TRANSCRIPTOMA DE *C. arabica* L. REVELA DIFERENÇAS NA EXPRESSÃO DE GENES ENVOLVIDOS NA BIOSSÍNTESE DA RAFINOSE[1]

Suzana Tiemi Ivamoto-Suzuki[2], Osvaldo Reis Junior[3], Douglas Silva Domingues[4]; Tiago Benedito dos Santos[5], Fernanda Freitas de Oliveira[6], Larissa Girotto[7]; Caroline Ariyoshi[8], David Pot[9], Thierry Leroy[10], Luiz Gonzaga Esteves Vieira[11], Marcelo Falsarella Carazzolle[12]; Gonçalo Amarante Guimarães Pereira[13]; Luiz Filipe Protasio Pereira[14]

[1]Trabalho financiado pelo Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café – Consórcio Pesquisa Café
[2]Pós-doutoranda, UNESP, Rio Claro, SP, Brasil, profsuzanatiemi@gmail.com,
[3]Bioinformata, LaCTAD, Unicamp, Campinas, Brasil, osvaldoreisss@gmail.com
[4]Pesquisador, UNESP, Rio Claro, SP, Brasil, doug@unesp.com.br
[5]Professor, Unoeste, Presidente Prudente, Brasil, tiagobio02@yahoo.com.br
[6]Doutoranda, UEL, Londrina, PR, Brasil, fernandafreitasdeoliveira92@gmail.com
[7]Doutoranda, UEL, Londrina, PR, Brasil, larissagirotto@yahoo.com.br
[8]Doutoranda, UEL, Londrina, PR, Brasil, carolinearyoshi@hotmail.com
[9]Pesquisador, CIRAD, Montpellier, França, david.pot@cirad.fr
[10]Pesquisador, CIRAD, Montpellier, França, thierry.leroy@cirad.fr
[11]Professor, Unoeste, Presidente Prudente, Brasil, luizvie@gmail.com
[12]Pesquisador, Unicamp, Campinas, Brasil, mcarazzo51@gmail.com
[13]Pesquisador, Unicamp, Campinas, Brasil, goncalo@unicamp.br
[14]Pesquisador, Embrapa Café, Londrina, Brasil, filipe.pereira@embrapa.br

RESUMO: *Coffea arabica* L. é uma cultura importante em vários países em desenvolvimento. Apesar de sua importância econômica, os dados do transcriptoma mínimo estão disponíveis para tecidos de frutas, especialmente o perisperma, onde vários compostos relacionados à qualidade do café são produzidos. Para compreender os aspectos moleculares relacionados ao desenvolvimento de frutos e grãos de café, relatamos uma análise transcriptoma em larga escala de folhas, flores e frutos. O sequenciamento da Illumina (RNA-seq) resultou em 41.881.572 sequências trimadas com alta qualidade. A montagem de novo gerou 65.364 unigenes com um comprimento médio de 1.264 pb. Um total de 24.548 unigenes foram anotados como genes codificadores de proteínas, dos quais 12.560 sequências eram completas para esta codificação (*full lenghts*). No processo de anotação, identificamos nove genes candidatos relacionados à biossíntese de oligossacarídeos da família da rafinose (RFOs). Estes açúcares conferem osmoproteção e são acumulados durante o desenvolvimento inicial dos frutos. Quatro genes dessa via tiveram o seu padrão transcricional validado pela reação em cadeia da polimerase via transcrição reversa quantitativa em tempo real (RT-qPCR). Este atlas transcriptômico de *C. arabica* fornece um importante passo para a identificação de genes candidatos relacionados a diversas vias metabólicas do café, especialmente aquelas relacionadas à composição química dos frutos e a qualidade da bebida. Nossos resultados são o ponto de partida para aumentar o conhecimento sobre as proteínas do café que são produzidas durante os estádios de floração e desenvolvimento inicial dos frutos.

