

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA



RAMON OLIVEIRA VIDAL

**AVALIAÇÃO *IN SILICO* DO TRANSCRIPTOMA DO CAFÉ:
IDENTIFICAÇÃO DE SNPS E INFERÊNCIA DE MECANISMOS
DE REGULAÇÃO DA EXPRESSÃO GÊNICA**

Este exemplar corresponde à redação final da tese defendida pelo(a) candidato (a) <u>Ramon Oliveira Vidal</u> e aprovada pela Comissão Julgadora.
--

Tese apresentada ao Instituto de
Biologia para obtenção do Título de
Doutor em Genética e Biologia
Molecular, na área de Bioinformática .

Orientador(a): Prof(a). Dr(a). Gonçalo Amarante Guimarães Pereira

Campinas, 2010

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP**

V667a	Vidal, Ramon Oliveira Avaliação in silico do transcriptoma do café: identificação de SNPs e inferência de mecanismos de regulação da expressão gênica / Ramon Oliveira Vidal. – Campinas, SP: [s.n.], 2010. Orientador: Gonçalo Amarante Guimarães Pereira. Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia. 1. Aloploidia. 2. <i>Coffea arabica</i>. 3. Homeólogos. I. Pereira, Gonçalo Amarante Guimarães. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título. (rcdt/ib)
--------------	---

Título em inglês: In silico analysis of coffee transcriptome: identification of SNPs and inference of mechanisms of gene expression regulation.

Palavras-chave em inglês: Allopoliploidy; *Coffea arabica*; Homeologs.

Área de concentração: Bioinformática.

Titulação: Doutor em Genética e Biologia Molecular.

Banca examinadora: Gonçalo Amarante Guimarães Pereira, Pierre Marraccini, Marcelo Mendes Brandão, Paulo Mazzafera, Michel Eduardo Beleza Yamagishi.

Data da defesa: 16/12/2010.

Programa de Pós-Graduação: Genética e Biologia Molecular.

Campinas, 16 de dezembro de 2010

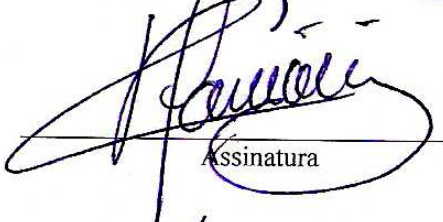
BANCA EXAMINADORA

Prof(a). Dr(a) Gonçalo A. G. Pereira (Orientador(a))



Assinatura

Prof(a). Dr(a). Pierre Marraccini



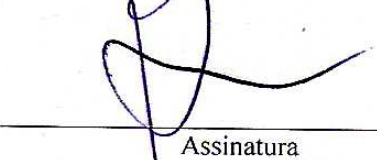
Assinatura

Prof(a). Dr(a) Marcelo Mendes Brandão



Assinatura

Prof(a). Dr(a) Paulo Mazzafera



Assinatura

Prof(a). Dr(a) Michel Eduardo Beleza Yamagishi



Assinatura

Prof(a). Dr(a) Paula Kuser Falcão

Assinatura

Prof(a). Dr(a) Fabio Tebaldi Silveira Nogueira

Assinatura

Prof(a). Dr(a) Johana Rincones

Assinatura

Resumo

O café é uma das culturas mais importantes do mundo, sendo consumido mundialmente e com significativa participação na economia em países em desenvolvimento. *Coffea arabica* e *Coffea canephora* são responsáveis por 70% e 30% da produção comercial, respectivamente. Análise citogenética indicou que *C. arabica* é uma planta alotetraploide autógama formada por uma hibridação (1 milhão de anos atrás) entre os diplóides *C. canephora* e *Coffea eugenoides*. *C. eugenoides* é uma espécie silvestre que cresce em maiores altitudes próximo das bordas de florestas e produz poucas e pequenas sementes com baixo teor de cafeína. Por outro lado, *C. canephora* é alógama e cresce melhor em terras baixas, é também caracterizada por maior produtividade, maior tolerância a pragas e maior teor de cafeína, mas tem uma bebida considerada de qualidade inferior em comparação com a produzida por *C. arabica*. Durante a última década, algumas iniciativas de pesquisa têm sido lançadas para produzir dados genômicos e transcritômicos de algumas espécies de café. Esta coleção de ESTs representa uma boa visão do transcriptoma de *C. arabica* e *C. canephora*, sendo um importante recurso para análise molecular dessas espécies. Este trabalho teve como objetivo obter mais informações sobre algumas espécies do gênero *Coffea*, incluindo a estrutura dos genes, análise de expressão e identificação de genes e famílias gênicas que são específicos ou expandidos em café. Além disso, também foi proposto estudar a regulação da expressão gênica nos genes homeólogos da alotetraploide *C. arabica*. A fim de investigar estes conjuntos de dados de EST foram realizadas duas montagens: (i) a primeira montagem com cada espécie individualmente, com o objetivo de fazer uma análise comparativa entre as *C. arabica*, *C. canephora* e outras culturas, e (ii) com as duas espécies de café juntas, permitindo a identificação de SNPs entre *C. arabica* e *C. canephora*, e avaliar questões evolutivas em *C. arabica*. A identificação dos transcritos diferencialmente expressos e novas famílias gênicas foram utilizados como ponto de partida para a correlação de características de desenvolvimento e de perfis de expressão gênica em *Coffea* sp.. Domínios de proteínas e análises de Gene Ontology sugerem diferenças significativas entre os dados das espécies de café analisadas, principalmente em relação a síntese de açúcares, ligação de proteínas a nucleotídeos, retrotransposons e proteínas de resposta a estresse. A ferramenta OrthoMCL identificou as famílias de proteínas específicas ou predominante de café quando comparado com

outras cinco espécies de plantas. Usando as discrepâncias de alta qualidade encontradas em ESTs sobrepostos de *C. arabica* e *C. canephora*, os perfis de diversidade de sequência foram avaliados em ambas as espécies e utilizados para deduzir a contribuição de *C. canephora* e *C. eugenioides* na transcrição de *C. arabica*. A identificação de genes homeologous de *C. arabica* aos genomas ancestrais permitiu analisar as contribuições de expressão gênica de cada subgenoma. Nós sugerimos que este fenômeno tem uma questão importante na expressão dos genes e fisiologia de *Coffea*.

Palavras-chave: Aloploidia, *Coffea arabica*, Homeólogos.

Abstract

Coffee is one of the most important crops in the world, being worldwide consumed and having significant participation in under development economies. *Coffea arabica* and *Coffea canephora* are responsible for 70% and 30% of commercial production, respectively. Cytogenetic analysis established that *C. arabica* is an autogamous allotetraploid formed by a recent (1 mya) hybridization between the diploids *C. canephora* and *Coffea eugenioides*. *C. eugenioides* is a wild species which grows in higher altitudes near forest edges, and produces few berries with small beans of low caffeine content. On the other hand, *C. canephora* is allogamous and grows better in lowlands. It is also characterized by higher productivity, more tolerance to pests, and higher caffeine content, but it has an inferior beverage compared with *C. arabica*. During the last decade, research initiatives have been launched to produce genomic and transcriptomic data about *Coffea spp.* This EST collection represents a good overview of *C. arabica* and *C. canephora* transcriptome, being appropriate as a resource for *Coffea* molecular analysis. This work aimed to obtain further information about *Coffea spp.* gene structure and expression and to identify genes that are specific or expanded in coffee plants. Moreover, it also intended to study the homeologous gene expression regulation in the allotetraploid *C. arabica*. In order to investigate these data two different EST assemblies were performed: (i) with each species individually, aiming the comparative analysis between the *C. arabica*, *C. canephora* and other crops; and (ii) with both coffee species together, allowing the identification of SNPs between *C.*

arabica and one of its direct ancestors *C. canephora* and the examination of evolutive issues in *C. arabica*. The identification of differentially expressed transcripts and new gene families offered a starting point for the correlation of gene expression profiles and *Coffea sp.* development traits. Protein domain and Gene Ontology analyzes suggested significant differences between the data of coffee species analyzed, mainly in relation to complex sugar synthases, nucleotide binding proteins, retrotransposons and stress response. OrthoMCL tool identified specific or prevalent coffee protein families when compared with other five plant species. Using the high quality discrepancies, found in overlapped ESTs from *C. arabica* and *C. canephora*, sequence diversity profiles were evaluated within both species and used to deduce the transcript contribution of the *C. canephora* and *C. eugenioides* ancestors in the *C. arabica*. The assignment of the *C. arabica* homeologous genes to the ancestral genomes allowed us to analyze gene expression contributions of each subgenome. We suggest that this phenomenon has an important issue in *Coffea* gene expression and physiology.

Keywords: Allopoliploidy, *Coffea arabica*, Homeologs.

Agradecimentos

Ao meu orientador, Prof. Gonçalo Amarante Guimarães Pereira, sou grato pela orientação.

A todos os co-autores dos artigos, pelo apoio nos trabalhos.

Aos colegas da bioinformática do LGE, pelas críticas, discussões e sugestões.

Aos demais colegas, pela amizade.

Aos meus pais, minha irmã e demais familiares pelo apoio durante esta jornada.

À FAPESP, pelo apoio financeiro (processo #2007/51031-2).

Aos meus pais, irmã, avó e tios

Sumário

INTRODUÇÃO.....	10
POLIPLOIDIA E O <i>COFFEA ARABICA</i>	14
SNPs	16
SEQUENCIAMENTO DE ESTs.....	16
CAPÍTULO I - ANÁLISE DOS TRANSCRIPTOMAS DE <i>COFFEA ARABICA</i> E <i>COFFEA CANEPHORA</i>	19
CAPÍTULO II - EXPRESSÃO DIFERENCIAL DE HOMEÓLOGOS EM <i>COFFEA ARABICA</i> ..	71
DISCUSSÕES	86
CONCLUSÕES.....	89
REFERÊNCIAS.....	90
OUTROS TRABALHOS PUBLICADOS PELO AUTOR	94
ANEXO I - MATERIAL SUPLEMENTAR DO CAPÍTULO I.....	95
ANEXO II - MATERIAL SUPLEMENTAR DO CAPÍTULO II.....	127

INTRODUÇÃO

O café (*Coffea* sp.) é uma importante *commodity* agrícola produzida em mais de 60 países e muito consumido em todo o mundo. Em diversos países na África, Ásia e América Latina o café é responsável por uma parcela expressiva da sua economia, (Figura 1). O Brasil, Vietnam e Colômbia são responsáveis por aproximadamente 50% da produção do café no mundo, sendo que o Brasil responde por mais de um terço da produção e das exportações globais do café (Vieira *et al*, 2006).



Figura 1. Regiões onde o café é cultivado com destaque nos países principais produtores.

O gênero *Coffea* pertence à família Rubiaceae e existem cerca de 100 espécies desse gênero, sendo que a maioria delas cresce em baixa altitude nas florestas tropicais da África e Ásia (Bridson e Verdcourt, 1988). Apesar do grande número de espécies, apenas duas têm grande importância econômica mundial: *Coffea arabica* L. e *Coffea canephora* Pierre, correspondendo a aproximadamente 70% e 30% do mercado mundial de café, respectivamente (Fazuoli, 1986).

A família de plantas conhecidas mais estreitamente relacionada ao café é a das Solanaceae. Nesta família, bases de dados genômicos têm sido desenvolvidas para tomate, batata, pimentão, berinjela e petúnia (<http://www.sgn.cornell.edu/>). As Rubiaceae e Solanaceae pertencem ao clado das dicotiledôneas Euasterídeas, e baseados em provas fósseis existentes, eles devem ter se divergido a cerca de 50 milhões de anos atrás (Gandolfo *et al.*, 1998; Crepet *et al.*, 2004). Entre as afinidades taxonômicas mais próximas entre *Coffea ssp.* e as *Solanaceae* estão um grande número de semelhanças botânicas e genéticas, incluindo a produção de frutos carnosos, um conteúdo genômico similar ($C = 950$ e 640 Mb de tomate e café, respectivamente) (Hoeven *et al.*, 2002), número de cromossomos básicos semelhantes ($x = 12$ para tomate e maioria das *Solanaceae*; $x = 11$ para o café) e arquitetura cromossômica semelhante: com pericentro altamente condensado e heterocromatina e eucromatina descondensada na fase paquíteno da meiose (Rick, 1971; Pinto-Maglio e Cruz, 1998).

Coffea arabica (CA) é um alotetraplóide ($2n = 4x = 44$) recente (1 milhão de ano atrás) que tem seu principal centro de diversidade no sudoeste da Etiópia onde ainda crescem indivíduos selvagens (Sylvain 1955). Análises citogenéticas determinaram que *C. arabica* é um anfidiplóide formado por hibridação natural entre as espécies diplóides ($2n = 2x = 22$) *Coffea eugenioides* (CE) e *Coffea canephora* (CC) ou ecótipos relacionados a estas espécies, apresentando altos de níveis de autofecundação (Lashermes *et al.* 1999) (Figura 3). A qualidade da bebida derivada de *C. arabica* é considerada excelente, sendo conhecido no comércio como café suave.

C. canephora é cultivado em terras de baixa altitude e é mais bem adaptada ao clima equatorial quente e úmido. Produz grande quantidade de flores e frutos e é mais tolerante a doenças e pragas do que *C. arabica*. Essa espécie se multiplica por fecundação cruzada, principalmente pela ação do vento e insetos e possui uma grande variabilidade genética (Purseglove, 1968; Crane e Walker, 1983; Free, 1993). *C. canephora* tem um elevado teor de cafeína (1,7-4,0% de massa de sementes seca), mas a qualidade do produto (bebida) é bastante inferior em comparação com *C. arabica*, sendo utilizado em misturas de café solúvel.

C. eugenioides não é produzida em escala comercial, produz poucos frutos com grãos muito pequenos e tem baixo teor de cafeína (0,3-0,6% de massa de sementes seca). *C. eugenioides* cresce em altitudes mais elevadas e próximas a bordas florestais (Maurin *et al.* 2007).

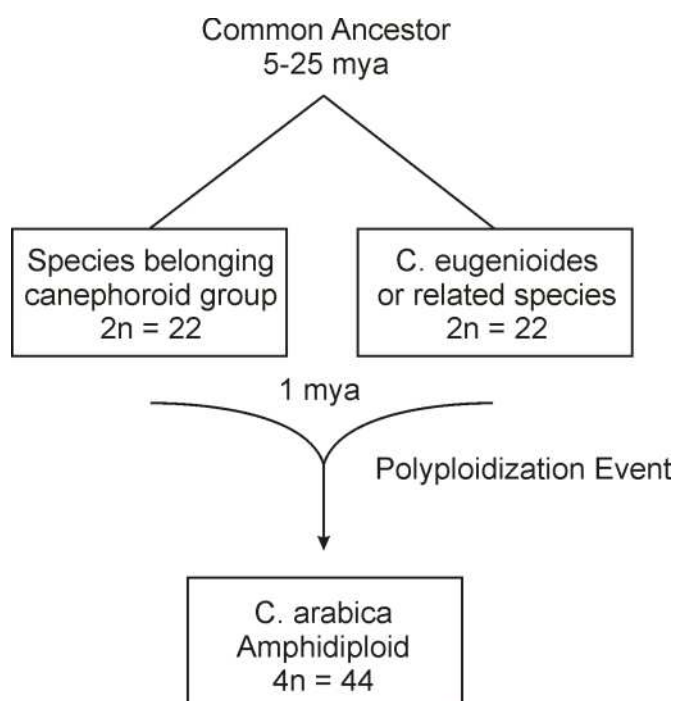


Figura 3. História evolutiva do *Coffea arabica* alotetraplóide. Os genomas progenitores de *C. arabica* estão representados pelos diplóides *C. eugenioides* e *C. canephora*. *C. arabica* formou-se a ~1-2 milhões de anos atrás pela fusão dessas duas espécies ou ecótipos relacionados. A divergência entre esses dois genomas é de 2,5 % em média.

Coffea arabica possui uma baixa diversidade genética atribuída à sua origem, biologia reprodutiva (autogamia) e pelo processo de evolução desta espécie (Anthony *et al.* 2002). Os cultivares de *C. arabica* mais produzidos são Caturra, Catuaí e Mundo Novo que foram selecionados a partir de duas populações de base chamadas comumente por Típica e Bourbon (Anthony *et al.* 2001). Caturra é um anão mutante do grupo Bourbon, enquanto Mundo Novo é um híbrido entre Bourbon e Típica. Catuaí é resultado de um cruzamento entre Caturra e Mundo

Novo. Essas cultivares de *C. arabica* são altamente produtivas e também produzem bebidas de alta qualidade (Figura 4).

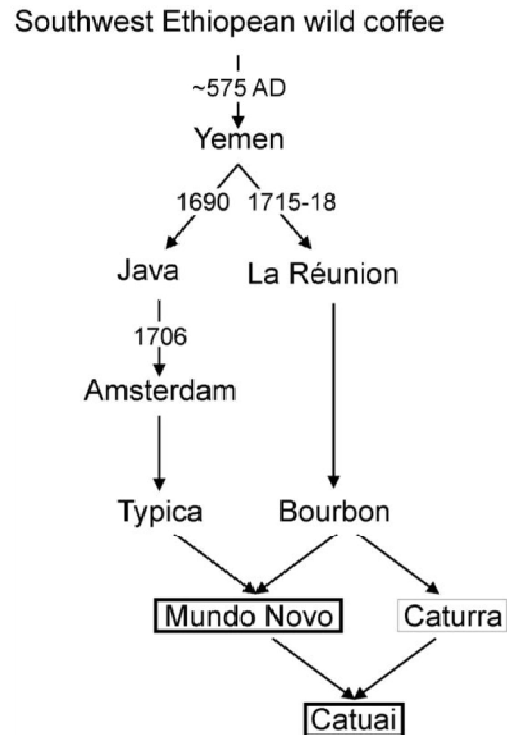


Figura 4. Histórico da seleção artificial de cultivares de *C. arabica*

Melhorias nas características agronômicas do *Coffea arabica*, como a sincronicidade no tempo floração, tamanho do grão, qualidade da bebida, teor de cafeína, resistência aos patógeno e insetos e tolerância a estresse à seca são bastante explorados pela comunidade do café. Entretanto, apesar dos esforços contínuos, o progresso na produção do café, usando abordagens convencionais, tem sido muito lento devido a diversos fatores tais como a estreita base genética do café cultivado, a falta de marcadores genéticos e a falta de ferramentas eficientes de seleção.

Poliploidia e o *Coffea arabica*

Entre as plantas angiospermas, 50 a 70% são poliplóides como *Coffea arabica* (Masterson 1994; Otto & Whitton 2000) e a maioria delas já passou por um evento de poliploidização em algum tempo durante a sua história evolutiva (Masterson 1994; Leitch & Bennett 1997). Muitas das espécies vegetais importantes para a agricultura são (i) autopoliplóides (alfafa e batata) ou (ii) alopoliplóides (trigo, aveia, algodão, canola e café). Outras culturas, como milho e soja parecem ter sofrido poliploidização na sua ascendência (paleopoliplóides), mas as evidências estão ocultas por rearranjos genômicos (Masterson 1994; Leitch & Bennett 1997). A compreensão das consequências da poliploidia sobre a organização do genoma, transcriptoma e da evolução é um ponto essencial para se levar em conta nas estratégias de melhoramento, conservação e reprodução de espécies vegetais.

Poliplóides freqüentemente apresentam novos fenótipos que não estão presentes em seus ancestrais diplóides (Osborn *et al.*, 2003). Em alopoliplóides algumas dessas características têm sido atribuída à expressão diferencial de homeólogos, que são os genes ortólogos das espécies ancestrais que compõem um poliplóide (Mochida *et al.*, 2004; Hovav *et al.*, 2008a; Hovav *et al.*, 2008b; Figura 5). Por exemplo, nos alopoliplóides *Triticum aestivum* (Trigo hexaplóide) e *Gossypium hirsutum* (algodão herbáceo), um subconjunto de genes apresentam homeólogos silenciados epigeneticamente em diferentes tecidos ou em diferentes estágios de desenvolvimento (Adams *et al.*, 2003; Mochida *et al.*, 2004; Adams, 2007; Liu e Adams, 2007; Hovav *et al.*, 2008). Esse fenômeno, conhecido como expressão particionado ou subfuncionalização (Doyle *et al.* 2008), tem o potencial de criar um transcriptoma que é diferente da soma dos valores das espécies ancestrais, permitindo assim a poliplóides ocupar novos nichos ecológico ou mostrar características úteis na agricultura (Osborn *et al.*, 2003, Adams e Wendel, 2005).

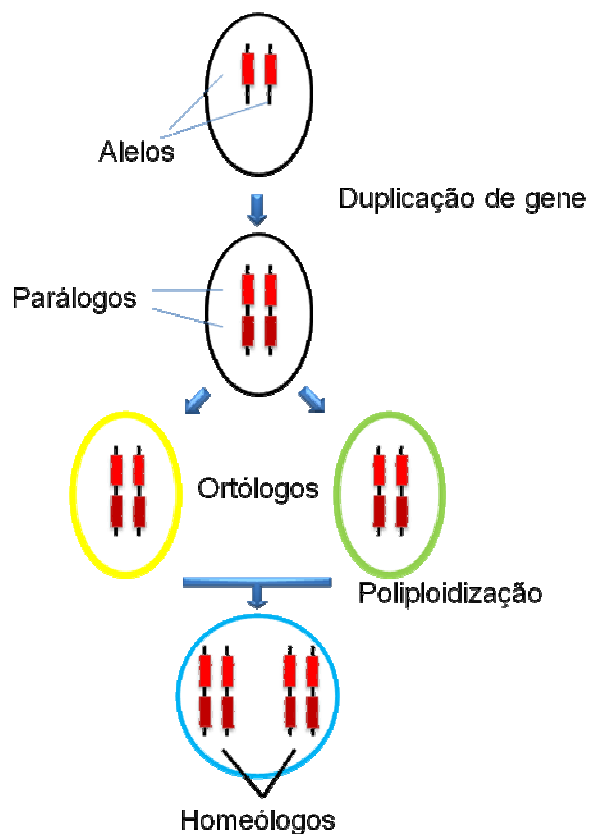


Figura 5. Esquema explicativo dos principais tipos de homologia: i) genes parálogos, que são aqueles que sofreram duplicação na mesma espécie; ii) ortólogos, são genes em diferentes espécies que tem origem em um mesmo gene ancestral; e iii) homeólogos, que são genes homólogos em uma poliplóide com origem em subgenomas diferentes.

A detecção de origem das seqüências de DNA derivadas de cada um dos pais é essencial para a análise do genoma poliplóide. As origens genéticas e diversidade de *C. arabica* foram estudadas anteriormente através da utilização de citogenética, RFLP convencional, AFLP e marcadores moleculares microssatélites (Lashermes *et al.*, 1999; Steiger *et al.*, 2002; Aggarwal *et al.*, 2007; Cubry *et al.*, 2008; Hendre *et al.*, 2008).

SNPs

A recente disponibilidade de seqüenciamento de DNA em larga escala permitiu estudos semelhantes utilizando polimorfismos de nucleotídeo único (SNPs). Polimorfismos que ocorrem em um único nucleotídeo numa mesma posição do gene (SNPs) e pequenas inserções/deleções (INDELS) são as mais frequentes variações que ocorrem em sequências de DNA. Elas podem ser responsáveis por importantes diferenças nas características fenotípicas entre indivíduos e espécie (Emahazion *et al.*, 2001; Sherry *et al.*, 2001). Análises de SNPs utilizando seqüências de ESTs a partir de culturas agrícolas foram empregadas para a criação de mapas genéticos de alta densidade e na identificação de regiões genômicas variáveis (Du *et al.*, 2002; Choi *et al.*, 2007; Novaes *et al.*, 2008; Pindo *et al.*, 2008; Duran *et al.*, 2009).

Uma das limitações dos SNPs, como um dos principais marcadores moleculares é o custo inicial associado com seu desenvolvimento. Entretanto, a utilização de métodos em bioinformática em larga escala para detecção de SNPs tem levado a uma inovação no seu uso como marcador molecular.

A simplicidade e a baixa taxa de mutação dos SNPs os tornam excelentes marcadores moleculares para o estudo de tratamentos genéticos complexos e como ferramenta para entender a evolução do genoma (Syvanen, 2001). Polimorfismos em regiões codificantes que resultam em mudanças nos resíduos de aminoácidos [i.e. non-synonymous SNPs(nsSNPs)] são importantes fontes de variação fenotípicas.

Além disso, SNPs presentes em regiões expressas também são úteis para identificar genes homeologos dos genomas ancestrais da alopoliplóide, bem como a seu nível de expressão relativa (Mochida *et al.*, 2004; Hovav *et al.* 2008b). Esta informação é essencial para compreender o novo fenótipo associado à expressão particionada.

Sequenciamento de ESTs

O desenvolvimento de novas tecnologias aplicadas à biologia vem gerando um vasto conhecimento na área de genômica de plantas. O seqüenciamento em larga escala de cDNA, visando a produção de ESTs (*Expressed sequenced tags*), fornece evidências diretas para todas as amostra de transcritos, permitindo uma rápida caracterização do conjunto do genes expressos. Os ESTs são geralmente o primeiro material a ser pesquisado quando se busca uma análise do transcriptoma e inferências sobre a estrutura genômica do organismo em estudo. São seqüências

curtas (200 a 800 nucleotídeos) que não sofrem edição, sendo selecionados de forma aleatória a partir de bibliotecas de cDNA. A comparação dos ESTs com seqüências do banco de dados gera informações a cerca da capacidade de codificação do genoma, e pode identificar novos genes com potencial biotecnológico de forma rápido e com baixo custo efetivo. Além disso, estas seqüências podem proporcionar a identificação de polimorfismos de nucleotídeo único (SNPs) em regiões codificantes uma vez que existe redundância para boa parte dos transcritos (Useche *et al*, 2001).

A fim de melhorar a compreensão das bases moleculares de características agronômicas e o desenvolvimento de plantas para programas de melhoramento do café, ESTs de duas espécies de café foram gerados. O Projeto Genoma Café Brasileiro (Vieira *et al*. 2006) seqüenciou cerca de 200.000 ESTs de 56 bibliotecas de *C. arabica* (Catuaí e Mundo Novo; 65%) e *C. canephora* (33%). Concomitantemente, Lin *et al*, 2005 seqüenciaram aproximadamente 47.000 ESTs de *C. canephora*. A distribuição das bibliotecas dessas duas espécies pode ser vista na Figura 6.

