Mol Genet Genomics
DOI 10.1007/s00438-015-1111-x

ORIGINAL ARTICLE

# Transcriptome analysis in *Coffea eugenioides*, an Arabica coffee ancestor, reveals differentially expressed genes in leaves and fruits

**Priscila Mary Yuyama[1,2] · Osvaldo Reis Júnior[3] · Suzana Tiemi Ivamoto[1,2] · Douglas Silva Domingues[2,6] · Marcelo Falsarella Carazzolle[3] · Gonçalo Amarante Guimarães Pereira[3] · Pierre Charmetant[4] · Thierry Leroy[4] · Luiz Filipe Protasio Pereira[2,5]**

Received: 16 March 2015 / Accepted: 24 August 2015


**Abstract** Studies in diploid parental species of polyploid plants are important to understand their contributions to the formation of plant and species evolution. *Coffea eugenioides* is a diploid species that is considered to be an ancestor of allopolyploid *Coffea arabica* together with *Coffea canephora*. Despite its importance in the evolutionary history of the main economic species of coffee, no study has focused on *C. eugenioides* molecular genetics. RNA-seq creates the possibility to generate reference transcriptomes and identify coding genes and potential candidates related to important agronomic traits. Therefore, the main objectives were to obtain a global overview of transcriptionally active genes in this species using next-generation sequencing and to analyze specific genes that were highly expressed in leaves and fruits with potential exploratory characteristics for breeding and understanding the evolutionary biology of coffee. A de novo assembly generated 36,935 contigs that were annotated using eight databases. We observed a total of ~5000 differentially expressed genes between leaves and fruits. Several genes exclusively expressed in fruits did not exhibit similarities with sequences in any database. We selected ten differentially expressed unigenes in leaves and fruits to evaluate transcriptional profiles using qPCR. Our study provides the first gene catalog for *C. eugenioides* and enhances the knowledge concerning the mechanisms involved in the *C. arabica* homeologous. Furthermore, this work will open new avenues for studies into specific genes and pathways in this species, especially related to fruit, and our data have potential value in assisted breeding applications.

**Keywords** *Coffea* · RNA-seq · Gene annotation · Differentially expressed genes · Homeologous

Communicated by S. Hohmann.

P. M. Yuyama, O. Reis Júnior and S. T. Ivamoto contributed equally to this work.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00438-015-1111-x) contains supplementary material, which is available to authorized users.

✉ Luiz Filipe Protasio Pereira
filipe.pereira@embrapa.br

[1] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Estadual de Londrina (UEL), Londrina, Paraná, Brazil

[2] Laboratório de Biotecnologia Vegetal, Instituto Agronômico do Paraná (IAPAR), Londrina, Paraná, Brazil

[3] Laboratório de Genômica e Expressão, Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, São Paulo, Brazil

[4] Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR AGAP, 34398, Montpellier, France

[5] Embrapa Café, Brasília, DF, Brazil

[6] Present Address: Department of Botany, Instituto de Biociências de Rio Claro, Universidade Estadual Paulista (UNESP), Rio Claro, São Paulo, Brazil

## Introduction

Polyploidization results from endopolyploidy or from the fusion of unreduced gametes. All seed plants are thought to have undergone at least one round of polyploidization

Published online: 03 September 2015

in their history (Jiao et al. 2011). Allopolyploidy is a process of the coexistence of two or more sets of divergent genomes that is associated with major changes at the genetic and epigenetic levels, with important consequences for the expression patterns of genes from both genomes and for the phenotype (Doyle et al. 2008; Jackson and Chen 2010). *Coffea arabica* ($2n = 4\times = 44$) is the only allotetraploid species of coffee in the genus and is derived from a recent hybridization event of two diploid species: *Coffea canephora* Pierre ex A. Froehner ($2n = 2\times = 22$) and *Coffea eugenioides* S. Moore ($2n = 2\times = 22$) (Lashermes et al. 1999). This hybridization event, which was most likely followed by genome duplication (Combes et al. 2012), occurred approximately 100–500 thousand years ago (Yu et al. 2011) or possibly more recently (between 10 and 50 thousand years ago) (Cenci et al. 2012).

The ancestors of *C. arabica* present different agroecological adaptations and characteristics. *C. canephora* grows better in lowlands and is characterized by higher productivity, tolerance to pests, drought stress and caffeine content, but its beverage is considered to be of lower quality compared with *C. arabica*. Therefore, it is used mostly by the instant coffee industry and/or in blends with *C. arabica* (Leroy et al. 2006). *C. eugenioides* grows in highlands and near forest edges in Central-East Africa and is not produced on a commercial scale due its low fruit production. *C. eugenioides* was included in breeding programs to reduce caffeine levels and improve cup quality because it produces small fruits with low caffeine content compared with both *C. arabica* and *C. canephora* (Mazzafera and Carvalho 1992). *C. arabica* can be grown in regions with marked variations in thermal amplitude and has a better cup quality compared with *C. canephora* (Leroy et al. 2006).

Most coffee transcriptome sequencing data relies on two major cultivated species (*C. arabica* and *C. canephora*). Sanger EST sequencing projects were developed for *C. arabica* (Vieira et al. 2006; Vidal et al. 2010; Mondego et al. 2011) and *C. canephora* (Lin et al. 2005; Poncet et al. 2006). Transcriptome analysis using next-generation sequence—NGS has been performed in studies investigating biotic and abiotic interactions (Fernandez et al. 2004; Combes et al. 2013), and recently a high draft of the *C. canephora* genome was generated (Denoeud et al. 2014). The two subgenomes contained in the *C. arabica* allotetraploid genome (subgenome *C. canephora*—CaCc and subgenome *C. eugenioides*—CaCe) do not contribute equally to the transcriptome (Vidal et al. 2010; Cotta et al. 2014).

Vidal et al. (2010) proposed that CaCc within the *C. arabica* transcriptome is more associated with the expression of genes encoding regulatory proteins, while CaCe expression appears to be more closely linked with basal processes. In another work, when the *CaWRKY1a* and *CaWRKY1b* genes that encode for transcription factors involved in the coffee defense responses to abiotic and biotic stresses were concomitantly expressed, both homeologous genes were found to contribute to the overall transcriptional level (Petitot et al. 2008). Given the regulation mechanisms between homeologous genes, these studies and others (Carvalho et al. 2014; Cotta et al. 2014) demonstrated the complexity of regulation in allopolyploids and indicated that genes useful for *C. arabica* breeding programs could be present in its genome but have become inactive due to partitioned expression. However, one of the drawbacks of these studies was the lack of *C. eugenioides* data, which was necessary to increase and improve the comparison of gene expression of these subgenomes.

Despite the strategic importance of understanding gene expression of the *C. arabica* ancestors, no studies have focused on *C. eugenioides* molecular genetics. Here, we present an overview of the *C. eugenioides* transcriptome as a potential model for future studies in *Coffea* that could explain the mechanisms involved in the expression of homeologous genes in *C. arabica*, thereby providing a comparative framework for *Coffea* gene expression in several species. We also employed quantitative real-time PCR (qPCR) to validate the results obtained from high throughput sequencing.

## Materials and methods

### Plant materials

Young leaves and mature fruits in the 'cherry' stage were harvested from four *C. eugenioides* plants from the same six- to eight-year-old parents maintained at COCARI, Mandaguari, PR, Brazil [latitude (S): 23°30′52″; longitude (W): 51°42′86″]. The region has a height of 650 m and a mean annual temperature of 22–23 °C. All samples were collected on July 6, 2011, between 9 and 11 am. After harvesting, the samples were immediately frozen in liquid nitrogen and stored at −80 °C for RNA extraction. Those plants are F3 generation from original accesses of *C. eugenioides* in the Institute Agronomic of Campinas, Campinas, SP, Brazil. SSRs analysis of F2 plants revealed that they belong to a Kenyan group (Philippe Cubry, personal communication).

### RNA extraction

Leaves and fruits of *C. eugenioides* were ground to a fine powder with liquid nitrogen using a cooled mortar and pestle. Total RNA was isolated (Chang et al. 1993) and the integrity of the RNA samples was examined by 1 % agarose gel electrophoresis following treatment of the samples with DNase (RNase-free). The quality and the concentration of extracted RNA samples were determined using

a NanoDrop® ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, USA). The absence of genomic DNA contamination in the RNA samples was confirmed by PCR using *GAPDH* primers.

### RNA sequencing

mRNA sequencing was performed at the High Throughput Sequencing Facility at the Carolina Center for Genome Sciences (University of North Carolina, Chapel Hill, NC, USA). For each organ, 10 μg of total RNA from a pool of four individuals was used to prepare the mRNAseq library according to the protocol provided by Illumina. Library quality control and quantification were performed using a Bioanalyzer Chip DNA 1000 series II (Agilent Technologies, Santa Clara, CA, USA). Libraries were tagged and multiplexed with the Illumina HiSeq™ 2000. The two libraries were sequenced in multiplex with 10 other libraries in one lane of the flow cell to generate 100 base-pair (bp) single-end sequences. The nucleotide sequence data are available at the sequence read archive (SRA) in the National Center for Biotechnology Information (NCBI) under accession number SRP052722.