PALAVRAS-CHAVE: Transcriptoma, café, rafinose, galactinol, RNA-seq, RT-qPCR

TRANSCRIPTOME ANALYSIS OF *C. arabica* L. REVEALS THE DIFFERENTIAL EXPRESSION OF GENES INVOLVED IN RAFFINOSE BIOSYNTHESIS

ABSTRACT: *Coffea arabica* L. is an important crop in several developing countries. Despite its economic importance, minimal transcriptome data are available for fruit tissues, especially the perisperm, where several compounds related to coffee quality are produced. To understand the molecular aspects related to coffee fruit and grain development, we report a large-scale transcriptome analysis of leaf, flower and fruits. Illumina sequencing yielded 41,881,572 high-quality filtered reads. *De novo* assembly generated 65,364 unigenes with an average length of 1,264 bp. A total of 24,548 unigenes were annotated as protein coding genes, including 12,560 full-length sequences. In the annotation process, we identified nine candidate genes related to the biosynthesis of raffinose family oligossacarides (RFOs). These sugars confer osmoprotection and are accumulated during initial fruit development. Four genes from this pathway had their transcriptional pattern validated by quantitative reverse transcription polymerase chain reaction (RT-qPCR). This *C. arabica* transcriptomic atlas provides an important step for identifying candidate genes related to several coffee metabolic pathways, especially those related to fruit chemical composition and, in one example, beverage quality. Our results are the starting point for enhancing our knowledge about the coffee proteins that are produced during the flowering and initial fruit development stages.

KEY-WORDS: Transcriptome, coffee, raffinose, galactinol, RNA-seq, RT-qPCR

INTRODUCÃO

O café representa uma das culturas mais importantes para os países tropicais em desenvolvimento. O gênero possui 124 espécies [1], mas apenas o alotetraplóide *Coffea arabica* L. e o diplóide *Coffea canephora* Pierre ex A. Froehner têm importância econômica. Apesar de sua importância econômica, o genoma de *C. arabica* ainda não foi publicado, mas o genoma de um ancestral diplóide de *C. arabica*, o *C. canephora*, foi recentemente publicado [2]. A identificação de genes candidatos relacionados a características agronômicas e seus respectivos perfis de expressão digital podem revelar uma nova hipótese sobre mecanismos genéticos que controlam a biossíntese de proteínas e metabólitos. Atualmente, técnicas de sequenciamento de mRNA de alto rendimento têm sido amplamente utilizadas em estudos de transcriptomas vegetais. Várias iniciativas estudaram o trancriptoma do cafeeiro [3-4], e o RNA-seq é considerado uma ferramenta molecular poderosa para investigar espécies não modelo que possuem pouca informação disponível para estudos genéticos [5].

A bebida do café é obtida do endosperma da semente moído; no entanto, a maioria dos dados de RNA-seq representa o transcriptoma de folhas. No café, a maioria dos metabólitos presentes nos frutos são sintetizados durante o desenvolvimento do perisperma. O perisperma é um tecido altamente ativo, com um metabolismo intenso e este tecido é substituído pelo endosperma durante o desenvolvimento do fruto [6].

O acúmulo de oligossacarídeos da família da rafinose (*Raffinose Family of Oligossacharide* - RFOs), como rafinose e estaquiose, foi observado anteriormente durante o desenvolvimento dos frutos do café [7]. Os RFOs são compostos envolvidos nos mecanismos de defesa da planta para aumentar a tolerância a estresses abióticos. As RFOs atuam como moléculas de sinal em resposta ao estresse [8] e estão relacionadas à tolerância contra a dessecação da semente em períodos de germinação [9]. Nas plantas de café, as RFOs estão relacionadas a osmoproteção contra estresses abióticos nas folhas [10], mas também podem ser possíveis doadores de esqueletos de carbono durante a síntese de polissacarídeos de armazenamento na parede celular (CWSPs). Uma análise baseada de *microarray* de endosperma do café mostrou que os níveis de transcrição do gene galactinol sintase (*GolS*) estavam significativamente correlacionados com a quantidade de CWSPs [7].

