arabica* (1.264 pb) (IVAMOTO *et al.,* 2017a), próximo ao descrito por Venturini *et al.* 2013, em *Vitis vinífera* (1.307 pb) e maior que o encontrado em *C. eugenioides* (701 pb) em uma análise de montagem *de novo* de transcriptoma (YUYAMA *et al.,* 2015). Nossas análises identificaram 16.743 transcritos, um número foi menor do que o observado por Yuyama *et al.* (2015) (36.935 *contigs*). A montagem *de novo* usando a técnica de RNA-Seq *single-end*, caso semelhante ao de Yuyama *et al.* (2015), normalmente gera um número alto de *contigs* (HAN *et al.*, 2013).

A anotação dos 16.743 *contigs* de *C. eugenioides* foram analisados com base na similaridade de suas sequências com o BlastX contra sequências públicas disponíveis no banco de dados NCBI (NCBI-nr). Destas, 91,49% apresentaram similaridade com pelo menos uma sequência do NCBI-nr. Além disso, foi realizada uma análise de anotação funcional dos unigenes com a ferramenta BlastX contra o banco de dados de proteínas curadas (UniProt – Swiss-Prot), onde foram anotados 12.277 contigs (73,32%) nessa análise.

Foi realizada também uma análise comparativa de similaridade entre os transcritos obtidos por este trabalho e os publicados na literatura: i) expressed sequence tags (ESTs) de *C. arabica* contendo 35.113 *contigs* (MONDEGO *et al.,* 2011); ii) *coding region sequences* (CDS) do genoma de *C. canephora* (DENOEUD *et al.,* 2014); iii) transcritos de *C. arabica* obtidos via RNA-seq contendo 65364 *unigenes* (IVAMOTO *et al.,* 2017); iv) 36935 unigenes de um transcriptoma *ab initio* de folhas e frutos de *C. eugenioides* obtidos via RNA-seq (YUYAMA *et al.,* 2015). Os resultados podem ser observados na tabela 2, onde 13740 transcritos apresentaram similaridade (82,06%) com os ESTs de *C. arabica* (MONDEGO *et al.*,2 011) (Tabela 4), 15300 (91,38%) o genoma de *C. canephora* (DENOUED *et al.*, 2014), 15885 com o transcriptoma de *C. arabica* obtidos via RNA-seq (IVAMOTO *et al.*, 2017) e 16421 (98,07%) com o transcriptoma de *Coffea eugenioides*..

**Tabela 2 –** Análise de similaridade entre os transcritos de *C. eugenioides* deste com banco de dados públicos de café

| Banco de dados | Contigs anotados | Porcentagem de anotação | Referência |
|---|---|---|---|
| EST *database* | 13.740 | 82,06 % | Mondego *et al*., 2011 |
| *C. canephora* | 15.300 | 91,38% | Denoeud *et al*,, 2014 |
| *C. arabica* | 15.885 | 94,87% | Ivamoto *et al*., 2017 |
| *C. eugenioides* | 16.421 | 98,07% | Yuyama *et al*., 2015 |

A última comparação realizada foi feita utilizando os unigenes do transcriptoma *ab initio* de *C. eugenioides* (YUYAMA *et al*., 2015), geradas com as mesmas bibliotecas utilizadas por este trabalho, e o resultado foi e 98,07% (16.421 *contigs*) de similaridade entre as sequências, onde apenas 322 *contigs* não apresentaram similaridade. Estes 322 contigs podem conter novos genes ainda não identificados para a espécie e configuram dados que podem ser explorados futuramente. Desses *contigs*, 286 também foram encontrados nos bancos de dados de *C. canephora* (DENOUED *et al.*, 2014) e de *C. arabica* (MONDEGO *et al.*, 2011), restando 36 *contigs* não anotados em nenhum dos dados de sequência de café analisados neste trabalho.

As análises comparativas realizadas até o presente momento, evidenciam que a montagem do transcriptoma de *C. eugenioides* utilizando o genoma de referência para o alinhamento dos *reads* mostrou-se muito mais eficiente em identificar genes que codificam proteínas funcionais do que o encontrado pelo transcriptoma *ab initio* (YUYAMA *et al*., 2015), o qual apresentou uma porcentagem menor de similaridade contra os mesmos bancos de dados utilizados neste trabalho.

A caracterização funcional dos *contigs* de *C. eugenioides* foi através da atribuição de termos GO (*gene ontology*) pelo software Blast2GO resultou em um total de 56.059 termos GOs distribuídos em 13.716 *contigs*. Destes termos, 38,39% estavam relacionados com processo biológico, 35,69% com função molecular e 25,91% com a categoria de

componentes celulares. Resultados semelhantes, em relação a distribuição dos termos GOs entre os 15 níveis, foram identificados em um trabalho anterior de transcriptoma de *C. arabica* (IVAMOTO *et al.*, 2017a). Os termos GO mais abundantes para categoria de função molecular foram *ATP binding*, *protein binding*, *metal ion binding* e *DNA binding*. Na categoria de processo biológico, *oxidation-reduction process* (GO:0055114), *protein phosphorylation*, *regulation of transcription*, *DNA-templated* e *serine family amino acid metabolic process* foram os termos atribuídos ao maior número de *contigs*. Por fim, na categoria de componente celular, os termos mais abundantes foram *integral component of membrane*, *nucleus*, *cytosol* e *cytoplasm* (Figura 1).