### RNA-seq data processing

To obtain high-quality clean read data for de novo assembly, the raw reads from mRNA-seq were filtered by discarding the reads with adaptor contamination and regions of low quality (quality <20). The processed reads of both organs were merged and assembled with Trinity assembler, 6-8-2012 version (Grabherr et al. 2011) using an optimized k-mer length of 25 for de novo assembly.

### Annotation and classification of contigs

For annotation, only the more representative isoform of each gene with a minimum length of 200 bp was considered. All contigs were compared using BLASTX against the NCBI non-redundant sequence database (nr) with an e-value cut-off of 1e-5. We also performed a BLASTN search against the *C. arabica* EST database (Mondego et al. 2011) and *C. canephora* coding sequences—CDS (Denoeud et al. 2014) with *e* value cut-offs of 1e−5 and >90 % identity. A BLASTX search with an e value cutoff of 1e−5 was performed against Swiss-Prot reference proteins (The UniProt Consortium 2014). Functional annotation of biological processes, molecular functions and cellular components was performed using BLAST2GO version v.2.7.0 (Conesa et al. 2005) considering GO Slim annotations provided by TAIR (2014). The InterProScan (Quevillon et al. 2005) and KEGG (Kanehisa and Goto 2000) databases were also used to identify protein domains and metabolic pathways, respectively.

### Differential expression analysis

Bowtie (Langmead et al. 2009) was used with default parameters to map the processed reads against the de novo assembled transcriptome allowing a maximum of three mismatches. Reads of each library were aligned separately, and only the reads that anchored in the reference were used. Differential expression analysis was performed from the raw read counts using the DESeq package (Anders and Huber 2010). Genes identified as differentially expressed were required to have a twofold change and $P \leq 0.05$. Additionally, BLAST2GO was used for GO functional enrichment analysis of exclusive contigs of leaves and fruits by performing Fisher's exact test.

### Phylogenetic analysis

Phylogenetic trees were built for the detailed annotation of differentially expressed BURP unigenes from *C. eugenioides* using protein sequences of *Populus trichocarpa* (Torr. & Gray) and *Arabidopsis thaliana* (L.) Heynh. We retrieved homologs of *P. trichocarpa* and *A. thaliana* BURP sequences using PLAZA v.2.5 (Van Bel et al. 2012). For analysis, we selected only sequences from the subfamilies PGI β-like, BNM2-like and class BURP V (Gan et al. 2011) and sequences that contained the BURP domain (PF03181) according to the Pfam database. Sequences were aligned with the MUSCLE tool of the MEGA package, and the resulting alignments were used to construct a neighbor-joining tree using the MEGA 6.0 software (Tamura et al. 2013). The confidence level was estimated using a bootstrap analysis of 10,000 replicates.

### qPCR validation and data analysis

Based on DESeq results for differential expression, 10 unigenes (Table 2) were selected for expression analysis by qPCR to validate the RNA-seq analysis. These genes were chosen based on a normalized log2 fold-change between the leaves and fruits. Complementary DNAs (cDNAs) of *C. eugenioides* leaves and fruits from the same pool used for sequencing were synthesized using SuperScript III Reverse Transcriptase (Invitrogen) following the manufacturer's instructions in a final volume of 20 μl including 5 μg of total RNA. Primers were designed using Primer Express v.3.0 (Applied Biosystems), and qPCRs were run in a 7500 Fast Real-Time PCR System (Applied Biosystems) using the SYBR Green PCR Master Mix (Applied Biosystems). We followed the basic qPCR procedure described in a previous publication on coffee plants (Carvalho et al. 2014).

The reaction mixture contained 12.5 µl of SYBR Green PCR Master Mix, 0.5 µl of each primer (5 µM), 1 µl of cDNA and 10.5 µl of Milli-Q water. The qPCR conditions were 95 °C for 5 min, followed by 40 cycles of 94 °C for 30 s, 62 °C for 60 s, 72 °C for 30 s and one last step of 72 °C for 10 min. Melting curves were analyzed to verify the presence of a single product including a negative control. All reactions were performed with three technical replicates and followed the information for publication of qPCR experiments (MIQE) (Bustin et al. 2009).

Data were analyzed to determine cycle threshold (Ct) values. The specificity of the PCR products generated for each set of primers was verified by analyzing the Tm (dissociation) of the amplified products. PCR efficiency ($E$) was determined using LinReg (Ramakers et al. 2003) using only qPCR reactions with an efficiency >94 %. Expression levels were calculated by applying the formula $(1+E)^{-\Delta\Delta Ct}$, where $\Delta Ct_{target} = Ct_{target\ gene} - Ct_{CaGAPDH}$ and $\Delta\Delta Ct = \Delta Ct_{\ target} - \Delta Ct_{reference\ sample}$ (Cotta et al. 2014). The gene with the lowest expression between the organs was used as reference sample to calculate the relative expression. Gene expression levels were normalized using the glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) gene as a reference; this gene has been largely used to calibrate expression in coffee qPCR experiments (Barsalobres-Cavallari et al. 2009; Cruz et al. 2009; Carvalho et al. 2013).

## Results

### Sequencing the *C. eugenioides* transcriptome

We produced a total of 8,435,413 raw reads (3,688,364 from leaves and 4,747,049 from fruits). Due to the absence of reference genomic sequences, a de novo RNA-seq assembly was performed using Trinity that resulted in 36,935 contigs with lengths >200 bp. The mean contig length was 701 bp, and more than 8000 contigs had lengths longer than 1000 bp. The contig distribution according to size is shown in Fig. S1 in the Online Resource.

### Functional characterization

To estimate the accuracy of the performed de novo assembly, the 36,935 contigs were annotated using eight databases. A summary is shown in Table 1. *C. eugenioides* contigs were analyzed based on similarity with BLASTX against available non-redundant sequences from the NCBI database (NCBI-nr). A total of 23,297 contigs (63.1 %) had at least one hit below 1e-5 (Table 1 and Table S1 in Online Resource). Among the NCBI-nr BLASTX top hits, 10,706 contigs had their first hit in *Vitis vinifera* L. proteins (29.4 %), followed by *Ricinus communis* L. (9.7 %), *P. trichocarpa* (8.7 %), *Nicotiana tabacum* L. (1.2 %) and *Glycine max* (L.) Merr. (1 %). The distribution of the 100 top-hit species is presented in Table S2 in the Online Resource.

To compare *C. eugenioides* contigs with public data for *Coffea* sp., a BLASTN search was performed between our dataset and two reference databases: an assembly of *C. arabica* ESTs with 35,113 contigs (Mondego et al. 2011) and the predicted coding sequences of the recently released *C. canephora* genome (Denoeud et al. 2014). Nearly half of the *C. eugenioides* contigs (18,567 contigs—50.2 %) had one homolog in *C. arabica* with >90 % identity and almost three-quarters of the *C. eugenioides* contigs (24,047 contigs—65.1 %) had a homolog in *C. canephora*. "No-hit" genes were annotated according to the NCBI-nr database and were listed in Table S3 in the Online Resource. To identify *C. eugenioides* contigs that potentially encoded proteins with known functions, a BLASTX analysis with a cut-off e-value of 1e-5 was performed using the Swiss-Prot protein databases. Based on this analysis, 16,902 contigs (45.8 %) were annotated (Table 1; Table S1 in the Online Resource).

Functional characterization of the *C. eugenioides* contigs was performed by assigning gene ontology (GO) annotation with BLAST2GO. To provide a general representation in plants, a classification of GO Slim was obtained. A total of 18,058 contigs (48.9 %) could be assigned to one or more ontologies (Table 1). The number of GO terms per contig ranged from 1 to 49. In total, 87,640 GO terms were retrieved, including 41.8 % in the biological process, 32.4 % in the molecular function and 25.8 % in the cellular component categories. A summary of contigs annotated in each GO Slim term in the main categories (biological process and molecular function) is shown in Fig. 1a.