Neste estudo, nós analisamos dados de transcriptoma *de novo* de folhas, flores e perisperma de frutos de café em cinco estágios de desenvolvimento, e identificamos genes que são expressos especificamente nesses órgãos. Também geramos um catálogo de possíveis elementos transponíveis transcricionalmente ativos e alvos de microRNAs (miRNA), que são componentes relevantes do transcriptoma e que são raramente estudados em abordagens transcriptômicas. Os genes relacionados à biossíntese de RFOs tiveram os seus padrões de expressão digital confirmados por RT-qPCR, o que demonstrou que nossos dados de transcriptoma em larga escala contêm informações valiosas para a descoberta de novos genes-chave envolvidos no metabolismo dos frutos do café e envolvidos na qualidade da bebida.

MATERIAL E MÉTODOS

Tecidos (folhas, flores, frutos) foram obtidos de plantas de *C. arabica* cv IAPAR59 com 20 anos de idade cultivadas no Instituto Agronômico do Paraná (Londrina - Brasil) em condições de pleno sol e com práticas padrão de irrigação e fertilização. As amostras de frutos foram coletadas mensalmente após a floração (30 a 150 dias a'pos a florada - DAF). Todas as amostras foram coletadas entre 9 e 11 horas da manhã, transferidas imediatamente para o nitrogênio líquido e armazenadas a -80 °C. O RNA total foi isolado com base no método de Chang et al. (1993). A pureza e integridade do RNA foi verificada com o espectrofotômetro NanoDrop® ND-1000 e o Bioanalyzer Chip DNA 1000 série II. O seqüenciamento de mRNA foi realizado na *High-Throughput Sequencing Facility*, no *Carolina Center for Genome Sciences Chapel Hill*, NC, EUA). Todas as bibliotecas foram marcadas e multiplexadas no equipamento Illumina HiSeq™ 2000, cujas sequências geradas foram de 100 pares de bases (pb). As sequências brutas foram trimadas com relação as suas respectivas qualidades (Q <20) usando um script *in house*. Sequências de alta qualidade foram alinhadas usando o software Trinity e possíveis sequências codificadoras de proteínas foram preditas usando o software Transdecoder. Todos os unigenes foram comparados com os bancos de dados do NCBI-nr, Swiss-Prot, *C. arabica* EST e genoma *C. canephora* usando a ferramenta BlastX, com um valor de corte de 1e-5.

A anotação funcional descrevendo os possíveis processos biológicos, funções moleculares e localização celular (componente celular) foi realizada utilizando as ferramentas disponíveis no software Blast2GO, bancos de dados do InterProScan e do KEGG. Usamos o software Bowtie com os parâmetros padrões para mapear todas as sequências no transcriptoma montado *de novo*. Os valores de RPKM foram normalizados para cada unigene. Comparações pareadas da análise dos dados de expressão entre folhas e flores e durante os estágios iniciais de desenvolvimento dos frutos do perisperma (30 a 150 DAF) foram usadas para caracterizar o perfil da expressão digital de genes (*digital gene expression* - DGE) com base nos resultados do pacote EdgeR. A análise de DGEs entre as bibliotecas foi realizada com um ponto de corte logarítmico, onde (Log2FC) ≥ 1 foi utilizado para genes superexpressos e Log2FC ≤ -1 para genes subexpressos. Os oligonucleotídeos foram desenhados usando o software Primer 3 e as suas respectivas eficiências foram calculadas com o auxílio do software LinRegPCR. Os cDNAs foram sintetizados usando um kit SuperScript III Reverse Transcriptase (Invitrogen). O RT-qPCR foi realizado no equipamento de PCR em tempo real 7500 (Applied Biosystems) e, seguindo os procedimentos e protocolos básicos previamente descritos para plantas de café [11]. A determinação da expressão relativa e o processo de normalização foram realizados através do software GenEX (MultiD,

Gotemburgo, Suécia). Os níveis de transcrição foram normalizados usando perfis de expressão gênica dos genes constitutivos: gliceraldeído-3-fosfato desidrogenase do café (GAPDH) e fator de alongamento 1 (EF1). Os dados foram analisados por ANOVA bidirecional e teste de Tukey ($p < 0,05$), através do software Assistat [12].