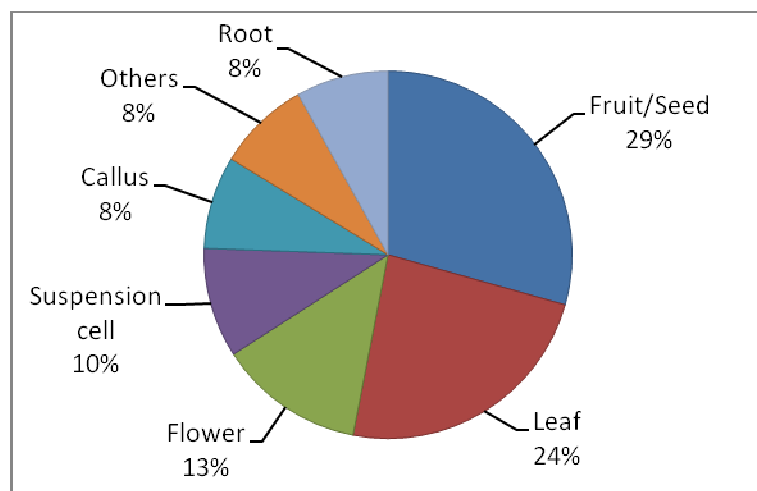


Figura 6. Distribuição das bibliotecas de *C. arabica* e *C. canephora* pelos principais tecidos

A partir desse conjunto de dados foram realizadas duas montagens, uma individual de cada espécie e outra híbrida entre as duas espécies de café. Essas montagens terão os objetivos distintos de: realizar um estudo comparativo entre *C. arabica* e *C. canephora* e identificar os

subgenomas *C. canephora* e *C. eugenioides* no transcriptoma de *C. arabica* através da análise de padrões de SNPs entre *C. arabica* e *C. canephora*.

CAPÍTULO I - Análise dos Transcriptomas de *Coffea arabica* e *Coffea canephora*

Artigo aceito para publicação na revista
BMC Plant Biology

An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*

Jorge Maurício Costa Mondego^{1,*}, Ramon Oliveira Vidal^{2,*}, Marcelo Falsarella Carazzolle^{2,3}, Eric Keiji Tokuda², Lucas Pedersen Parizzi², Gustavo Gilson Lacerda Costa², Luiz Filipe Protasio Pereira⁴, Ângela Metha⁵, Damares de Castro Monte⁵, Eduardo Romano⁵, Elionor Rita Pereira de Almeida⁵, Érika Valéria S. Albuquerque⁵, Felipe Rodrigues da Silva⁵, João Batista Teixeira⁵, Marcos Mota do Carmo Costa⁵, Maria de Fátima Grossi-de-Sá⁵, Luciana Beatriz Dutra Labuto⁵, Luiz Lehmann Coutinho⁶, Érika Cristina Jorge⁶, Marcos Antonio Machado⁷, Claudia Barros Monteiro-Vitorello⁸, Luis Eduardo Aranha Camargo⁹, Hamza Fahmi Ali El Dorry¹⁰, Helaine Carrer¹¹, Maria Helena S. Goldman¹², Ricardo Harakava¹³, Edna Teruko Kimura¹⁴, Éder Antônio Giglioti¹⁵, Marie-Anne Van Sluys¹⁶, Mariana Cabral de Oliveira¹⁶, Maria Inês T. Ferro¹⁷, Regina L.B.C. de Oliveira¹⁸, Paulo Arruda¹⁹, Celso Luis Marino²⁰, Walter José Siqueira¹, Haiko Enok Sawazaki¹, Eliana Gertrudes de Macedo Lemos²¹, Manoel Victor Franco Lemos²¹, Siu Mui Tsai²², Eiko Eurya Kuramae²³, Sonia Marli Zingaretti di Mauro¹⁷, Carlos Alberto Labate⁸, Mirian Therezinha Souza da Eira²⁴, João Paulo Kitajima²⁵, Eduardo Fernandes Formighieri²⁶, Oliveira Guerreiro Filho²⁷, Mirian Perez Maluf²⁸, Paulo Mazzafera²⁹, Alan Carvalho Andrade^{5,+}, Carlos Augusto Colombo^{1,+}, Luiz Gonzaga Esteves Vieira^{30,+}, Gonçalo Amarante Guimarães Pereira^{2,+,#}

* Both authors contributed equally to this work +

Coordinators of the initiative

Corresponding author

1 - Centro de Recursos Genéticos Vegetais, Instituto Agronômico de Campinas, CP 28, 13001-970, Campinas-SP, Brazil; 2 - Laboratório de Genômica e Expressão, Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas, CP 6109, 13083-970, Campinas, SP, Brazil; 3 - Centro Nacional de Processamento de Alto Desempenho em São Paulo, Universidade Estadual de Campinas, CP 6141, 13083-970, Campinas, SP, Brazil; 4 - Embrapa Café - Instituto Agronômico do Paraná, Laboratório de Biotecnologia Vegetal, CP 481, 86001-970, Londrina-PR, Brazil; 5 - Núcleo de Biotecnologia-NTBio, Embrapa Recursos Genéticos e Biotecnologia, Parque Estação Biológica, CP 02372, 70770-900, Brasília-DF, Brazil; 6 - Departamento de Zootecnia, Escola Superior de Agricultura Luiz de Queiroz, USP, 13418-900, Piracicaba, SP, Brazil; 7 - Centro APTA de Citros Sylvio Moreira, IAC, CP 04, 13490-970,

Cordeirópolis SP, Brazil; 9 - Departamento de Entomologia, Fitopatologia e Zoologia Agrícola, Escola Superior de Agricultura Luiz de Queiroz, USP, 13418-900, Piracicaba, SP, Brazil; 10 - Biology Department, American University in Cairo, Cairo, Egypt; 11 - Departamento de Ciências Biológicas, Escola Superior de Agricultura Luiz de Queiroz, USP, 13418-900, Piracicaba, SP, Brazil; 12 - Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, USP, 14040-901, Ribeirão Preto, SP, Brazil; 13 - Centro de Sanidade Vegetal, Instituto Biológico de São Paulo, 04014-002, São Paulo, SP, Brazil; 14 - Instituto de Ciências Biomédicas, USP, 05508-000, São Paulo, SP, Brazil; 15 - Faculdades Adamantinenses Integradas, Grupo de Pesquisa em Energia e Biotecnologia - Genebio, 17800-000, Adamantina, SP - Brazil; 16 - Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, 05508-900, São Paulo, SP, Brazil; 17 - Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, UNESP, 14884-900, Jaboticabal, SP, Brazil; 18 - Núcleo Integrado de Biotecnologia, Universidade de Mogi das Cruzes, 08780-911, Mogi das Cruzes, SP, Brazil; 19 - Centro de Biologia Molecular e Engenharia Genética, UNICAMP, CP 6010, 13083-970, Campinas, SP, Brazil; 20 - Departamento de Genética, Instituto de Biociências, UNESP, 18618-000, Botucatu SP, Brazil; 21 - Departamento de Biologia Aplicada à Agropecuária, Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal, UNESP, 14884-900, Jaboticabal, SP, Brazil; 22 - Centro de Energia Nuclear na Agricultura, USP, CP 96, 13400-970, Piracicaba, SP, Brazil; 23 - Departamento de Defesa Fitossanitária, Faculdade de Ciências Agrônomicas, UNESP, CP 237, 18603-970, Botucatu, SP, Brazil; 24 - Embrapa Café, Núcleo de Biotecnologia-NTBio, Embrapa Recursos Genéticos e Biotecnologia, CP 02372, 70770-900, Brasília, Brazil; 25 - Alellyx Applied Genomics, 13067-850, Campinas, SP, Brazil; 26 - Embrapa Agroenergia, Parque Estação Biológica, 70770-901, Brasília-DF, Brazil; 27 - Centro de Café Alcides Carvalho, Instituto Agrônomo de Campinas, 13012-970, Campinas, SP, Brazil; 28 - Embrapa Café, IAC, Centro de Café Alcides Carvalho, Campinas, 13012-970, Campinas, SP, Brazil; 29 - Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas, CP 6109, 13083-970, Campinas, SP, Brazil; 30 - Instituto Agrônomo do Paraná, Laboratório de Biotecnologia Vegetal, CP 481, CEP 86001-970, Londrina-PR, Brazil.

E-mails:

Jorge Maurício Costa Mondego - jmcomondego@gmail.com; jmcondago@iac.sp.gov.br; Ramon Oliveira Vidal - vidal@lge.ibi.unicamp.br; Marcelo Falsarella Carazzolle - mcarazzo@lge.ibi.unicamp.br; Eric Keiji Tokuda - keiji.eric@gmail.com; Lucas Pedersen Parizzi - lucas@lge.ibi.unicamp.br; Gustavo Gilson Lacerda Costa - glacerda@lge.ibi.unicamp.br; Luiz Filipe Protasio Pereira - lpereira@iapar.br; Ângela Metha - amehta@cenargen.embrapa.br; Damares de Castro Monte - damares@cenargen.embrapa.br; Eduardo Romano - romano@cenargen.embrapa.br; Elionor Rita Pereira de Almeida - elionor@cenargen.embrapa.br; Érika Valéria S. Albuquerque - erikavsa@cenargen.embrapa.br; Felipe Rodrigues da Silva - felipes@cenargen.embrapa.br; João Batista Teixeira - batista@cenargen.embrapa.br; Marcos Mota do Carmo Costa - mcosta@cenargen.embrapa.br; Maria de Fátima Grossi de Sá - fatimasa@cenargen.embrapa.br; Luiz Lehmann Coutinho - llcoutin@esalq.usp.br; Érika Cristina Jorge - ecjorge7@hotmail.com; Marcos Antonio Machado - marcos@centrodecitricultura.br; Claudia Barros Monteiro-Vitorello - cbmontei@esalq.usp.br; Luis Eduardo Aranha Camargo -

leacamar@esalq.usp.br; Hamza Fahmi Ali El Dorry - dorry@aucegypt.edu; Helaine Carrer – hecarrer@esalq.usp.br; Maria Helena S. Goldman - mgoldman@ffclrp.usp.br; Ricardo Harakava - harakava@biologico.sp.gov.br; Edna Teruko Kimura - etkimura@usp.br; Éder Antônio Giglioti - edergiglioti@fai.com.br; Marie-Anne Van Sluys - mavsluys@gmail.com; Mariana Cabral de Oliveira - mcdolive@ib.usp.br; Maria Inês T. Ferro - mitferro@fcav.unesp.br; Regina L.B.C. de Oliveira - biotec@umc.br; Paulo Arruda – parruda@unicamp.br; Celso Luis Marino - clmarino@ibb.unesp.br; Walter José Siqueira – walterjs@iac.sp.gov.br; Haiko Enok Sawazaki – henok@iac.sp.gov.br; Eliana Gertrudes de Macedo Lemos - egerle@fcav.unesp.br; Manoel Victor Franco Lemos – mvector@fcav.unesp.br; Siu Mui Tsai - tsai@cena.usp.br; Eiko Eurya Kuramae - kizioka@fca.unesp.br; Sonia Marli Zingaretti di Mauro - zingara@fcav.unesp.br; Mirian Therezinha Souza da Eira - Mirian.Eira@embrapa.br; João Paulo Kitajima - joao.kitajima@alellyx.com.br; Eduardo Fernandes Formighieri - eduformi@gmail.com; Carlos Alberto Labate - calabate@esalq.usp.br; Oliveira Guerreiro Filho - oliveiro@iac.sp.gov.br; Mirian Perez Maluf - maluf@iac.sp.gov.br; , Paulo Mazzafera – pmazza@unicamp.br; Alan Carvalho Andrade – alan@cenargen.embrapa.br; Carlos Augusto Colombo – ccolombo@iac.sp.gov.br; Luiz Gonzaga Esteves Vieira – lvieira@iapar.br; Gonçalo Amarante Guimarães Pereira – gonçalo@unicamp.br

ABSTRACT

Background: Coffee is one of the world's most important crops; it is consumed worldwide and plays a significant role in the economy of producing countries. *Coffea arabica* and *C. canephora* are responsible for 70 and 30% of commercial production, respectively. *C. arabica* is an allotetraploid from a recent hybridization of the diploid species, *C. canephora* and *C. eugenioides*. *C. arabica* has lower genetic diversity and results in a higher quality beverage than *C. canephora*. Research initiatives have been launched to produce genomic and transcriptomic data about *Coffea* spp. as a strategy to improve breeding efficiency.

Results: Assembling the expressed sequence tags (ESTs) of *C. arabica* and *C. canephora* produced by the Brazilian Coffee Genome Project and the Nestlé-Cornell Consortium revealed 32,007 clusters of *C. arabica* and 16,665 clusters of *C. canephora*. We detected different GC3 profiles between these species that are related to their genome structure and mating system. BLAST analysis revealed similarities between coffee and grape (*Vitis vinifera*) genes. Using KA/KS analysis, we identified coffee genes under purifying and positive selection. Protein domain and gene ontology analyses suggested differences between *Coffea* spp. data, mainly in relation to complex sugar synthases and nucleotide binding proteins. OrthoMCL was used to identify specific and prevalent coffee protein families when compared to five other plant species. Among the interesting families annotated are new cystatins, glycine-rich proteins and RALF-like peptides. Hierarchical clustering was used to independently group *C. arabica* and *C. canephora* expression clusters according to expression data extracted from EST libraries, resulting in the identification of differentially expressed genes. Based on these results, we emphasize gene annotation and discuss plant defenses, abiotic stress and cup quality-related functional categories.

Conclusion: We present the first comprehensive genome-wide transcript profile study of *C. arabica* and *C. canephora*, which can be freely assessed by the scientific community at www.lge.ibi.unicamp.br/coffee. Our data reveal the presence of species-specific/prevalent genes in coffee that may help to explain particular characteristics of these two crops. The identification of differentially expressed transcripts offers a starting

point for the correlation between gene expression profiles and *Coffea* spp. developmental traits, providing valuable insights for coffee breeding and biotechnology, especially concerning sugar metabolism and stress tolerance.

INTRODUCTION

Coffee is the most important agricultural commodity in the world and is responsible for nearly half of the total exports of tropical products [1]. Indeed, coffee is an important source of income for many developing tropical countries. Brazil, Vietnam and Colombia account for > 50% of global coffee-production. In addition, coffee is also important to many non-tropical countries that are highly involved in coffee industrialization and commerce and are intensive consumers of coffee beverages.

Two species of the genus *Coffea* are responsible for almost all coffee bean production: *C. arabica* and *C. canephora* (approximately 70 and 30% of worldwide production, respectively). *C. arabica* is an autogamous allotetraploid (amphidiploid; $2n = 4x = 44$) species originating from a relatively recent cross (1 mya) between *C. canephora* (or a canephoroide-related species) and *C. eugenioides*, which occurred in the plateaus of Central Ethiopia [2, 3]. As a consequence of its autogamy and evolutionary history, “Arabica” coffee plants have a narrow genetic basis. This problem is amplified in the main cultivated genotypes (i.e., Mundo Novo, Catuai and Caturra), which were selected from only two base populations: Typica and Bourbon [4]. Conversely, *C. canephora* is a diploid ($2n = 2x = 22$), allogamous and more polymorphic *Coffea* species. In contrast to *C. arabica*, which is grown in highland environments, *C. canephora* is better adapted to warm and humid equatorial lowlands. *C. arabica* is regarded as having a better cup quality, which seems to depend on the quality and amount of compounds stored in the seed endosperm during bean maturation [5-7]. Conversely, *C. canephora* is considered more resistant to diseases and pests and has a higher caffeine content than *C. arabica* [8]. Other important differences are related to fruit maturation. Though *C. canephora* blossoms earlier, its fruit maturation is delayed in comparison to *C. arabica* [9]. Improvements in the agronomic characteristics of coffee (e.g., cup quality, pathogen and insect resistance and drought stress tolerance) are long-sought by the coffee farming-community. However, the

introduction of a new trait into an elite coffee variety via conventional breeding techniques is a lengthy process due to the narrow genetic basis of *C. arabica* [4, 10] and the long seed-to-seed generation cycle.

Expressed sequence tags (ESTs) provide a source for the discovery of new genes and for comparative analyses between organisms. Many EST sequencing efforts have successfully provided insights into crop plants development [11-18]. EST sequencing allows quantitative expression analyses by correlating EST frequency with the desirable traits of plant species. It also constitutes an interesting tool for the detection of tissue/stress specific promoters and genetic variation that may account for specific characteristics. Furthermore, EST analyses can provide targets for transgenesis, an interesting tool for genetic improvement of such a long generation time crop as coffee. In fact, data in coffee genetic transformation indicate the potential of this approach in molecular breeding [19, 20].

Research on coffee genomics and transcriptomics has gained increasing attention recently. A Brazilian consortium (Brazilian Coffee Genome Project; BCGP) [21] was developed to investigate coffee traits by sequencing cDNA derived from a series of tissues of *C. arabica*, *C. canephora* and *C. racemosa*, a coffee species used in breeding programs for the introgression of resistance against coffee leaf miner. Concomitantly, an initiative from the Nestlé Research Center and the Department of Plant Biology at Cornell University sequenced ESTs from *C. canephora* farm-grown in east Java, Indonesia. This research group compared the EST repertoires of *C. canephora*, *Solanum lycopersicum* (tomato) and *Arabidopsis thaliana* [22, 23]. Based on their analysis, it was verified that *C. canephora* and tomato have a similar assembly of genes, which is in agreement with their similar genome size, chromosome karyotype, and chromosome architecture [22]. In addition, an important platform for functional genomics that can be applied to coffee was carried out by the SOL Genomics Network (SGN; <http://sgn.cornell.edu>), a genomics information resource for the Solanaceae family and related families in the Asterid clade, such as *Coffea spp.* and other Rubiaceae species [23].

The availability of EST data from both of the commercially most important *Coffea*

spp. prompted us to perform a wide bioinformatics analysis. In this report, we surveyed the coffee transcriptome by analyzing ESTs from *C. arabica* and *C. canephora*. Resources developed in this project provide genetic and genomic tools for *Coffea* spp. evolution studies and for comparative analyses between *C. arabica* and *C. canephora*, regarding gene families' expansion and gene ontology. We also identified *Coffea*-specific/prominent gene families using automatic orthology analysis. Additionally, we describe the annotation of differentially expressed genes according to *in silico* analysis of EST frequencies.

RESULTS AND DISCUSSION

Overall *Coffea* spp. EST libraries data

To evaluate ESTs from *Coffea* spp. we collected 187,412 ESTs derived from 43 cDNA libraries produced by the Brazilian Coffee Genome Project initiative [21]. The *C. arabica* libraries represent diverse organs, plant developmental stages and stress treatments from Mundo Novo and Catuaí cultivars, excluding germinating seeds (cv Rubi) (Additional File 1). In the case of *C. canephora*, 62,823 ESTs from six cDNA libraries of the Nestlé and Cornell *C. canephora* sequencing initiative [22] and 15,647 *C. canephora* ESTs from three cDNA libraries constructed by the Brazilian Coffee Genome Project initiative [21] were collected yielding a total of 78,470 ESTs (Additional File 1). All ESTs were produced by the Sanger method, and cDNA clones were subjected only to 5' sequencing. The pipeline of *C. arabica* and *C. canephora* EST analysis is described in Figure 1.

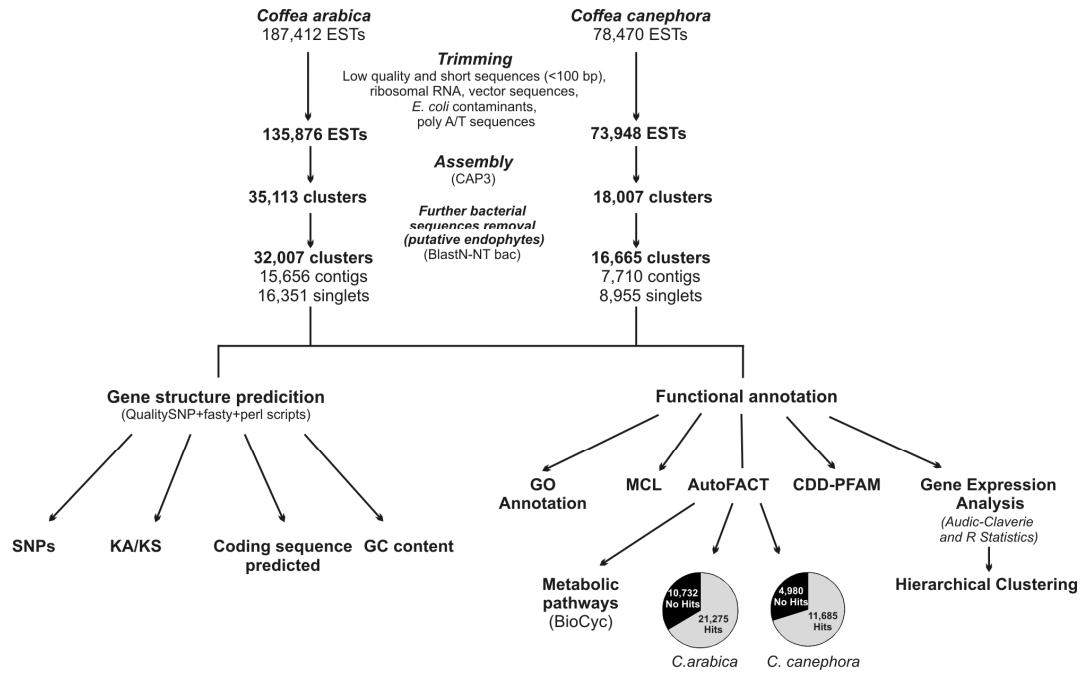


Figure 1. Flow diagram of bioinformatics procedures applied in *C. arabica* and *C. canephora* transcriptomic analyses.

After trimming (i.e., vector, ribosomal, short, low quality and *E. coli* contaminant sequences removal), 135,876 *C. arabica* ESTs were assembled into 17,443 contigs and 17,710 singlets (35,113 clusters; Figure 1), and the *C. canephora* ESTs were assembled into 8,275 contigs and 9,732 singlets (18,007 clusters; Figure 1). After manual annotation, we detected some clusters similar to bacterial sequences that were not identified during trimming. Clusters were then evaluated using BLASTN against a version of NT-bac and BLASTX against the NR database. Sequences similar to bacteria were removed from further analyses. These sequences are likely derived from endophytes of coffee plants. After their removal from the dataset, the final number of clusters was 32,007 (15,656 contigs and 16,351 singlets) from *C. arabica* and 16,665 (7,710 contigs and 8,955 singlets) from *C. canephora* (Table 1). The average length of *C. canephora* and *C. arabica* clusters in the dataset was 662 bp (ranging from 100 to 3,584 bp) and 663 bp (ranging from 100 to 2,988 bp), respectively (Table 1). The number of ESTs in the *C. canephora* and *C. arabica* contigs ranged from 2 to 1,395 and 2 to 493, respectively (Figure 2). In both cases, approximately 63% were composed of ≤ 20 ESTs, and 98% of the contigs contained < 50

ESTs. We also verified the distribution of ESTs in contigs across multiple libraries. Nineteen percent of *C. arabica* contigs and 4% of *C. canephora* contigs were found in only one library (Additional File 2). The majority of *C. arabica* contigs (32%) have only two ESTs, each one from a different EST library. Due to the limited depth of sequencing and the variety of tissue samples used to construct the *C. arabica* libraries, a smoother distribution of contigs *per* library was observed in comparison with *C. canephora* (Additional File 2).

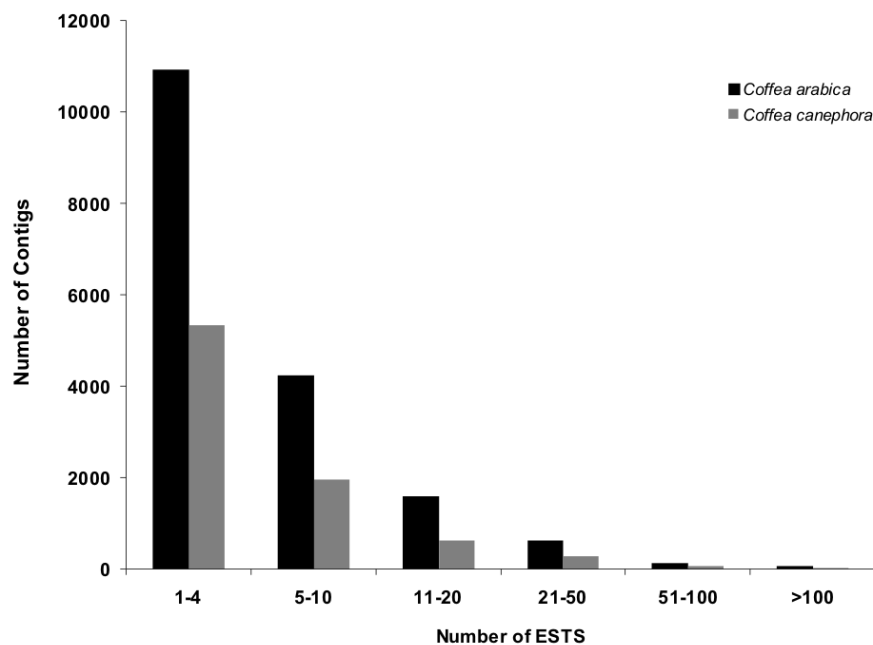


Figure 2. Distribution of the number of ESTs in contigs of *C. arabica* and *C. canephora* after the assembly process.

Table 1. Summary of *Coffea* spp. cluster datasets.

	Contigs	Average contig length	Singlets	Average singlet length	Clusters	Average cluster length
<i>C. arabica</i>	15,656	868 bp	16,351	459 bp	32,007	662 bp (ranging from 100 to 3,584 bp)
<i>C. canephora</i>	7,710	832 bp	8,955	494 bp	16,665	663 bp (ranging from 100 to 2,988 bp)

Evaluation of GC content, SNPs and sequence similarity with other

species

We evaluated the structure of *Coffea* contigs to identify the percentage of coding sequences (CDS) in our dataset using the QualitySNP program tools [24]. The mode and median length of CDS and 5' and 3' UTRs were similar to both species (Table 2). We also inspected the amount of full length CDS in our dataset, resulting in 1,189 contigs in *C. arabica* (8%) and 518 contigs in *C. canephora* (7%; Table 2).

Table 2. Evaluation of CDS, 5'UTR and 3'UTR of *Coffea* spp.

