GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

CoffeebEST: an integrated resource for *Coffea* spp expressed sequence tags

A.R. Paschoal[1], E.D.M. Fernandes[1,2], J.C. Silva[1,2], F.M. Lopes[1], L.F.P. Pereira[2,3] and D.S. Domingues[2]

[1]Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, Brasil
[2]Laboratório de Biotecnologia Vegetal, Instituto Agronômico do Paraná, Londrina, PR, Brasil
[3]Embrapa Café, Brasília, DF, Brasil

Corresponding author: D.S. Domingues
E-mail: doug@iapar.br

Genet. Mol. Res. 13 (4): 10913-10920 (2014)
Received May 13, 2014
Accepted September 25, 2014
Published December 19, 2014
DOI http://dx.doi.org/10.4238/2014.December.19.13

ABSTRACT. Coffee is one of the most important commodities in the world, and its production relies mainly on two species, *Coffea arabica* and *Coffea canephora*. Although there are diverse transcriptome datasets available for coffee trees, few research groups have exploited the potential knowledge contained in these data, especially with respect to fruit and seed development. Here, we present a comparative analysis of the transcriptomes of *Coffea arabica* and *Coffea canephora* with a focus on fruit development using publicly available expressed sequence tags (ESTs). Most of the fruit and seed EST data has been obtained from *C. canephora*. Therefore, we performed a fruit EST analysis of the 5 developmental stages of this species (18, 22, 30, 42, and 46 weeks after flowering) comprising 29,009 sequences. We compared *C. canephora* fruit ESTs to reference unigenes of *C. canephora* (7710 contigs and 8955 singletons) and *C. arabica* (15,656 contigs and 16,351 singletons). Additional analyses included functional annotation based on Gene Onthology, as well as an annotation using PlantCyc, a curated plant protein database. The Coffee Bean EST (CoffeebEST) is a public database available at http://bioinfo-02.cp.utfpr.edu.br/. This database

represents an additional resource for the coffee scientific community, offering a user-friendly collection of information for non-specialists in coffee molecular biology to support experimental research on comparative and functional genomics.

Key words: Fruit; *Coffea arabica*; *Coffea canephora*; Bioinformatics; Transcriptome; Expressed sequence tag

INTRODUCTION

Coffee is one of the most important commodities in the world, with more than 80 million people depending on the coffee chain for their income (Dereeper et al., 2013). Brazil is the major producer and exporter of coffee, and is its second-highest consumer. The *Coffea* genus has more than 120 species (Davis et al., 2011), but global production relies mainly on two species, *Coffea arabica* (65%) and *C. canephora* (35%) (ICO, www.ico.org). *C. arabica* (CA) is a tetraploid species (2n = 4x = 44) that is probably derived from a recent (< 1 million years) hybridization between *C. canephora* and *C. eugenioides* (Vidal et al., 2010; Yu et al., 2011), and *C. canephora* (CC) is an allogamous diploid species (2n = 2x = 22). Although, there have been efforts to study the *Coffea* species genome composition and transcriptional patterns (Lin et al., 2005; Vidal et al., 2010; Mondego et al., 2011; Combes et al., 2013; Dereeper et al., 2013), few research groups have exploited the potential knowledge contained in these data, especially with regard to fruit and grain development. The use of bioinformatics through the development of programs and databases can help in this task not only to analyze the available data but also to generate new knowledge. One example is the standardization of studies of molecular markers in coffee (Plechakova et al., 2009). Most of the large-scale transcriptomic studies in *Coffea* are focused on stress responses (Bardil et al., 2011; Carazzolle et al., 2011; Marraccini et al., 2012; Combes et al., 2013), and seed development studies are mostly focused on selected candidate genes (Lepelley et al., 2007; Joët et al., 2009; Budzinski et al., 2011). Understanding these large-scale studies would help to identify genes involved in the chemical composition of the bean and the sensorial quality of the coffee beverage. This study presents the development of a website and a bioinformatic system that incorporates the publicly available expressed sequence tags (ESTs) of CC and CA, which were previously reported by Mondego et al. (2011). Our study aimed to explore this transcriptome analysis to develop a user-friendly website including functional annotation and basic local alignment search tool (BLAST) results and focusing on fruit development in CC. Based on these data, scripts implemented in Java and PostgreSQL enabled the evaluation of CC- and CA-specific transcripts. This system also allows the application of diverse filters to retrieve stage-specific or species-specific transcripts. Overall, we offer a public repository of sequence and functional annotation of *Coffea* unigenes with special emphasis on understanding the functional differences among the repertoire of expressed *Coffea* genes in 5 stages of fruit development. The database is available at http://bioinfo-02.cp.utfpr.edu.br/.

