TRANSCRIPTOMA DE FRUTOS DE *Coffea arabica* L. AO LONGO DO SEU DESENVOLVIMENTO INICIAL[1]

Suzana Tiemi Ivamoto[2]; Osvaldo Reis Junior[3]; Marcelo Falsarella Carazzolle[4]; Douglas Silva Domingues[5]; Luiz Filipe Protasio Pereira[6]

1 Trabalho financiado pela Financiadora de Estudos e Projetos - FINEP
2 Doutoranda em Genética e Biologia Molecular, UEL, Londrina-PR, suzanatiemi@yahoo.com.br
3 Mestrando em Genética e Biologia Molecular, Unicamp, Campinas-SP, osvadoreisss@gmail.com
4 Pesquisador, MS, Unicamp, Campinas-SP, mcarrazo@lge.ibiunicamp..com
5 Pesquisador, DSc, IAPAR, Londrina-PR, doug@iapar.br
6 Pesquisador, PhD, Embrapa Café, Londrina-PR, filipe.pereira@embrapa.br

RESUMO: Neste trabalho foi realizado o seqüenciamento de RNA em larga escala de 12 bibliotecas de *Coffea arabica* cv. IAPAR59: folha, flor, frutos ao longo do seu desenvolvimento e frutos tratados com o indutor de metabolismo secundário metiljasmonato. Foram gerados no total 84.798.835 sequências (Illumina, HiSeq2000), que foram montadas em 127.600 contigs com tamanho médio de 1264bp, dos quais 65480 foram considerados como variantes de *splicing* únicos (unigenes). Neste trabalho, reportamos a montagem *de novo* do transcriptoma realizada com as sequências destas bibliotecas e os dados preliminares de anotação funcional e categorização. Trata-se do primeiro trabalho de sequenciamento Illumina com frutos de cafeeiro, que pode trazer importantes informações sobre genes candidatos relacionados a produção de compostos-chave envolvidos na qualidade do café e maturação de frutos.

PALAVRAS-CHAVE: RNA-Seq, Transcriptoma, *Coffea arabica*, Illumina

TRANSCRIPTOMIC RESOURCES FOR EARLY-STAGE FRUIT DEVELOPMENT OF *Coffea arabica* L.

ABSTRACT: In this work, 12 RNA-Seq libraries were obtained for *Coffea arabica* cv. IAPAR59: leaves, flowers and fruits along development and treated with methyljasmonate. A total of 84.798.835 sequences were generated using Illumina, HiSeq2000. After clusterization, 127.600 contigs were formed with an average size of 1264bp. From those, 65480 were considered unique splicing variants (unigenes). Here we report a *de novo* transcriptome assembly from those libraries and their initial functional annotation and categorization. This is the first work with Illumina sequencing of coffee fruits, which can provide important information on key genes related to enzymes and metabolites involved in fruit ripening and cup quality.

KEYWORDS: RNA-Seq, Transcriptome, *Coffea arabica*, Illumina

INTRODUÇÃO

A identificação conjunta de genes assim como de padrões de expressão gênica envolvidos em vias metabólicas durante o desenvolvimento de frutos de cafeeiro são importantes visando um maior o conhecimento destes na qualidade da bebida. Atualmente, estratégias de sequenciamento de alto desempenho do transcriptoma (RNA-Seq) permitem gerar um grande volume de dados, a baixo custo e em um curto período de tempo. Estas informações permitem a realização de estudos de perfis transcricionais e redes gênicas em uma compreensão aprofundada de diversas vias metabólicas. Além disso, a técnica de RNA-Seq pode ser utilizada para espécies que não possuem muitas informações genômicas como *C. arabica*. Além disto, não existem dados sobre a atividade transcricional de genes durante a fase inicial do desenvolvimento do fruto, no qual o tecido mais abundante é o perisperma e não o endosperma. Com o objetivo de caracterizar o transcriptoma desta importante espécie foi realizado o RNA-Seq de 12 bibliotecas diferentes com a tecnologia Illumina/Solexa (HiSeq2000; 100pb). Foram analisados: folha, flor, fruto ao longo do seu desenvolvimento e sob a ação do indutor de metabólitos secundário (metil jasmonato). Este trabalho pode aumentar as informações e o número de genes descritos para esta espécie, assim como desenvolver de um atlas global de unigenes ativos nessas bibliotecas.