**Table 1** Annotation summary of *C. eugenioides* contigs in eight databases

| Database | BLAST | Annotations[a] | %[b] |
|---|---|---|---|
| NCBI-nr | BLASTX | 23,297 | 63.1 |
| *Coffea* EST database | BLASTN[c] | 18,567 | 50.2 |
| *C. canephora* CDS | BLASTN[c] | 24,047 | 65.10 |
| Gene ontology (GO Slim) | BLASTX | 18,058 | 48.9 |
| Swiss-Prot | BLASTX | 16,902 | 45.8 |
| InterProScan | BLASTX | 12,834 | 34.7 |
| PlantCyc | BLASTX | 7669 | 20.8 |
| KEGG | BLASTX | 802 | 2.2 |

[a] Number of contigs annotated

[b] Percentage of annotated contigs from *C. eugenioides* considering a total of 36,935 contigs

[c] Identity >90 %

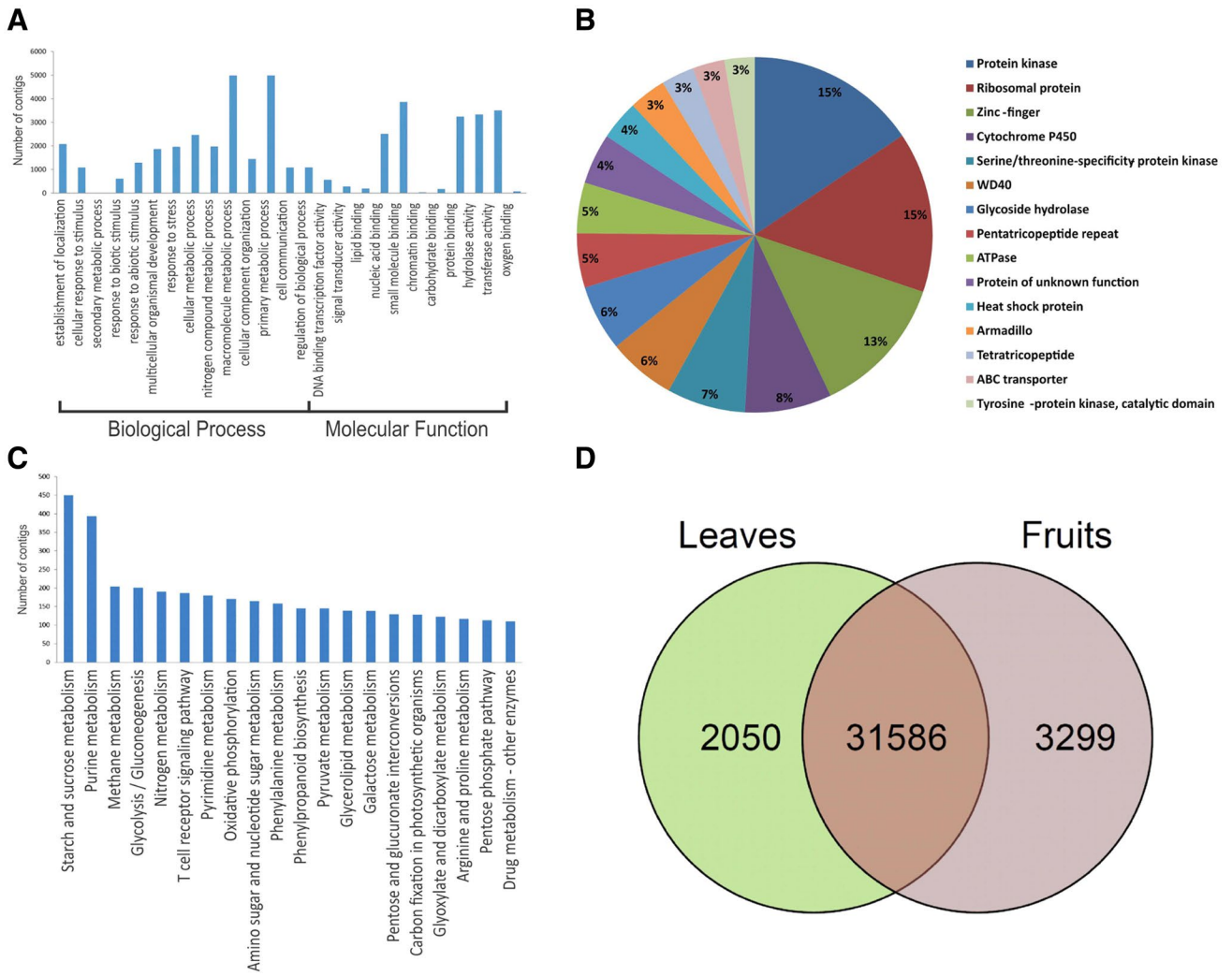


**Fig. 1** Contig numbers in each functional category based on gene ontology classification and the 15 most abundant InterProScan categories in *C. eugenioides*. **a** Contigs were classified into different functional groups based on a set of GO slims in the biological process and molecular function categories. **b** The most frequent domains of proteins present in the *C. eugenioides* contigs according to InterProScan. **c** Top 20 biological pathways obtained in the *C. eugenioides* transcriptome. A total of 802 predicted pathways were characterized based on KEGG database analysis using the BLAST2GO tool. **d** Shared and unique contigs of *C. eugenioides* based on Bowtie mapping onto the reference transcriptome. A total of 31,586 sequences are expressed in leaves and fruits, with 2050 contigs exclusively expressed in leaves and 3299 in fruits

We used GO annotations to assign each contig to a set of GO Slims in the biological process and molecular functions categories. Primary metabolic process (GO:0044238), macromolecule metabolic process (GO:0043170), cellular metabolic process (GO:0044237), nitrogen compound metabolic process (GO:0006807) and response to stress (GO:0006950) were among the most highly represented groups under the biological process category. Under the molecular function category, assignments were mainly to small molecule binding (GO:0036094), transferase activity (GO:0016740), hydrolase activity (GO:0016787), protein binding (GO:0005515) and nucleic acid binding (GO:0003676).

Conserved domains in *C. eugenioides* contigs were identified using the InterProScan database. A total of 12,834 contigs were annotated within 4961 different domains/families (Fig. 1b). InterProScan categories were ranked according to the number of contigs; the 15 most abundant domains/families are represented in Fig. 1b. Details of the InterProScan characterization are shown in Table S4 in the Online Resource. The 15 most frequent domains were protein kinase with 3762 contigs (15 %), ribosomal protein with 3532 contigs (15 %), zinc-finger with 3090 contigs (13 %), cytochrome P450 with 1907 contigs (8 %), and serine/threonine dual-specificity protein kinase and catalytic domain with 1727 contigs (7 %).

The distribution of *C. eugenioides* into various metabolic pathways was verified using the PlantCyc and KEGG databases. BLASTX analyses against the PlantCyc database resulted in the annotation of 7669 contigs (20.8 %) (Table S1). From KEGG, 802 contigs (2.2 %) were assigned to 142 pathways and 374 enzymes (Fig. 1c). Starch and sucrose metabolism was the most abundant category with 450 members/contigs (Fig. S2 in Online Resource), followed by purine metabolism (393 members), methane metabolism (204 members) and glycolysis/gluconeogenesis (201 members) (Fig. 1c).

### Differential gene expression

Our work showed that 2050 contigs were considered specific to leaves, 3299 contigs were exclusively present in fruits, and 31,586 contigs were expressed in both organs (Fig. 1d). A BLASTX hit in the NCBI-nr and Swiss-Prot databases was observed for 20 genes that were specific for each organ (Table S5 in Online Resource). BLAST2GO was used for GO functional enrichment analysis of genes exclusively expressed in one organ (leaves or fruits—Fig. 2). In leaves, the biological process category showed mainly terms related with phosphorylation (GO:0016310), protein phosphorylation (GO:0006468), carbohydrate catabolic process (GO:0016052), hexose metabolic process (GO:0019318) and single-organism carbohydrate catabolic process (GO:0044724) (Fig. 2a). In the molecular function category, the most common terms were catalytic activity (GO:0003824), small molecule binding (GO:0036094), nucleotide binding (GO:0000166), nucleoside phosphate binding (GO:1901265) and transferase activity (GO:0016740) (Fig. 2a).

In fruits, the biological process category was mainly related to metabolic process (GO:0008152), organic substance metabolic process (GO:0071704), primary metabolic process (GO:0044238), single-organism metabolic process (GO:0044710) and biosynthetic process (GO:0009058) (Fig. 2b). The most represented functions related to the molecular function category were oxidoreductase activity (GO:0016491), structural molecule activity (GO:0005198), structural constituent of ribosome (GO:0003735), transporter activity (GO:0005215) and transmembrane transporter activity (GO:0022857) (Fig. 2b).

### Differential expression validation by qPCR

We used qPCR to validate the transcriptional pattern of ten selected unigenes with high expression levels in leaves and fruits. All of the genes displayed a transcriptional pattern that was in agreement with the DESeq analysis. The *Ce14433*, *Ce15205*, *Ce2770*, *Ce14847* and *Ce10671* unigenes were mostly expressed in leaves, and the *Ce13100*, *Ce9246*, and *Ce13525* unigenes exhibited preferential expression in fruits (Fig. 3; Fig. S3 in the Online Resource). Two unigenes (*Ce14834* and *Ce13451*) had amplification signals only in fruits, as demonstrated by their Ct (cycle threshold) values (Fig. S3 in the Online Resource) of 24.77 (*Ce14834*) and 26.21 (*Ce13451*).