MATERIAL AND METHODS

Coffea transcriptome resources

The entire Sanger EST public dataset available for *Coffea* spp, from the studies by

Modego et al. (2011) (http://www.lge.ibi.unicamp.br/coffea) and Lin et al. (2005) (http://solgenomics.net/content/coffee.pl) were downloaded. The assembly of these ESTs represents the *Coffea* unigene set mostly used for transcriptomic analyses (Vidal et al., 2010; Bardil et al., 2011; Combes et al., 2013). These sequences comprise 35,153 CA contigs and 18,007 CC contigs. The cDNA reads are derived from several organs and developmental stages of coffee plants. CoffeebEST comprises assembled and unassembled data, annotation of individual sequences, and functional analysis focused on fruit transcriptome data. The CC fruit transcriptome dataset was obtained from the Sol Genomics network, and it is composed of 726 sequences of 18-week fruits, 9113 sequences of 22-week fruits, 10,077 sequences of 30-week fruits, 210 sequences of 42-week fruits, and 8883 sequences of 46-week fruits. These datasets are all available for download at the CoffeebEST database.

Bioinformatic comparative analysis

Fruit ESTs were used as queries for sequence similarity analysis against EST contigs of CA and CC, which were assembled by Mondego et al. (2011) as the subject. Nucleotide BLAST was used with the following parameters: dust filter and E-value of 10^{-10}. Results were filtered using ≤ 85% alignment coverage in the query sequence and ≤ 85% identity as inclusion criteria.

Functional annotation associations

All *Coffea* sequences were mapped and quantified using Gene Ontology (Ashburner et al., 2000) terms. Coffee datasets were annotated and mapped for the gene ontologies "biological process" and "molecular function" (only level 3). All sequences were functionally classified using PlantCyc annotation (ftp://ftp.plantcyc.org/Pathways/BLAST_sets/; Zhang et al., 2010) and coffee EST contig annotation (www.lge.ibi.unicamp.br/coffea). Results were filtered using ≤ 85% alignment coverage in the query sequence and ≤ 85% identity as inclusion criteria.

Biological analysis of data from CoffeebEST

The 30 most expressed CC fruit clusters and 161 CC-exclusive clusters were annotated using translated nucleotide BLAST (threshold of e-05) and the TRAPID tool (Van Bel et al., 2014). Results of the translated nucleotide BLAST for the most expressed clusters are shown in **Table S1**.

CoffeebEST website and database

The CoffeebEST website was built in NetBeans 7.2 IDE by adopting Java Server Pages technology, the Apache Tomcat 7 server, and the PostgreSQL database (**Figure S1**). The dataset was parsed and treated using PERL scripts to insert this information in a database. Finally, a local National Center for Biotechnology Information (NCBI) BLAST (BLASTN program, version 2.2.25) was implemented to align sequences against *Coffea* databases available in CoffeebEST. Once a hit is identified, the user can use the relational database to inspect the sequence/annotation. The website is hosted on a Linux Ubuntu Server and is available at

http://bioinfo-02.cp.utfpr.edu.br/.

RESULTS AND DISCUSSION

Comparative analysis of *Coffea* species

We parsed Sanger EST sequences as described in Figure 1 (see details in the Material and Methods section).