RESULTADOS E DISCUSSÃO

Obtivemos um total de 41.881.572 sequências de mRNA de alta qualidade. Como *C. arabica* não possui um genoma de referência, optamos por fazer uma montagem *de novo* do transcriptoma. Um total de 127.600 contigs foram gerados, e 78.794 contigs foram preditos como possíveis sequências codantes de proteínas. Um total de 65.364 transcritos foram considerados como unigenes (variantes únicas de *splicing*) com tamanho >200 pb. O comprimento médio desses 65.364 contigs foi de 1.264 pb, com variação de 201 a 12.891 pb. Alcançamos um N50 de 2.118 pb, e o conteúdo médio de GC foi de 41,13% (Tab. 1). Aproximadamente 60% dos contigs tinham 200 a 500 pb, 16% entre 501 e 1.000 pb, 12% entre 1001 e 2.000 pb e 4% com mais de 3000 pb.

Tabela 1. Resumo do alinhamento *de novo* de *C. arabica*

Informações do alinhamento	Valores
Sequências com alta qualidade	41,881,572
Némero de sequências mapeadas	65%
Conteúdo de GC	41.13%
N50	2,118 pb
Número total de contigs	127,600
Número total de unigenes (>200 bp)	65,364
Número de Proteínas codantes	24,548
Número proteínas codantes completes (*full lenght*)	12,560
Tamanho médio dos unigenes	1,264 pb

O processo de anotação automática foi realizado para identificar sequências de domínio conservadas e obter mapas das vias metabólicas do KEGG para caracterizar o nosso conjunto de dados de transcriptoma de café. Um total de 24.548 unigenes foram anotados com sucesso através da ferramenta BlastX, dentro desse total, 12.560 codificavam a sequência completa das proteínas (Tab. 1). *Vitis vinifera* (40,64%) foi a espécie com maior semelhança com sequências de café seguida por *Populus trichocarpa* (11,13%), *Ricinus communis* (10,89%) e *Glycine max* (4,24%). Também investigamos a contribuição de novos transcritos para estudos de transcriptoma de café. Comparamos nossa montagem com os 35.153 unigenes de café Arábica disponíveis no banco de dados CafESTs [3-4], 25.574 unigenes do genoma de *Coffea canephora* [2] e dados de transcriptoma de *Coffea eugenioides* (36.935 unigenes) [30]. Um total de 26.176 unigenes correspondiam aos contigs de CafEST, 24.798 unigenes correspondiam a CDS de genoma de *C. canephora* e 20.542 unigenes correspondiam ao transcriptoma de *C. eugenioides* (Tab. 2).

Tabela 2. Análise de similaridade entre as sequências transcritas depositadas em 3 banco de dados públicos de café

Reference Database	Hits	No Hits	Referencial bibliográfico
C. arabica (ESTs)	26.176	39.188	Mondego et al., 2011
C. canephora (genoma)	24.798	40.566	Denoued et al., 2014
C. eugenioides (*de novo* transcriptoma)	20.542	44.822	Yuyama et al., 2015

Identificamos nove unigenes relacionados à biossíntese de RFOs em nosso processo de anotação funcional e manual de genes (Fig 1). Os genes galactinol sintase (GolS), rafinose sintase (RS) e estacquiose sintase (STS) foram selecionados para análises posteriores (Tab. 3). Para cada unigene anotado como um possível gene das RFOs em *C. arabica*, foi realizada a identificação do seu respectivo gene ortólogo em *Arabidopsis thaliana*, no banco de dados de CafEST (*expressed sequence tags*) de *C. arabica* [4] e no genoma de *C. canephora* [2]. O processo de anotação via Blast2GO (Tab. 3) nos permitiu identificar os domínios conservados (pfam) para cada um dos genes anotados como pertencentes a RFOs. Além disso, os genes candidatos de galactinol, rafinose e estaquiose sintase foram mapeados na via metabólica das RFOs (metabolismo da galactose; MAP00052) disponível no banco de dados KEGG.