MATERIAL E MÉTODOS

Foram utilizados frutos, folhas e flores de plantas de *Coffea arabica* L. cv. IAPAR59, cultivadas no Instituto Agronômico do Paraná (Londrina-PR). As amostras foram coletadas em campo e imediatamente congeladas em

nitrogênio líquido para posterior extração de RNA total. Os tecidos dos frutos foram separados (perisperma, endosperma e polpa) e apenas o perisperma foi selecionado para análise.
O RNA foi extraído com o protocolo descrito por Chang et al. (1993) com adaptações e as 12 bibliotecas de cDNA foram construídas a partir de diferentes tecidos e condições: folha, flor, frutos em cinco fases de desenvolvimento (30, 60, 90, 120 e 150 dias após florada-DAF), frutos tratados com metiljasmonato (durante 24 e 48 horas), bem como seus controles (0, 24 e 48 horas com água). A qualidade da extração de RNA foi analisada por eletroforese em gel de agarose denaturante (1,4%) e não-denaturante (1,0%). O RNA total foi quantificado pelo espectofotômetro NanoDrop® ND-100 (Thermo Scientific).
O seqüenciamento foi realizado pela "High Throughput Sequencing Facility" da "Carolina Center for Genome Sciences" (Universidade da Carolina do Norte - EUA). Para cada amostra, 5 µg de RNA total foram utilizadas para preparo da biblioteca de mRNAseq de acordo com o protocolo fornecido pela empresa Illumina. Para cada biblioteca, foi realizado seqüenciamento single-end, com sequências de 100pb de comprimento, em uma linha do Illumina Genome Analyzer IIx (Solexa/Illumina, HiSeq2000). Os adaptadores das sequências brutas foram removidos por *scripts* Perl e *reads* com alta qualidade (>20) foram selecionados com a ferramenta FastQC. Estes foram utilizados na montagem *de novo* dos transcritos com o *software* Trinity (Grabherr et al., 2011). Os unigenes identificados pelo Trinity foram analisados pelo programa Blast2GO para fins de anotação e caracterização funcional (Götz et al., 2008).

RESULTADOS E DISCUSSÃO

A seleção das amostras foi baseada em dados anteriores de quantificação de diterpenos para diferentes tecidos, estágios de desenvolvimento do fruto e frutos sob tratamento com o indutor de metabólitos secundários metil jasmonato através de cromatografia líquida de alta precisão (Ivamoto, 2012; Kitzberger et al., 2013).
O número de sequências (originais e pós limpeza) resultantes do RNA-Seq para as 12 bibliotecas estão descritas na tabela 1 de acordo com os seus respectivos tecidos. O total de sequências brutas geradas foi de 84.798.835 e após o processo de trimagem, 41.881.572 sequências foram utilizadas para a montagem dos contigs pelo software Trinity.

Tabela 1. Número de leituras para as 12 bibliotecas de mRNAseq

Bibliotecas	Tecidos	Número de Sequências (dado Bruto)	Número de Sequências (qualidade>20)
1	Fruto Controle 0h	3440068	1357155
2	Fruto Controle 24h	11053426	6277406
3	Fruto Controle 48h	17444494	6993057
4	Fruto MJ 24h	829876	691859
5	Fruto MJ 48h	4938522	3028506
6	Flor	6471863	3221890
7	Folha	3499743	2124179
8	Fruto 30DAF	16324779	7049563
9	Fruto 60DAF	2530615	1893293
10	Fruto 90DAF	9207895	4398819
11	Fruto 120DAF	6724365	3608856
12	Fruto 150DAF	2333189	1236989
	TOTAL	84798835	41881572

Obteve-se um total de 127.600 contigs acima de 200bp (tabela 2). Considerando apenas a maior variante de *splicing* de cada gene, o resultado final foi de 65.480 unigenes.
Os dados de transcriptômica publicados até o presente momento para *C. arabica* foram obtidos com as tecnologias GS FLX Titanium – 454 (Vidal et al., 2011) e Sanger (Mondego et al., 2011) obtiveram um total de 55.578 e 32.007 contigs, respectivamente. Comparando os dados deste trabalho com os da literatura, este possui aproximadamente 15% mais unigenes em relação à metodologia 454 da Roche e 50% com relação à tecnologia Sanger de seqüenciamento e montagem.
O maior número de unigenes apresentados pelas tecnologias de RNA-Seq provavelmente é uma consequência do maior volume de sequências obtidas através da tecnologia Illumina e 454, em relação à metodologia Sanger de seqüenciamento, além do fato de que as montagens foram feitas com estratégias diferentes. As sequências de 454 foram montadas com o software MIRA (Chevreux et al., 2004) e as sequências Sanger com o programa CAP3 (Huang et al., 1999). Além disto, o tamanho médio dos contigs obtidos por este trabalho foi quase o dobro do obtido por seqüenciamento Sanger (tabela 2).