Annotation of these unigenes (Table 2) using BLASTX against the TAIR database and the *C. canephora* coding sequence (Denoeud et al. 2014) allowed the identification of several highly expressed genes in leaves: a UDP-glycosyltransferase superfamily protein (*Ce14433*); a germin-like and/or auxin-binding protein ABP20 (*Ce15205*); a glucose-6-phosphate/phosphate translocator 2 (*Ce2770*); a chitinase (*Ce14847*); and a BURP domain-containing protein (*Ce10671*). In fruits, we also identified a differentially expressed BURP domain-containing protein (*Ce14834*), a serine carboxypeptidase-like 29 (*Ce13100*), a cytochrome P450 and a phenylalanine *N*-monooxygenase (*Ce13451*), a transcription factor-related protein (*Ce9246*), and an oxidoreductase (*Ce13525*).

## Discussion

### Assembly and annotation of the *C. eugenioides* transcriptome

This report represents the first large scale study of the *C. eugenioides* transcriptome using NGS technology. We report a comprehensive annotation of its transcriptome in several databases to determine genes that are differentially expressed in leaves and fruits. Contig lengths ranged from 200 bp to 10 kb, with a mean length of approximately 700 bp; this result is similar to recent analyses in *Panicum maximum* Jacq. (758 bp) (Toledo-Silva et al. 2013) and *Cocos nucifera* L. (752 bp) (Fan et al. 2013) using Illumina sequencing and is more reliable than those for *Camellia sinensis* (L.) Kuntze (355 bp) (Shi et al. 2011) and *Cymbidium sinense* (G. Jackson ex H. C. Andrews) Willdenow (612 bp) (Zhang et al. 2013).

We found that 63.1 % of the *C. eugenioides* contigs were similar to proteins annotated in the NCBI-nr database. The species with the highest similarity was *V. vinifera* with 10,706 hits. Similar results were also reported when analyzing *Coffea* sp. EST contigs (Mondego et al. 2011). It was expected that *Coffea* sequences would share more similarity with plants from the Asteridae clade (e.g., Solanaceae species) and with *V. vinifera* in terms of high functional synteny because they were possibly derived from the same ancestral genome (Mondego et al. 2011; Guyot et al. 2012). A higher level of conservation observed between coffee and grapevines could be due to the function of the grapevine (a rosid), which is a conservative species in terms

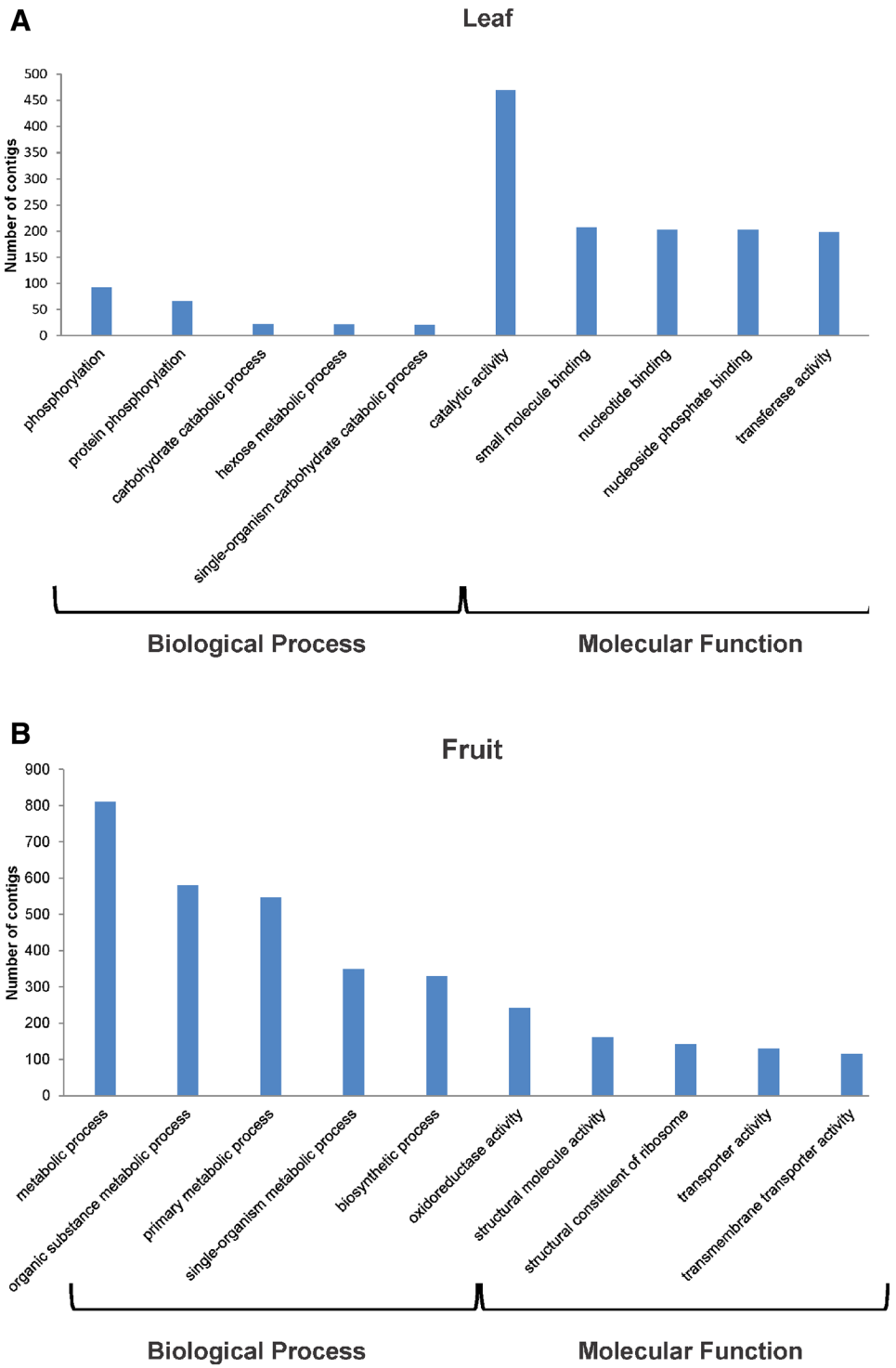


**Fig. 2** GO term distribution of differentially expressed contigs in *C. eugenioides*. **a** Top-hit GO distribution contigs exclusively expressed in leaves. **b** Top-hit GO distribution contigs exclusively expressed in fruits

of the integrity of its general chromosomal structure; coffee exhibits less gene-order divergence compared to other rosids (Jaillon et al. 2007; Denoeud et al. 2014). Interestingly, we observed that 9.7 % of contigs had *R. communis* as a top hit.

To investigate the contributions of the present catalog to the discovery of new genes, *C. eugenioides* contigs were analyzed for sequence similarity against *C. arabica* and *C. canephora* putative unigenes; matches were found for between 50 and 65 %, respectively. This result was expected because there is no public reference genome for *C. arabica* and a large amount of data is available for *C. canephora* (Denoeud et al. 2014). A large number of no hit genes were observed even after comparison with the

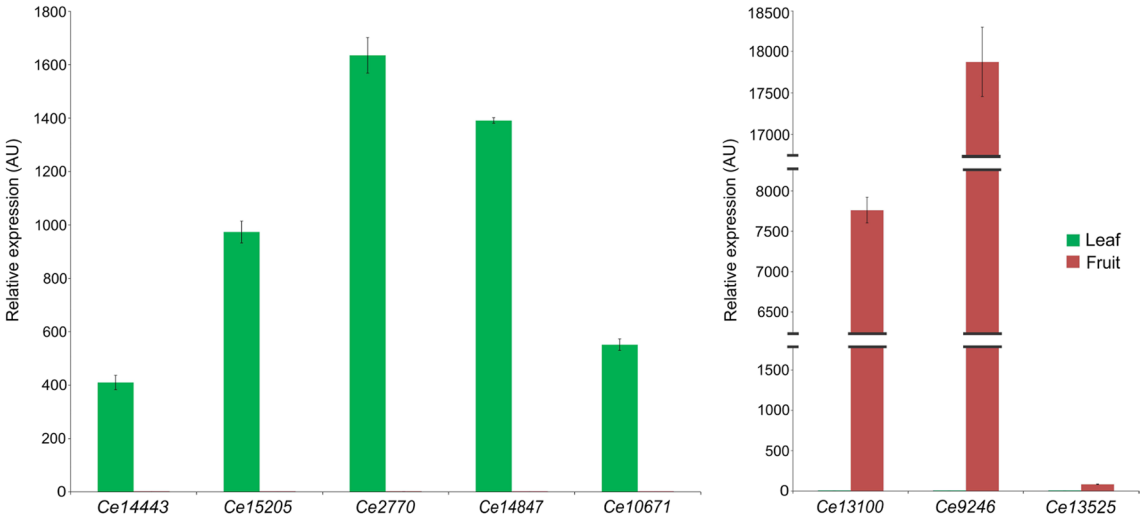


**Fig. 3** Relative expression values of unigenes up-regulated in leaves (**a**) and fruits (**b**) by qPCR. Relative quantification of each transcript was normalized against *GAPDH*. The gene with the lowest expression was used to calibrate the relative value between the organs. The *Ce14834* and *Ce13451* unigenes had amplification signals only in fruits and therefore were not represented in the figure. *Bars* represent the standard deviation values

two *Coffea* species. These results reinforce the suggestion that our study probably contains new potential genes in *Coffea* (Table S3 in the Online Resource).