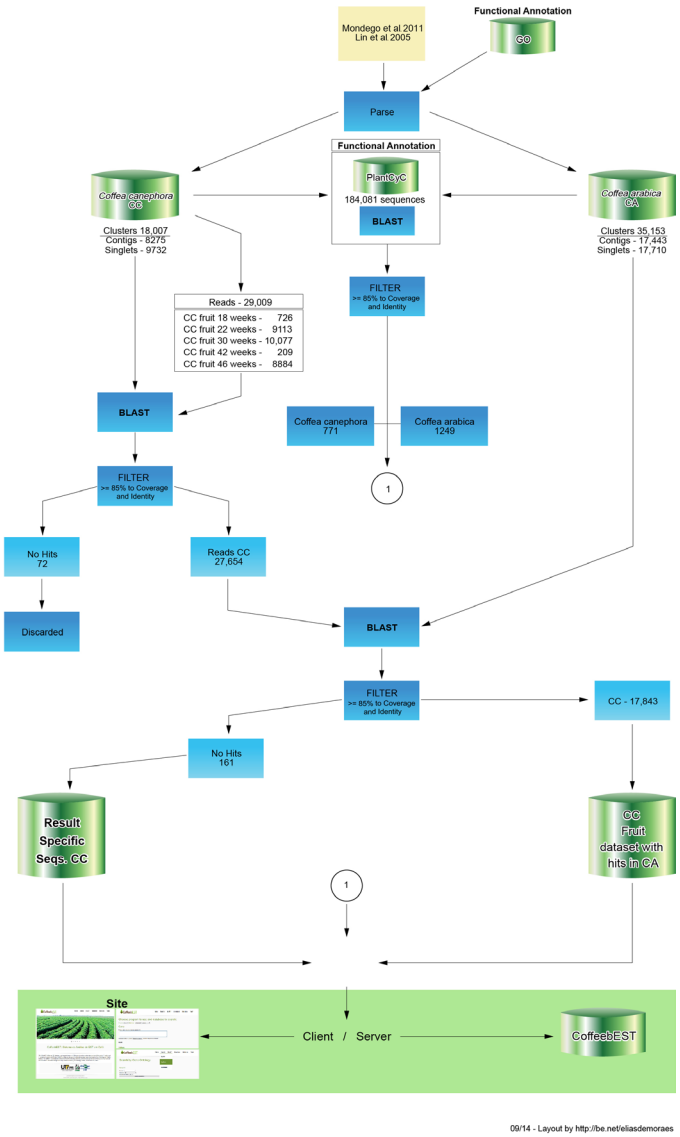

Figure 1. Pipeline of the workflow for comparative analysis of *Coffea* species.

First, a total of 29,009 CC fruit EST reads were mapped to CC contigs using BLAST. In order to remove sequences derived from contamination and retrieve the most expressed genes in CC fruits, a filter step was performed. A total of 30 CC clusters were mapped with >100 fruit ESTs representing the most expressed contigs among CC data (**Table S1**).

Secondly, the 27,654 CC fruit reads aligned in the previous step were then compared with 35,153 CA clusters. Considering our criteria, in this second BLAST, we could infer that 17,843 sequences are conserved between CA and CC while 161 CC fruit reads are specific to CC. This comparative analysis was used to build the CoffeebEST website.

Functional annotation analysis

To analyze the functional annotation information of the *Coffea* transcriptome dataset, we extracted and parsed all PlantCyc annotations. PlantCyc proteins were used as a database to query the sequences of the CC and CA datasets (Figure 1). We obtained functional annotation information for 771 CC clusters and 1249 CA clusters. We also extracted and parsed the functional information from the Gene Ontology classification based on the method described by Mondego et al. (2011). This information was used to assign gene ontology terms to the *Coffea* species transcriptome data.

CoffeebEST: a resource of coffee ESTs

CoffeebEST is mainly designed to further characterize fruit-specific *Coffea* spp. transcript sequences by comparative analysis with publicly available CA and CC EST resources. The website is a relational database implemented in PostgreeSQL to organize, store, and retrieve normalized information about fruit-related genes in coffee (**Figure S1**).

Because fruit ESTs are available only from CC, the analysis relied on the discovery of fruit ESTs that are exclusively found in CC or in CA homologs. We parsed and compared the gene ontology data with the PlantCyc functional annotation information to functionally annotate all these sequences and also to compare CC and CA fruit ESTs. All comparison results were integrated in a user-friendly website to help retrieve the information (Figure 2).

Figure 2. CoffeebEST website portal with all of the options available to the user.