	Full length CDS sequences	5'UTR length (median)	CDS length (median)	CDS length (mode)	3'UTR length (median)
<i>C. arabica</i>	1,189	160 bp	836 bp	479 bp	240 bp
<i>C. canephora</i>	518	134 bp	708.5 bp	476 bp	229.5 bp

Based on the annotation of CDS, we evaluated the GC content in coding regions. In general, the GC and GC3 profiles (i.e., the GC level at the third codon position) of *C. canephora* and *C. arabica* are similar to Arabidopsis and tomato. The unimodal GC distribution is a common feature of dicotyledons (Figure 3), whereas bimodal distribution is common in monocotyledons [17, 25]. Nevertheless, *Coffea* spp. and Arabidopsis have a slightly higher proportion of genes with high GC content than tomato and have a more accentuated peak shift in GC3 content (Figure 3). This difference between Arabidopsis and tomato was found previously [25] and was attributed to differences in the gene samples, such as the presence of intron-retained transcripts (differentially spliced transcripts) in tomato. A more detailed inspection revealed that *C. arabica* has only one GC3 peak, while *C. canephora* has two close peaks: the first similar to that found for *C. arabica* and the other positioned toward the “GC-rich content area”. This *C. canephora* pattern may be related to its outcrossing mating system because allogamous species tend to accumulate more polymorphism in the third codon position and to be more GC-rich than autogamous species

[26], as is the case of Arabica coffee, tomato and Arabidopsis.

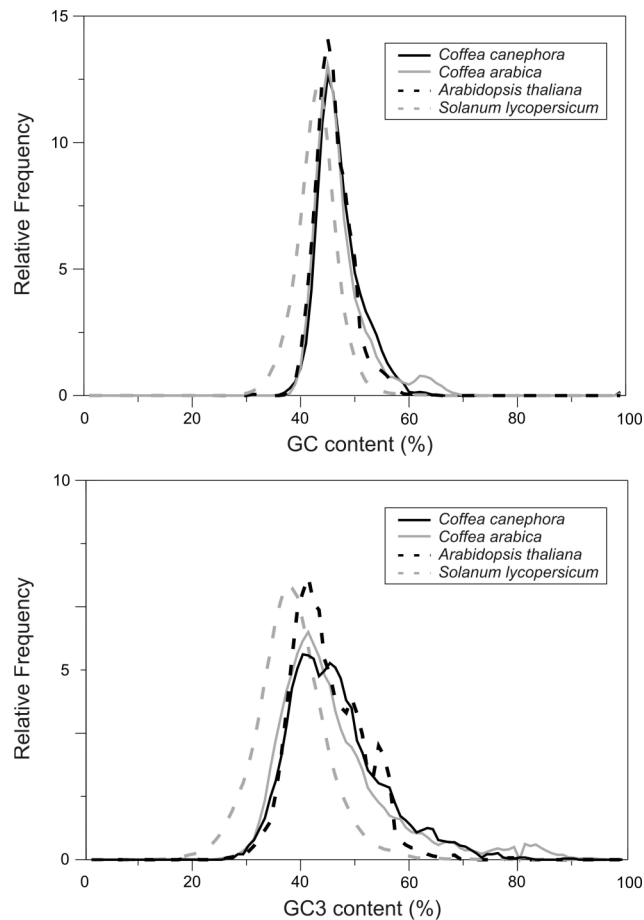


Figure 3. Distribution of GC in the coding regions of *Arabidopsis thaliana*, *Solanum lycopersicum*, *C. arabica* and *C. canephora*.

We also used QualitySNP to calculate SNPs present in *C. arabica* and *C. canephora* contigs. In the case of *C. arabica*, we selected contigs containing at least four reads, which in theory provide two copies for each allele, yielding 8,514 *C. arabica* and 3,832 *C. canephora* contigs. Approximately 53% (4,535) of the *C. arabica* contigs and 52% (2,000) of the *C. canephora* contigs were found to contain SNPs (Additional File 3). Similar to other reports [27-29], more transitions than transversions were found for both species (Additional File 3), likely reflecting the high frequency of cytosine to thymine mutation after methylation.

The frequency of SNPs in *C. arabica* was 0.35 SNP/100 bp, almost double the *C. canephora* SNP frequency (0.19 SNP/100 bp). Similarly, Lashermes et al. [3] and Vidal et al. [30] indicated that Arabica has a level of internal genetic variability almost twice that present in *C. canephora*. The majority of polymorphisms found in both species was bi-allelic (99.8% for *C. arabica* and 99.5% for *C. canephora*), with a low percentage of tri-allelic and no tetra-allelic SNPs (Additional File 3)

We next used AutoFACT [31] to evaluate the putative functions of the two *Coffea* datasets. The results of BLASTX against the non-redundant protein sequence database (NR; E-value cutoff of $1e^{-10}$) available at AutoFACT were inspected to evaluate the similarity of *Coffea* clusters with proteins deposited in GenBank. Approximately 68% of *C. arabica* and 71% of *C. canephora* clusters have significant sequence similarity (E-value $\leq 1e^{-10}$) with genes in the databank. The remaining clusters represented sequences with lower E-value scores (E-value $> 1e^{-10}$) designated as “no-hits” (Table 3). Because *C. arabica* and *C. canephora* are species from the Rubiaceae family, which have few sequences deposited in the NR database, we expected that sequences from other species in the Asteridae clade (e.g., members of the Solanaceae family *S. lycopersicum*, *S. tuberosum* and *Nicotiana tabacum*) would be the most similar to *Coffea* sequences. However, the majority of *Coffea* clusters have higher similarity with *Vitis vinifera* sequences (~40%), a species from the Rosids clade, followed by the other rosids *Arabidopsis* (~5.5%) and *Populus trichocarpa* (~3.5%). The top hits of Coffee sequences with Solanaceae range from 1 to 2% (Table 3). We then compared the *Coffea* sequences with a database containing contigs from the plant EST databank TIGR, the plant transcript database (<http://plantta.jcvi.org>) and GeneIndex Plants (<http://compbio.dfci.harvard.edu/tgi/plant.html>), which have a higher amount of Solanaceae data. For both *C. arabica* and *C. canephora*, *N. tabacum* was the species with more top hits (11.15 and 11.59%, respectively), followed by *V. vinifera* (10.34 and 10.03%), *S. lycopersicum* (6.5 and 5%) and *S. tuberosum* (5 and 4.8%; data not shown). We believe that the most parsimonious hypothesis for these results is related to phylogenetic issues. Grape is basal to the rosids clade and did not undergo whole genome duplication (WGD)

events, such as Arabidopsis, thus being theoretically more similar to the rosids paleohexaploid ancestor [32, 33]. Analysis of genomic sequences from the asterid common monkey flower (*Mimulus guttatus*) revealed extensive synteny with grape, suggesting that paleohexaploidy antedates the divergence of the rosid and asterid clades [33]. Notably, recent data prove that there is a high level of collinearity between diploid *Coffea* and *V. vinifera* genomic regions [34], and that these species derive from the same paleo-hexaploid ancestral genome [35]. Intensive genomic analyses are currently underway to more deeply compare the genomes of rosids and asterids species.

Table 3. Predicted *C. arabica* and *C. canephora* gene comparisons.

<i>Coffea arabica</i>		
Species	# Hits*	% Hits
<i>Vitis vinifera</i>	13,855	43.29%
<i>Arabidopsis thaliana</i>	1,846	5.77%
<i>Populus trichocarpa</i>	1,161	3.63%
<i>Oryza sativa</i>	643	2.01%
<i>Nicotiana tabacum</i>	641	2.00%
<i>Solanum tuberosum</i>	428	1.34%
<i>Solanum lycopersicum</i>	392	1.22%
<i>Medicago truncatula</i>	149	0.47%
<i>Catharanthus roseus</i>	115	0.36%
<i>Glycine max</i>	104	0.32%
Others	1,941	6.06%
No hits	10,732	31.66%
<i>Coffea canephora</i>		
Species	# Hits	% Hits
<i>Vitis vinifera</i>	7,427	44.57%
<i>Arabidopsis thaliana</i>	972	5.83%
<i>Populus trichocarpa</i>	639	3.83%
<i>Oryza sativa</i>	372	2.23%
<i>Nicotiana tabacum</i>	362	2.17%
<i>Solanum tuberosum</i>	232	1.39%
<i>Solanum lycopersicum</i>	225	1.35%
<i>Medicago truncatula</i>	105	0.63%
<i>Solanum demissum</i>	64	0.37%
<i>Catharanthus roseus</i>	56	0.32%
Others	1,231	7.39%
No hits	4,980	29.88%

* Each coffee cluster was compared to all of the proteins from the organisms listed. The BLASTX score was defined as $1e^{-10}$.

To gain insight into the molecular evolution of protein coding genes in the two *Coffea* species analyzed, we estimated the rates of synonymous (KS, silent mutation) and non-synonymous (KA, amino-acid altering mutation) substitutions generated by QualitySNP analysis, and performed the KA/KS test for positive selection of each hypothetical gene. KA/KS is a good indicator of selective pressure at the sequence level. Theoretically, a KA/KS >1 indicates that the rate of evolution is higher than the neutral rate. Conversely, a gene with KA/KS <1 has a rate of evolution less than the neutral rate [36]. As in other plant species [37, 38], most genes in *C. arabica* and *C. canephora* appear to be under purifying selection (KA/KS <1), indicating that the majority of protein-coding genes are conserved over time as a result of selection against deleterious variants.

The correlation between AutoFACT annotations with KA/KS analysis allowed the detection of genes with low KA/KS ratios, such as those encoding proteins involved in photosynthesis, morphogenetic development and translation (Additional File 4). The majority of these proteins have been shown to be highly conserved and to suffer strong purifying selection [37]. Analyzing the genes with the highest KA/KS, we identified effector proteins and transcription factors related to biotic and abiotic stress and proteins involved in oxidative respiration (Additional File 4). These results are in accordance with previous reports, which show that genes acting in response to stress are often positively selected for diversification due to the competition with the evolving effector proteins of pathogens [37, 39].

Metabolic Pathways

We constructed hypothetical metabolic maps for both *C. arabica* and *C. canephora* using BioCyc [40]. After manual annotation, 345 pathways in *C. arabica* and 300 pathways in *C. canephora* were detected. *C. arabica* pathways included 3,366 enzymes in 1,807 enzymatic reactions. In the case of *C. canephora*, 1,889 enzymes were present in 1,653 enzymatic reactions. The almost two-fold difference in the number of enzymes between the two coffee species is related to the number of ESTs annotated for each species. Therefore, assigning the presence/absence of a pathway in one *Coffea* species relative to the other

should be done carefully. Further, the number of *C. arabica* enzymatic reactions may be underestimated due to duplicated genes in *C. arabica*, each one most likely derived from a different ancestor (*C. canephora* and *C. eugenoides*), because that two enzymatic reactions in *C. arabica* may be annotated as only one. The data for the fully annotated pathways are available at the website: <http://www.lge.ibi.unicamp.br/biocyc/cafe>.

Protein Domains

We performed a comparison of *C. arabica* and *C. canephora* gene clusters with the CDD-PFAM databank to catalog the protein domains present in the *Coffea* EST datasets. The submission of the clusters to RPS-BLAST resulted in 30% (9,886) of *C. arabica* and 32% (5,478) of *C. canephora* clusters containing an assigned domain. To compare the prevalence of protein domains in *Coffea* species, the number of clusters assigned to each domain was normalized by dividing by the total number of clusters containing a domain. Serine threonine kinases (Pfam00069), cytochrome P450 monooxygenases (Pfam00067), tyrosine kinases (Pfam07714) and proteins containing RNA recognition motifs (RRM; Pfam00076) are among the top 20 PFAM families in *Coffea* species (Additional File 5). Next, we plotted the percentage of protein domains in *Coffea* datasets in a comparative histogram. Protein domain analysis revealed significant differences between the two species datasets (Figure 4). For example, *C. arabica* contains more cytochrome P450 monooxygenases, tyrosine kinases, extensin-like proteins, glycine-rich proteins, sugar transporters, UDP glucosyl- transferases, NAD-dependent epimerases, DNA-J proteins, NB-ARC proteins, cellulose synthases, raffinose synthases, D-mannose-binding lectins and flavin amine oxidoreductases than *C. canephora* (Figure 4). In contrast, the *C. canephora* dataset contains a higher percentage of transcripts coding for proteins containing RRM motifs, ubiquitin conjugation enzymes, ABC transporters, Ras/Rab/Rac proteins, 2-OG oxygenases, cupin proteins, HSP20s, HSP70s, ADP-ribosylation factors, dehydrins, glutenins and seed maturation proteins (Figure 4). Despite these dissimilarities between datasets may be caused by the different tissues used for constructing the *C. arabica* and *C. canephora* cDNA libraries, such results offer clues for further comparative research.

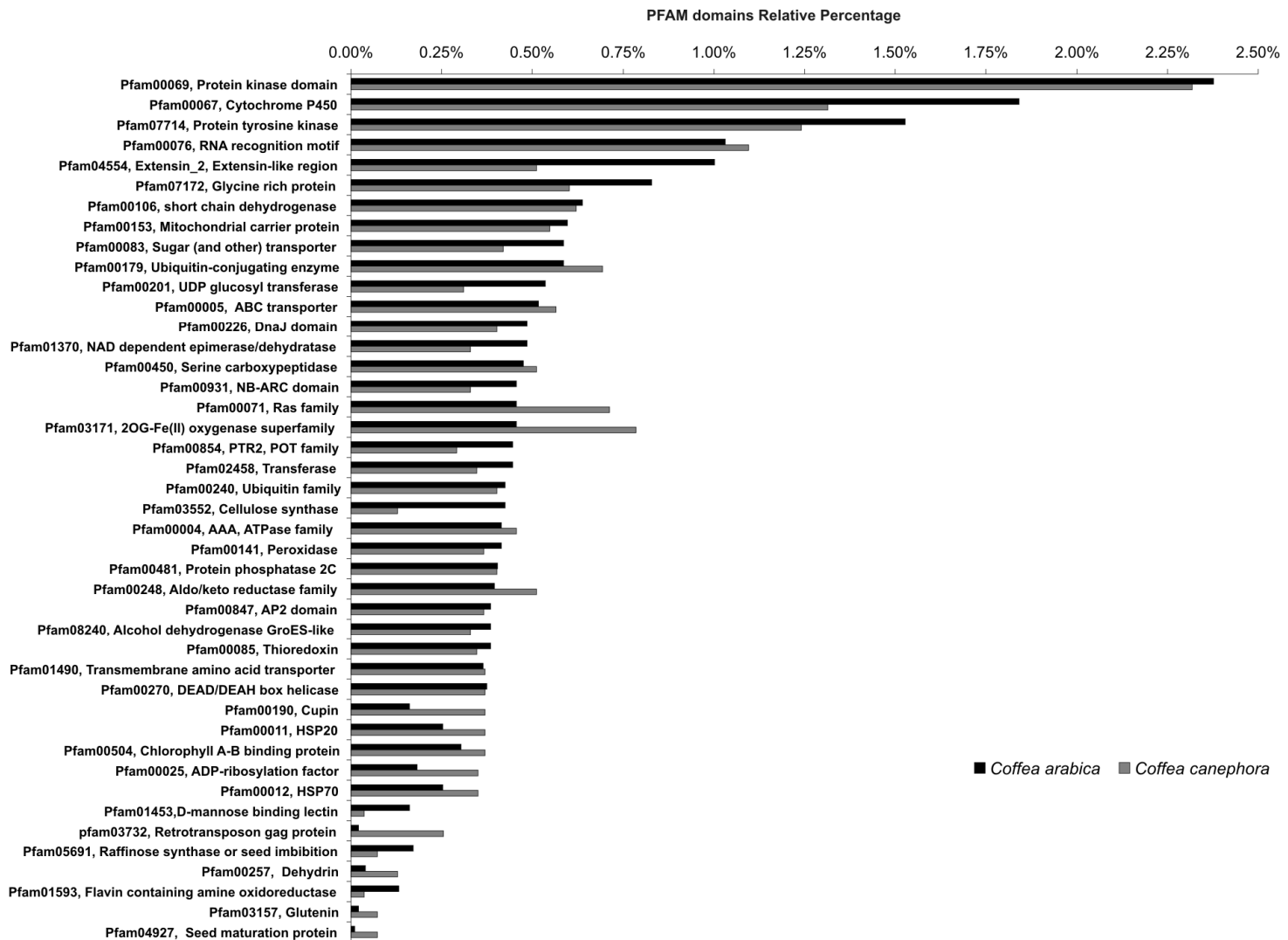


Figure 4. Comparative chart between the relative percentage of Pfam domains in *C. arabica* and *C. canephora* EST databases.

One noteworthy difference between domains is the greater percentage of proteins containing the retrotransposon gag protein domain (Pfam03732) in *C. canephora* (0.26%) than in *C. arabica* (0.02%). This domain is found in LTR-retrotransposons, the most widespread transposable element (TE) family in plants [41]. Lopes et al. [42] found that *Coffea* species harbor fewer TE-cassettes (> 0.04%) than would be expected from the translation of TE-containing transcripts (0.23%). These authors hypothesized that such

incongruence may either be a consequence of the exonization/exaptation of TE fragments or an indication of the tolerance of alternatively spliced “TE-invaded” mRNAs that do not encode functional proteins. A more detailed investigation is in progress to explore the diversity and differences between *Coffea* spp. TEs (F.R. Lopes, M.F. Carazzolle, G.A.G. Pereira, C.A. Colombo, C.M.A. Carareto; unpublished data).

Gene Ontology Analysis and Annotation

A functional annotation was performed by mapping contigs assembling onto gene ontology (GO) structures [43]. Approximately 38% of *C. arabica* and 49% of *C. canephora* clusters were mapped with a biological process, and 43 and 55% were mapped with a molecular function. These differences reflect the greater amount of *C. arabica* ESTs in the libraries compared to *C. canephora* and are likely related to the fact that some tissues used in *C. arabica* libraries (i.e., callus) were not extensively studied, resulting in genes with unassigned ontologies. To compare the gene ontologies, the amount of sequences associated with each term was normalized (see methods), and then hypergeometric statistics were applied [44]. To compare GO data with our other protein-related analysis, we focused our evaluation on molecular activity ontology. We observed that *C. arabica* has a greater amount of transcripts coding for proteins with catalytic activity, transferase activity and transporter activity than *C. canephora* (Figure 5). In accordance, the CDD-PFAM analyses showed that *C. arabica* had a greater percentage of cellulose synthases, raffinose synthases, UDP-glucuronosyl transferases, secondary metabolism-related transferases, ABC transporters and sugar transporters (Figure4; Additional File 5). The evidence that transcripts coding for proteins related to sugar metabolism and transport are more prevalent in *C. arabica* than in *C. canephora* may be related to the high content of sugars (especially sucrose) in fruits of Arabica plants, one of the traits that provides a better cup quality (see below). In contrast to *C. arabica*, *C. canephora* has more proteins annotated as containing binding activity, which is extended for the binding activity branch child terms of nucleic acid binding, DNA and RNA binding activities, transcription regulation and transcription factor activities (Figure 5). These data are also in agreement with our domain analysis (Figure 4;

Additional File 5), indicating a higher percentage of Ras/Rac/Rab GTPase proteins, including regulators of vesicle biogenesis in intracellular traffic, ADP-ribosylation factors and proteins containing RRM and G-patch motifs, involved in RNA binding activity [45].

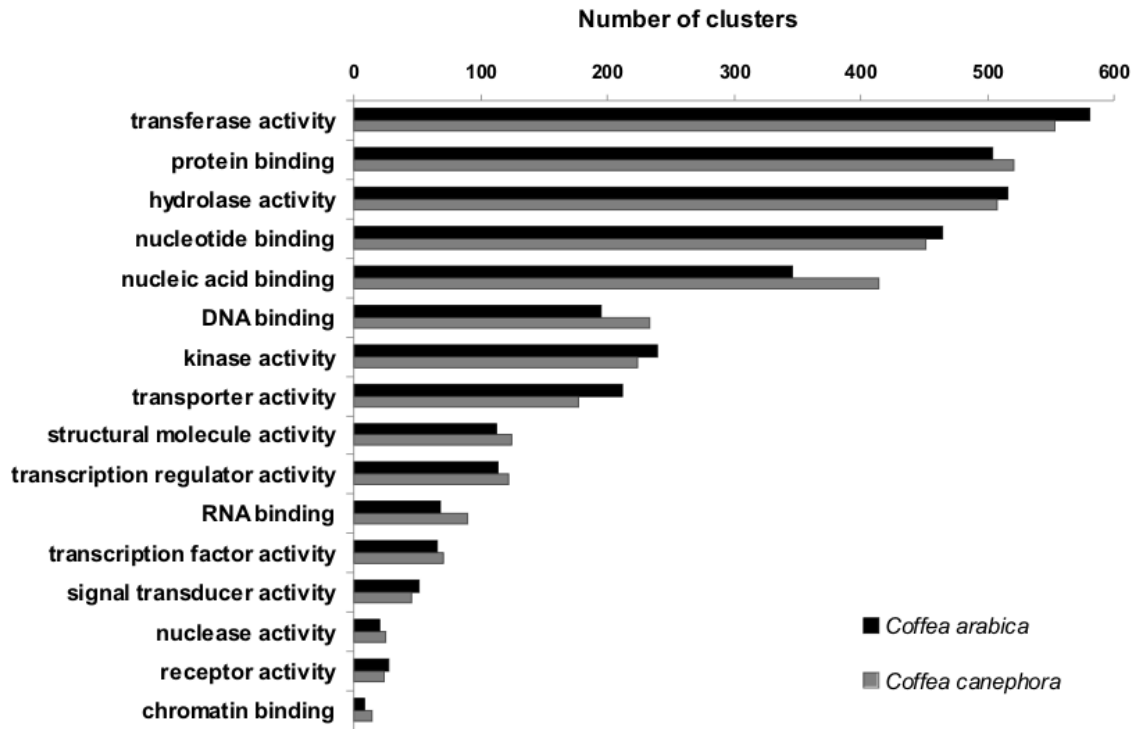


Figure 5. Distribution of *C. arabica* and *C. canephora* clusters with putative functions assigned through annotation using molecular function gene ontology.

Orthologous Family Clustering: Searching for Coffee-Specific Families

To identify proteins that are hypothetically specific or at least prominent in *Coffea* spp. in comparison to other species, we applied OrthoMCL, a graph-clustering algorithm designed to identify homologous proteins based on sequence similarity [46, 47]. Two different types of datasets were used in this analysis: i) the annotated proteins from the available complete genomes of *A. thaliana*, *V. vinifera*,

O. sativa, *Ricinus communis* and *Glycine max* and ii) the proteins predicted by FrameDP

software [48] from the available ESTs assemblies for *C. arabica*, *C. canephora* and *S. lycopersicum*. Based on the fact that some genes are not picked in EST libraries, the evaluation of *Coffea* spp. gene family retraction was not performed (i.e., the absence of a gene does not mean that it is not present in the genome but rather that it is expressed in a minor amount).

We identified 24,577 different families using the eight aforementioned species. The majority of families were ubiquitous, being present in all analyzed species. The top three OrthoMCL families in *Coffea* spp. are: i) a family composed of serine/threonine kinases (family 1), ii) pentatricopeptide repeat-containing proteins (family 2) and iii) cytochrome P450 monooxygenases (family 6; Table 4). The analysis was focused on the annotation of families that appeared to be specific from *Coffea* species or that are prominent in those EST datasets. In *C. arabica*, we highlight family 544, which contains proteins similar to the cysteine proteinase inhibitors cystatins. This family includes 21 members in *C. arabica*, six in *C. canephora* and only one member in the grape genome (Table 4). Two other proteins families composed of cystatin-like proteins (families 2703 and 11594) are also prominent in coffee plants. Other protein families that appear to be prominent/specific in *C. arabica* include small secreted glycine-rich proteins similar to *Panax ginseng* [49] (families 1231, 4031 and 11588), NBS-LRR resistance proteins (families 453, 3289 and 2722), Pin2-like serine proteinase inhibitors (families 7241 and 10273), conserved proteins of unknown function (families 10956, 11617, 12384, 12386, 11626 and 13353), proteins not previously described (no hits; families 14110 and 14413), etc. (Table 4). In *C. canephora*, the “species-specific/prominent” gene families include those encoding miraculin-like proteins (family 14813), *C. canephora*-specific invertase inhibitors (family 14814), small secreted glycine-rich proteins (family 11055), Ty3 Gypsy-like retrotransposons (family 10952), kelch repeat phosphatases (family 14392), 2S albumin storage proteins (family 14392), etc. (Table 4). Five families are specific or prominent in both *C. arabica* and *C. canephora* when compared to the other species analyzed. Two of these contain proteins not previously described (no hits, families 10281 and 12375). The other three include proteins similar to rapid alkalization factor (RALF, family 8498), GTP binding proteins (family 9023) and proline-rich extensins (family 12371; Table 4).

Table 4. OrthoMCL analysis of *C. arabica* and *C. canephora*, highlighting prominent and specific families in *Coffea* spp.