Several databases of functional annotation were used to predict potential genes and their biological functions. We restricted GO information only to the TAIR database, which represents robust databases for plant studies. Nearly half of the *C. eugenioides* contigs found annotations onto GO structures. These numbers may indicate genes not described in the database similar to those observed in the NCBI-nr and Swiss-Prot annotations.

GO annotations in the *C. eugenioides* transcriptome found terms associated with macromolecule metabolic process and primary metabolic process in the biological process category. These terms could reflect expression associated with basal processes, thereby reinforcing information observed in *C. arabica* (Vidal et al. 2010) that identified the contributions of subgenome CaCe in *C. arabica* proteins associated with the citric acid cycle, pentose-phosphate shunt and photosynthesis.

Small molecule binding, transferase activity, hydrolase activity, protein binding and nucleic acid binding were annotated in the molecular function category. A previous study in *C. arabica* and *C. canephora* reported similar categories (Mondego et al. 2011), demonstrating the high similarity between *C. eugenioides* and the other *Coffea* species gene catalog. However, top domains such as small molecule binding and transferase activity may indicate processes associated with sugar synthesis and transport; these terms were found with higher prevalence in *C. arabica* than *C. canephora* (Mondego et al. 2011). Furthermore, these terms could indicate proteins related to sugar metabolism (especially sucrose), which is in agreement with the annotation of the starch and sucrose metabolism pathway by KEGG analysis.

Sucrose has an important role in determining coffee cup quality. It is one of the major resources of the free reducing sugars that participate in the Maillard reaction, which occurs during the roast of the coffee grain and generates a significant number of properties, such as caramel, sweetness and dark colors (Holscher and Steinhart 1995). *C. canephora* accumulates less sucrose than *C. arabica* because *C. canephora* presents proteins activities that participate mainly in two stages (early in grain development to prevent the accumulation of sucrose and in the final stage of grain) and have less capacity for sucrose re-synthesis (Geromel et al. 2006; Privat et al. 2008). However, *C. arabica* produces and accumulates more sucrose during grain development compared with *C. canephora* (Joët et al. 2009) and CaCe could contribute this trait in the allopolyploid *C. arabica* gene expression.

InterProScan analysis of *C. eugenioides* RNA-seq showed results similar to those previously observed in ESTs from *C. arabica* and C. *canephora* (Lin et al. 2005; Mondego et al. 2011). They showed several families in common, such as protein kinases, cytochrome P450, serine threonine kinases and pentatricopeptide repeats, which were the most predicted protein families annotated for *Coffea*.

## Characterization of differentially expressed genes

BLAST was performed for *C. eugenioides* unigenes against the TAIR database and *C. canephora* CDS, which showed

**Table 2** First hit annotation in TAIR database and *C. canephora* coding sequences of *C. eugenioides* unigenes and primers designed for candidate genes validation by qPCR analysis

| Unigenes | BLAST TAIR | Annotation TAIR | BLAST *C. canephora* genome | Annotation *C. canephora* CDS | Primer sequence |
|---|---|---|---|---|---|
| *Ce14433* | AT1G22400.1 | UDP-Glycosyltransferase superfamily protein | Cc10_g06970 | UDP-Glycosyltransferase 85A1 | F: 5′ GCCAAGCTCCTC-CACCAAA 3′<br>R: 5′ GCATCAGGAC-CGCTGGAT 3′ |
| *Ce15205* | AT5G20630.1 | Germin 3 | Cc06_g12080 | Auxin-binding protein ABP20 | F: 5′ CTCCAGGGTGCGT-GTGAAA 3′<br>R: 5′ CGTTCCCTGGTGT-GAATGG 3′ |
| *Ce2770* | AT1G61800.1 | Glucose-6-phosphate/phos-phate translocator 2 | Cc05_g04890 | Glucose-6-phosphate/ phosphate translocator 2, chloroplastic | F: 5′ GCATTGAGGAC-CTTCTTGTTGTAG 3′<br>R: 5′ TGCAGCGCAGAA-GCTTAAGAT 3′ |
| *Ce14847* | AT5G24090.1 | Chitinase A | Cc05_g00780 | Acidic endochitinase | F: 5′ GGCCAAACACCG-GAACTG 3′<br>R: 5′ CAGGCTCTG-GCAAACCTCTATC 3′ |
| *Ce10671* | AT1G23760.1 | BURP domain-containing protein | Cc03_g13730 | Putative BURP | F: 5′ ACGCGTCCAACCAT-CAATT 3′<br>R: 5′ TTCAAAAACTGCCA-TAGGTGACA 3′ |
| *Ce14834* | AT1G23760.1 | BURP domain-containing protein | Cc03_g13730 | Putative BURP | F: 5′ CCCACTAAAACTCTC-CGCTAAAAT 3′<br>R: 5′ TTTTCTCAACATCGC-CTTTTGA 3′ |
| *Ce13100* | AT4G30810.1 | Serine carboxypeptidase-like 29 | Cc11_g02270 | Serine carboxypeptidase-like 29 | F: 5′ GAGGGCTTGTTTAG-GCTTGTGT 3′<br>R: 5′ GAGGATG-GACTCAGCAGTATGAAG 3′ |
| *Ce13451* | AT4G39950.1 | Cytochrome P450, family 79, subfamily B, polypeptide 2 | Cc02_g05380 | Putative Phenylalanine *N*-monooxygenase | F: 5′ TGTGCCGGAAAAT-GAAGGA 3′<br>R: 5′ ACGTGGGCTGGCAT-GTG 3′ |
| *Ce9246* | AT4G18650.1 | Transcription factor-related | Cc02_g15140 | Transcription factor-related | F: 5′ GAAAGGAGTGTTG-GATGTGTTGAA 3′<br>R: 5′ CTTTTCTTCCCC-CATTTTCTCA 3′ |
| *Ce13525* | AT4G21580.1 | Oxidoreductase, zinc-binding dehydrogenase family protein | Cc09_g10020 | Putative quinone oxidoreduc-tase PIG3 | F: 5′ TGGAGTGAACTTTTT-GAACCAGAAT 3′<br>R: 5′ TTTACGAATCCC-CATGGATCTT 3′ |
| *GAPDH* | AT1G13440 | Glyceraldehyde-3-phosphate dehydrogenase | | | F: 5′ AGGCTGTTGGGAAA-GTTCTTC 3′ |

similar annotations or correlated functions. Because the ultimate goal of this study was to identify potential candidates for studies in coffee (mainly in fruit), some genes were characterized from the comparisons with databases and selected for validation using qPCR. In this study, *Ce14443* presented high expression in leaves and was annotated as a UDP-glycosyltransferases (UGTs) superfamily protein. Glycosyltransferases (GTs) are a ubiquitous group of enzymes that catalyze the transfer of a sugar moiety from an activated donor molecule onto saccharide or non-saccharide acceptors, including molecules involved in secondary metabolism (Caputi et al. 2012). GTs utilize UDP-activated sugars as the major donor molecule and have a conserved UDP-glycosyltransferase (UGT)-defining motif. Several studies showed that UGTs were associated with the glycosylation of terpenoids (Rivas et al. 2013), benzoates (Osmani et al. 2009), flavonoids (Veljanovski and Constabel 2013), saponins (Shibuya et al. 2010) and

plant hormones (Poppenberger et al. 2005). UGTs also play a pivotal role in the detoxification and deactivation of xenobiotics (Brazier-Hicks et al. 2007) and in plant–pathogen interactions (Langenbach et al. 2013). The large number of UGTs has been identified in plants, and the detection of their expression in different tissues (Sharma et al. 2014) indicates important roles in growth and development and could assist in the selection of candidate genes for various genetic applications.