Tabela 2. Resultados comparativos de RNA-Seq (Illumina e 454) e seqüenciamento Sanger de *C. arabica*

Informações (*C. arabica*)	RNA-Seq (Illumina)	RNA-Seq (454)	Sanger (ESTs)
Número Total de Sequências	41881572	740627	135876
Tamanho Médio de cada read	100 pb	256pb	Nd
Número Total de Contigs	65480	55578	32007
Tamanho Médio dos Contigs	1264 pb	Nd	663 pb

Nd: informações não disponíveis.

As anotações funcionais dos 65480 unigenes encontrados estão em desenvolvimento pelo programa Blast2GO. Análises de BlastX contra o banco de dados do NCBI-nr, UniProt e TAIR encontraram 24.548 unigenes com similaridade a sequências previamente anotadas. A espécie que resultou maior número de sequências similares com os dados deste trabalho foi *Vitis vinifera* com 40.6% dos *hits*.
Foram encontrados 74.108 termos GOs (gene ontology) para as 3 diferentes classes: processos biológicos (P), localização celular (C) e função molecular (F). A maior parte dos contigs encontra termos anotados nos níveis 7 e 8 para as 3 grandes classes P, C e F (Figura 1).

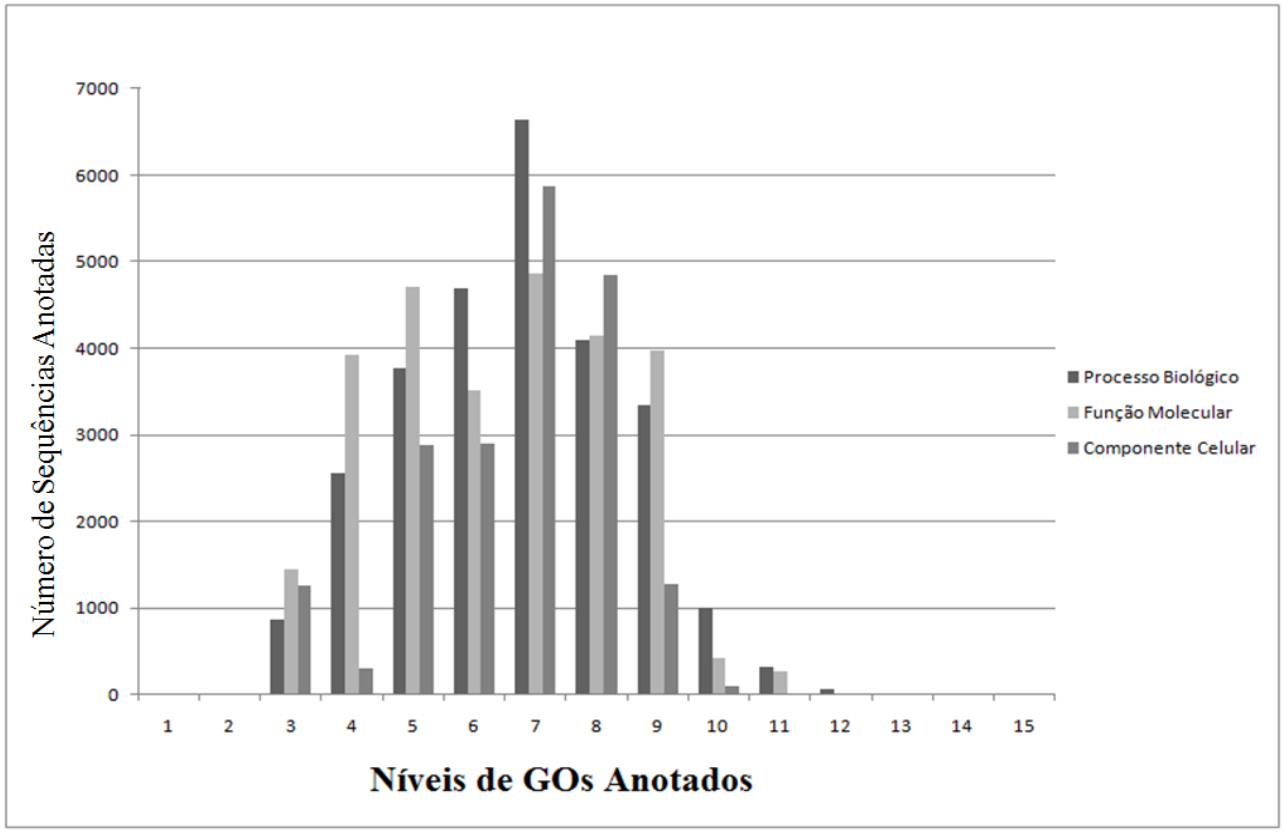

Figura 1. Anotação dos níveis de GOs encontrados através das análises realizadas no Blast2GO.