OrthoMCL family ID	<i>Coffea arabica</i>	<i>Coffea canephora</i>	<i>Vitis Vinifera</i>	<i>Solanum lycopersicum</i>	<i>Glycine max</i>	<i>Ricinus communis</i>	<i>Oryza sativa</i>	<i>Arabidopsis thaliana</i>	Manual Annotation*
1	446	189	1402	808	2532	1378	813	847	Serine-threonine kinase
2	152	51	580	212	967	461	478	447	PPR repeat protein
6	84	41	193	123	226	99	101	108	Cytochrome P450
544	21	6	1	-	-	-	-	-	Cystatin
453	14	4	1	7	3	1	1	1	NBS LRR resistance protein
1231	13	5	-	-	-	-	-	-	Small secreted glycine-rich protein
4031	10	-	-	-	-	-	-	-	Glycine-rich protein
1510	7	1	1	-	2	1	1	3	UDP-glucosyltransferase
2703	6	3	-	1	1	-	1	-	Cysteine proteinase inhibitor like protein
3289	6	-	1	-	2	-	2	-	NBS LRR resistance protein
5056	6	1	-	1	-	-	-	-	Alcohol dehydrogenase
2306	5	1	-	2	1	1	2	-	Cytochrome P450
2722	5	1	-	1	1	2	1	1	NBS LRR resistance protein
3294	5	-	1	-	3	-	1	1	Poly-A binding protein
3303	5	1	2	1	-	-	-	1	NADPH-dependent cinnamyl alcohol dehydrogenase
3305	5	2	1	2	-	-	-	-	Specific tissue protein 2
4049	5	2	1	1	-	-	1	-	Sugar transport protein
4070	5	-	1	1	3	-	-	-	Cytochrome P450
7241	5	1	1	-	-	-	1	-	Potato type II serine proteinase inhibitor family
10956	5	-	-	-	-	-	-	-	Hypothetical protein
7610	4	1	-	1	-	-	-	1	Ubiquitin-conjugating enzyme
7611	4	1	-	1	1	-	-	-	P-glycoprotein ABC
7613	4	-	-	2	1	-	-	-	Hexose transporter
9014	4	1	-	-	-	1	-	-	GH3 family protein/Indole-3-acetic acid-amido synthetase
10273	4	1	-	-	-	-	-	-	Potato type II serine proteinase inhibitor family
11588	4	-	-	-	-	-	-	-	Small secreted glycine-rich protein
11617	4	-	-	-	-	-	-	-	Hypothetical protein
12384	4	-	-	-	-	-	-	-	Hypothetical protein
12385	4	-	-	-	-	-	-	-	Defensin/gamma thionin
12386	4	-	-	-	-	-	-	-	Hypothetical protein
7324	3	2	-	-	2	-	-	-	Helix-loop-helix DNA-binding protein
9019	3	-	-	1	-	1	-	-	Zinc/iron transporter
9830	3	-	3	-	-	-	-	-	Eukaryotic initiation

									factor (eIF1)/SU1
10271	3	1	-	-	-	-	1	-	Metallothionein
10276	3	-	-	-	-	1	-	1	SEC14 cytosolic factor family protein
10293	3	-	-	1	1	-	-	-	ABC transporter
10300	3	1	-	-	1	-	-	-	Phytochrome B/histidine kinase
10309	3	1	-	1	-	-	-	-	Oxidoreductase
11058	3	-	1	1	-	-	-	-	ATP-binding cassette transporter
11594	3	-	-	-	-	-	-	1	<i>A. thaliana</i> -related cystatin
11600	3	-	-	-	-	-	1	-	Alcohol dehydrogenase
11607	3	1	-	-	-	-	-	-	CAAX amino-terminal protease
11626	3	1	-	-	-	-	-	-	Hypothetical protein
13353	3	-	-	-	-	-	-	-	Hypothetical protein
13392	3	-	-	-	-	-	-	-	GDP-D-mannose 4,6-dehydratase
14410	3	-	-	-	-	-	-	-	No hits found
14413	3	-	-	-	-	-	-	-	No hits found
14414	3	-	-	-	-	-	-	-	Aspartate aminotransferase superfamily protein
14418	3	-	-	-	-	-	-	-	HAT transposase element
14420	3	-	-	-	-	-	-	-	Protein translation factor SUI1
8498	2	5	-	-	-	-	-	-	Rapid ALKalinization Factor (RALF)-like protein
9023	2	3	-	-	-	1	-	-	GTP binding protein
10281	2	3	-	-	-	-	-	-	No hits found
12371	2	2	-	-	-	-	-	-	Hydroxyproline-rich glycoprotein/extensin
12375	2	2	-	-	-	-	-	-	No hits found
1715	-	4	1	2	1	8	-	-	Viroid polyprotein ORF4 protein
6375	-	4	2	1	1	-	-	-	NBS LRR resistance protein
9679	-	3	1	-	1	1	-	-	Replication factor A 1
10952	-	3	1	-	-	1	-	-	LTR retrotransposon
11055	-	5	-	-	-	-	-	-	Small glycine-rich protein
14392	-	3	-	-	-	-	-	-	Kelch repeat-containing phosphatase
14397	-	3	-	-	-	-	-	-	Albumin/sulfur-rich seed storage protein
14809	-	3	-	-	-	-	-	-	Hypothetical protein
14813	-	3	-	-	-	-	-	-	Miraculin-like protein
14814	-	3	-	-	-	-	-	-	Invertase inhibitor

* Annotation based on BLASTX-NR (E-value $1e^{-5}$).

In silico* Evaluation of Gene Expression in *C. arabica* and *C. canephora

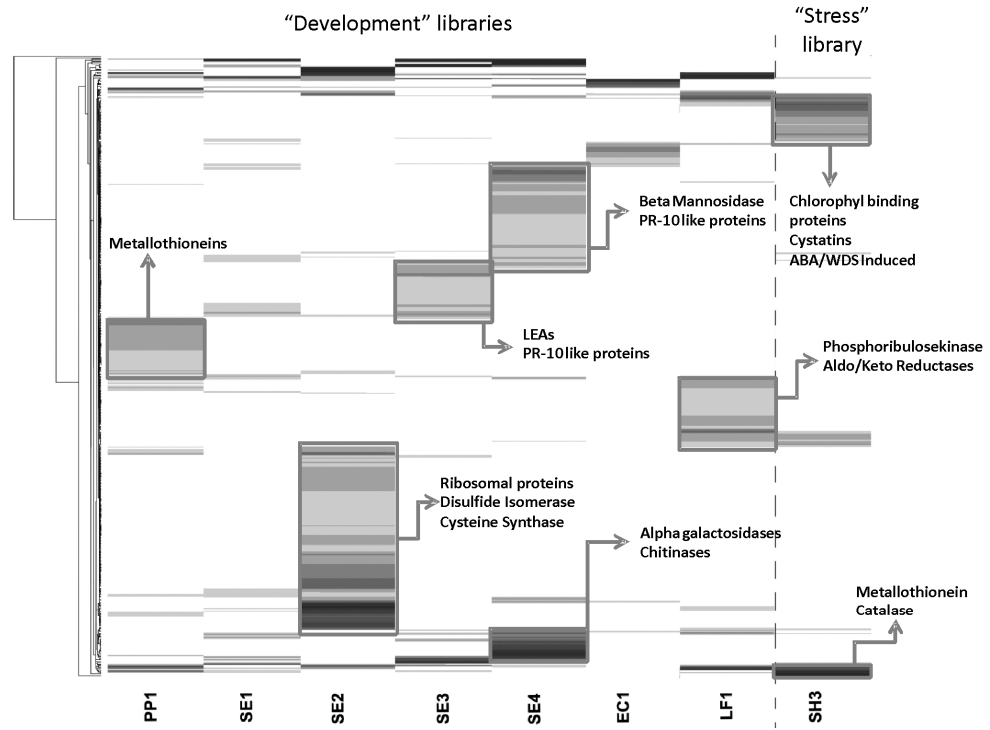
We correlated the AutoFACT annotation results with the distribution of contigs in the *C. arabica* and *C. canephora* libraries (Additional Files 6 and 7). The majority of the most widely distributed genes is related to RNA processing, translation, protein turnover and protein folding. This was an expected result because these biological processes are ubiquitous and indispensable for cellular homeostasis (Additional File 6). In *Arabica*, the most widely expressed contigs encode a papain-like cysteine (cys) proteinase (234 ESTs) and a polyubiquitin (207 ESTs), each one distributed among 30 libraries, followed by glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*; 162 ESTs) and a heme-containing peroxidase (245 ESTs), both distributed among 29 libraries (Additional File 6). Both polyubiquitin and *GAPDH* were previously tested as suitable reference genes for qPCR expression analysis in *C. Arabica* [50-52], which reinforces the accuracy of our bioinformatics analyses. The data presented here provide additional genes to be tested for normalization of qPCR, an essential procedure to avoid misinterpretation when measuring gene expression [53]. The lack of libraries from diverse tissues does not allow reliable inferences about the ubiquity of genes in *C. canephora*. However, the most widely expressed contig (22 ESTs in nine libraries) encodes a putative VTC2 protein, a GDP-D-glucose phosphorylase involved in ascorbic acid biosynthesis [54], suggesting the synthesis of ascorbate throughout fruit development in *C. canephora*.

The evaluation of the contigs distribution in *Coffea* libraries also revealed the contigs containing the most redundant (most highly expressed) ESTs (Additional File 7). In *C. arabica*, a contig encoding a Rubisco small subunit was found to be the most highly expressed gene, followed by a contig encoding a putative class III chitinase (Additional File 7). Among the top 20 most expressed ESTs are genes involved in detoxification and reactive oxygen species (ROS) tolerance and genes related to biotic and abiotic stress. These annotations may be biased by the significant amount of ESTs derived from biotic or

abiotic stressed tissues (Additional File 1). Two genes encoding seed storage proteins (2S albumin and 11S globulin) were the most highly expressed genes in the *C. canephora* dataset, a result similar to that described by Lin et al. [22] (Additional File 7). The use of regulatory elements of these highly expressed genes may be an excellent tool for conferring strong expression to a target gene in transgenesis approaches.

To identify genes uniquely or preferentially expressed in specific coffee EST libraries, R statistics [55] and Audic Claverie (AC) statistics [56] were used through IDEG6, a web tool for the statistical analysis of gene expression data [57]. Libraries containing < 300 ESTs were discarded from these analyses, because libraries with a small amount of ESTs tend to disturb the prediction of differentially expressed genes. After some manual clusterization, we observed that several libraries derived from the same tissues (EA1, IA1 and IA2; EM1 and SI3; LV4, LV5, LV8 and LV9; FB1 and FB4; and FR1 and FR2) present the same set of genes differentially expressed in comparison to the other libraries. Thus, they were combined for further analyses. After evaluating statistical data, the merging of AC and R statistical analyses resulted in 331 contigs from *C. arabica* and 443 contigs from *C. canephora*. Thereafter, hierarchical clustering was applied to this data using a correlation matrix constructed from EST frequencies for differentially expressed *C. arabica* and *C. canephora* contigs (Figure 6; Additional File 8). The clustering results indicated that the differences among *C. canephora* libraries were more evident than in *C. arabica*, likely due to the small number of libraries of the former (Figure 6A and B).

a)



b)

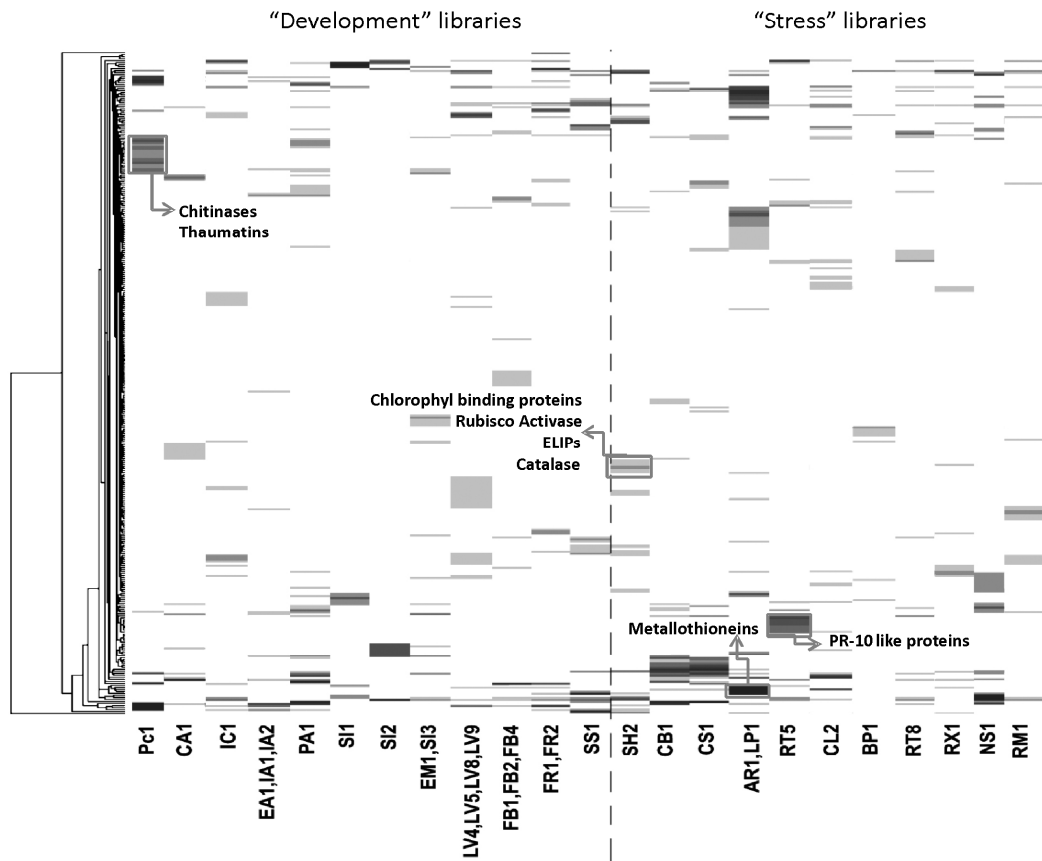


Figure 6. Hierarchical clustering of coffee cDNA libraries and clusters based on EST distribution. a) *C. canephora* hierarchical clustering of 443 clusters differentially expressed vs. the eight cDNA library assemblies. b) *C. arabica* hierarchical clustering of 331 clusters differentially expressed vs. the 23 cDNA library assemblies. Hierarchical clustering was performed using a correlation matrix constructed from EST frequencies for differentially expressed *C. arabica* and *C. canephora* contigs. Black intensity designates relative transcript abundance in a given library, as inferred from EST frequency within each contig. Library abbreviations correspond to the following descriptions:

C. canephora: LF; young leaves, PP1; pericarp, all developmental stages; SE1;_whole cherries, 18 and 22 weeks after pollination; SE2, whole cherries, 18 and 22 weeks after pollination; SE3: endosperm and perisperm, 30 weeks after pollination SE4; endosperm and perisperm, 42 and 46 weeks after pollination; EC1: embryogenic calli; SH1: leaves from water deficit stressed plants; and SH3: leaves from water deficit stressed plants (drought resistant clone).

C. arabica: PC1, *C. arabica*_non-embryogenic cell line induced with 2,4-D; CA1, _non-embryogenic calli; IC1, *C. arabica* non-embryogenic cell line without 2,4-D; EA; EA2, *C. arabica* embryogenic calli; IA2, *C. arabica* embryogenic cell line induced with 2,4-D; PA1, primary embryogenic *C. arabica* calli; EM1, zygotic embryo from mature germinating seeds; SI3, germinating whole seeds; LV4, young leaves from orthotropic branches; LV5, young leaves from orthotropic branches; LV8, mature leaves from plagiotropic branches; LV9, mature leaves from plagiotropic branches; FB1, floral buds at developmental stages 1 and 2; FB2, floral buds at developmental stages 1 and 2; FB4, floral buds at developmental stages 3 and 4; FR1, floral buds, pinhead fruits, fruit developmental stages 1 and 2; FR2, floral buds, pinhead fruits, fruit developmental stages 1 and 2; SS1, well-watered field plant tissues; SH2, water-stressed plant tissues; CB1, suspension cells treated with acibenzolar-S-methyl and brassinosteroids; CS1, suspension cells under osmotic stress; AR1, leaves treated with arachidonic acid; LP1, plantlets treated with arachidonic acid; RT5, roots with acibenzolar-S-methyl; CL2, hypocotyls treated with acibenzolar-S-methyl; BP1, suspension cells treated with acibenzolar-S-methyl; RT8, root suspension cells under aluminum stress; RX1, *Xylella* spp.-infected stems; NS1, nematode-infected roots; and RM1, leaves infected with leaf miner and coffee leaf rust.

The libraries were manually separated into two groups: “development” libraries, derived from tissues that did not suffer stress; and “stress” libraries that were constructed using RNA from plants challenged with biotic or abiotic stress-triggering factors. This expression “fingerprinting” provides a guideline for the isolation of promoters that regulate expression in specific tissues or stress conditions. Brandalise et al. [58] applied a similar strategy in the isolation of a *C. arabica* promoter that drives stress-responsive expression in leaves. Some genes with agronomical importance or with interesting expression profiles depicted in Figure 6 are discussed in more detailed in the following section. The full annotation of differentially expressed genes can be assessed at

www.lge.ibi.unicamp.br/coffeea.

Functional Classification of Differentially Expressed Genes and

Prevalent Protein Families in *C. arabica* and *C. canephora*.

Based on the results of protein domain annotation, GO analysis, OrthoMCL data and Expression Hierarchical Clustering, we established functional categories to elucidate putative gene expression and its consequences in coffee development and environmental adaptations.

Genes related to plant defense

Pathogenesis related proteins (PR)

PRs are a heterogeneous group of plant proteins, inducible by biotic stresses [59, 60]. Some of these proteins are effectors against pathogens and insects, while others are involved in reestablishing homeostasis after the stress [59].

Defensins or gamma-thionins (PR-12) are small, cationic, Cys-rich proteins structurally and functionally related to biocide defensins previously characterized in mammals and insects [61]. All EST reads that compose contigs encoding gamma-thionins from OrthoMCL family 12385 were expressed in tissues treated with benzothiadiazole - BTH (BP1, CL2) or infected with nematodes (NS1). This OrthoMCL family was *C. arabica*-specific (Table 4), perhaps due to the lack of EST libraries from *C. canephora* plants treated with BTH. However, their specificity in Arabica suggests that these proteins rapidly evolved in *Coffea* spp., acquiring specific structural traits important for *Coffea* adaptation to pathogens.

The PR-10 protein family is a large group of PR proteins that are considered allergenic and exert ribonuclease activity, which is paralleled with cytokinin binding and anti-pathogenic roles [62]. In *C. arabica*, a PR-10 was found to be highly expressed in an

incompatible reaction against the causative agent of coffee leaf rust, the biotrophic fungus *Hemileia vastatrix* [63]. A PR-10 from *C. arabica* (CaContig15067) was predicted to be more expressed in suspension cells treated with aluminum (Additional File 8). Concerning *C. canephora*, we observed an expression prevalence of PR-10 genes in late stages of fruit development (SE3 e SE4; Additional File 8). A proteomic analysis indicated that a *C. arabica* PR-10 was expressed only in the endosperm but not in zygotic embryos [64]. This result is similar to that found by Botton et al. [65], who reported the accumulation of a peach PR-10 during the fruit ripening stage.

One interesting result was the presence of a relatively large amount of chitinases (four contigs) and thaumatins (six contigs) in *C. arabica* calli libraries (PC1, EA1, IA1, IA2 and PA1; Additional File 8; Figure 6B). Several reports indicate the participation of these PR proteins not only in plant defense but also during somatic embryogenesis [66-69]. The chitinases are hypothesized to have signaling functions during embryogenesis, because these proteins are able to rescue somatic embryos beyond globular stage [70]. Moreover, arabinogalactan proteins (AGPs), chitinases and thaumatins secreted in suspension-culture cells can promote the production of somatic embryos [69, 71]. Our data strongly indicate a role for these PRs during coffee embryogenesis.

Resistance Genes

Most of the disease resistance genes (R genes) in plants encode nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins. They are engaged in the recognition of pathogens, being considered specific determinants of the plant immune response [72, 73]. Upon annotation of OrthoMCL gene families, we detected 91 clusters and 36 clusters of CC-NBS-LRR proteins in *C. arabica* and *C. canephora*, respectively. In addition, some CC-NBS-LRR families were prevalent in *C. arabica* (Families 453, 3289, 2722) and in *C. canephora* (Family 6375; Table 4). The majority of clusters have higher identity with the PRF protein from tomato (with the exception of CaContig16622, which is more similar to RPP8 and LOV1 proteins). In a seminal report concerning the evaluation of resistance genes in coffee, 43 resistance gene analogues (RGAs) from both *C. arabica* and *C. canephora* were isolated, and it was verified that all RGAs are from the CC-NBS-LRR

subfamily [74]. Nevertheless, we identified a *C. arabica* contig analogous to TIR-NBS-LRR proteins (CaContig7327), with similarity to the nematode resistance potato proteins Gro1 [75] and Arabidopsis TAO1 protein [76]. The extensive retraction (almost disappearance) of *Coffea* spp. TIR-NBS-LRR proteins is similar to that described in cereals and sugar beet [77, 78] and likely resulted from independent gene loss events in such different plant lineages [74, 77, 78]. The implications of the loss of TIR-type NBS-LRR genes and diversification of CC-NBS-LRRs deserve special attention in the understanding of coffee defense mechanisms.

Genes Related to Abiotic Stress and Detoxification

Genes related to abiotic stresses are potentially important in the recent scenario of harsh environmental changes, such as the increase of extreme temperatures and drought periods. Coffee plantations are threatened by global warming due to coffee's susceptibility to high temperatures and drought when these stresses occurs during flowering and fruit development [79]. The understanding of the relationship between tolerance/susceptibility mechanisms and abiotic stress is essential for the prospection of biotechnological and crop management strategies in coffee.

We inspected the genes that were more expressed in *C. arabica* drought stressed plants (SH2) in comparison to well-watered plants conditions (SS1). Genes encoding Rubisco activases (CaContig 5581 and 14729), a putative photosystem II type I chlorophyll a/b-binding (CAB) protein (CaContig5621) and a PSI-E subunit of photosystem I (CaContig5564) were preferentially expressed in the SH2 library (Additional File 8; Figure 6). Cramer et al. [80] also found similar expression patterns with RuBisCo activase and CAB proteins during water and salinity stresses in grapevines. In drought stress, RuBisCo activase augments RuBisCo activity that is diminished as a consequence of a lower stomatal conductance caused by diffusion limitations through stomata and mesophyll [80]. Damages in PSII proteins are associated with the decrease of PSII chemistry caused by ROS [81]. The increase of photosystem I and II genes (CAB and PSI-E subunit) may be a mechanism to sustain photosystems susceptible to ROS attack [80]. These results indicate that the activation

of the photosynthetic apparatus is a mechanism of drought stress mitigation in coffee plants.

Catalase controls H_2O_2 concentrations by dismuting H_2O_2 to water and oxygen. Montavon and Bortlik [82] detected increasing of catalase activity throughout coffee grain maturation. Among genes preferentially expressed in SH2 (Additional File 8; Figure 6A), CaContig13838 has similarity to Arabidopsis catalase 2, which is activated by drought stresses [83], supporting its involvement in the dehydration response in *C. arabica*. Another contig preferentially expressed in the SH2 library (CaContig13998) is similar to early light-induced proteins (ELIPs), thylakoid-target proteins that are similar to light harvesting complex (LHC) proteins (Additional File 8; Figure 6B). ELIPs are reported to be up-regulated during various environmental stresses, such as cold and drought, and during fruit ripening [84, 85]. ABA/WDS are proteins C-terminally enriched in His and Lys and are induced during ripening in pummel [86] and under water deficit stress in loblolly pine [87]. CaContig1691 appears to be one of the most expressed in water deficit stressed plants (Additional File 8; Figure 6B).

Other genes encoding proteins related to drought stress, such as dehydrins, metallothioneins and LEAs, were not differentially expressed in the SH2 library. However, we detected interesting profiles for these genes, especially for dehydrins and LEAs during fruit maturation and for metallothioneins preferentially expressed in libraries from plants treated with arachidonic acid, a polyunsaturated fatty acid present in pathogens (further details in Additional File 9).

Plant Hormones: Auxin Regulation Genes and RALF-like Peptides

Plant hormones (phytohormones) are crucial for a series of developmental mechanisms, such as organ initiation and development, resistance to stress and reproduction. Auxins are the most studied class of phytohormones, being implicated in cell division, cell elongation and cell differentiation [88]. Using OrthoMCL analysis, we identified a family of GH3-like proteins that is expanded in *C. arabica* (Family 9014; Table 4). GH3 enzymes conjugate amino acids to the auxin indole-3-acetic (IAA), decreasing the

concentration of free auxin [89]. This mechanism is important in the regulation of IAA availability in plants. We also detected a family of Aux/IAA proteins that is prominent in *C. arabica* (Family 770; Table 4). Aux/IAA proteins have been shown to function as negative regulators of gene expression mediated by auxin response factor (ARF). A gene similar to auxin receptor TIR1 that promotes ubiquitin (Ub)-mediated degradation of Aux/IAA repressors was identified in *C. arabica* (CaContig 593). In addition, we also detected another putative auxin receptor in *C. arabica*, ABP1 (CaContig16576), a cupin-like protein that is implicated in early auxin responses [90].

Together with small lipophilic “classical phytohormones,” small peptides have been described as factors involved in plant growth regulation [91]. Rapid alkalization factor (RALF) is a small peptide initially isolated in tobacco that induces a rapid alkalization in cell suspension and inhibits root growth in tomato and Arabidopsis seedlings [92]. Based on BLAST searching, we found a family of RALF peptides in *C. arabica* (two members) and *C. canephora* (five members). However, the evaluation of OrthoMCL families revealed that coffee has a particular family of small peptides slightly similar to RALFs (Family 8498; Table 4). These proteins contain the four cysteines in their C-termini required for RALF activity but are richest in Trp. Further, some members do not contain the conserved dibasic site (Additional File 10), which is essential for processing tomato and Arabidopsis RALFs [92-94]. The isolation and functional analysis of these coffee proteins/peptides constitute an important approach in order to verify whether they exert the same growth retarding effect as RALFs.

Glycine-Rich Proteins

The glycine-rich protein (GRP) superfamily is a large complex of plant proteins that share the presence of glycine-rich domains arranged in (Gly) n -X repeats [95]. Generally considered as involved in protein-protein interactions, GRPs have diverse functions and structural domains [96]. Evaluating hierarchical clusterization data, we found that several GRPs are preferentially expressed in suspension cells treated with BTH, brassinosteroids and NaCl, as well as in embryogenic calli (Additional File 8). Those genes encode GRPs from Class I, which may contain a signal peptide for secretion followed by a glycine-rich

region with GGGX repeats [95]. Other GRPs (CaContigs 1089, 3317, 10126) were found to be differentially expressed in plantlets and leaves treated with arachidonic acid (Additional File 8). These genes encode proteins containing signal peptides and are similar to class II GRPs, which contain a peptide motif rich in cysteine and tyrosine residues located in their C-termini [95]. However, a deeper annotation revealed that these coffee GRPs contain 12 cysteines instead of the six cysteines of the aforementioned class II GRPs (Additional File 11). These cysteine-rich domain proteins, such as class II AtGRP-3 and NtTLRP, were shown to interact with receptor protein kinase WAK1 [97] and to mediate the cross-linking of proteins to the cell wall [98]. We also detected the presence of some “specific” GRP OrthoMCL families in coffee (Table 4). Family 1231 is composed of class I GRPs, while family 4011 has GRPs from class II that contain six to 10 cysteines (Additional File 11). The diversification of GRPs in coffee is quite remarkable, especially in Class II and is probably important to coffee cell wall dynamics and signal transduction.

Proteinase Inhibitors (PIs)

The phytocystatins (PhyCys) are 12- to 16-kDa plant proteinaceous inhibitors of Cys-proteases of the papain C1A family [99, 100]. All cystatins contain three motifs involved in the interaction with their target enzymes: the reactive site QxVxG, one or two glycine residues in the N-terminal part of the protein, and an A/PW located downstream of the reactive site. In addition, PhyCys contain a consensus sequence ([LVI]-[AGT]-[RKE]-[FY]-[AS]-[VI]-x-[EDQV]-[HYFQ]-N) that conforms to a predicted secondary-helix structure [99]. Family 544 of hypothetical PhyCys was prevalent in coffee plants, containing 21 members in *C. arabica* and six members in *C. canephora* (Table 4). Proteins from family 544 are 10 kDa, contain a variation of the LARFAV-like domain and do not contain the canonical reactive site QxVxG but have a GG-X-YY motif (Additional File 12). Other OrthoMCL families (2703 and 942) were annotated as containing putative cystatins prevalent in coffee (Table 4; Additional File 12). All members of those three families have low but significant identities (30-40%) with hypothetical cystatins from *Arabidopsis* (At5g47550), grape (XP_002274494.1) and *Brassica oleracea* (ABD64972). Two *C. canephora* members from

those families (CcContigs 7844 and 3825) were highly expressed in leaves from water deficit stressed plants (SH3; Additional File 8; Figure 6A). The majority of these new coffee cystatins do not have signal peptides (Additional File 12), likely being responsible for the regulation of endogenous protein turnover as hypothesized for alfalfa and barley cystatins [101, 102]. In a recent phylogenomic analysis, it was proposed that cystatins had undergone a complex and dynamic evolution through gene losses and duplications [103]. This assignment may explain the expansion of cystatins in coffee and may indicate functional diversification of these proteins.