Tabela 3. Genes candidatos da família dos oligossacarídeos de café

Gene	Atividade enzimática	TAIR database	*C. canephora* (genoma)	CDD *database*	Pfam	Tamanho da proteina
CaGolS2	galactinol sintase	At1G56600	*Cc03_g00450*	PLN00176	pfam01501	345 aa
CaGolS3	galactinol sintase	At1G09350	*Cc02_g35350*	PLN00176	pfam01501	335 aa
CaGolS4	galactinol sintase	At1G60470	*Cc11_g15250*	PLN00176	pfam01501	339 aa
CaGolS8	galactinol sintase	At3G28340	*Cc11_g14010*	PLN00176	pfam01501	389 aa
CaGolS9	galactinol sintase	At3G06260	*Cc11_g10580*	PLN00176	pfam01501	350 aa
CaRS1	rafinose sintase	At1G55740	*Cc05_g15530*	PLN02355	pfam05695	678 aa
CaRS5	rafinose sintase	At5G40390	*Cc07_g01840*	PLN02355	pfam05695	782 aa
CaRS6	rafinose sintase	At5G20250	*Cc06_g08070*	PLN02355	pfam05695	870 aa
CaSTS	estaquiose sintase	At4G01970	*Cc01_g21600*	PLN02355	pfam05695	879 aa

Os perfis de expressão digital de genes (DGE) dos genes relacionados aos RFOs foram obtidos a partir dos valores de RPKM deste trabalho e dos disponíveis publicamente para *C. canephora* [49]. Observamos valores mais altos de RPKM nas folhas do que nos tecidos do perisperma para os genes *GolS2*, *GolS3*, *RS5* e *STS* nas duas espécies de *Coffea*. Por outro lado, observamos uma alta expressão de *RS1* no perisperma em comparação com as folhas em ambas as espécies. *CaGolS4* foi altamente expresso em folhas, comparado ao perisperma em *C. arabica*, em oposição ao observado em *C. canephora* (*CcGolS4*). Perfis de expressão semelhantes foram obtidos para *GolS8* no perisperma das duas espécies; no entanto, nas folhas, uma expressão mais alta foi detectada em *C. canephora* (*CcGolS8*) em comparação com *C. arabica*. *GolS9* e *RS6* exibiram perfis de expressão semelhantes em ambas as espécies de café, com pequenas diferenças entre folhas e perisperma (Fig. 1).