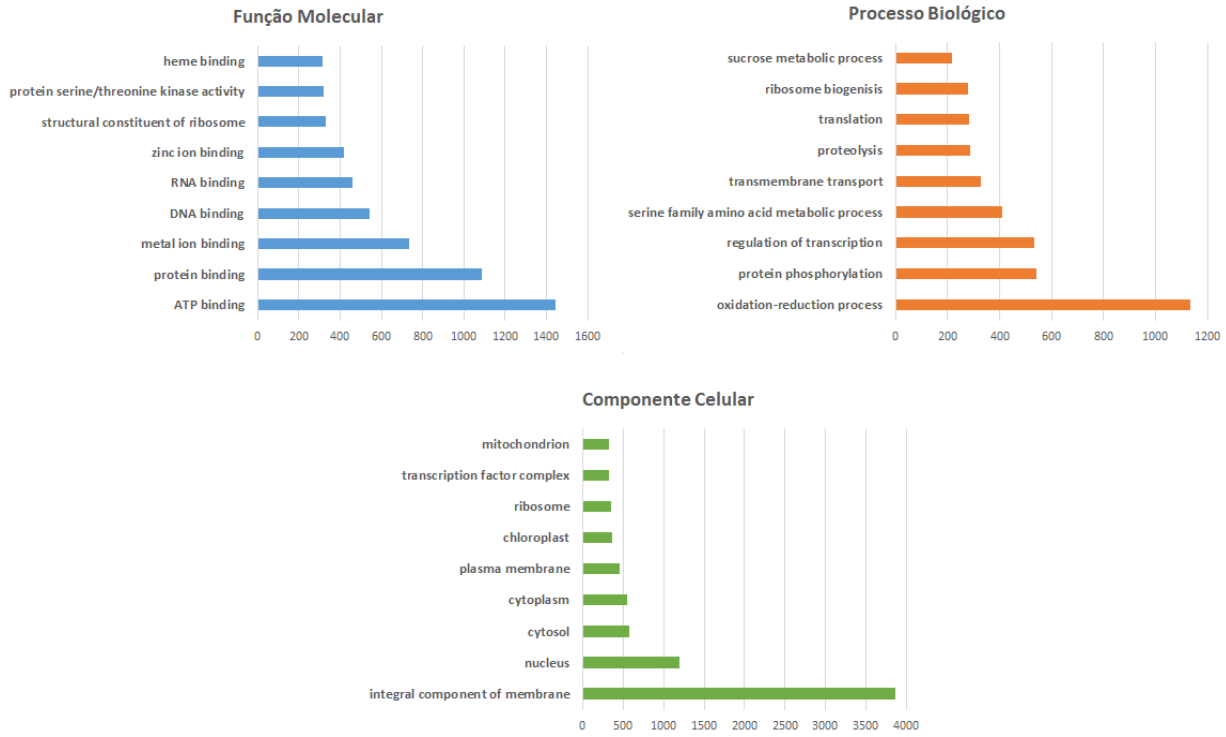


**Figura 1 –** Os termos GO mais abundantes para cada categoria.

Posteriormente, os *contigs* foram mapeados nos mapas de vias metabólicas do banco de dados do KEGG, através de uma ferramenta do software Blast2GO. Um total de 1.249 *contigs* (7,45%) foram distribuídos em 149 vias metabólicas e representam 777 enzimas diferentes. A categoria mais abundante foi o metabolismo de purina, seguida pela de metabolismo de tiamina e biossíntese de antibióticos (Figura 2).