Members of the Potato type II (PotII) inhibitors (Pin2) family are PIs restricted to plants that belong to the MEROPS inhibitor family I20, clan IA [104]. Several Pin2 proteins have a multi-domain structure. However, sequences from coffee-prevalent proteins of OrthoMCL families 7241 and 10273 appear to be uni-domain Pin2 proteins (Additional File 13). Although we did not find any of the coffee *Pin2* genes preferentially expressed in EST libraries of stressed plants, predicted coffee Pin2 proteins contain signal peptides and, additionally, have 30-40% identity with a Pin2 protein of tobacco that confers tolerance to NaCl and resistance against herbivorous insects in transgenic plants [105]. In addition to the fact that PI expansions may be related to biotic stress regulation, PIs may also have an important role in proteolysis during coffee fruit development because the peptides and amino acids are precursors of coffee flavor and aroma (see below).

Coffee Cup Quality Related Genes

Coffee cup quality is a complex trait that is being unraveled. The components of coffee endosperm are the source of the precursors of aroma and flavor after roasting. The degradation of sucrose and cell wall polysaccharides generate reducing sugars, which react with amino acids during roasting through Maillard glycation reactions. This reaction gives rise to aromatic products, such as pyrazines, furans and aliphatic acids, which are associated with pleasant flavor and aroma [106]. Conversely, the bitterness of coffee is related to caffeine and chlorogenic acid content in coffee beans [107]. During our annotation, we give a panorama of genes related to coffee cup quality that were, by some

means, emphasized in at least one of our bioinformatics analyses.

Genes Related to Carbohydrate Metabolism

Due to the importance of the amount and composition of carbohydrates to the final quality of the coffee beverage, the study of coffee bean carbohydrate synthesis and degradation is intense [5-7, 108-112]. Coffee bean cell walls are mainly made of galactomannans, arabinogalactans and cellulose [108]. One interesting finding in our analysis was the prevalence of cellulose synthase superfamily proteins (pfam 03552; CesA) in *C. arabica* in relation to *C. canephora* (Figure 5, Additional File 5). CesA proteins interact in a cellulose synthase complex, and it is believed that each cell type contains three types of CesA subunits in a single complex [113]. Therefore, the broader origin of *C. arabica* ESTs may be the reason for the prevalence of *C. arabica* CesAs in comparison to *C. canephora*. The CesA family includes the “true” cellulose synthase genes and eight other families named ‘cellulose synthase-like’ genes *CsIA-CsIH* [114]. It was verified that some *CsIA* proteins act in the synthesis of mannans and xyloglucans [112, 115, 116]. The orthologs of these *CsI* genes were found in our *C. arabica* EST data (CaContigs 3405 and 11680).

It is considered that the role of carbohydrates in the differences in cup quality between *C. arabica* and *C. canephora* is related to low molecular weight carbohydrate content, especially sucrose [117]. Arabica grains have a higher amount of sucrose (7.3–11.4%) than *C. canephora* grains (4–5%). Though sucrose is almost completely degraded during coffee bean roasting (0.4–2.8% dry weight), sucrose remains are thought to improve coffee sweetness and cup quality [118]. Privat et al. [6] found that the synthesis of sucrose phosphate synthase (SPS) was higher in late stages of *C. arabica* grains than in *C. canephora*, and invertase activity was lower in Arabica, likely due to the higher expression of invertase inhibitors in this species, justifying the higher sucrose content in *C. arabica* beans. Based on BLAST and OrthoMCL analysis, we found that Invertase Inhibitor 3 (Invl3) is part of a *Coffea* spp.-specific protein family (Family 14814; Table 4). These proteins have 20-30% identity to *Zea mays* invertase inhibitors from the pectin-methylesterase family [6, 119, 120]. We did not detect *C. arabica* ESTs encoding Invl3, likely due to the low coverage

of fruit/seed libraries of this species. The presence of such a particular Inv1 in coffee may indicate new molecular mechanisms of invertase regulation.

The raffinose family oligosaccharides (RFOs) are soluble galactosyl-sucrose carbohydrates such as raffinose, stachyose and verbascose. Their participation in coffee seed development was assessed by Joet et al. [7], who indicated that RFOs were transiently present during the storage phase and remobilized during mid-stages of development to supply the extensive demand for galactose in galactomannan synthesis. Raffinose synthases (RS; EC 2.4.1.82) catalyze the synthesis of raffinose from sucrose and galactinol [121]. Our CDD-PFAM analysis indicated that *C. arabica* has a larger amount of RS than *C. canephora* (Figure 5). Such data seem to corroborate biochemical analyses that showed that grains from *C. canephora* contain reduced raffinose levels in comparison to Arabica [122, 123]. A more careful inspection of RS *C. arabica* clusters revealed that these sequences were derived from diverse tissue libraries. The presence of more EST libraries from stressed plants in *C. arabica* may be the cause of such bias, because RFO accumulation has been associated with responses to abiotic stresses, protecting cellular metabolism from oxidative damage and drought [124, 125]. Indeed, a recent analysis indicated that three *C. arabica* RFO synthase transcripts are induced by drought and saline stress (T.B. Santos, I. G. Budzinski, C.J. Marur, C.L. Petkowicz, L.F. Pereira, L.G. Vieira; unpublished results). Therefore, raffinose may exert dual functions in coffee: galactose reservoirs in coffee grains and protective roles in vegetative development.

It is assumed that the RFOs decrease in late stages of coffee bean development are caused by α -D-Galactosidase (α -Gal; EC 3.2.1.22) activity. We identified three α -Gal-encoding genes as more expressed in the late stages of *C. canephora* seed development (CcContigs 2650, 3171, 7083; Additional File 8; Figure 6A), data that agree with previous findings verifying increased α -Gal activity during *in vitro* germination of coffee beans [126]. Together with α -Gal, β -mannosidases (EC 3.2.1.25) and Endo β -mannanase (EC 3.2.1.78) are enzymes involved in the degradation of galactomannans during germination of seeds. Despite the fine analysis of *C. arabica* β -mannanases and α -Gal [109, 126], there is no biochemical analysis of β -mannosidases activity in coffee of

which we are aware. We found that β -mannosidases are preferentially expressed in germinating seeds of *C. arabica* and *C. canephora* (CaContig 3009, CcContig6678; Figure 6; Additional File 8), a similar pattern in comparison to α -Gal from *C. canephora* (CcContig 6678; Additional File 8).

Amino Acid Content: Storage Protein Synthesis and Protease Expression

As cited above, proteins and amino acids are also fundamental for the generation of flavor and aroma-related Maillard-end products. In effect, the level of protein synthesis during early fruit stages, the amount of seed storage proteins (SSPs) in the endosperm and the relationship between proteinases and their inhibitors during seed development are all factors that determine the amino acid content in mature beans. Examining the expression profile of the SE2 library, we found a series of ribosomal proteins expressed in this stage of seed maturation (Figure 6A; Additional File 8), indicating an intense cellular effort in translation. Many SSPs are enriched in cysteines, which confer high stability to these proteins, an important factor for storage proteins. These cysteines are also a source of sulfur used in seed germination. Two genes involved in cysteine metabolism, protein folding and sulfur metabolism were preferentially expressed in the early stage of *C. canephora* seed maturation (SE2 library; Figure 6A). CcContigs 7827 and 99 encode a cysteine synthase (O-acetylserine (thiol) lyase) (EC 4.2.99.8), an enzyme that synthesizes cysteine [127], and a protein disulfide isomerase (PDI), an enzyme that catalyzes the formation and breakage of disulfide bonds between cysteine residues within proteins as they fold [128], respectively.

In coffee, the Cupin family protein 11S globulin represents 45% of the total protein in the endosperm (corresponding to 5-7% of coffee bean dry weight) [129] and is probably one of the main sources of nitrogen during coffee bean roasting. Our expression hierarchical clustering analysis indicated that two 11S globulin genes were preferentially expressed in *C. arabica* fruit libraries (CaContigs 12252 and 13966; Additional File 8), and one was more highly expressed in the late stages of *C. canephora* seed development (i.e., 42 weeks after pollination) (CcContig 4069; Additional File 8). This contig was the second most abundant in the *C. canephora* database (Additional File 7) after a 2S albumin

(CcContig1385; Additional File 7). We also identified a cysteine and an aspartic protease preferentially expressed in the last phase of Arabica seed maturation (CaContigs 7768 and 8165; Additional File 8). The coincidence of expression profiles of important storage proteins such as 11S globulin and 2S albumin together with proteinases is an indication that the release of free amino acids or small peptides that contribute to coffee cup quality can occur in the final stage of coffee maturation.

Secondary Metabolism: Caffeine, Trigoneline and Chlorogenic Acid

Other precursors of flavor and aroma in coffee are secondary metabolites, such as alkaloids (caffeine and trigoneline) and phenylpropanoid chlorogenic acid (CGA). These three components, together with sucrose, seem to be the main factors influencing coffee quality, because sucrose and trigoneline enhance coffee quality, while CGA and caffeine confer bitter taste [7, 107, 130-133]. The comparison between the two coffee species showed that *C. arabica* has more trigoneline and sucrose, and *C. canephora* contains more CGA and caffeine [131]. Despite intense annotation, our data did not reveal any outstanding results concerning the differential expression of the genes in the metabolic pathways of these compounds during fruit development or any interesting difference between *C. arabica* and *C. canephora* plants.

CONCLUSION

We assembled ESTs from *C. arabica* and *C. canephora* and applied a diverse array of bioinformatics tools to extract information about gene content features, transcriptome changes and novel genes and gene families. The results concerning the prevalence of proteins related to sugar metabolism in *C. arabica* and signal transduction in *C. canephora* can be correlated with agronomical characteristics of each species due to the better cup quality of *C. arabica* and the high tolerance to specific stresses in *C. canephora* plants. Despite knowing that comparisons between these *Coffea* species data should be carefully inspected, our initiative established possible transcriptomic elements that could guide the coffee scientific community in unraveling the molecular mechanisms

that distinguish these two extremely important *Coffea* species. In addition, the annotation of coffee-specific/prominent genes adds new elements to genomic initiatives that are searching for traits that could differentiate coffee from other Asteridae species. In a recent report, Vidal et al. [30] showed that *C. arabica* displays differential expression of homeologous genes and suggested that *C. arabica* ancestral subgenomes encode proteins involved in different physiological mechanisms, adding a new element of investigation concerning gene expression regulation in coffee plants.

All data presented here are available at www.lge.ibi.unicamp.br/coffea. We believe that such data are a valuable aid to the interpretation of coffee development, and provide insights that could help coffee breeding programs and indicate potential targets for functional analysis and biotechnology products of such socially and economically important species.

METHODS

EST assembly and trimming

ESTs from *C. arabica* (187,142) and *C. canephora* (78,470) were derived from 51 libraries collected by the BCGP and from the *C. canephora* EST sequencing initiative of the Nestlé Research Center. The Brazilian project sources were two *C. arabica* genotypes (Catuai and Mundo Novo) and one *C. canephora* genotype (Conillon). The Cornell-Nestlé project EST sources were five different varieties of *C. canephora* [22]. Sequences were trimmed using BDTrimmer to remove ribosomal sequences, polyA/T tails, low quality sequences, vector sequences (UniVec database) and *E. coli* contaminants [134]. EST assembling was executed using the CAP3 program, with a minimum similarity threshold of 90% and a minimum overlap of 40 bases. ESTs from each species were assembled separately, and the genotypes were assembled together into the same species. After the assembly, nucleic acid contamination from bacterial organisms that were not removed during trimming analysis (putative endophytes of coffee) was detected using BLASTN against a version of the NT database containing only bacteria (NT-bac) and BLASTX against the NR database. The results against NT-bac with E-values $> 1e^{-40}$ and the

percent of identical nucleotides > 80% were considered bacterial contamination. In addition, hits against NR with a percent of identity > 30% and all of the hits against bacteria were considered bacterial contamination. All of the BCGP ESTs were submitted to GenBank with accession numbers GT640310-GT640366, GT669291- GT734396, GW427076 - GW492625 (*C. arabica*) and GT645618-GT658452 (*C. canephora*).

Single Nucleotide Polymorphism (SNP) analyses and GC content

QualitySNP [24] was used to analyze polymorphisms present in *C. arabica* and *C. canephora*. QualitySNP uses three quality filters for the identification of reliable SNPs. The first filter screens for all potential SNPs. False SNPs caused by sequencing errors are identified by the chromatogram quality given by Phred. The second filter is the core filter, which uses a haplotype-based strategy to detect reliable SNPs. The clusters with potential paralogs are identified using the differences in SNP number between potential haplotypes of the same contig. All potential haplotypes consisting of only one sequence are removed, and singleton SNPs that are not linked to other polymorphisms are not considered. This may lead to an underestimation of nucleotide diversity but assures that false positives will be discarded. The last filter screens SNPs by calculating a confidence score based on sequence redundancy and base quality. To label each polymorphism as synonymous or non-synonymous, the correct open reading frame (ORF) of each sequence was identified by looking for similarity calculated with the FASTA algorithm against the Uniprot databank (<http://www.uniprot.org>) using an E-value threshold of -05. The alignments were analyzed with QualitySNP script GetnonsySNPfasty, which corrects frame shifts and attempts to expand the 3' end until the next stop codon and the 5' end until the next ATG codon. This script identifies if the polymorphism changes the amino acid, labeling each polymorphism as non-synonymous (KA) or synonymous (KS). This information was used to calculate KA/KS ratios for positive selection using kaks calculator software [135]. All of the ORFs predicted in QualitySNP were used to calculate the GC content of *C. arabica* and *C. canephora*. A total of 1,380 full length sequences > 200 bp of *Arabidopsis thaliana* were extracted from Genbank. Sequences of *Solanum lycopersicum* were also randomly retrieved from the Kazusa (<http://www.kazusa.or.jp/jsol/microtom/indexj.html>) and SGN

databanks [23]. Total GC and GC3 were calculated for each sequence and plotted in a histogram graph with 100 classes, which were smoothed by using the average of each three sets of classes.

Automatic Functional Annotation, Metabolic Pathways and Evaluation of

Protein Domains

The complete set of ESTs from *C. arabica* and *C. canephora* were automatically annotated using the AutoFACT program [31]. AutoFACT summarizes results of BLAST similarity searches against nucleotide, protein and domain databases in functional annotation. The databases used were Uniref100, Uniref90, NCBI-nr, KEGG and CDD (E-value $\leq 1\text{E-}5$). The annotation was submitted to the Pathologic module of the Pathway Tools program (version 13.0) in order to generate metabolic maps. Pathologic module looks at the product name and E.C. number of annotations and imports the pathways likely to be present from the reference database (MetaCyc). The *C. arabica* and *C. canephora* metabolic maps were compared with PlantCyc, which contains curated information about pathways present in > 250 plant species. The divergence among the maps was manually annotated to eliminate false positives. To evaluate protein domains, ESTs were submitted to similarity searches against the CDD-PFAM database using RPS-BLAST (E-value $\leq 1\text{e}^{-10}$). Data were normalized by dividing the number of clusters from each CDD-PFAM by the total number of hits from each species against CDD-PFAM.

Gene Ontology Analyses

Coffee datasets were annotated and mapped for the gene ontologies “Biological Process” and “Molecular Function” (only level 3) by Blast2go [43]. Blast2go lists all gene ontology terms found in biological processes and molecular functions found in each dataset and associates the amount of sequences with each term. These data were normalized to the total number of sequences that were labeled with a gene ontology term. Hypergeometric distribution statistical analysis [44] was applied in the datasets from fruit

and leaf to find the sub- and over-estimated GO terms in each species.

Orthologous Clustering (Ortho-MCL)

The Ortho-MCL algorithm [47] was applied to generate orthologous groupings. Two different datasets were used: i) the annotated proteins from the available complete genomes of *A. thaliana* (27,379 proteins), *O. sativa* (56,797 proteins), *Ricinus communis* (31,221 proteins) and *Glycine max* (66,210 proteins) and ii) the proteins predicted by FrameDP software [48] from the available EST assemblies for *C. arabica* (28,585 predicted proteins), *C. canephora* (16,477 predicted proteins) and *S. lycopersicum* (52,437 predicted proteins). All proteins were compared (all against all) using BLASTP, and a score for each pair of proteins (u,v) with significant BLAST hits was assigned (E-value $1e^{-5}$; with at least 50% of similarity). Based on these scores, the MCL algorithm was applied to find clusters in this graph. The protocol used is described at http://lge.ibi.unicamp.br/Ortho_MCL_UserGuide.txt.

Gene Expression Hierarchical Clustering Analysis

For *in silico* expression analysis, contig and singlet frequencies across the libraries were obtained from the dataset derived from the CAP3 assembly. The frequency of a contig over a library represents its transcript abundance. Only contigs containing more than two ESTs were used for transcript profiling. Differentially expressed contigs were identified using two statistical tests, R [55] and AC [56], with the webtool IDEG6 [57]. In R statistics, a threshold p-value of 0.05 (95% confidence) was used with Bonferroni correction. AC statistics were calculated for pairwise combinations of all libraries. Under this criterion, a contig was considered of significant interest if the AC statistics of at least one library against all of the other libraries were lower than the threshold 0.05. The resulting differentially expressed contigs were obtained with the union of the two sets above. Each library frequency was then normalized by the frequency of the contig.

In an attempt to cluster elements that are similar (in some sense), hierarchical clustering [136] of the differentially expressed contigs was performed using MatlabR2009a (The Mathworks). Hierarchical algorithms attempt to group the differentially expressed contigs based on the expression profile of these contigs in the libraries. The clustering of the rows (contigs) was performed, generating a heat map and a dendrogram. The libraries were manually sorted according to tissue sources and stress conditions, visually creating two libraries groups: “development” libraries and “stress” libraries.

AUTHOR’S CONTRIBUTIONS

Core Manuscript Team

JMCM: ESTs and cluster re-annotations, conception of bioinformatics analyses, evaluation and interpretation of GC content, Ortho-MCL, CDD-PFAM, Gene Ontology and Hierarchical clustering data, conception and writing of the manuscript; ROV: EST assembly, conception of bioinformatics analyses, GC content and SNP analyses and evaluation, Gene Ontology and Ortho-MCL analysis; MFC: EST assembly, conception of bioinformatics analyses and CDD-PFAM analysis; EKT: Hierarchical clustering analysis; LPP: AutoFACT and Metabolic Pathways analysis; GGLC: EST assembly and Ortho-MCL analysis; LFPP: SNP evaluation and revision of the manuscript;

BCGP Team

MPM, OGF, PM: Plant material; AM, CAL, CBMV, CLM, DCM, EAG, ECJ, EGML, ER, ERPA, ETK, EUK, EVSA, HC, HES, HFEAD, JBT, LEAC, LLC, MAM, MAVS, MCO, MFGS, MHSH, MITF, MTSE, MVFC, PA, RH, RLBCO, SMT, SMZM, WJS: EST library construction and sequencing effort; EFF, FRS, JPK, MMCC:

Annotation and bioinformatics;

Coordinators of BCGP

ACA: Coordination of EMBRAPA EST libraries and revision of the manuscript; CAC: Coordination of AEG/FAPESP EST libraries and revision of the manuscript; LGEV: Coordination of the EST consortium and revision of the manuscript; GAGP: coordination of the bioinformatics group and elaboration of the final manuscript.

ACKNOWLEDGEMENTS

We especially thank all sequencing and annotation technician teams for their excellent work and support, and Paulo José Teixeira (LGE, UNICAMP) for comments about the manuscript. ROV obtained a PhD fellowship from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; 2007/51031-2). MFC and GGLC received TT-4 Information Technology fellowships from the Applied and Environmental Genomes (AEG) initiative from FAPESP. This work was sponsored by Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café (CBP&D Café),

Empresa Brasileira de Pesquisa Agropecuária (Embrapa) and FAPESP.

REFERENCES

1. Pay E: **The market for organic and fair-trade coffee**. *FAO Rome* 2009.
2. Charrier A, Berthaud J: **Botanical classification of coffee**. In: *Coffee: botany, biochemistry, and production of beans and beverage*. Edited by Clifford MN, Wilsson KC. New York; 1985: 13-47.
3. Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A: **Molecular characterisation and origin of the *Coffea arabica* L. genome**. *Mol Gen Genet* 1999, **261**(2):259-266.
4. Anthony F, Combes C, Astorga C, Bertrand B, Graziosi G, Lashermes P: **The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers**. *Theor Appl Genet* 2002, **104**(5):894-900.
5. Geromel C, Ferreira LP, Guerreiro SM, Cavalari AA, Pot D, Pereira LF, Leroy T, Vieira LG, Mazzafera P, Marraccini P: **Biochemical and genomic analysis of sucrose metabolism during coffee (*Coffea arabica*) fruit development**. *J Exp Bot* 2006, **57**(12):3243-3258.
6. Privat I, Foucrier S, Prins A, Epalle T, Eychenne M, Kandalaft L, Caillet V, Lin C,

- Tanksley S, Foyer C *et al*: **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* 2008, **178**(4):781-797.
7. Joet T, Laffargue A, Salmona J, Doulebeau S, Descroix F, Bertrand B, de Kochko A, Dussert S: **Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study.** *New Phytol* 2009, **182**(1):146-162.
 8. Leroy T, Marraccini P, Dufour M, Montagnon C, Lashermes P, Sabau X, Ferreira LP, Jourdan I, Pot D, Andrade AC *et al*: **Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes.** *Theor Appl Genet* 2005, **111**(6):1032-1041.
 9. Wintgens JN: **Coffee: growing, processing, sustainable production.** Weinheim; 2004.
 10. Maluf MP, Silvestrini M, Ruggiero LMD, Guerreiro O, Colombo CA: **Genetic diversity of cultivated *Coffea arabica* inbred lines assessed by RAPD, AFLP and SSR marker systems.** *Scientia Agricola* 2005, **62**(4):366-373.
 11. Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL *et al*: **Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane.** *Genome Res* 2003, **13**(12):2725-2735.
 12. da Silva FG, Iandolino A, Al-Kayal F, Bohlmann MC, Cushman MA, Lim H, Ergul A, Figueroa R, Kabuloglu EK, Osborne C *et al*: **Characterizing the grape transcriptome. Analysis of expressed sequence tags from multiple *Vitis* species and development of a compendium of gene expression during berry development.** *Plant Physiol* 2005, **139**(2):574-597.
 13. Verza NC, TR ES, Neto GC, Nogueira FT, Fisch PH, de Rosa VE, Jr., Rebello MM, Vettore AL, da Silva FR, Arruda P: **Endosperm-preferred expression of maize genes as revealed by transcriptome-wide analysis of expressed sequence tags.** *Plant Mol Biol* 2005, **59**(2):363-374.
 14. Ramirez M, Graham MA, Blanco-Lopez L, Silvente S, Medrano-Soto A, Blair MW, Hernandez G, Vance CP, Lara M: **Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics.** *Plant Physiol* 2005, **137**(4):1211-1227.
 15. Sakurai T, Plata G, Rodriguez-Zapata F, Seki M, Salcedo A, Toyoda A, Ishiwata A, Tohme J, Sakaki Y, Shinozaki K *et al*: **Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response.** *BMC Plant Biol* 2007, **7**:66.
 16. Argout X, Fouet O, Wincker P, Gramacho K, Legavre T, Sabau X, Risterucci AM, Da Silva C, Cascardo J, Allegre M *et al*: **Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao* L. generated from various tissues and under various conditions.** *BMC Genomics* 2008, **9**:512.
 17. Alexandrov NN, Brover VV, Freidin S, Troukhan ME, Tatarinova TV, Zhang H, Swaller TJ, Lu YP, Bouck J, Flavell RB *et al*: **Insights into corn genes derived from large-scale cDNA sequencing.** *Plant Mol Biol* 2009, **69**(1-2):179-194.
 18. Marques MC, Alonso-Cantabrana H, Forment J, Arribas R, Alamar S, Conejero

- V, Perez-Amador MA: **A new set of ESTs and cDNA clones from full-length and normalized libraries for gene discovery and functional characterization in citrus.** *BMC Genomics* 2009, **10**:428.
19. Leroy T, Henry AM, Royer M, Altosaar I, Frutos R, Duris D, Philippe R: **Genetically modified coffee plants expressing the *Bacillus thuringiensis* cry1Ac gene for resistance to leaf miner.** *Plant Cell Rep* 2000, **19**(4):382-389.
 20. Ogita S, Uefuji H, Yamaguchi Y, Koizumi N, Sano H: **Producing decaffeinated coffee plants.** *Nature* 2003, **423**(6942):823.
 21. Vieira LGE, Andrade AC, Colombo CA, Moraes AHA, Metha A, Oliveira AC, Labate CA, Marino CL, Monteiro-Vitorello CB, Monte DC *et al*: **Brazilian coffee genome project: an EST-based genomic resource.** *Brazil J Plant Physiol* 2006, **18**:95-108.
 22. Lin C, Mueller LA, Mc Carthy J, Crouzillat D, Petiard V, Tanksley SD: **Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts.** *Theor Appl Genet* 2005, **112**(1):114-130.
 23. Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y *et al*: **The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond.** *Plant Physiol* 2005, **138**(3):1310-1317.
 24. Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA: **QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species.** *BMC Bioinformatics* 2006, **7**:438.
 25. Carels N, Bernardi G: **Two classes of genes in plants.** *Genetics* 2000, **154**(4):1819-1825.
 26. Glemin S, Bazin E, Charlesworth D: **Impact of mating systems on patterns of sequence polymorphism in flowering plants.** *Proc Biol Sci* 2006, **273**(1604):3011-3019.
 27. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M: **Milling SNPs from EST databases.** *Genome Res* 1999, **9**(2):167-174.
 28. Batley J, Hayes PK: **Development of high throughput single nucleotide polymorphism genotyping for the analysis of *Nodularia* (Cyanobacteria) population genetics.** *J Phycol* 2003, **39**(1):248-252.
 29. Dantec LL, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio JM, Chaumeil P, Leger P, Garcia V *et al*: **Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences.** *Plant Mol Biol* 2004, **54**(3):461-470.
 30. Vidal RO, Mondego JM, Pot D, Ambrosio AB, Andrade AC, Pereira LF, Colombo CA, Vieira LG, Carazzolle MF, Pereira GA: **A high-throughput data mining of SNPs in *Coffea* spp ESTs suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*.** *Plant Physiol* 2010, DOI **10.1104/pp.110.162438**.
 31. Koski LB, Gray MW, Lang BF, Burger G: **AutoFACT: an automatic functional annotation and classification tool.** *BMC Bioinformatics* 2005, **6**:151.
 32. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral**

- hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**(7161):463-467.
33. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D *et al*: **Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids.** *Plant Physiol* 2008, **148**(4):1772-1781.
 34. Guyot R, de la Mare M, Viader V, Hamon P, Coriton O, Bustamante-Porras J, Poncet V, Campa C, Hamon S, de Kochko A: **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* 2009, **9**:22.
 35. Cenci A, Combes MC, Lashermes P: **Comparative sequence analyses indicate that *Coffea* (Asterids) and *Vitis* (Rosids) derive from the same paleo-hexaploid ancestral genome.** *Mol Genet Genomics* 2010, **283**(5):493-501.
 36. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends Ecol Evol* 2000, **15**(12):496-503.
 37. Roth C, Liberles DA: **A systematic search for positive selection in higher plants (Embryophytes).** *BMC Plant Biol* 2006, **6**:12.
 38. Novaes E, Drost DR, Farmerie WG, Pappas GJ, Jr., Grattapaglia D, Sederoff RR, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome.** *BMC Genomics* 2008, **9**:312.
 39. Stukenbrock EH, McDonald BA: **Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions.** *Mol Plant Microbe Interact* 2009, **22**(4):371-380.
 40. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N: **Expansion of the BioCyc collection of pathway/genome databases to 160 genomes.** *Nucleic Acids Res* 2005, **33**(19):6083-6089.
 41. Vitte C, Panaud O: **LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model.** *Cytogenet Genome Res* 2005, **110**(1-4):91-107.
 42. Lopes FR, Carazzolle MF, Pereira GA, Colombo CA, Carareto CM: **Transposable elements in *Coffea* (Gentianales: Rubiaceae) transcripts and their role in the origin of protein diversity in flowering plants.** *Mol Genet Genomics* 2008, **279**(4):385-401.
 43. Conesa A, Gotz S: **Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics.** *Int J Plant Genomics* 2008, **2008**:619832.
 44. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: **GOToolBox: functional analysis of gene datasets based on Gene Ontology.** *Genome Biol* 2004, **5**(12):R101.
 45. Frenal K, Callebaut I, Wecker K, Prochnicka-Chalufour A, Dendouga N, Zinn-Justin S, Delepierre M, Tomavo S, Wolff N: **Structural and functional characterization of the TgDRE multidomain protein, a DNA repair enzyme from *Toxoplasma gondii*.** *Biochemistry* 2006, **45**(15):4867-4874.
 46. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9):2178-2189.
 47. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS: **OrthoMCL-DB: querying a**

- comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34**(Database issue):D363-368.
48. Gouzy J, Carrere S, Schiex T: **FrameDP: sensitive peptide detection on noisy matured sequences.** *Bioinformatics* 2009, **25**(5):670-671.
 49. Luo ZY, Lu QH, Liu SP, Chen XH, Luo JQ, Tan LJ, Hu WX: **Screening and identification of novel genes involved in biosynthesis of ginsenoside in *Panax ginseng* plant.** *Acta biochim biophys Sinica* 2003, **35**(6):554-560.
 50. Salmons J, Dussert S, Descroix F, de Kochko A, Bertrand B, Joet T: **Deciphering transcriptional networks that govern *Coffea arabica* seed development using combined cDNA array and real-time RT-PCR approaches.** *Plant Mol Biol* 2008, **66**(1-2):105-124.
 51. Barsalobres-Cavallari CF, Severino FE, Maluf MP, Maia IG: **Identification of suitable internal control genes for expression studies in *Coffea arabica* under different experimental conditions.** *BMC Mol Biol* 2009, **10**:1.
 52. Cruz F, Kalaoun S, Nobile P, Colombo C, Almeida J, Barros LMG, Romano E, Grossi-de-Sa MF, Vaslin M, Alves-Ferreira M: **Evaluation of coffee reference genes for relative expression studies by quantitative real-time RT-PCR.** *Mol Breed* 2009, **23**(4):607-616.
 53. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F: **Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.** *Genome Biol* 2002, **3**(7):RESEARCH0034.
 54. Linster CL, Gomez TA, Christensen KC, Adler LN, Young BD, Brenner C, Clarke SG: **Arabidopsis VTC2 encodes a GDP-L-galactose phosphorylase, the last unknown enzyme in the Smirnoff-Wheeler pathway to ascorbic acid in plants.** *J Biol Chem* 2007, **282**(26):18879-18885.
 55. Stekel DJ, Git Y, Falciani F: **The comparison of gene expression from multiple cDNA libraries.** *Genome Res* 2000, **10**(12):2055-2061.
 56. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7**(10):986-995.
 57. Romualdi C, Bortoluzzi S, D'Alessi F, Danieli GA: **IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments.** *Physiol Genomics* 2003, **12**(2):159-162.
 58. Brandalise M, Severino FE, Maluf MP, Maia IG: **The promoter of a gene encoding an isoflavone reductase-like protein in coffee (*Coffea arabica*) drives a stress-responsive expression in leaves.** *Plant Cell Rep* 2009, **28**(11):1699-1708.
 59. van Loon LC, Rep M, Pieterse CM: **Significance of inducible defense-related proteins in infected plants.** *Annu Rev Phytopathol* 2006, **44**:135-162.
 60. Sels J, Mathys J, De Coninck BM, Cammue BP, De Bolle MF: **Plant pathogenesis-related (PR) proteins: a focus on PR peptides.** *Plant Physiol Biochem* 2008, **46**(11):941-950.
 61. Carvalho Ade O, Gomes VM: **Plant defensins-prospects for the biological functions and biotechnological properties.** *Peptides* 2009, **30**(5):1007-1020.
 62. Zubini P, Zambelli B, Musiani F, Ciurli S, Bertolini P, Baraldi E: **The RNA hydrolysis and the cytokinin binding activities of PR-10 proteins are differently performed by two isoforms of the Pru p 1 peach major allergen**

- and are possibly functionally related. *Plant Physiol* 2009, **150**(3):1235-1247.
63. Ramiro DA, Escoute J, Petitot AS, Nicole M, Maluf MP, Fernandez D: **Biphasic haustorial differentiation of coffee rust (*Hemileia vastatrix* race II) associated with defence responses in resistant and susceptible coffee cultivars.** *Plant Pathology* 2009, **58**(5):944-955.
 64. Koshino LL, Gomes CP, Silva LP, Eira MT, Bloch C, Jr., Franco OL, Mehta A: **Comparative proteomical analysis of zygotic embryo and endosperm from *Coffea arabica* seeds.** *J Agric Food Chem* 2008, **56**(22):10922-10926.
 65. Botton A, Andreotti C, Costa G, Ramina A: **Peach (*Prunus persica* L. Batsch) allergen-encoding genes are developmentally regulated and affected by fruit load and light radiation.** *J Agric Food Chem* 2009, **57**(2):724-734.
 66. Helleboid S, Hendriks T, Bauw G, Inze D, Vasseur J, Hilbert JL: **Three major somatic embryogenesis related proteins in *Cichorium* identified as PR proteins.** *J Exp Bot* 2000, **51**(348):1189-1200.
 67. Yasuda H, Nakajima M, Ito T, Ohwada T, Masuda H: **Partial characterization of genes whose transcripts accumulate preferentially in cell clusters at the earliest stage of carrot somatic embryogenesis.** *Plant Mol Biol* 2001, **45**(6):705-712.
 68. Rojas-Herrera R, Loyola-Vargas VM: **Induction of a class III acidic chitinase in foliar explants of *Coffea arabica* L. during somatic embryogenesis and wounding.** *Plant Sci* 2002, **163**(4):705-711.
 69. Borderies G, le Behec M, Rossignol M, Lafitte C, Le Deunff E, Beckert M, Dumas C, Elisabeth MR: **Characterization of proteins secreted during maize microspore culture: arabinogalactan proteins (AGPs) stimulate embryo development.** *Eur J Cell Biol* 2004, **83**(5):205-212.
 70. Kragh KM, Hendriks T, de Jong AJ, Lo Schiavo F, Bucherna N, Hojrup P, Mikkelsen JD, de Vries SC: **Characterization of chitinases able to rescue somatic embryos of the temperature-sensitive carrot variant ts 11.** *Plant Mol Biol* 1996, **31**(3):631-645.
 71. van Hengel AJ, Guzzo F, van Kammen A, de Vries SC: **Expression pattern of the carrot EP3 endochitinase genes in suspension cultures and in developing seeds.** *Plant Physiol* 1998, **117**(1):43-53.
 72. Belkhadir Y, Subramaniam R, Dangl JL: **Plant disease resistance protein signaling: NBS-LRR proteins and their partners.** *Curr Opin Plant Biol* 2004, **7**(4):391-399.
 73. McHale L, Tan X, Koehl P, Michelmore RW: **Plant NBS-LRR proteins: adaptable guards.** *Genome Biol* 2006, **7**(4):212.
 74. Noir S, Combes MC, Anthony F, Lashermes P: **Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (*Coffea* L.).** *Mol Genet Genomics* 2001, **265**(4):654-662.
 75. Paal J, Henselewski H, Muth J, Meksem K, Menendez CM, Salamini F, Ballvora A, Gebhardt C: **Molecular cloning of the potato Gro1-4 gene conferring resistance to pathotype Ro1 of the root cyst nematode *Globodera rostochiensis*, based on a candidate gene approach.** *Plant J* 2004, **38**(2):285-297.
 76. Eitas TK, Nimchuk ZL, Dangl JL: **Arabidopsis TAO1 is a TIR-NB-LRR protein that contributes to disease resistance induced by the *Pseudomonas***

- syringae* effector AvrB. *Proc Natl Acad Sci USA* 2008, **105**(17):6475-6480.**
77. Pan Q, Wendel J, Fluhr R: **Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J Mol Evol* 2000, **50**(3):203-213.**
 78. Tian Y, Fan L, Thureau T, Jung C, Cai D: **The absence of TIR-type resistance gene analogues in the sugar beet (*Beta vulgaris* L.) genome. *J Mol Evol* 2004, **58**(1):40-53.**
 79. DaMatta FM, Ramalho JDC: **Impacts of drought and temperature stress on coffee physiology and production: a review. *Brazil J Plant Physiol* 2006, **18**:55-81.**
 80. Cramer GR, Ergul A, Grimplet J, Tillett RL, Tattersall EA, Bohlman MC, Vincent D, Sonderegger J, Evans J, Osborne C *et al*: **Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Funct Integr Genomics* 2007, **7**(2):111-134.**
 81. Lu CM, Zhang JH: **Effects of water stress on photosystem II photochemistry and its thermostability in wheat plants. *J Exp Bot* 1999, **50**(336):1199-1206.**
 82. Montavon P, Bortlik K: **Evolution of robusta green coffee redox enzymatic activities with maturation. *J Agric Food Chem* 2004,**52**(11):3590-3594.**
 83. Du YY, Wang PC, Chen J, Song CP: **Comprehensive functional analysis of the catalase gene family in *Arabidopsis thaliana*. *J Integr Plant Biol* 2008, **50**(10):1318-1326.**
 84. Montane MH, Petzold B, Kloppstech K: **Formation of early-light-inducible-protein complexes and status of xanthophyll levels under high light and cold stress in barley (*Hordeum vulgare* L.). *Planta* 1999, **208**(4):519-527.**
 85. Bruno AK, Wetzal CM: **The early light-inducible protein (ELIP) gene is expressed during the chloroplast-to-chromoplast transition in ripening tomato fruit. *J Exp Bot* 2004, **55**(408):2541-2548.**
 86. Canel C, Bailey-Serres JN, Roose ML: **Pummelo fruit transcript homologous to ripening-induced genes. *Plant Physiol* 1995,**108**(3):1323-1324.**
 87. Padmanabhan V, Dias DM, Newton RJ: **Expression analysis of a gene family in loblolly pine (*Pinus taeda* L.) induced by water deficit stress. *Plant Mol Biol* 1997, **35**(6):801-807.**
 88. Teale WD, Paponov IA, Palme K: **Auxin in action: signalling, transport and the control of plant growth and development. *Nat Rev Mol Cell Biol* 2006, **7**(11):847-859.**
 89. Staswick PE, Serban B, Rowe M, Tiriyaki I, Maldonado MT, Maldonado MC, Suza W: **Characterization of an *Arabidopsis* enzyme family that conjugates amino acids to indole-3-acetic acid. *Plant Cell* 2005, **17**(2):616-627.**
 90. Dahlke RI, Luethen H, Steffens B: **ABP1: An auxin receptor for fast responses at the plasma membrane. *Plant Signal Behav* 2010, **5**(1).**
 91. Matsubayashi Y, Sakagami Y: **Peptide hormones in plants. *Annu Rev Plant Biol* 2006, **57**:649-674.**
 92. Pearce G, Moura DS, Stratmann J, Ryan CA, Jr.: **RALF, a 5-kDa**

- ubiquitous polypeptide in plants, arrests root growth and development.** *Proc Natl Acad Sci USA* 2001, **98**(22):12843-12847.
93. Matos JL, Fiori CS, Silva-Filho MC, Moura DS: **A conserved dibasic site is essential for correct processing of the peptide hormone AtRALF1 in *Arabidopsis thaliana*.** *FEBS Lett* 2008, **582**(23-24):3343-3347.
 94. Srivastava R, Liu JX, Guo H, Yin Y, Howell SH: **Regulation and processing of a plant peptide hormone, AtRALF23, in *Arabidopsis*.** *Plant J* 2009, **59**(6):930-939.
 95. Fusaro AF, Sachetto-Martins G: **Blooming time for plant glycine-rich proteins.** *Plant Signal Behav* 2007, **2**(5):386-387.
 96. Mangeon A, Junqueira RM, Sachetto-Martins G: **Functional diversity of the plant glycine-rich proteins superfamily.** *Plant Signal Behav* 2010, **5**(2): 99-104.
 97. Park AR, Cho SK, Yun UJ, Jin MY, Lee SH, Sachetto-Martins G, Park OK: **Interaction of the *Arabidopsis* receptor protein kinase Wak1 with a glycine-rich protein, AtGRP-3.** *J Biol Chem* 2001, **276**(28):26688-26693.
 98. Domingo C, Sauri A, Mansilla E, Conejero V, Vera P: **Identification of a novel peptide motif that mediates cross-linking of proteins to cell walls.** *Plant J* 1999, **20**(5):563-570.
 99. Margis R, Reis EM, Villeret V: **Structural and phylogenetic relationships among plant and animal cystatins.** *Arch Biochem Biophys* 1998, **359**(1):24-30.
 100. Martinez M, Diaz I: **The origin and evolution of plant cystatins and their target cysteine proteinases indicate a complex functional relationship.** *BMC Evol Biol* 2008, **8**:198.
 101. Rivard D, Girard C, Anguenot R, Vezina LP, Trepanier S, Michaud D: **MsCYS1, a developmentally-regulated cystatin from alfalfa.** *Plant Physiol Biochem* 2007, **45**(6-7):508-514.
 102. Martinez M, Cambra I, Carrillo L, Diaz-Mendoza M, Diaz I: **Characterization of the entire cystatin gene family in barley and their target cathepsin L-like cysteine-proteases, partners in the hordein mobilization during seed germination.** *Plant Physiol* 2009, **151**(3):1531-1545.
 103. Kordis D, Turk V: **Phylogenomic analysis of the cystatin superfamily in eukaryotes and prokaryotes.** *BMC Evol Biol* 2009, **9**:266.
 104. Rawlings ND, Barrett AJ, Bateman A: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2010, **38**(Database issue):D227-233.
 105. Srinivasan T, Kumar KR, Kirti PB: **Constitutive expression of a trypsin protease inhibitor confers multiple stress tolerance in transgenic tobacco.** *Plant Cell Physiol* 2009, **50**(3):541-553.
 106. De Maria CAB, Trugo LC, Aquino Neto FR, Moreira RFA, Alviano CS: **Composition of green coffee water-soluble fractions and identification of volatiles formed during roasting.** *Food Chem* 1996, **55**:203-207.
 107. Leloup V, Louvrier A, Liardon R: **Degradation mechanisms of chlorogenic acids during roasting.** *Proc Internat Congr ASIC* 1995, **16**:192-198.
 108. Fischer M, Reimann S, Trovato V, Redgwell RJ: **Polysaccharides of green *Arabica* and *Robusta* coffee beans.** *Carbohydr Res* 2001, **330**(1):93-101.
 109. Marraccini P, Rogers WJ, Allard C, Andre ML, Caillet V, Lacoste N, Lausanne F, Michaux S: **Molecular and biochemical characterization of endo-beta-**

- mannanases from germinating coffee (*Coffea arabica*) grains. *Planta* 2001, **213**(2):296-308.
110. Redgwell RJ, Trovato V, Curti D, Fischer M: **Effect of roasting on degradation and structural features of polysaccharides in Arabica coffee beans.** *Carbohydr Res* 2002, **337**(5):421-431.
 111. Kasai N, Konishi A, Iwai K, Maeda G: **Efficient digestion and structural characteristics of cell walls of coffee beans.** *J Agric Food Chem* 2006, **54**(17):6336-6342.
 112. Pre M, Caillet V, Sobilo J, McCarthy J: **Characterization and expression analysis of genes directing galactomannan synthesis in coffee.** *Ann Bot* 2008, **102**(2):207-220.
 113. Somerville C: **Cellulose synthesis in higher plants.** *Annu Rev Cell Dev Biol* 2006, **22**:53-78.
 114. Richmond TA, Somerville CR: **The cellulose synthase superfamily.** *Plant Physiol* 2000, **124**(2):495-498.
 115. Liepman AH, Nairn CJ, Willats WG, Sorensen I, Roberts AW, Keegstra K: **Functional genomic analysis supports conservation of function among cellulose synthase-like a gene family members and suggests diverse roles of mannans in plants.** *Plant Physiol* 2007, **143**(4):1881-1893.
 116. Cocuron JC, Lerouxel O, Drakakaki G, Alonso AP, Liepman AH, Keegstra K, Raikhel N, Wilkerson CG: **A gene from the cellulose synthase-like C family encodes a beta-1,4 glucan synthase.** *Proc Natl Acad Sci USA* 2007, **104**(20):8550-8555.
 117. Arya M, Rao LJ: **An impression of coffee carbohydrates.** *Crit Rev Food Sci Nutr* 2007, **47**(1):51-67.
 118. Chahan Y, Jordon A, Badoud R, Lindinger W: **From the green bean to the cup of coffee: investing coffee roasting by on-line monitoring of volatiles.** *Eur Food Res Technol* 2002, **214**:92-104.
 119. Helentjaris T, Bate NJ, Allen SM: **Novel invertase inhibitors and methods of use.** In., vol. Patent WO/2001/058939. USA; 2001.
 120. Bate NJ, Niu X, Wang Y, Reimann KS, Helentjaris TG: **An invertase inhibitor from maize localizes to the embryo surrounding region during early kernel development.** *Plant Physiol* 2004, **134**(1):246-254.
 121. Lehle L, Tanner W: **The function of myo-inositol in the biosynthesis of raffinose. Purification and characterization of galactinol:sucrose 6-galactosyltransferase from *Vicia faba* seeds.** *Eur J Biochem* 1973, **38**(1):103-110.
 122. Rogers WJ, Michaux S, Bastin M, Bucheli P: **Changes to the content of sugars, sugar alcohols, myo-inositol, carboxylic acids and inorganic anions in developing grains from different varieties of Robusta (*Coffea canephora*) and Arabica (*C. arabica*) coffees.** *Plant Sci* 1999, **149**(2):115-123.
 123. Chabrilange N, Dussert S, Engelmann F, Doulebeau S, Hamon S: **Desiccation tolerance in relation to soluble sugar contents in seeds of ten coffee (*Coffea* L.) species.** *Seed Sci Res* 2000, **10**(3):393-396.
 124. Peters S, Mundree SG, Thomson JA, Farrant JM, Keller F: **Protection mechanisms in the resurrection plant *Xerophyta viscosa* (Baker): both**

- sucrose and raffinose family oligosaccharides (RFOs) accumulate in leaves in response to water deficit.** *J Exp Bot* 2007, **58**(8):1947-1956.
125. Nishizawa A, Yabuta Y, Shigeoka S: **Galactinol and raffinose constitute a novel function to protect plants from oxidative damage.** *Plant Physiol* 2008, **147**(3):1251-1263.
 126. Marraccini P, Rogers WJ, Caillet V, Deshayes A, Granato D, Lausanne F, Lechat S, Pridmore D, Petiard V: **Biochemical and molecular characterization of alpha-D-galactosidase from coffee beans.** *Plant Physiol Biochem* 2005, **43**(10-11):909-920.
 127. Gruber CW, Cemazar M, Heras B, Martin JL, Craik DJ: **Protein disulfide isomerase: the structure of oxidative folding.** *Trends Biochem Sci* 2006, **31**(8):455-464.
 128. Alvarez C, Calo L, Romero LC, Garcia I, Gotor C: **An O-acetylserine(thiol)lyase homolog with L-cysteine desulphydrase activity regulates cysteine homeostasis in Arabidopsis.** *Plant Physiol* 2010, **152**(2):656-669.
 129. Marraccini P, Deshayes A, Petiard V, Rogers WJ: **Molecular cloning of the complete 11S seed storage protein gene of *Coffea arabica* and promoter analysis in transgenic tobacco plants.** *Plant Physiol Biochem* 1999, **37**(4):273-282.
 130. Aerts RJ, Baumann TW: **Distribution and Utilization of Chlorogenic Acid in *Coffea* Seedlings.** *J Exp Bot* 1994, **45**(273):497-503.
 131. Ky CL, Louarn J, Dussert S, Guyot B, Hamon S, Noirot M: **Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica* L. and *C. canephora*.** *Food Chem* 2001, **75**:223-230.
 132. Stadler RH, Varga N, Milo C, Schilter B, Vera FA, Welti DH: **Alkylpyridiniums. 2. Isolation and quantification in roasted and ground coffees.** *J Agric Food Chem* 2002, **50**(5):1200-1206.
 133. Mazzafera P: **Catabolism of caffeine in plants and microorganisms.** *Front Biosci* 2004, **9**:1348-1359.
 134. Baudet C, Dias Z: **New EST Trimming Procedure Applied to SUCEST Sequences.** *Advances in Bioinformatics and Computational Biology, Proceedings* 2007:57-68.
 135. Nei M, Gojobori T: **Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions.** *Mol Biol Evol* 1986, **3**(5):418-426.
 136. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**(25):14863-14868.

CAPÍTULO II - Expressão diferencial de homeologos em *Coffea arabica*

Artigo aceito para publicação na
Plant Physiology

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*^{1[W]}

Ramon Oliveira Vidal², Jorge Maurício Costa Mondego², David Pot², Alinne Batista Ambrósio, Alan Carvalho Andrade, Luiz Filipe Protasio Pereira, Carlos Augusto Colombo, Luiz Gonzaga Esteves Vieira, Marcelo Falsarella Carazzolle, and Gonçalo Amarante Guimarães Pereira*

Laboratório de Genômica e Expressão, Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia (R.O.V., A.B.A., M.F.C., G.A.G.P.), and CENAPAD-SP, Centro Nacional de Processamento de Alto Desempenho em São Paulo (M.F.C.), Universidade Estadual de Campinas, CEP 13083–970, Campinas-SP, Brazil; Centro de Recursos Genéticos Vegetais, Instituto Agronômico de Campinas, CEP 13001–970, Campinas-SP, Brazil (J.M.C.M., C.A.C.); Centre de Coopération Internationale en Recherche Agronomique pour le Développement, UMR Développement et Amélioration des Plantes, 34398 Montpellier cedex 5, France (D.P.); Laboratório de Genética Molecular-Núcleo Temático de Biotecnologia, Laboratório de Genética Molecular, Embrapa Recursos Genéticos e Biotecnológicos, Brasília-DF 70770–917, Brazil (A.C.A.); and Embrapa Café, Instituto Agronômico do Paraná (L.F.P.P.), and Instituto Agronômico do Paraná (L.G.E.V.), Laboratório de Biotecnologia Vegetal, CEP 86001–970, Londrina-PR, Brazil

Polyploidization constitutes a common mode of evolution in flowering plants. This event provides the raw material for the divergence of function in homeologous genes, leading to phenotypic novelty that can contribute to the success of polyploids in nature or their selection for use in agriculture. Mounting evidence underlined the existence of homeologous expression biases in polyploid genomes; however, strategies to analyze such transcriptome regulation remained scarce. Important factors regarding homeologous expression biases remain to be explored, such as whether this phenomenon influences specific genes, how paralogs are affected by genome doubling, and what is the importance of the variability of homeologous expression bias to genotype differences. This study reports the expressed sequence tag assembly of the allopolyploid *Coffea arabica* and one of its direct ancestors, *Coffea canephora*. The assembly was used for the discovery of single nucleotide polymorphisms through the identification of high-quality discrepancies in overlapped expressed sequence tags and for gene expression information indirectly estimated by the transcript redundancy. Sequence diversity profiles were evaluated within *C. arabica* (Ca) and *C. canephora* (Cc) and used to deduce the transcript contribution of the *Coffea eugenioides* (Ce) ancestor. The assignment of the *C. arabica* haplotypes to the *C. canephora* (CaCc) or *C. eugenioides* (CaCe) ancestral genomes allowed us to analyze gene expression contributions of each subgenome in *C. arabica*. In silico data were validated by the quantitative polymerase chain reaction and allele-specific combination TaqMAMA-based method. The presence of differential expression of *C. arabica* homeologous genes and its implications in coffee gene expression, ontology, and physiology are discussed.

Coffee (*Coffea* spp.) is one of the most important agricultural commodities, being widely consumed in the entire world. This crop is produced in more than 60

countries and represents a major source of income to many developing nations. Commercial coffee production relies on two main species, *Coffea arabica* (Ca) and *Coffea canephora* (Cc), which are responsible for approximately 70% and 30% of the global crop, respectively. *C. canephora* grows better in lowlands than *C. arabica*. It is also characterized by higher productivity, tolerance to pests and drought stress, and caffeine content. Despite these agronomic advantages, its resulting beverage is considered inferior; therefore, *C. canephora* is consumed mostly in the instant coffee industry and in blends with *C. arabica*.

Cytogenetic analysis established that *C. arabica* is an amphidiploid (allotetraploid; $2n = 4x = 44$) formed by a recent (approximately 1 million years) natural hybridization between the diploids *C. canephora* and *Coffea eugenioides* ($2n = 2x = 22$; Sylva, 1955;

¹ This work was supported by the Fundação de Amparo a Pesquisa do Estado de São Paulo (grant no. 07/51031–2 to R.O.V.), the Consórcio Pesquisa Café, Conselho Nacional de Desenvolvimento Científico e Tecnológico, and the Agronomical and Environmental Genomes/Fundação de Amparo a Pesquisa do Estado de São Paulo (grant no. 00/10154–5).