Another gene with increased expression in the leaves was *Ce15205*, which exhibited similarity to the germin 3 protein according to TAIR and was related to an auxin-binding protein ABP20 in the *C. canephora* CDS. Germin-like proteins (GLPs) have been identified in various plants and constitute a large and highly diverse family of ubiquitous plant proteins (El-Sharkawy et al. 2010). Their function is unclear and variable, but includes restructuring of the cell wall, salt, cold and heavy metal response and plant defense (Hurkman and Tanaka 1996; Bae et al. 2003). Members of GLP subfamily 3, which includes germin 3, are characterized as not having oxalate oxidase activity (Ohmiya 2002). They are specifically expressed in leaves, and their mRNA levels undergo circadian oscillations (Staiger et al. 1999). ABP19/20 are two sequence homologues that were isolated from the shoot apices of a peach. They were identified as proteins that specifically bound auxins and were highly homologous to the GLPs in subfamily 3 (Ohmiya et al. 1998).

The *Ce2770* unigene demonstrated similarity with glucose-6-P/phosphate translocator 2 (GPT2). This gene product is involved in the transport of glucose 6-phosphate across plastid membranes in return for inorganic phosphate (Niewiadomski et al. 2005). GPT2 was associated with impaired carbon metabolism or the presence of increased soluble sugar concentrations (Kunz et al. 2010), as well as senescence (Pourtau et al. 2006) and an increase in carbon fixation due to increased light (Athanasiou et al. 2010). In the last study, the expression of GPT2 was demonstrated to be required for photosynthesis acclimation according to light intensity. This finding implies that its plays a role in sugar perception and affects the balance of metabolites in cellular compartments.

*Ce14847* was annotated as a chitinase gene that catalyzed the hydrolysis of *N*-acetylglucosamine (GlcNAc) 1,4-linkages in chitin (Passarinho and Vries 2002). All classes of chitinase possess some conserved amino acid residues in their catalytic domains, but the different enzyme activity for each class indicates that some unidentified residues may also contribute to substrate specificity (Sasaki et al. 2006). In terms of function, chitinases are classic pathogenesis-related proteins that are involved in non-host-specific defense (Stintzi et al. 1993). Plant chitinase genes can be induced by plant pathogens and other biotic and abiotic stress responses. They may also be associated with normal plant growth and development, such as cellulose biosynthesis and root expansion (Hermans et al. 2011; Wu et al. 2012; Chen et al. 2014).

*Arabidopsis thaliana* class III chitinases (*AtChiA* gene) are also up-regulated under environmental stresses, especially to salt and wounds (Takenaka et al. 2009). Allosamidin, known as an effective inhibitor of chitinases, appeared to enhance *AtChiA* expression to stress tolerance (mainly heat and strong light stresses), probably through crosstalk between the two pathways for biotic and abiotic stress responses (Takenaka et al. 2009). Chitinase activity was detected in *C. arabica* leaves and indicated the participation of these proteins in plant defense against coffee leaf rust disease (Guerra-Guimarães et al. 2009). Higher expression in *C. eugenioides* may also be a mechanism to improve biotic tolerance against pathogens.

*Ce10671* had a higher expression in leaves, and *Ce14834* was exclusively expressed in fruits (Fig. 3; Fig. S3 in Online Resource) according to our qPCR analysis. Both unigenes showed high similarity with *Arabidopsis* gene AT1G23760 and were characterized in agreement with the *C. canephora* CDS as a BURP domain-containing protein. BURP domain proteins are broadly distributed in plants, possess plant-specific functions, are involved in a variety of functions, and have been found in different plant organs (Van Son et al. 2009). BURP proteins are characterized by an N-terminal hydrophobic domain with a signal peptide, followed by a highly conserved region, a variable internal region that is unique to each member and consists of repeated units, and a cysteine-histidine pattern known as a C-terminal BURP domain (Van Son et al. 2009).

BURP-domain proteins were initially classified into four subfamilies: BNM2-like, USP-like, RD22-like and PG1b-like (Granger et al. 2002). Advances in genomic data led to the discovery of new members of the BURP-domain family (Gan et al. 2011). A phylogenetic tree was developed to verify whether the two unigenes belonged to the same BURP subfamily (Fig. S4 in Online Resource). The identification of members from the BURP gene family with organ-specific patterns in *C. eugenioides* warrants further analyses of this gene family in coffee, as observed in other species (Granger et al. 2002; Gan et al. 2011).

The *Ce10671* and *Ce14834* unigenes were included in the PGI β-like subfamily in our phylogenetic analysis. *C. eugenioides* BURPs are distinguishable from the other BURPs by a 14 repeat amino acid sequence and the presence of a β subunit of polygalacturonase. A cotton protein with BURP domains demonstrated a relationship with plant growth, fiber length and seed mass (Xu et al. 2013). In this sense, *Ce14834* could be a possible target gene for future studies to increase yield because its expression was found only in fruit.

With increased expression in fruits, the *Ce13100* unigene was annotated as a serine carboxypeptidase (SCP) that is a member of the α/β hydrolase family of proteins. The enzymes in this family have a catalytic triad composed of a nucleophilic residue, an acidic (Glu or Asp) residue and a His. These proteins have diverse catalytic functions (i.e., hydrolases) and non-catalytic functions, but the majority of members have not yet been characterized experimentally (Lenfant et al. 2013). Several serine carboxypeptidase-like (SCPL) proteins has been related with peptidases, but they also have shown other functions, such as acyltransferases and lyases (Wajant et al. 1994; Shirley et al. 2001). Our qPCR results suggest that some SCPL genes are expressed in an organ-specific pattern, whereas others are transcribed in a wide range of tissue types. Taken together, these data suggested that the SCPL gene family encodes a diverse group of enzymes whose functions are likely to extend to protein degradation and processing and include activities such as the production of secondary metabolites (Fraser et al. 2005).

*Ce13451* was described as a cytochrome P450, family 79 and a phenylalanine *N*-monooxygenase and was expressed exclusively in fruits. CYP79A2, a cytochrome P450-dependent monooxygenase, was reported to catalyze the conversion of L-phenylalanine into phenylacetaldoxime during the biosynthesis of benzylglucosinolate in *A. thaliana*, demonstrating that CYP79 homologues are involved in the biosynthesis of glucosinolates (Wittstock and Halkier 2000). Glucosinolates are related to cyanogenic glucosides, which are widely distributed in the plant kingdom, and the biosynthesis of both classes of secondary plant products (Bak et al. 1998; Wittstock and Halkier 2000). Several CYP79 homologues were previously identified in glucosinolate-producing plants (Bak et al. 1998) with functions that have not been determined, but some species seen to function in defense towards herbivores and pathogens (Textor and Gershenzon 2009).

The transcription factor *Ce9246* was highly expressed in fruits. This transcription factor was related to locus AT4G18650.1 in *Arabidopsis* that encoded a *DOG1*-like 4 (*DOGL4*) transcription factor. *DOGL4* is a member of a small gene family that includes *DOG1*, which functions in the control of seed dormancy (Heisel et al. 2013). However, the function of *DOGL4* is currently unknown (Bentsink et al. 2006).

*Ce13525* was described with high expression in fruits and was annotated as an oxidoreductase in the zinc-binding dehydrogenase family. It was demonstrated in *Arabidopsis* that the phytohormone abscisic acid (ABA) caused an increase in the expression of this gene (Raghavendra et al. 2010). ABA is a regulator of plant development and stress responses, including seed dormancy, germination, stomatal aperture regulation and drought resistance responses (Raghavendra et al. 2010). Binding of ABA to the receptor leads to reaction cycle activation, ion transport regulation and up and down modulation of gene expression in *Arabidopsis* (Böhmer and Schroeder 2011).

This is the first gene catalog for *C. eugenioides* using RNA-seq and the first large report on the analysis of leaf and fruit genes for this species. We identified genes with expression patterns directly related to basal processes and sugar metabolism, which corroborates with previous information concerning the differential homeologous expression of the CaCe subgenome in *C. arabica*. Furthermore, our study identified and analyzed the expression of genes with variable functions, such as participation in secondary metabolic pathways, sugar mobilization pathways, defense against pathogens and seed development. These genes could represent potential candidates for various biotechnological applications in coffee because they showed considerable expression, especially in fruits. Thus, they could be important candidates for an improved functional characterization and help to develop plants with certain adaptive characteristics to improve coffee quality and production.

In allopolyploids, the duplication of genes displays a homoelog expression bias. This bias has been studied in several allopolyploid plant species, such as *Gossypium hirsutum* L. and *Triticum aestivum* L. (Mochida et al. 2003; Flagel et al. 2009). In coffee, the expression variation of homeologous genes was demonstrated to depend on the gene, the organ and the growth condition (Combes et al. 2012). Complex genetic and epigenetic regulatory mechanisms determine genome-specific expression biases in allopolyploid species and may be sensitive to environmental conditions (Dong and Adams 2011). In this sense, this suite of experiments and results presents new genes that could aid in the understanding of the contribution of *C. eugenioides* to the development, adaptation and evolution of *C. arabica* and provide the basis for further gene expression studies in the *Coffea* genus.