Os unigenes encontrados foram relacionados a 139 vias metabólicas do KEGG (*Kyoto Encyclopedia of Genes and Genomes*, avaliado pelo Blast2GO). Um total de 106.266 domínios conservados (DC) foram encontrados através da análise do InterProScan. Na figura 2 estão representados os 25 domínios mais encontrados (22.399) entre os 65.480 unigenes.
Baseado nestes resultados *in silico*, as próximas análises serão com relação aos genes diferencialmente expressos entre as 12 bibliotecas. Com a elaboração deste atlas de genes transcricionalmente ativos e diferencialmente expressos, pretende-se identificar genes específicos para determinados tecidos e em diferentes rotas metabólicas, e propiciar inferências sobre genes relacionados ao desenvolvimento de frutos e com a qualidade da bebida. .

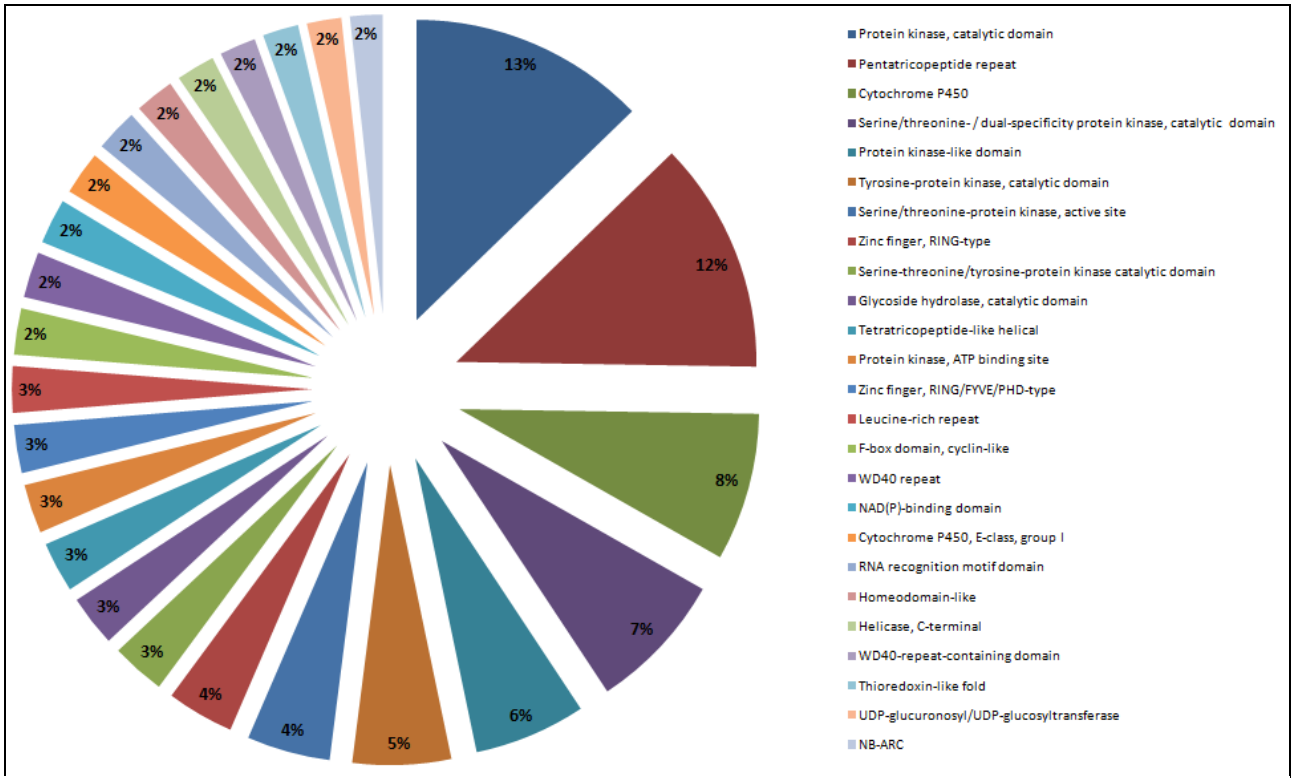

Figura 2. Resultado do Blast2GO para os 25 domínios conservados mais preponderantes para os 65.480 unigenes anotados.