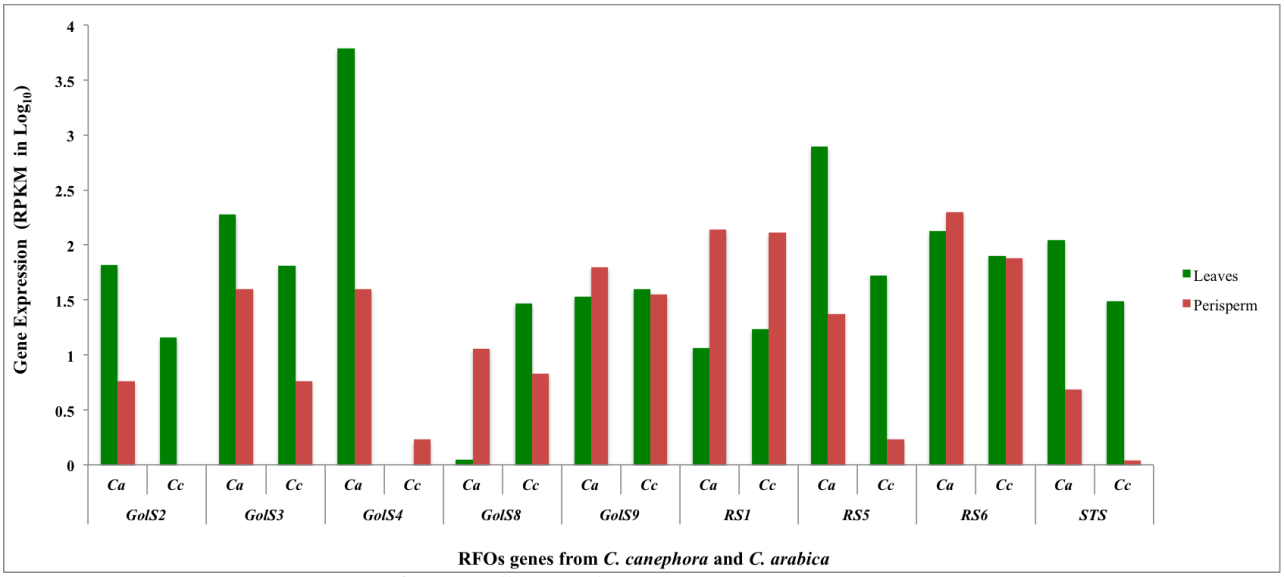

Figura 1. DGE dos genes relacionados a biossíntese da raffinose em folhas e perisperma. Valores de RPKM estão representados em escala de Log10. Folhas em verde, e perisperma en vermelho. *Ca= C. arabica*. *Cc= C. canephora. C. canéfora*. Os valores de RPKM de C. canephora foram obtidos do banco de dados *Coffee Genome Hub* [13].

Para validar o perfil de expressão digital dos dados de RNA-Seq, escolhemos quatro genes: *CaGolS2*, *CaGolS3*, *CaGolS4* e *CaRS1*. Os resultados do RT-qPCR seguiram o nosso padrão de expressão digital para todos os genes da RFO (Fig 2). Os genes *GolS* foram mais expressos nas folhas do que no perisperma dos frutos em todas as emostras avaliadas. O padrão oposto foi observado para o gene *CaRS1*, onde a expressão foi regulada positivamente no perisperma (90 a 150 DAF) em comparação com as folhas.