**Acknowledgments** We are grateful to thank João Batista Gonçalves Dias da Silva (COCARI) for providing the *C. eugenioides* leaves and fruits used in this study and Juliana Costa Silva (UTFPR-Cornélio Procópio) for bioinformatics assistance. This work was funded by CAPES/Agropolis (1002-02 PHEGECO), CNPq, INCT-Café, FINEP (01.05.0665-00) and Fundação Araucária. We acknowledge the scholarships from CAPES (Priscila M. Yuyama, Suzana T. Ivamoto). Luiz Filipe P. Pereira received a research fellowship from CNPq. We also acknowledge the Center for Computational Engineering and Sciences at UNICAMP, SP, Brazil (FAPESP/CEPID project #2013/08293-7).

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** The experiments in this manuscript comply with the current laws of the country in which they were performed. This article does not contain any studies with human participants or animals performed by any of the authors.

## References


Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106

Athanasiou K, Dyson BC, Webster RE, Johnson GN (2010) Dynamic acclimation of photosynthesis increases plant fitness in changing environments. Plant Physiol 152:366–373

Bae MS, Cho EJ, Choi EY, Park OK (2003) Analysis of the *Arabidopsis* nuclear proteome and its response to cold stress. Plant J 36:652–663

Bak S, Nielsen HL, Halkier BA (1998) The presence of CYP79 homologues in glucosinolate-producing plants shows evolutionary conservation of the enzymes in the conversion of amino acid to aldoxime in the biosynthesis of cyanogenic glucosides and glucosinolates. Plant Mol Biol 38:725–734

Barsalobres-Cavallari CF, Severino FE, Maluf MP, Maia IG (2009) Identification of suitable internal control genes for expression studies in *Coffea arabica* under different experimental conditions. BMC Mol Biol 10:1

Bentsink L, Jowett J, Hanhart CJ, Koornneef M (2006) Cloning of *DOG1*, a quantitative trait locus controlling seed dormancy in *Arabidopsis*. PNAS USA 103:17042–17047

Böhmer M, Schroeder JI (2011) Quantitative transcriptomic analysis of abscisic acid-induced and reactive oxygen species-dependent expression changes and proteomic profiling in *Arabidopsis* suspension cells. Plant J 67:105–118

Brazier-Hicks M, Offen WA, Gershater MC, Revett TJ, Lim EK, Bowles DJ, Davies GJ, Edwards R (2007) Characterization and engineering of the bifunctional *N*- and *O*-glucosyltransferase involved in xenobiotic metabolism in plants. PNAS USA 104:20238–20243

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 55:611–622

Caputi L, Malnoy M, Goremykin V, Nikiforova S, Martens S (2012) A genome-wide phylogenetic reconstruction of family 1 UDP-glycosyltransferases revealed the expansion of the family during the adaptation of plants to life on land. Plant J 69:1030–1042

Carvalho K, Bespalhok Filho JC, Dos Santos TB, de Souza SGH, Vieira LGE, Pereira LFP, Domingues DS (2013) Nitrogen starvation, salt and heat stress in coffee (*Coffea arabica* L.): identification and validation of new genes for qPCR normalization. Mol Biotechnol 53:315–325

Carvalho K, Petkowicz CL, Nagashima GT, Bespalhok Filho JC, Vieira LG, Pereira LF, Domingues DS (2014) Homeologous genes involved in mannitol synthesis reveal unequal contributions in response to abiotic stress in *Coffea arabica*. Mol Genet Genomics 289:951–963

Cenci A, Combes MC, Lashermes P (2012) Genome evolution in diploid and tetraploid *Coffea* species. Plant Mol Biol 78:135–145

Chang S, Puryear J, Cairney J (1993) A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep 11:113–116

Chen PJ, Senthilkumar R, Jane WN, He Y, Tian Z, Yeh KW (2014) Transplastomic *Nicotiana benthamiana* plants expressing multiple defence genes encoding protease inhibitors and chitinase display broad-spectrum resistance against insects, pathogens and abiotic stresses. Plant Biotechnol J 12:503–515

Combes MC, Cenci A, Baraille H, Bertrand B, Lashermes P (2012) Homeologous gene expression in response to growing temperature in a recent allopolyploid (*Coffea arabica* L.). J Hered 103:36–46

Combes MC, Dereeper A, Severac D, Bertrand B, Lashermes P (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

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) BLAST2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676

Cotta MG, Barros LMG, Almeida JD, De Lamotte F, Barbosa EA, Vieira NG, Alves GS, Vinecky F, Andrade AC, Marraccini P (2014) Lipid transfer proteins in coffee: isolation of *Coffea* orthologs, *Coffea arabica* homeologs, expression during coffee fruit development and promoter analysis in transgenic tobacco plants. Plant Mol Biol 85:11–31

Cruz F, Kalaoun S, Nobile P, Colombo C, Almeida J, Barros LM, Romano E, Grossi de Sá MF, Barros LMG, Alves-Ferreira M (2009) Evaluation of coffee reference genes for relative expression studies by quantitative real-time RT-PCR. Mol Breed 23:607–616

Denoeud F, Carretero-Paulet L, Dereeper A et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184

Dong S, Adams KL (2011) Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. New Phytol 190:1045–1057

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42:443–461

El-Sharkawy I, Mila I, Bouzayen M, Jayasankar S (2010) Regulation of two germin-like protein genes during plum fruit development. J Exp Bot 61:1761–1770

Fan H, Xiao Y, Yang Y, Xia W, Mason AS, Xia Z, Qiao F, Zhao S, Tang H (2013) RNA-seq analysis of *Cocos nucifera*: transcriptome sequencing and de novo assembly for subsequent functional genomics approaches. PLoS One 8:e59997

Fernandez D, Santos P, Agostini C, Bon MC, Petitot AS, Silva MC, Guerra-Guimarães L, Ribeiro A, Argout X, Nicole M (2004) Coffee (*Coffea arabica* L.) genes early expressed during infection by the rust fungus (*Hemileia vastatrix*). Mol Plant Pathol 5:527–536

Flagel LE, Chen L, Chaudhary B, Wendel JF (2009) Coordinated and fine-scale control of homoeologous gene expression in allotetraploid cotton. J Hered 100:487–490

Fraser CM, Rider LW, Chapple C (2005) An expression and bioinformatics analysis of the *Arabidopsis* serine carboxypeptidase-like gene family. Plant Physiol 138:1136–1148

Gan D, Jiang H, Zhang J, Zhao Y, Zhu S, Cheng B (2011) Genome-wide analysis of BURP domain-containing genes in Maize and Sorghum. Mol Biol Rep 38:4553–4563

Geromel C, Ferreira LP, Guerreiro SMC, Cavalari AA, Pot D, Pereira LFP, Leroy T, Vieira LGE, Mazzafera P, Marraccini P (2006) Biochemical and genomic analysis of sucrose metabolism during coffee (*Coffea arabica*) fruit development. J Exp Bot 57:3243–3258

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren BW, Chad Nusbaum, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644–652

Granger C, Coryell V, Khanna A, Keim P, Vodkin L, Shoemaker RC (2002) Identification, structure, and differential expression of members of a BURP domain containing protein family in soybean. Genome 45:693–701

Guerra-Guimarães L, Silva MC, Struck C, Loureiro A, Nicole M, Rodrigues CJ Jr, Ricardo CPP (2009) Chitinases of *Coffea arabica* genotypes resistant to orange rust *Hemileia vastatrix*. Biol Plant 53:702–706

Guyot R, Lefebvre-Pautigny F, Tranchant-Dubreuil C, Rigoreau M, Hamon P, Leroy T, Hamon S, Poncet V, Crouzillat D, Kochko A (2012) Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum* sp.) and rosid (*Vitis vinifera*) clades. BMC Genomics 13:103

Heisel TJ, Li CY, Grey KM, Gibson SI (2013) Mutations in *HISTONE ACETYLTRANSFERASE1* affect sugar response and gene expression in *Arabidopsis*. Front Plant Sci 4:245

Hermans C, Porco S, Vandenbussche F, Gille S, De Pessemier J, Van Der Straeten D, Verbruggen N, Bush DR (2011) Dissecting the role of *CHITINASE-LIKE1* in nitrate-dependent changes in root architecture. Plant Physiol 157:1313–1326

Holscher W, Steinhart H (1995) Aroma compounds in green coffee. Dev Food Sci 37:785–803

Hurkman WJ, Tanaka CK (1996) Effect of salt stress on germin gene expression in barley roots. Plant Physiol 110:971–977

Jackson S, Chen ZJ (2010) Genomic and expression plasticity of polyploidy. Curr Opin Plant Biol 13:153–159

Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jiao Y, Wickett NJ, Ayyampalayam S et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100

Joët T, Laffargue A, Salmona J, Doulbeau S, Descroix F, Bertrand B, de Kochko A, Dussert S (2009) Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study. New Phytol 182:146–162