² These authors contributed equally to the article.

* Corresponding author; e-mail gonçalo@unicamp.br.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Gonçalo Amarante Guimarães Pereira (gonçalo@unicamp.br).

^[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.110.162438

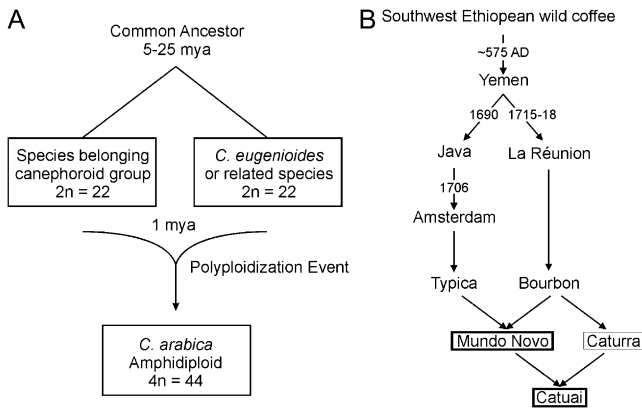


Figure 1. Evolutionary history of allotetraploid *C. arabica*. A, Origin of *C. arabica*. The progenitor genomes are represented by diploid *C. eugenioides* and *C. canephora*. *C. arabica* arose 1 to 2 million years ago (mya) from the fusion of *C. canephora* (or related species) and *C. eugenioides*. B, Origin of cultivated cultivars of *C. arabica* (based on Anthony et al., 2002).

Lashermes et al., 1999; Fig. 1A). *C. eugenioides* is a wild species that grows in higher altitudes near forest edges and produces few berries with small beans of low caffeine content (Maurin et al., 2007).

The narrow diversity observed in *C. arabica* is believed to be a consequence of its reproductive biology, origin, and evolution (Cros et al., 1998; Lashermes et al., 1999; Anthony et al., 2001). In contrast to its ancestors, *C. arabica* is an autogamous species (self-pollinating). Moreover, most commercial *C. arabica* cultivars, including Caturra, Mundo Novo, and Catuai, were selected from only two base populations: Bourbon and Typica (Anthony et al., 2002). The Caturra cultivar is a dwarf mutant of the Bourbon group, whereas Mundo Novo is a hybrid between Bourbon and Typica. The Catuai cultivar resulted from a cross between Mundo Novo and Caturra (Fig. 1B). Each of these three cultivars displays specific plant architecture and physiological properties. *C. arabica* breeding programs have aimed to obtain new cultivars with improved traits, such as flowering time synchronicity, bean size, beverage (cup) quality, caffeine content, resistance to pests, and drought stress tolerance. However, the limited genetic diversity in the base populations has hindered success in those efforts.

Polyploids often display novel phenotypes that are not present or that exceed the range of those found in their diploid ancestors (Osborn et al., 2003). In allopolyploids, some of these traits have been attributed to differential expression of homeologs, which are the orthologous genes from the ancestral species that compose a polyploid (Mochida et al., 2003; Hovav et al., 2008a, 2008b). For example, in the allopolyploids *Triticum aestivum* (hexaploid wheat) and *Gossypium hirsutum* (upland cotton), a subset of the homeologous genes exhibit epigenetic silencing in different tissues or at different developmental stages (Adams et al., 2003; Mochida et al., 2003; Adams, 2007; Liu and

Adams, 2007; Hovav et al., 2008b). This phenomenon, known as partitioned expression or subfunctionalization (Doyle et al., 2008), has the potential to create a transcriptome that is different from the sum of those of the ancestral species, therefore allowing polyploids to occupy new ecological niches or to display traits useful in agriculture (Osborn et al., 2003; Adams and Wendel, 2005).

The detection of variation between the DNA sequences derived from each of the ancestors is essential for the analysis of polyploid genome architecture. The genetic origins and diversity of *C. arabica* have been studied previously through the use of cytogenetics, conventional RFLP, amplified fragment length polymorphism, and microsatellite molecular markers (Lashermes et al., 1999; Steiger et al., 2002; Aggarwal et al., 2007; Cubry et al., 2008; Hendre et al., 2008). The recent availability of high-throughput DNA sequencing data has enabled similar studies based on highly informative single nucleotide polymorphisms (SNPs). SNP analyses using large EST sequence data sets from agricultural crops have been employed for the generation of high-density genetic maps and the identification of variable genomic regions (Du et al., 2003; Choi et al., 2007; Novaes et al., 2008; Pindo et al., 2008; Duran et al., 2009). Furthermore, SNPs present within expressed regions are also useful to identify homeologous genes from ancestral genomes in allopolyploids as well as their relative expression levels (Mochida et al., 2003; Hovav et al., 2008b). This information is essential to understand the novel phenotypes associated with the differential expression of homeologous genes.

Despite increasing amounts of data about the presence of homeologous expression biases in polyploid genomes, some questions remain to be answered. Are there specific gene classes affected by this phenomenon? How are different paralogs affected by genome doubling? Does the variability of homeologous expression bias contribute to the phenotypic differences between cultivars of the same species?

As part of the Brazilian Coffee Genome Project (Vieira et al., 2006), we generated nearly 267,533 ESTs from nonnormalized cDNA libraries of *C. arabica* and *C. canephora* using the Sanger sequencing method. Another initiative resulted in the sequencing of approximately 47,000 ESTs from *C. canephora* (Lin et al., 2005). In this study, we conducted an integrated analysis of these data sets, on the basis of which we assembled sequencing reads and inspected the detected SNPs to identify homeologous genes. We were able to examine the relative contributions of the ancestor species to the *C. arabica* transcriptome, implicating differential homeolog expression mechanisms as a major source of expression plasticity in *C. arabica*.

Among the specific results describe here are (1) the development of in silico strategies for *C. arabica* subgenome detection and differential homeologous gene evaluation, both of which were confirmed by experimental validation; (2) the Gene Ontology (GO) assess-

ment that *C. arabica* may have specific physiological contributions derived from specific ancestors; and (3) the evidence that paralogs display differential expression in *C. arabica*, which seems to be maintained in relation to the subgenome ancestors.

RESULTS

The Pipeline for SNP Discovery

A total of 267,533 coffee ESTs (78,182 from *C. canephora* and 189,351 from *C. arabica*) from 53 libraries (Supplemental Table S1) were analyzed through a pipeline for SNP discovery and annotation (Fig. 2). The *Coffea* libraries were constructed from a variety of tissues and organs (Lin et al., 2005; Vieira et al., 2006), with most ESTs being produced from seeds/berries, leaves, and flowers. A detailed description of the construction of the *C. arabica* cDNA libraries and sources of plant material is presented in Supplemental Table S1.

All sequences were retrieved in FASTA format with Phred software. Prior to assembly, sequencing reads were trimmed to remove vector and ribosomal sequences, poly(A/T) tails, and low-quality sequences, reducing the number of ESTs to 198,986. These se-

quences were then assembled with the CAP3 program using a conservative approach (Wang et al., 2004) to align ESTs and form the consensus; this was done by aligning ESTs that shared at least 100 bp with at least 95% similarity. Using this conservative approach, the homeologs from *C. arabica* and the same alleles from *C. arabica* and *C. canephora* were expected to coalesce into the same contig. The assembly resulted in 62,195 sequences formed by 23,019 contigs and 39,176 singletons. Only the contigs were analyzed further. BLASTN against the nucleotide database of GenBank (NT) was applied to the 23,019 contigs, removing 1,434 possible contaminant contigs (mainly bacterial sequences). In the remaining 21,585 coffee contigs, 64% of the contigs had ESTs from the two species and 85% had EST members from more than one library.

The protocol for SNP discovery was based on QualitySNP software (Tang et al., 2006, 2008). Throughout this paper, we have two different sources for the polymorphisms: (1) the segregating polymorphisms ("real SNPs") and polymorphisms between the subgenomes that we labeled as "sgSNPs" (for subgenome SNPs). As the first polymorphism detection was performed in a "blind" way, it was not possible to define the source of the polymorphism, those being labeled as xSNP. Then, using the sequence information

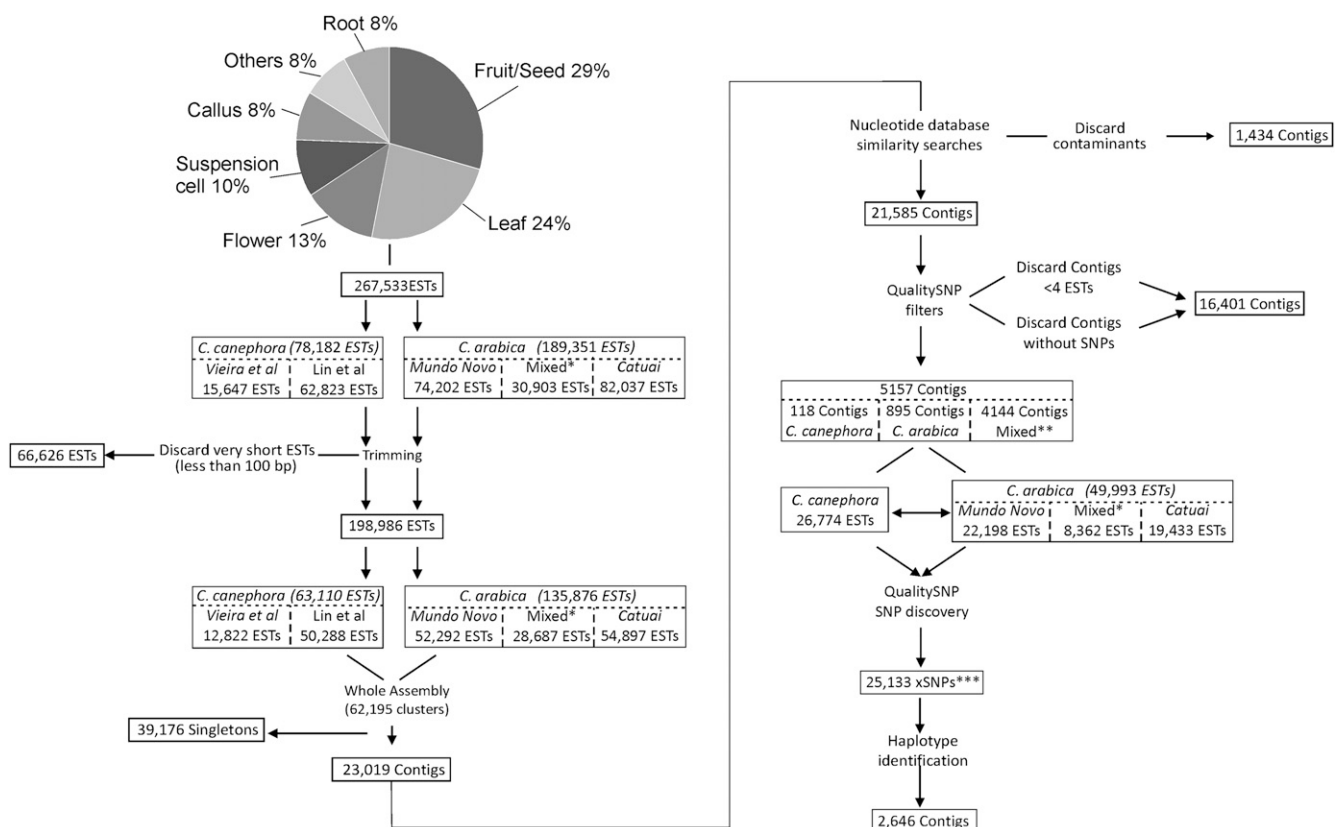


Figure 2. Flow diagram of the pipeline for data cleaning, EST assembly, SNP discovery, and analysis in *Coffea* species. * It is not possible to know the origin of these sequences because the libraries were constructed from cv Catuai and cv Mundo Novo. ** Contigs with ESTs from both *C. arabica* and *C. canephora*. *** xSNPs are sequence polymorphisms of unknown source.

from both *Coffea* species, it was possible to characterize the xSNPs more accurately, making the difference between SNPs (real polymorphisms that are variable between genotypes) and polymorphisms between subgenomes, sgSNPs (for details, see below and Table I).

xSNPs were called only when at least two reads were found in the contigs with the same base for each noncoincident position. Overall, 25,133 xSNPs (0.45 xSNPs per 100 bp) were found in 5,157 contigs. These were composed of 118 contigs (128.5 kb) of *C. canephora*-only ESTs, 895 contigs (895.7 kb) of *C. arabica*-only ESTs, and 4,144 contigs (80%; 4,989.9 kb) of ESTs from both species, corresponding to a total of 6,014 kb of unique sequence. The contigs were formed by 26,774 *C. canephora* and 49,993 *C. arabica* ESTs (22,198 were derived from cv Mundo Novo, 19,433 from cv Catuai, and 8,362 from mixed libraries).

C. arabica Subgenome Identification

We organized the 5,157 contigs in subsets to identify the xSNPs within species. We found 0.1694 SNPs per 100 bp within *C. canephora* and 0.3934 xSNPs per 100 bp within *C. arabica* (Table I). Within the *C. arabica* reads, nearly half of the sequences were highly similar to the *C. canephora* reads. This was consistent with the hypothesis that *C. arabica* is an allotetraploid species formed by an ancestor from the canephoroid group.

In order to assign the *C. arabica* reads to their two ancestral subgenomes (i.e. *C. canephora* and *C. eugenoides* genomes), a haplotype analysis based on the QualitySNP software was performed. Briefly, this analysis allows the identification of haplotypes that correspond to different combinations of alleles from multiple loci. About 80% of contigs with *C. canephora* ESTs had one or two "QualitySNP haplotypes" (Fig. 3A). The analysis of haplotypes in *C. arabica* contigs shows that in most cases two QualitySNP haplotypes per contig were identified (72%; Fig. 3B), a pattern consistent with the fact that this species is an autogamous allotetraploid and with the results presented

above regarding the assignment of the *C. arabica* reads to their subgenomes of origin (one of these haplotypes corresponding to the *C. canephora* ancestor and the other to the *C. eugenoides* ancestor). A smaller number of contigs had only one haplotype (16%) or more than two haplotypes (12%; Fig. 3B). The detection of only one haplotype can reflect a low divergence of these genes between the two subgenomes or specific expression of only one of them. On the other hand, the observation of more than two haplotypes for *C. arabica* reflects the existence of different haplotypes within at least one of the subgenomes.

Contrary to the usual definition of haplotypes, the ones defined by QualitySNP can include more than one real haplotype, as sequences harboring low divergence (similarity higher than 80% considering exclusively the polymorphic sites) will be assigned to the same haplotype. This strategy avoided the separation of reads caused by sequencing artifacts and made it possible for haplotypes with low divergences from *C. arabica* and *C. canephora* to come together as one. Therefore, within one QualitySNP haplotype, it is possible to have more than one real haplotype. According to this haplotype definition strategy, *C. arabica* reads belonging to the same haplotypes as *C. canephora* reads were designated CaCc (i.e. belonging to the subgenome of the canephoroid ancestor in the *C. arabica* genome). As a corollary of this assumption, the reads that did not match this pattern were considered as originating from the second ancestor species, *C. eugenoides*, and were labeled as CaCe (subgenome of the *C. eugenoides*-related ancestor). A schematic representation of this strategy is shown in Figure 4A.

We identified the 2,646 contigs for which the composing reads could be assigned to the corresponding ancestor genome; these contigs contained reads of both species, with at least four reads originating from *C. arabica* and at least two from *C. canephora*. From these 2,646 contigs, 2,069 have at least four reads from one of the subgenomes. Consequently, the analysis of CaCc and CaCe read frequency in each of these 2,069 contigs may reflect the contribution of each homeolo-

Table I. Polymorphism frequency (xSNP per 100 bp) in *Coffea* species and in *C. arabica* subgenomes calculated from 5,157 contigs

Level of Analysis ^a	No. of Contigs Analyzed and (Total Length)	No. of Contigs with xSNPs	No. of xSNPs	No. of xSNPs per 100 bp
Species				
Cc (SNP)	3,544 (4,301 kb)	1,717	4,449	0.1694
Ca (xSNP)	4,113 (4,994 kb)	3,409	14,866	0.3934
Ca subgenomes				
CaCc (SNP)	2,646 (3,396 kb)	113	589	0.0409
CaCe (SNP)	2,646 (3,396 kb)	71	371	0.0249
CaCc × CaCe (sgSNP)	2,646 (3,396 kb)	843	5,507	0.3596

^aDepending on the data set considered, the single nucleotide change detected corresponded to different types. At the species level, the SNP detected in Cc corresponded to SNPs that are polymorphic between genotypes, whereas the xSNPs detected in Ca encompass sgSNPs and SNPs within subgenomes. In the data sets corresponding to the Ca subgenomes, the CaCc and CaCe polymorphisms correspond to SNPs, whereas the CaCc × CaCe polymorphisms correspond to sgSNPs.

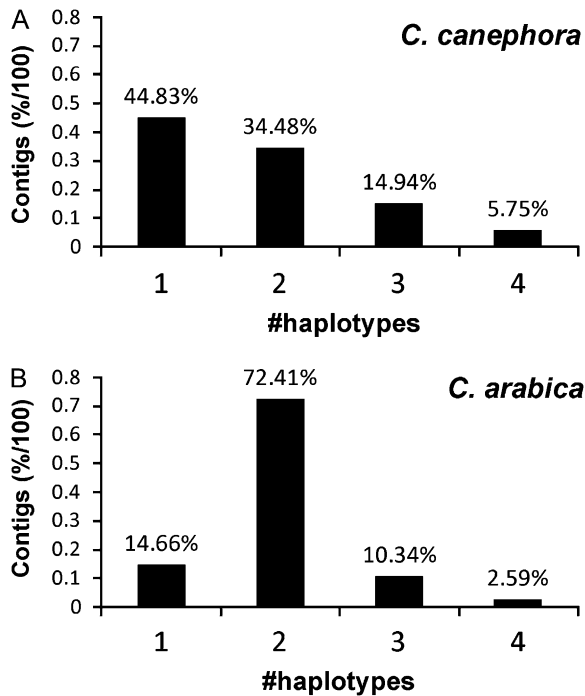


Figure 3. Variability of the number of haplotypes per contig in *C. arabica* (A; only the contigs with at least eight reads were considered) and *C. canephora* (B; only contigs with at least four reads were considered).

gous gene to the *C. arabica* transcriptome (considering the tissues indicated in Fig. 2). Considering a mix of all the tissues analyzed, we estimated that the *C. arabica* transcriptome is composed of roughly equal contributions from the two ancestors (48% of reads from the CaCc subgenome and 52% of reads from the CaCe

subgenome; Fig. 4). However, in a subset of genes, this balance was significantly biased toward one ancestor over the other. For instance, when analyzing the contigs formed by these reads, we see that in some cases these contigs are formed mainly, or only, by reads from one of those subgenomes, which provides evidence for the differential expression of homeologous genes (see below).

To confirm the homeologous gene separation performed using the subtractive method, we used two strategies. First, we sequenced some *C. eugenoides* ESTs and mapped them in the assembly. It was possible to map 18 *C. eugenoides* ESTs in 16 of those 2,646 contigs. *C. eugenoides* reads presented haplotypes consistent with the CaCe subgenome identified (Contig15883, Contig5092, Contig4585, Contig19759, Contig19359, Contig18072, Contig17875, Contig17654, Contig1667, Contig17447, Contig15020, Contig13941, Contig12228, Contig10821, Contig5097, Contig1924), with the exception of two contigs (Contig5097 and Contig1924) at which *C. eugenoides* and *C. canephora* have the same SNP pattern (no divergence between the two ancestral genomes). In addition, sequencing of several gene fragments (6.7 kb) from a small set of genes was performed in *C. eugenoides*. For all the genes analyzed, the *C. eugenoides* sequences clustered together with the CaCe haplotypes. These data confirm the accuracy of the subtractive method of homeologous gene identification.

Polymorphisms in the *C. arabica* Subgenomes

Within the 2,646 contigs in which the composing reads could be assigned to the ancestor genomes, SNPs within the *C. arabica* subgenomes (i.e. between the reads that were assigned to a particular subgenome) were identified (Table I). In CaCc, the frequency

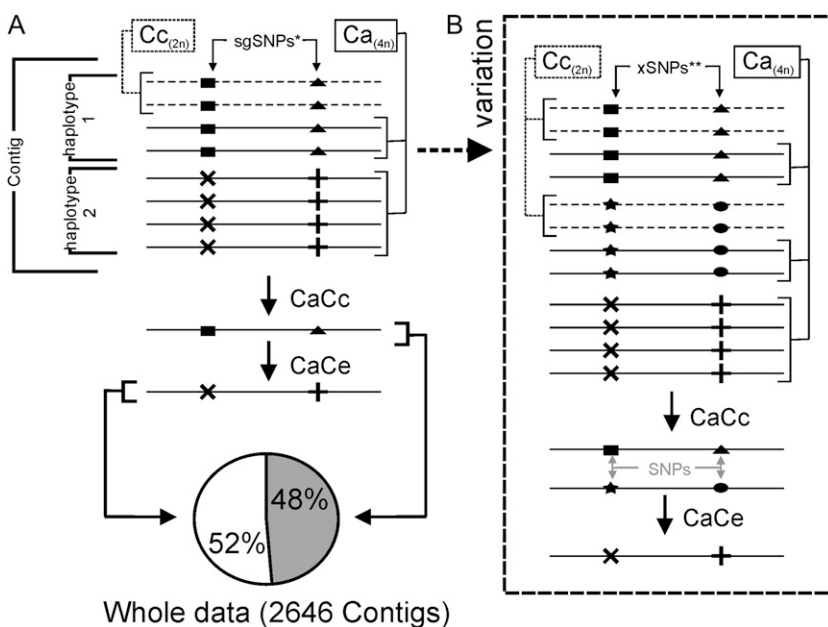


Figure 4. Identification of homeologous genes. A, Scheme showing the assembly of *C. canephora* ESTs (Cc) with *C. arabica* ESTs (Ca) into the same haplotype in the same contig. ESTs from *C. arabica* presenting the same pattern as *C. canephora* were labeled as derived from the CaCc subgenome, and the remaining ESTs were labeled as derived from the CaCe subgenome (for details, see "Materials and Methods"). From all contigs in which *C. arabica* subgenomes were identified, 52% of ESTs from *C. arabica* were transcribed from the CaCe subgenome and 48% from the CaCc subgenome. B, A variation of homeologous gene identification. In some contigs, it was possible to find more than one haplotype for each subgenome.

obtained was 0.0409 SNPs per 100 bp, corresponding to a total of 589 SNPs in 113 contigs. In CaCe, we also found a low SNP frequency (0.0249 SNPs per 100 bp; 371 SNPs), almost similar to that found in CaCc (Table I). The low levels of polymorphism observed within the CaCc and CaCe genomes are consistent with the autogamous reproductive regime of *C. arabica* and with the reduced panel of diversity analyzed in this study (only two genotypes with low genotypic diversity between them). Notably, 589 SNPs detected within the CaCc subgenome coincide with *C. canephora* polymorphisms (Fig 4B; Table I).

The frequency of sgSNPs found by comparison between CaCc and CaCe subgenomes was 0.3596 sgSNPs per 100 bp, a number very close to that calculated for the polymorphism within *C. arabica* (0.3934 xSNPs per 100 bp; Table I). Thus, differences between subgenomes represented the main source of the *C. arabica* single nucleotide changes. According to our limited sample of genotypes analyzed, it appears quite clear that the genetic diversity between genotypes is extremely reduced, whereas the genetic divergence between the subgenomes is quite large.

Differential Homeologous Expression

We then analyzed the total of 2,069 contigs that contained at least four ESTs of one of the subgenomes (Fig. 5); most of those (approximately 78%) had a balanced number of ESTs from each origin. The remaining contigs had a greater than 2-fold excess of ESTs from one ancestor over the other; and the *P* values for those imbalanced contigs were highly significant ($P < 0.005$; Fig. 6). Approximately 10% of contigs had more ESTs from CaCc than CaCe (6% with CaCc only), and approximately 12% had more ESTs from CaCe than CaCc (9% with CaCe only). A representative list of genes displaying this pattern of gene expression regulation is shown in Supplemental Table S2. We interpreted this bias as a result of the differential contribution of homeologs to the pool of tran-

scripts from each of these genes in the analyzed tissues.

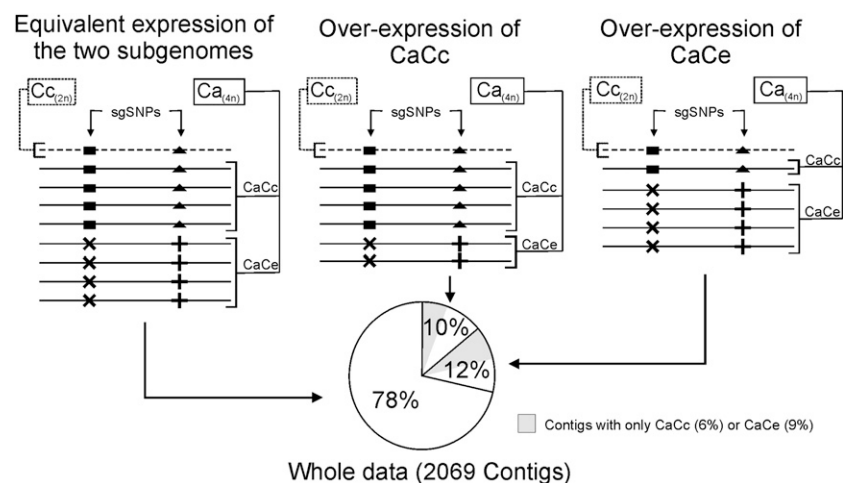
Due to the fact that low coverage contigs tend to push the results toward overestimating equivalent expression among homeologs, we compared the bias of the differential expression of homeologs in four subsets, limiting the minimum coverage (Supplemental Table S3). We observed that in higher coverage contigs there is a greater ability to detect biased expression than in low coverage contigs (Supplemental Table S3). However, we decided to maintain a global selection (low coverage contigs + high coverage contigs) in our analysis, since the assortment of only high coverage contigs would lead to the loss of a significant portion of genes that should be interesting for functional annotation analysis (GO; see below).

The contigs with differential subgenome read frequency were inspected for biological processes (GO; Table II). We observed a tendency of contigs with more CaCe ESTs to encode genes related to photosynthesis, carbohydrate metabolic processes, aerobic respiration, and phosphorylation. In contrast, contigs with a higher CaCc EST content encoded mostly genes related to regulatory processes, such as response to hormone stimuli (mainly auxin), GTP signal transduction, translation, ribosome biogenesis, proteasome activity, and vesicle-mediated transport (Supplemental Table S4). This pattern suggested that *C. arabica* may have specific physiological contributions derived from specific ancestors.