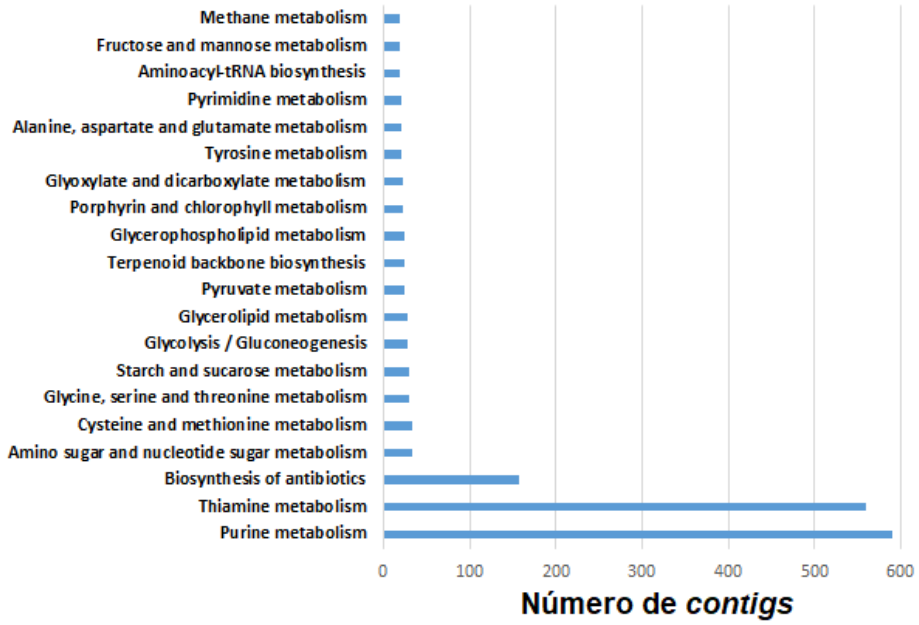


**Figura 2 –** Principais categorias de vias metabólicas obtidas pelo KEGG.

A análise de identificação de enzimas em suas respectivas vias metabólicas, realizada com a utilização do banco de dados do KEGG, apresentou números superiores (1249) aos encontrados por um estudo anterior (802) Yuyama *et al.* (2015), sendo que deste total, 777 e 374 enzimas possuíam funções diferentes, respectivamente. Em relação as categorias mais abundantes, ambos os trabalhos apresentaram semelhanças, como por exemplo, o metabolismo de purina, metabolismo de amido e sacarose e glicólise/gliconeogênese. Resultados semelhantes obtendo como categorias mais abundantes o metabolismo de purina e metabolismo de tiamina foi encontrado por Tran *et al.* (2018), em *C. arabica*.

## CONCLUSÃO

No presente trabalho, de reanálise de dados de RNA-Seq de *C. eugenioides*, foram montados 16.743 transcritos, sendo identificados 322 transcritos novos em relação ao trabalho original de Yuyama *et al.* (2015), o que levanta a possibilidade de novos genes descritos para a espécie *C. eugenioides*. Apesar do número de transcritos ter sido menor em comparação Yuyama *et al.* (2015), esta pesquisa apresentou uma porcentagem maior de anotação em todos os bancos de dados utilizados em ambos os trabalhos em relação ao número de transcritos. Também foram identificados 36 genes ainda não descritos. O número de genes diferencialmente expressos também foi maior em relação a Yuyama *et al.* (2015), sendo 923 genes contra 79.

## AGRADECIMENTOS

Ao Consórcio Pesquisa Café, e as instituições Universidade Estadual de Londrina (UEL), Universidade Estadual Paulista (UNESP – câmpus Rio Claro), Instituto Agronômico do Paraná (IAPAR) e Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA CNPTIA) colaboradoras deste projeto. O presente trabalho também foi realizado com apoio do a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Genoma de C. eugenioides gentilmente cedido pelo Consórcio Genoma Café Arabica (ACGC) e pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – Processo n. 2017/01455-2.

## REFERÊNCIAS BIBLIOGRÁFICAS

ABIC. Associação Brasileira da Indústria de Café. 2018. Disponível em: <http://www.abic.com.br>. Acesso em 20 de janeiro de 2019.

CECAFE. Conselho dos Exportadores de Café do Brasil. 2018. Disponível em: <http://www.cecafe.com.br>. Acesso em: 12 de agosto de 2018.

CARDOSO, D. C. *et al.* 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**, v. 15, n. 1, p. 66, 2014.

MOFATTO, L. S. *et al.* Identification of candidate genes for drought tolerance in coffee by high-throughput sequencing in the shoot apex of different *Coffea arabica* cultivars. **BMC plant biology**, v. 16(1), p. 1, 2016.
IVAMOTO ST, et al., Transcriptome analysis of leaves, flowers and fruits perisperm of *Coffea arabica* L. reveals the differential expression of genes involved in raffinose biosynthesis. PloS one, v. 12, n. 1, p. e0169595, 2017.
DENOEUD F, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science, v. 345, n.6201, p.1181-1184, 2014.
DEREEPER, A. *et al.* The coffee genome hub: a resource for coffee genomes. **Nucleic acids research**, v. 43, n. D1, p. D1028-D1035, 2014.
YUYAMA, Priscila Mary et al. Transcriptome analysis in Coffeaeugenioides, an Arabica coffee ancestor, reveals differentially expressed genes in leaves and fruits. **Molecular genetics and genomics**, v. 291, n. 1, p. 323-336, 2016.
WANG *et al.*, 2009.
ANDREWS, S. FastQC: A Quality Control tool for High Throughput Sequence Data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> 2010.
BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. **Bioinformatics**, v. 30, p. 2114–2120, 2014.
KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature methods**, v. 12, n. 4, p. 357-360, 2015.
PERTEA, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. **Nature biotechnology**, v. 33, n. 3, p. 290-295, 2015.
BRAY, Nicolas L. et al. Near-optimal probabilistic RNA-seq quantification. **Nature biotechnology**, v. 34, n. 5, p. 525, 2016.
CONESA, Ana et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674-3676, 2005.
KANEHISA, Minoru et al. New approach for understanding genome variations in KEGG. **Nucleic acids research**, v. 47, n. D1, p. D590-D595, 2018.
MONDEGO, J. M. C. *et al.* An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. **BMC plant biology**, v. 11, n. 1, p. 30, 2011.
LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome biology**, v. 15, n. 12, p. 550, 2014.