CoffeebEST allows users to: i) search the database for information; ii) perform BLAST against CA and CC datasets; iii) download all sequences available on the website; and iv) perform comparative analysis of CA, CC, and annotation data. The search option on each page presents the main functions that allow users to search for information about the CoffeebEST database. This tool gives the user 4 search options. The first search is by the dataset that was used in the comparative analysis between CA and CC (Figure 3); simultaneously, the user can also choose the parameters for coverage and identity. The results are given in a table with a list of contigs/reads. The user can choose a record result to see the details page. On the details page, all of the specific information about the CA and CC results and functional annotation from Gene Ontology and PlantCyc are presented. The sequence of each alignment hit is also presented for the user with a download option. The second search option is by gene ontology terms against the same analysis, where the user can also choose one of the main ontology categories. The third option is to search by gene; with this option, the user needs to type at least some key words to find the right information. The last search option is to search by ID code.

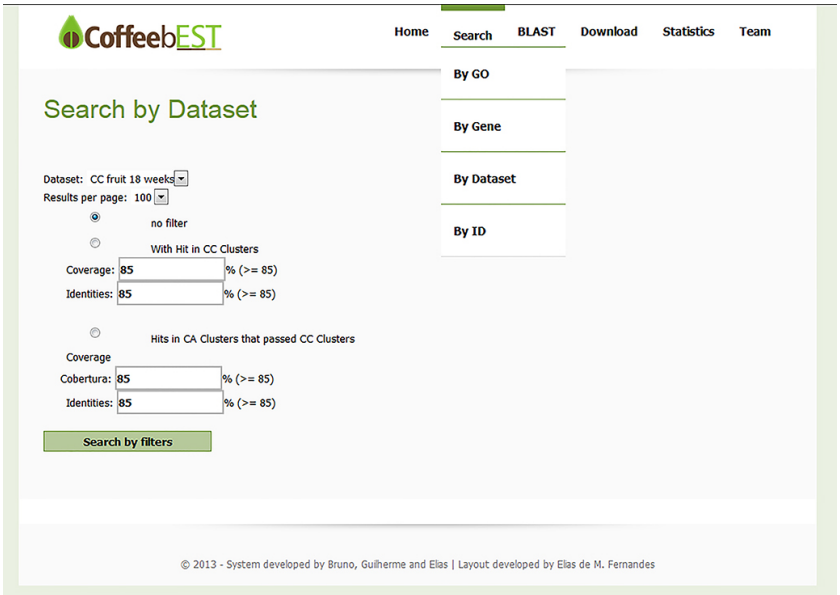

Figure 3. Search pages in CoffeebEST with 4 search options available: by dataset, by Gene Ontology (GO), by gene, and by ID.

Biological analysis of data from CoffeebEST

In order to demonstrate the usage of CoffeebEST, we focused our analysis on two data subsets: i) the 30 most-expressed genes in CC and ii) the 161 clusters that are exclusive to CC without a homolog in CA according to our criteria.

Clusters with more than 100 mapped reads comprise the 30 most-expressed genes in CC. Among these clusters, 7 do not have an Interpro domain. A detailed annotation and expression pattern are assigned in **Table S1**, and they are in agreement with a previous report

of Lin et al. (2005). Most clusters are preferentially expressed in fruits 46 weeks after flowering (WAF), followed by 30 WAF. The most-expressed gene in 46 WAF fruits, Contig4069, is an 11S globulin. This globulin represents the most important storage protein in coffee seeds (Rogers et al., 1999). Transcriptional data in CA identified a peak at 30 WAF with mRNA downregulation in later ripening stages (Rogers et al., 1999). However, in CA, we identified reads associated with 11S globulin in fruits 42 and 46 WAF, suggesting that the globulin profile is not the same in CC and CA.

The most-expressed gene in fruits 30 WAF, Contig5887, was 2S albumin that is also an important storage protein in seeds (Tai et al., 1999). In fruits 22 WAF, the most-expressed cluster does not have any hits in the NCBI and Interpro databases, indicating that unknown proteins may have important roles in coffee seed development.

Among the 161 clusters that were exclusively found in CC fruits, 53 (32.9%) have an Interpro domain. The most represented domain is IPR005636 (DTW). This domain is present in 7 clusters, and its function remains unknown despite the presence of orthologs in several plant species. Soybean was the species that had the best similarity search hit for CC-exclusive transcripts (22.8%), which contrasts with most large-scale analysis in coffee, where grape is usually the species with the closest homologs of *Coffea* (Guyot et al., 2009; Mondego et al., 2011). A total of 10 clusters had a significant hit (> 400 nt) against long terminal repeat retrotransposons (**Table S1**), indicating that a significant part of CC-exclusive clusters may represent transcriptionally active transposable elements.