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kunz HH, Häusler RE, Fettke J, Herbst K, Niewiadomski P, Gierth M, Bell K, Steup M, Flügge UI, Schneider A (2010) The role of plastidial glucose-6-phosphate/phosphate translocators in vegetative tissues of *Arabidopsis thaliana* mutants impaired in starch biosynthesis. Plant Biol 12:115–128

Langenbach C, Campe R, Schaffrath U, Goellner K, Conrath U (2013) UDP-glucosyltransferase UGT84A2/BRT1 is required for *Arabidopsis* nonhost resistance to the Asian soybean rust pathogen *Phakopsora pachyrhizi*. New Phytol 198:536–545

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterisation and origin of the *Coffea arabica* L. genome. Mol Gen Genet 261:259–266

Lenfant N, Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A (2013) ESTHER, the database of the a/b-hydrolase fold superfamily of proteins: tools to explore diversity of functions. Nucleic Acids Res 41:D423–D429

Leroy T, Ribeyre F, Bertrand B, Charmetant P, Dufour M, Montagnon C, Marraccini P, Pot D (2006) Genetics of coffee quality. Braz J Plant Physiol 18:229–242

Lin C, Mueller LA, Mc Carthy J, Crouzillat D, Petiard V, Tanksley SD (2005) Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. Theor Appl Genet 112:114–130

Mazzafera P, Carvalho A (1992) Breeding for low seed caffeine content of coffee (*Coffea* L.) by interspecific hybridization. Euphytica 59:55–60

Mochida K, Yamazaki Y, Ogihara Y (2003) Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. Mol Genet Genomics 270:371–377

Mondego JMC, Vidal RO, Carazzolle MF 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

Niewiadomski P, Knappe S, Geimer S, Fischer K, Schulz B, Unte US, Rosso MG, Ache P, Flügge UI, Schneider A (2005) The *Arabidopsis* plastidic glucose 6-phosphate/phosphate translocator GPT1 is essential for pollen maturation and embryo sac development. Plant Cell 17:760–775

Ohmiya A (2002) Characterization of ABP19/20, sequence homologues of germin-like protein in *Prunus persica* L. Plant Sci 163:683–689

Ohmiya A, Tanaka Y, Kadowaki K, Hayashi T (1998) Cloning of genes encoding auxin-binding proteins (ABP19/20) from peach: significant peptide sequence similarity with germin-like proteins. Plant Cell Physiol 39:492–499

Osmani SA, Bak S, Moller BL (2009) Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. Phytochemistry 70:325–347

Passarinho PA, Vries SC (2002) *Arabidopsis chitinases*: a genomic survey. Arabidopsis Book 1:e0023

Petitot AS, Lecouls AC, Fernandez D (2008) Sub-genomic origin and regulation patterns of a duplicated *WRKY* gene in the allotetraploid species *Coffea arabica*. Tree Genet Genomes 4:379–390

Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, de Kochko A, Hamon P (2006) SSR mining in coffee tree EST databases: potential use of EST–SSRs as markers for the *Coffea* genus. Mol Genet Genomics 276:436–449

Poppenberger B, Fujioka S, Soeno K, George GL, Vaistij FE, Hiranuma S, Seto H, Takatsuto S, Adam G, Yoshida S, Bowles D (2005) The UGT73C5 of *Arabidopsis thaliana* glucosylates brassinosteroids. PNAS USA 102:15253–15258

Pourtau N, Jennings R, Pelzer E, Pallas J, Wingler A (2006) Effect of sugar-induced senescence on gene expression and implications for the regulation of senescence in *Arabidopsis*. Planta 224:556–568

Privat I, Foucrier S, Prins A, Epalle T, Eychenne M, Kandalaft L, Caillet V, Lin C, Tanksley S, Foyer C, McCarthy J (2008) Differential regulation of grain sucrose accumulation and metabolism in *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta) revealed through gene expression and enzyme activity analysis. New Phytol 178:781–797

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120

Raghavendra AS, Gonugunta VK, Christmann A, Grill E (2010) ABA perception and signalling. Trends Plant Sci 15:395–401

Ramakers C, Ruijter JM, Deprez RH, Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339:62–66

Rivas F, Parra A, Martinez, Garcia-Granados A (2013) Enzymatic glycosylation of terpenoids. Phytochem Rev 12:327–339

Sasaki C, Varum KM, Itoh Y, Tamoi M, Fukamizo T (2006) Rice chitinases: sugar recognition specificities of the individual subsites. Glycobiology 16:1242–1250

Sharma R, Rawat V, Suresh CG (2014) Genome-wide identification and tissue-specific expression analysis of UDP-glycosyltransferases genes confirm their abundance in *Cicer arietinum* (Chickpea) genome. PLoS One 9:e109715

Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T, Wan XC (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics 12:131

Shibuya M, Nishimura K, Yasuyama N, Ebizuka Y (2010) Identification and characterization of glycosyltransferases involved in the biosynthesis of soyasaponin I in *Glycine max*. FEBS Lett 584:2258–2264

Shirley AM, McMichael CM, Chapple C (2001) The *sng2* mutant of *Arabidopsis* is defective in the gene encoding the serine

carboxypeptidase-like protein sinapoylglucose:choline sinapoyltransferase. Plant J 28:83–94

Staiger D, Apel K, Trepp G (1999) The Atger3 promoter confers circadian clock-regulated transcription with peak expression at the beginning of the night. Plant Mol Biol 40:873–882

Stintzi A, Heitz T, Prasad V, Wiedemann-Merdinoglu S, Kauffmann S, Geoffroy P, Legrand M, Fritig B (1993) Plant 'pathogenesis-related' proteins and their role in defense against pathogens. Biochimie 75:687–706

TAIR—The Arabidopsis Information Resource (2014). http://www.arabidopsis.org/aboutarabidopsis.html. Accessed 25 Feb 2014

Takenaka Y, Nakano S, Tamoi M, Sakuda S, Fukamizo T (2009) Chitinase gene expression in response to environmental stresses in *Arabidopsis thaliana*: chitinase inhibitor allosamidin enhances stress tolerance. Biosci Biotechnol Biochem 73:1066–1071

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Textor S, Gershenzon J (2009) Herbivore induction of the glucosinolate–myrosinase defense system: major trends, biochemical bases and ecological significance. Phytochem Rev 8:149–170

The UniProt Consortium (2014) Activities at the universal protein resource (UniProt). Nucleic Acids Res 42:D191–D198

Toledo-Silva G, Cardoso-Silva CB, Jank L, Souza AP (2013) De novo transcriptome assembly for the tropical grass *Panicum maximum* Jacq. PLoS One 8:e70781

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol 158:590–600

Van Son L, Tiedemann J, Rutten T, Hillmer S, Hinz G, Zank T, Manteuffel R, Bäumlein H (2009) The BURP domain protein AtUSPL1 of *Arabidopsis thaliana* is destined to the protein storage vacuoles and overexpression of the cognate gene distorts seed development. Plant Mol Biol 71:319–329

Veljanovski V, Constabel CP (2013) Molecular cloning and biochemical characterization of two UDP-glycosyltransferases from poplar. Phytochemistry 91:148–157

Vidal RO, Mondego JM, Pot D, Ambrósio AB, Andrade AC, Pereira LF, Colombo CA, Vieira LG, Carazzolle MF, Pereira GA (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

Vieira LGE, Andrade AC, Colombo CA et al (2006) Brazilian coffee genome project: an EST-based genomic resource. Braz J Plant Physiol 18:95–108

Wajant H, Mundry K, Pfizenmaier K (1994) Molecular cloning of hydroxynitrile lyase from *Sorghum bicolor* (L.). Homologies to serine carboxypeptidases. Plant Mol Biol 26:735–746

Wittstock U, Halkier BA (2000) Cytochrome P450 CYP79A2 from *Arabidopsis thaliana* L. catalyzes the conversion of L-phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. J Biol Chem 275:14659–14666

Wu B, Zhang B, Dai Y, Zhang L, Shang-Guan K, Peng Y, Zhou Y, Zhu Z (2012) *Brittle Culm15* encodes a membrane-associated chitinase-like protein required for cellulose biosynthesis in rice. Plant Physiol 159:1440–1452

Xu B, Gou JY, Li FG, Shangguan XX, Zhao B, Yang CQ, Wang LJ, Yuan S, Liu CJ, Chen XY (2013) A cotton BURP domain protein interacts with α-expansin and their co-expression promotes plant growth and fruit production. Mol Plant 6:945–958

Yu Q, Guyot R, de Kochko A, Byers A, Navajas-Pérez R, Langston BJ, Dubreuil-Tranchant C, Paterson AH, 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*). Plant J 67:305–317

Zhang J, Wu K, Zeng S, da Silva JAT, Zhao X, Tian CE, Xia H, Duan J (2013) Transcriptome analysis of *Cymbidium sinense* and its application to the identification of genes associated with floral development. BMC Genomics 14:279