CONCLUSÕES

Neste estudo, reportamos dados preliminares de uma análise global do transcriptoma de *Coffea arabica* cv. IAPAR59, o qual servirá de base para os futuros estudos de expressão gênica desta espécie. O aumento do conhecimento sobre a atividade transcricional nestes tecidos irá proporcionar uma maior compreensão de como eles influenciam a síntese e degradação de diversos compostos bioquímicos. Este painel de unigenes é uma importante ferramenta de informação que auxiliará a direção de projetos futuros de caracterização e validação de genes que contribuam para compreensão de processos fisiológicos e bioquímicos relacionados ao desenvolvimento dos frutos e da qualidade da bebida.

AGRADECIMENTOS

Aos órgãos financiadores do projeto: FINEP-Genocafé; Consórcio Pesquisa Café; CNPq. A CAPES-Embrapa pela concessão da bolsa de doutorado à STI.

REFERÊNCIAS BIBLIOGRÁFICAS

Carazzolle, M.F., Rabello, F.R., Martins, N.F., de Souza, A.A., do Amaral, A.M., Freitas-Astua, J., ... & Mehta, A. (2011). Identification of defence-related genes expressed in coffee and citrus during infection by Xylella fastidiosa. European journal of plant pathology, 130(4), 529-540.

Chang, S, Puryear, J., Cairney, J.A. (1993). A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology, 11, 113-116.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A. J., Müller, W. E., Wetter, T., & Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome research, 14(6), 1147-1159.

Dudareva, N., Klempien, A., Muhlemann, J.K., & Kaplan, I. (2013). Biosynthesis, function and metabolic engineering of plant volatile organic compounds. New Phytologist.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., ... & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-3435.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit,I., ... & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, 29(7), 644-652.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program.Genome research, 9(9), 868-877.

Ivamoto, S.T. (2012). Diterpenos em *Coffea arabica*: aspectos bioquímicos e caracterização *in silico* e transcricional de genes de Cyt P450s candidatos na biossíntese e cafestol e caveol. Dissertação de Mestrado em Genética e Biologia Molecular – Universidade Estadual de Londrina, Londrina-PR. 94p.

Kitzberger, C.S.G., Scholz, M.B.D.S., Pereira, L.F.P., Vieira, L.G.E., Sera, T., Silva, J.B.G.D., Benassi, M.D.T. (2013). Diterpenes in green and roasted coffee of Coffea arabica cultivars growing in the same edapho-climatic conditions. Journal of Food Composition and Analysis.

Lulin, H., Xiao, Y., Pei, S., Wen, T., & Shangqin, H. (2012). The first Illumina-based *de novo* transcriptome sequencing and analysis of safflower flowers. PloS one, 7(6), e38653.

Mondego, J.M., Vidal, R.O., Carazzolle, M.F., Tokuda, E.K., Parizzi, L.P., Costa, G.G., ... & Pereira, G.A. (2011). An EST-based analysis identifies new genes and reveals distinctive gene expression features of Coffea arabica and Coffea canephora. BMC plant biology, 11(1), 30.

Shi, C.Y., Yang, H., Wei, C.L., Yu, O., Zhang, Z.Z., Jiang, C.J., ... & Wan, X.C. (2011). Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC genomics, 12(1), 131.

Vidal, R., Leroy, T., De Bellis, F., Pot, D., Rodrigues, G.C., Pereira, G.A.G., Andrade, A.C., Marraccini, M. (2011). Construção do perfil de expressão gênica da resistência à seca do café a partir de dados de seqüenciamento de segunda geração. In: VII Simpósio de Pesquisa dos Cafés do Brasil, 2011, Araxá-MG. VII Simpósio de Pesquisa dos Cafés do Brasil. Brasília-DF. Embrapa Café.

Ziemann, M., Kamboj, A., Hove, R.M., Loveridge, S., El-Osta, A., & Bhave, M. (2013). Analysis of the barley leaf transcriptome under salinity stress using mRNA-Seq. Acta Physiologiae Plantarum, 1-10.