CONCLUSION

Here, using a bioinformatic approach we compared the transcriptome of 2 *Coffea* species: CA and CC. CoffeebEST is a website and repository of EST comparisons focused on giving a user-friendly analysis of bean EST data. We believe that this database will help improve the current understanding of *Coffea* species and their molecular datasets.

ACKNOWLEDGMENTS

Research supported by Conselho Nacional de Pesquisa e Desenvolvimento Tecnológico (CNPq), Fundação Araucária, and the Diretoria de Pesquisa e Pós-Graduação (DIRPPG) of Universidade Tecnológica Federal do Paraná (UTFPR) - Câmpus Cornélio Procópio. This project is part of the collaboration between UTFPR and Instituto Agronômico do Paraná.

Supplementary material

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25-29.

Bardil A, de Almeida JD, Combes MC, Lashermes P, et al. (2011). Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. *New Phytol.* 192: 760-774.

Budzinski IG, Santos TB, Sera T, Pot D, et al. (2011). Expression patterns of three alpha-expansin isoforms in *Coffea arabica* during fruit development. *Plant Biol.* 13: 462-471.

Carazzolle MF, Rabello FR, Martins NF and Souza AA (2011). Identification of defence-related genes expressed in coffee and citrus during infection by *Xylella fastidiosa*. 529-540.

Combes MC, Dereeper A, Severac D, Bertrand B, et al. (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 AP, Tosh J, Ruch N and Fay MF (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*. 357-377.

Dereeper A, Guyot R, Tranchant-Dubreuil C, Anthony F, et al. (2013). BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. *Plant Mol. Biol.* 83: 177-189.

Guyot R, de la Mare M, Viader V, Hamon P, et al. (2009). Microcollinearity in an ethylene receptor coding gene region of the *Coffea canephora* genome is extensively conserved with *Vitis vinifera* and other distant dicotyledonous sequenced genomes. *BMC Plant Biol.* 9: 22.

Joët T, Laffargue A, Salmona J, Doulbeau S, et al. (2009). Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study. *New Phytol.* 182: 146-162.

Lepelley M, Cheminade G, Tremillon N and Simkin A (2007). Chlorogenic acid synthesis in coffee: an analysis of CGA content and real-time RT-PCR expression of *HCT*, *HQT*, *C3H1*, and *CCoAOMT1* genes during grain development in *C. canephora*. 978-966.

Lin C, Mueller LA, Mc CJ, Crouzillat D, et al. (2005). Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theor. Appl. Genet.* 112: 114-130.

Marraccini P, Vinecky F, Alves GS, Ramos HJ, et al. (2012). Differentially expressed genes and proteins upon drought acclimation in tolerant and sensitive genotypes of *Coffea canephora*. *J. Exp. Bot.* 63: 4191-4212.

Mondego JM, Vidal RO, Carazzolle MF, Tokuda EK, et al. (2011). An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. *BMC Plant Biol.* 11: 30.

Plechakova O, Tranchant-Dubreuil C, Benedet F, Couderc M, et al. (2009). MoccaDB - an integrative database for functional, comparative and diversity studies in the Rubiaceae family. *BMC Plant Biol.* 9: 123.

Rogers WJ, Bézard G, Deshayes A and Meyer I (1999). Biochemical and molecular characterization and expression of the 11S-type storage protein from *Coffea arabica* endosperm. *Plant Physiol. Bioch.* 37: 261-272.

Tai SS, Wu LS, Chen EC and Tzen JT (1999). Molecular cloning of 11S globulin and 2S albumin, the two major seed storage proteins in sesame. *J. Agric. Food Chem.* 47: 4932-4938.

Van Bel M, Proost S, Van Neste C, Deforce D, et al. (2013). TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. *Genome Biol.* 14: R134.

Vidal RO, Mondego JM, Pot D, Ambrosio AB, et al. (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.

Yu Q, Guyot R, de Kochko A, Byers A, et al. (2011). Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*). *Plant J.* 67: 305-317.

Zhang P, Dreher K, Karthikeyan A, Chi A, et al. (2010). Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.* 153: 1479-1491.