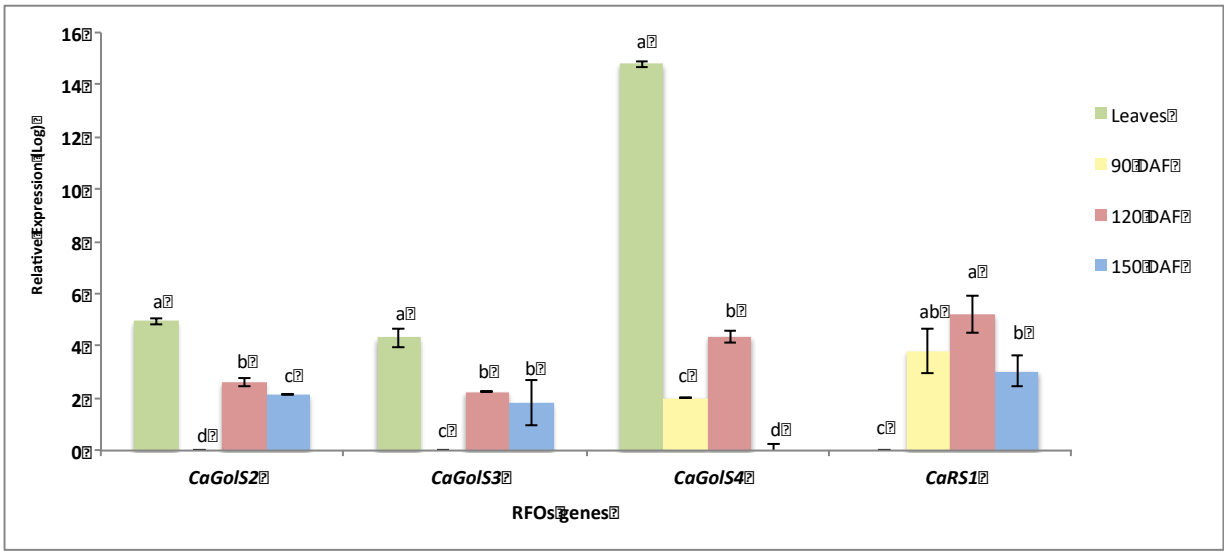

Figura 2. Análise de RT-qPCR dos genes RFOs candidatos. Folhas em verde e perisperma em amarelo (90 DAF), vermelho (120 DAF) e azul (120 DAF). Lower-case letters, from *a* to *d*, represent statistically significant differences for each RFO gene among coffee tissues.

CONCLUSÃO

Esse foi o primeiro trabalho de análise de transcriptoma em larga escala realizado para os tecidos: folhas, flores e frutos *C. arabica* durante os estágios iniciais de seu desenvolvimento com a metodologia de RNA-seq. Nossos dados revelaram um maior número de sequências gênicas completas para a codificação de proteínas do que os publicados anteriormente, assim como, identificamos genes específicos para os diferentes tecidos e estádios de desenvolvimento dos frutos. Fornecemos um conjunto de dados robustos para estudos futuros de transcriptoma com foco nos mecanismos genéticos que podem regular o desenvolvimento inicial dos frutos e a biossíntese de compostos bioquímicos presentes dos grãos de café. Além de permitir a identificação de genes transcricionalmente ativos em tecidos de café que são importantes para a produção e a qualidade da bebida. Este estudo está disponível publicamente na revista PloSOne Journal (DOI: 10.1371/journal.pone.0169595).

AGRADECIMENTOS

Nós gostaríamos de agradecer ao Consórcio Pesquisa Café, EMBRAPA, IAPAR, UNESP, CAPES, CNPq e FAPESP.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Davis AP, 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*. **Bot J Linn Soc**, 167:357–377, 2011.
2. Denoeud F, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. **Science**, 345(6201):1181-1184, 2014.
3. Vieira LGE, et al. Brazilian coffee genome project: an EST-based genomic resource. **Braz J Plant Physiol**, 18(1):95-108, 2006.
4. Mondego JMC, et al. An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. **BMC Plant Biol**, 11: 30, 2011.
5. Mutz KO, et al. Transcriptome analysis using next-generation sequencing. Curr Opin Biotech, 24(1):22-30, 2013.
6. Geromel C, et al. Biochemical and genomic analysis of sucrose metabolism during coffee (*Coffea arabica*) fruit development. **J Exp Bot**, 57(12): 3243-3258, 2006.
7. Joët T, et al. Regulation of galactomannan biosynthesis in coffee seeds. **J Exp Bot**, 65(1):323-337, 2014.
8. Sengupta S, et al. Significance of galactinol and raffinose family oligosaccharide synthesis in plants. **Front Plant Sci**, 2015.
9. de Souza Vidigal D, et al. Galactinol as marker for seed longevity. **Plant Sci**, 246: 112-118, 2016.
10. dos Santos TB, et al. Galactinol synthase transcriptional profile in two genotypes of *Coffea canephora* with contrasting tolerance to drought. **Genet Mol Biol**, 38(2):182-190, 2015.
11. 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**, 12(1): e0169595, 2017.
12. Silva FAS, Azevedo CAV. Principal Components Analysis in the Software Assistat-Statistical Attendance. In: World Congress on Computers in Agriculture. Reno-NV-USA: **Amer Soc Agric Biolog Eng**, 2009.
13. Dereeper A, Bocs S, Rouard M, Guignon V, Ravel S, Tranchant-Dubreuil C, et al. The coffee genome hub: a resource for coffee genomes. **Nucleic Acids Res**, 43(D1): D1028-D1035, 2015.