Validating in Silico Homeologous Differential Expression Detected by Quantitative PCR

In order to perform a biological validation of our bioinformatics approach of homeolog identification and inference of differential homeologous expression, we applied a method based on TaqMAMA (Li et al., 2004), which combines the quantitative nature of real-time quantitative PCR (qPCR) with the allele-specific PCR mismatch amplification mutation assay, known

Figure 5. Variability of homeologous gene frequency in the contigs. The left panel shows that in 78% of contigs, the frequency of CaCc and CaCe ESTs was equivalent. The middle and right panels show that in 10% of contigs, the frequency of CaCc was higher than that of CaCe, while in 12% of contigs, the frequency of CaCe was higher than that of CaCc, indicating that *C. arabica* displays partitioning expression of homeologous genes.



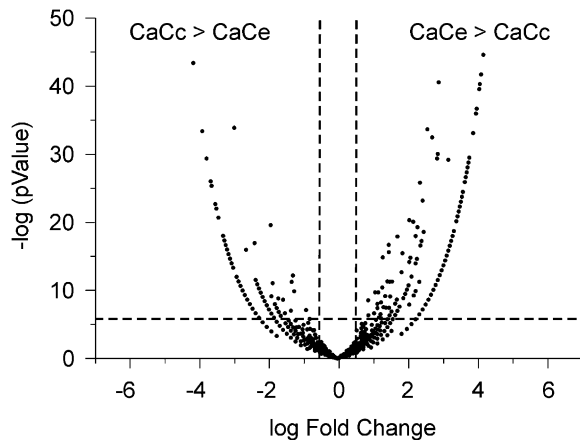


Figure 6. Volcano plot with the 2,645 contigs with CaCc and CaCe ESTs identified. The x axis corresponds to the fold change values calculated according to the following formulas: if the number of CaCe ESTs (#CaCe) is higher than the number of CaCc ESTs (#CaCc), the fold change is $(\#CaCe + 1)/(\#CaCc + 1)$; if #CaCc is greater than #CaCe, the fold change is $-(\#CaCc + 1)/(\#CaCe + 1)$, having negative values. The y axis represents the *P* value (differential expression of the two subgenomes) based on the Audic-Claverie function. Above $-\log 0.005$ (horizontal dashed line), the frequency of one subgenome is significantly higher than the other subgenome. The two vertical dashed lines delimit the area where one subgenome is two times more frequent than the other.

as MAMA (Cha et al., 1992). We chose six genes (Contig21552, Contig11105, Contig10821, Contig10284, Contig17875, Contig18072) that presented high numbers of ESTs from leaves and that presented “higher expression” of one of the *C. arabica* subgenomes (at least two times more reads from one subgenome). This expression ratio was calculated by counting reads from all EST libraries and from only leaf EST libraries (LV4, LV5, LV8, LV9). Leaf was chosen in experimental validation because this was the most representative tissue in EST libraries.

Primers were designed containing the sgSNP in the last 3' nucleotide and a mismatch before it to increase the allele (homeolog) discrimination (Supplemental Table S5). The amplification of the homeologous genes had similar efficiency compared with the reference primers (primers without sgSNP and mismatches that will lead to the amplification of both homeologous genes), indicating that the primer modification did not change the reaction efficiency. All the amplifications were specific, showing allele discrimination, observed by melting curves and by cycle threshold variation between the alleles and the reference reaction (Supplemental Fig. S1). As shown on the amplification plots, the alleles (homeologous genes) tested present differential expression (Supplemental Fig. S1). As expected, the expression of the alleles was lower than the detected expression from the reference primer, which theoretically represents the combination of both alleles in gene expression. Thereafter, we inspected whether these gene expression profiles concurred with in silico

data. From the six contigs tested by the TaqMAMA-based approach, five have similar profiles of homeologous differential gene expression (Table III), which confirms the application of our in silico strategy to analyze homeologous gene expression biases.

Differential Homeologous Expression in Paralogous Genes

We analyzed in detail five distinct paralogous gene sets (homologous genes separated by a duplication event occupying two different positions in the same genome) with their respective homeologous genes. If those proteins contain similar functions (similar results in BLASTX) and have at least 30% identity (BLAST2seq analysis), they were considered paralogs. They were found among the genes with high differential homeologous gene expression (Table IV). The number of reads was not equivalent for the different paralogs in *C. canephora*, indicating that a paralog can be more expressed than another within this species, despite encoding equivalent proteins. For example, for osmotin, whose paralogs have 98% identity at the protein level, there were 45 reads from paralog A and only 17 reads from paralog B. Most relevant for differential homeologous expression, this pattern seems to be maintained among the homeologs of the paralogous genes in the *C. arabica* subgenomes. For example, for paralog A of osmotin, we found 39 reads from CaCc and none from CaCe. For paralog B, there was a complete inversion of this pattern: 21 reads from CaCe and none from CaCc. A similar situation was found for genes FLP (for Frigida-like protein), MLP (for Miraculin-like protein), and SAMDC (for S-adenosyl-Met decarboxylase), all of them presenting high similarity between the paralogs (greater than 65%). However, for Thiazole Biosynthetic Enzyme1 (THI1), which has only 44% similarity between the paralogs, this pattern was not observed: reads from paralog B were more frequent in Cc, whereas reads of paralog A were dominant in CaCc.

We made a further analysis to evaluate the differential expression of these paralogs in *C. arabica* tissues. By counting the reads per tissue composing each contig of the “homeologs-paralogs,” we have found that sometimes one homeolog (i.e. CaCc) is recruited to be expressed in all tissues while the other (i.e. CaCe) is “silenced.” However, when examining the paralogs of genes first analyzed, the homeolog expression is inverted: when the CaCc homeolog is silenced, the CaCe is expressed (Supplemental Fig. S2). This occurs with MLPs (in leaves and bud flowers), FLPs (in leaves), osmotin (in leaves), and SAMDC (in flower buds). In other examples, we found expression of only one “paralog-homeolog” in a specific tissue (osmotin in callus and seed, FLPs in callus and flower buds). We have also found more extreme expression patterns. For instance, in the case of THI1, only one CaCc paralog 1 is expressed in leaves, while CaCe paralog 2 is expressed in seeds. A similar pattern occurs in

Table II. GO of contigs with homeologous genes differentially expressed in the *C. arabica* genome

GO Term	Contigs with High Frequency of CaCc ESTs	Contigs with High Frequency of CaCe ESTs	CaCc ESTs ^a	CaCe ESTs ^a
Translational elongation	4	0	76	10
Signal transduction	8	1	114	7
Auxin-mediated signaling pathway	3	0	50	2
Vesicle-mediated transport	3	0	56	3
Nucleotide biosynthetic process	2	0	26	6
Multicellular organismal process	2	0	8	2
Small GTPase-mediated signal transduction	4	1	60	6
Response to hormone stimulus	4	1	62	7
Biological regulation	14	5	217	31
Ser family amino acid metabolic process	3	1	24	6
Response to auxin stimulus	3	1	50	7
Ribosome biogenesis and assembly	3	1	57	11
Protein catabolic process	5	2	43	14
Homeostatic process	2	1	17	4
Nitrogen compound metabolic process	10	7	137	65
Translation	22	15	286	96
External encapsulating structure organization and biogenesis	4	3	73	18
Organic acid metabolic process	11	13	137	92
Lipid metabolic process	6	7	90	50
Cellular component assembly	2	2	31	21
Biopolymer modification	6	9	111	89
Biogenic amine metabolic process	1	2	28	17
Carbohydrate biosynthetic process	2	5	41	38
Carbon utilization by fixation of carbon dioxide	1	3	13	24
Reductive pentose-phosphate cycle	1	3	13	24
Dicarboxylic acid metabolic process	1	3	9	16
Vitamin metabolic process	1	4	33	27
Photosynthesis, dark reaction	1	4	14	33
Protein import	0	2	1	9
Phosphorylation	1	4	24	30
Secondary metabolic process	0	3	1	19
Cofactor metabolic process	1	7	8	34
Aerobic respiration	0	4	1	17
Coenzyme metabolic process	0	5	3	22

^aNormalized number of ESTs taking into account the total number of ESTs from all contigs used from each data set (CaCe EST more expressed data set and CaCc EST more expressed data set).

SAMDC in roots and suspension cells when compared with seeds (Supplemental Fig. S2).

Diversity in *C. arabica* Cultivars

Analysis of the nucleotide diversity between the two *C. arabica* cultivars (Mundo Novo or Catuai) did not allow the detection of polymorphism between them. Polymorphisms within subgenomes (589 in CaCc and 371 in CaCe; Table I) were not specific to one of the genotypes. In all cases, these polymorphisms were present in both cultivars (data not shown), suggesting the maintenance of a residual subgenome heterozygosity.

DISCUSSION

In this report, we explored EST data sets from *C. arabica* and *C. canephora*, performing an assembly of

sequencing reads and identifying SNPs and sgSNPs throughout these species. We were able to develop an in silico methodology to detect subgenomes inside allotetraploid *C. arabica*. This method helped us to analyze the differential expression of homeologous genes and estimate expression bias according to gene function. We also detected hints about the expression regulation of *C. arabica* paralogs correlated with ancestor origin and variability of expression bias according to *C. arabica* genotypes.

Coffee is an important agricultural commodity and has great economic impact on producing and consuming countries alike. Although *C. arabica* is the main cultivated *Coffea* species (approximately 70%), it has a narrow genetic basis. This low level of diversity is presumably one of the contributing factors to the high susceptibility to pathogens and pests often observed in *C. arabica*. For instance, coffee leaf rust devastated *C. arabica* crops in the 19th century (Staples, 2000).

Table III. Comparison between *in silico* differential expression of homeologous genes and results obtained by qPCR analysis

For *in silico* data, evaluation of the differential expression of homeologous genes was based on a subtractive strategy; for qPCR data, evaluation of the differential expression of homeologous genes was based on the TaqMAMA method. Contig21552, Cys proteinase; Contig11105, histone H3; Contig10821, lipoxygenase; Contig10284, NADPH-protochlorophyllide oxidoreductase; Contig17875, Ala aminotransferase; Contig18072, myo-inositol phosphate synthase; ESTs, total number of ESTs in each contig; ESTsCa, number of *C. arabica* ESTs in each contig; ESTsCc, number of *C. canephora* ESTs in each contig; ESTsCaCc, number of ESTs labeled as derived from the CaCc subgenome; ESTsCaCe, number of ESTs labeled as derived from the CaCe subgenome; CaCc/CaCe, fold change between ESTsCaCc and ESTsCaCe; Leaves CaCc, number of ESTs labeled as derived from the CaCc subgenome expressed in leaves; Leaves CaCe, number of ESTs labeled as derived from the CaCe subgenome expressed in leaves; L-CaCc/L-CaCe, fold change between Leaves CaCc and Leaves CaCe.

Contig	In Silico Data					qPCR Data					
	ESTs	ESTsCa	ESTsCc	ESTsCaCc	ESTsCaCe	CaCc/ CaCe	Leaves CaCc	Leaves CaCe	L-CaCc/ L-CaCe	sgSNP Position	CaCc/ CaCe
Contig21552	58	28	30	26	0	26	6	0	6	377	1.11
Contig11105	55	40	15	38	0	38	17	0	17	247	6.73
Contig10821	56	53	3	10	42	-4.2	10	23	-2.30	1,433	-21
Contig10284	76	60	16	12	48	-4	8	26	-3.25	193	-50
Contig17875	65	39	26	16	23	-1.5	3	12	-4	521	-30
Contig18072	63	41	22	41	0	41	16	0	16	1,297	1.9

C. canephora is one of the main sources of disease resistance genes for *C. arabica* breeding programs, but it produces an inferior cup quality. Therefore, the beverage characteristics of disease-resistant hybrids between *C. canephora* and *C. arabica* can be inferior to that of parental *C. arabica*. This limitation underscores the need for an understanding of the genetic mechanisms underlying the phenotypic variability between *C. arabica* and *C. canephora*, which may support alternative strategies for breeding and guiding selection. Therefore, the findings described here are particularly interesting in low-diversity species such as *C. arabica*.

The cDNA sequences derived from two transcriptomic initiatives (Lin et al., 2005; Vieira et al., 2006) provided us a source for the identification of 25,133 SNPs within *Coffea* EST databases. We describe here a high-throughput evaluation of these SNPs in an EST

assembly based on the allopolyploid species (*C. arabica*) and one of its diploid ancestors (*C. canephora*). The assembly between *C. arabica* and *C. canephora* together with a SNP-based haplotype identification strategy allowed us to analyze the two *C. arabica* subgenomes. *C. arabica* reads presenting the same SNP pattern as *C. canephora* were labeled as derived from *C. canephora* (CaCc), whereas the reads that did not match this pattern were considered as originating from the second ancestor species, *C. eugenioides* (CaCe; Lashermes et al., 1999). Alternatively, a subset of the ESTs considered as CaCe could be *C. arabica* ESTs belonging to the original CaCc subgenome that suffered a rapid nucleotide evolution that led to a high divergence from the original *C. canephora* ancestral genome. Even though such cases may exist, they would not be expected to be present at a frequency that would invalidate our

Table IV. Paralogous genes with expression differences in homeologous genes

Gene	Functional Annotation	Paralog	Identity ^a	ESTsCc	ESTsCa	CaCc	CaCe
1	Osmotin	A	98%	45	39	39	0
		B		17	21	0	21
		Total		62	60	39	21
2	FLP	A	90%	23	32	30	2
		B		3	15	1	14
		Total		26	47	31	16
3	MLP	A	80%	12	56	56	0
		B		3	34	0	34
		Total		15	90	56	34
4	SAMDC	A	65%	13	40	40	0
		B		8	22	0	22
		Total		21	62	40	22
5	THI1	A	44%	5	55	47	8
		B		9	22	1	21
		Total		14	77	48	29

^aProtein identity between the paralogs is as follows: 1A = Contig5325; 1B = Contig12695; 2A = Contig6035; 2B = Contig18336; 3A = Contig11687; 3B = Contig6853; 4A = Contig21736; 4B = Contig164135; 5A = Contig 12496; 5B = Contig21264.

interpretation of the results. As mentioned above, we validated the relevance of the *in silico* methods through an analysis of a small panel of *C. eugenioides* ESTs and resequencing of some *C. eugenioides* genes. These data confirm the efficiency of the *in silico* method and show that the subtractive strategy described here provided an indirect, yet robust, way of identifying the complementary ancestor genome of *C. arabica*.

ESTs were obtained from a mix of two *C. arabica* cultivars and six *C. canephora* genotypes. While *C. arabica* is autogamous, *C. canephora* is allogamous and therefore was expected to display higher levels of nucleotide diversity. Nevertheless, the analysis of polymorphisms showed that *C. arabica* exhibited a higher polymorphism frequency (0.393 xSNPs per 100 bp) than *C. canephora* (0.169 SNPs per 100 bp; Table I), a result consistent with a previous RFLP-based analysis (Lashermes et al., 1999). In that report, the authors observed that *C. arabica* has a level of internal genetic variability roughly twice that present in diploid species with high heterozygosity. To explain this observation, the presence of two subgenomes in *C. arabica* was evoked (Sylvain, 1955; Lashermes et al., 1999). The use of SNPs in our work confirmed this hypothesis by means of a more robust analysis. In this study, we determined that the *C. arabica* polymorphism frequency (0.393 xSNPs per 100 bp) was similar to that found between CaCc and CaCe (0.359 sgSNPs per 100 bp). We also observed that SNP frequency within each *C. arabica* subgenome was around 0.035 SNPs per 100 bp, indicating that the sequence diversity between, and not within, subgenomes is the major source of genetic variability in the most cultivated coffee species. We also found that the few cases of polymorphisms within subgenomes (589 in CaCc and 371 in CaCe) were not specific from one of the *C. arabica* cultivars (Mundo Novo and Catuai), which suggests that those are ancestral polymorphisms that have not been fixed yet. Intriguingly, several SNPs found within the CaCc subgenome are coincident with *C. canephora* polymorphisms (Fig 4B; Table I). Some hypotheses can be proposed regarding this observation (i.e. gene flow occurred between *C. arabica* and *C. canephora*; polymorphisms result from several events of hybridization between *C. canephora* and *C. eugenioides*, suggesting multiple origins of *C. arabica*; the existence of a selective pressure favoring the heterozygote). However, due to the low diversity of *C. arabica* data used in this report, we can not affirm the cause of this result. Further studies dedicated to evolutionary aspects of *Coffea* species are indicated to unravel the origin and maintenance of such "residual ancestral heterozygosity."

The divergence between subgenomes may indicate that there is a mechanism to prevent *C. arabica* genome homogenization by avoiding the recombination between CaCc and CaCe. Previous studies indicated that despite the minor differentiation among the two constitutive genomes, the chromosomes of *C. arabica* only

pair homogenetically (Pinto-Maglio and Cruz, 1998; Lashermes et al., 2000). These authors hypothesized that homeologous chromosomes do not pair in *C. arabica*, probably due to the functioning of pairing-regulating factors.

Since our DNA sequence data were derived from ESTs, the analysis of each individual sequence frequency allowed us to make inferences about the composition of the *C. arabica* transcriptome. In contigs containing reads of both species (*C. arabica* and *C. canephora*), it was possible to assign 48% of the *C. arabica* ESTs as transcribed from the *C. canephora* subgenome (CaCc). As a consequence, the remaining sequences (52%) would have been transcribed from the *C. eugenioides* subgenome (CaCe). An inspection of the contigs showed that in 29% of the *C. arabica* genes there was a higher contribution of one subgenome in comparison with the other: 13% of the contigs had more ESTs from CaCc and 16% of contigs had more ESTs from CaCe. Therefore, our work showed that *C. arabica* displays differential expression of homeologous genes. This phenomenon has been reported for other allopolyploid species such as wheat (Mochida et al., 2003) and mainly in upland cotton (Udall et al., 2006; Hovav et al., 2008a, 2008b). It was demonstrated that 80% of the genes from hexaploid wheat, formed by three diploid species, showed biased expression for specific subgenomes and that the preferentially expressed homeolog could vary between tissues (Mochida et al., 2003). In addition, these authors observed that the gene expression or silencing among homeologs was not regulated at the chromosome or genome level but at the level of individual genes (Mochida et al., 2003). It is possible that a similar differential expression between tissues also exists in coffee, but our data set was not extensive enough to conclusively test this hypothesis. The differential expression of homeologs during allotetraploid cotton fiber development using allele-specific microarray platforms was evaluated (Udall et al., 2006; Hovav et al., 2008a, 2008b). These authors suggested that domestication increased the modulation of homeologous gene expression and that 30% of the homeologs are biased toward A or D cotton subgenomes. This percentage is not far from the 22% of differentially expressed *C. arabica* homeologs detected in our analysis. Although aware that using only high coverage contigs we would find more biases in homeolog differential expression, this would result in the selection of only highly expressed genes, leading to missing some interesting genes (which do not have such high levels of expression) for functional analyses. It is likely that a larger portion of the contigs present differential expression of the homeologs. Thus, despite these analysis limitations, the phenomenon of homeolog differential expression in *C. arabica* is consistent with our experimental validation (see below).

Our inference of homeolog differential expression based on an *in silico* subtractive strategy was validated in five of the six genes tested (Table III; Supplemental Fig. S1) using a TaqMAMA-based method (Li et al.,

2004). To the best of our knowledge, this is the first report of homeolog differential expression analysis using this method. The values of CaCc/CaCe homeolog expression observed in TaqMAMA assays are similar to those found by the in silico strategy (Table III), indicating that our bioinformatics approach was accurate. Although the ratios of “wet” and “dry” methods were not precisely equal, both follow the same tendency (i.e. they agree with the induction or repression of the CaCc homeolog in comparison with the CaCe homeolog) when assessing global EST data and leaf-only EST libraries. We believe that this biological experimentation validates our homeolog expression findings using the in silico strategy.

We also analyzed the putative functions of genes displaying differential expression of homeologs (Table II; Supplemental Table S4). The GO analysis suggested that auxin metabolism proteins (auxin-binding proteins, AUX/IAA-responsive proteins) appeared to be preferentially expressed from the CaCc subgenome. The CaCc subgenome also had a higher contribution for a set of GTP-binding proteins (Ras, Rac, Rab GTP-binding proteins), elongation and initiation translation factors (EF1- β , EF1- γ , EIF5a, EIF4a), ribosomal proteins, vesicular protein transport (ARF1, synaptobrevin), and proteosome subunits. Thus, the CaCc transcriptome seems to fine-tune *C. arabica* gene expression by the regulation of protein turnover and signal transduction. In contrast, CaCe subgenome expression appears to be more closely associated with basal processes. For example, proteins of the citric acid cycle (malate dehydrogenase, citrate synthase, succinate dehydrogenase), pentose-phosphate shunt (transaldolase, glyceraldehyde-3-phosphate dehydrogenase), and light and dark reactions of photosynthesis (chlorophyll *a/b*-binding protein, NADPH: protochlorophyllide oxidoreductases, phosphoglycerate kinase, phosphoribulokinase) had higher contributions from CaCe (Supplemental Table S4). These data suggested that the CaCe subgenome may provide the foundations for basal *C. arabica* metabolism.

As mentioned above, *C. eugenoides* has been used in breeding programs to reduce caffeine levels (Mazzafera and Carvalho, 1991) and in cup quality breeding (Carvalho, 2008). We believe that the result indicating that the *C. eugenoides* subgenome contributes to particular biological processes of *C. arabica* can provide further strategies to *C. arabica* breeding programs. For instance, the fact that the *C. arabica* photosynthetic apparatus is more similar to *C. eugenoides* can be a clue to guide the shade management of *C. arabica* coffee plantations.

Besides the presence of homeolog differential expression in *C. arabica*, we found another level of gene expression regulation involving paralogous genes. We detected that in five *C. arabica* genes, for each paralog a specific homeolog had been recruited, being much more expressed than the other. It is worth noting that for each member of a pair of paralogs, the two homeologs may be partitioned in opposite directions.

For example, while in one paralog the CaCc homeolog was more frequently expressed, in the other one it was the CaCe homeolog that was overrepresented. In addition, the expression difference between the homeologous genes in paralogous pairs was very pronounced (Table IV). We observed that in the case of FLPs, MLPs, SAMDC, and osmotin, the paralog more expressed in *C. canephora* continued to be the more expressed in *C. arabica* (CaCc subgenome; Table IV), showing a conservation of expression patterns. Inversely, the TH11 paralog gene more expressed in *C. canephora* was the least expressed in *C. arabica* (Table IV). Homeolog expression analysis revealed that such paralogs display differential expression in *C. arabica*, which, in most cases, seems to be maintained in relation to the *C. canephora* ancestor.

Furthermore, the evaluation of tissue expression profiles of these homeologs revealed another type of gene expression regulation. We have found in some cases that apparently one homeolog (i.e. CaCc) is recruited to be expressed in the analyzed tissue, whereas the other (i.e. CaCe) is silenced. More intriguingly is that the paralogs of genes first analyzed have an inverted expression profile: when the CaCc homeolog is silenced, the CaCe homeolog is expressed (Supplemental Fig. S2). This event cannot be named as subfunctionalization, as it implies that one homeolog is expressed in a specific tissue but the other is expressed in another one. However, we consider that we have detected another level of homeologous differential expression that is related to paralogs. As far as we know, this level of gene expression regulation was not reported previously and suggests a functional relevance for the coordination of paralog transcription in polyploids.

The genetic diversity observed between the two *C. arabica* genotypes analyzed (Mundo Novo and Catuai) in this study is narrow, and the results are in accordance with studies performed with other markers on larger sets of genotypes. The limited diversity observed hinders the identification of genes/alleles that provide resistance to biotic/abiotic stress, making the search for new sources of *Coffea* species genome diversity still essential. Therefore, wide crosses with the ancestor *C. eugenoides* and other *Coffea* species is the foremost direction for long-term breeding programs aiming to increase *C. arabica* variability. Regarding *C. canephora*, we have identified 4,449 SNPs that can be a good base to perform fine-mapping and initiate association studies. Such resources can be very interesting for *C. canephora* genetics studies (i.e. structure analysis, whole genome association mapping) and for the recently launched *C. canephora* genome sequencing initiative.

Our SNP discovery pipeline and the homeologous gene identification strategy described here are efficient tools to study diversity and evolution in recent allopolyploids. Moreover, our data show *C. arabica* as one of the polyploid species that displays differential expression of homeologous genes, indicating that

this phenomenon is indeed pervasive in polyploids. Such a phenomenon is very relevant to transcriptome regulation and can be a key factor to understanding gene expression in a perennial species such as *C. arabica* and provide the basis for breeding strategies. This result implies that genes useful for *C. arabica* breeding programs may already be present in its genome but are inactive due to partitioned expression. Methods that cause genome rearrangements (i.e. induced mutagenesis, somatic hybridization) may be an alternative to the conventional hybridization of parent lines by activating silenced genes and therefore generating new phenotypes that can provide traits to be selected by *C. arabica* breeders.

MATERIALS AND METHODS

EST Data Collection

A total of 267,533 ESTs, 78,182 from *Coffea canephora* and 189,351 from *Coffea arabica*, derived from 53 nonnormalized libraries were collected from the Brazilian Coffee Genome Project (Vieira et al., 2006) and from the *C. canephora* EST sequencing initiative (Lin et al., 2005; Supplemental Table S1). Two *C. arabica* cultivars originating from several generations of selfing were used to generate ESTs from the Brazilian coffee project: cv Catuai Vermelho IAC 144 for berry and leaf libraries and cv Mundo Novo IAC 388 for berry, leaf, root, and cell culture libraries. Six different genotypes were used for *C. canephora*, one genotype (Conilon) in the Brazilian Coffee Genome Project and five (collected in the east of Java Island) in the analysis performed by Lin et al. (2005). No information regarding cultivar origin of each EST library is available for the latter EST data set.

Assembly Procedures

Before the assembly, the sequences were trimmed (Baudet and Dias, 2007). This was done to remove ribosomal sequences, vector, poly(A/T) tails, and low-quality regions. After these alterations, the sequences with less than 100 bp remaining were discarded (Baudet and Dias, 2007).

The EST assembly was performed using the CAP3 program (Huang and Madan, 1999), whose parameters were adjusted to minimize the occurrence of type II assembly error (a minimum similarity threshold of 95% with a minimum overlap of 100 bases; Wang et al., 2004), preventing different genes of the same family, such as paralogs, from assembling in the same contig. Furthermore, using these parameters, alleles of the different homeologous genes were expected to coalesce in the same contig (Udall et al., 2006). To verify if such parameters were accurate in the assembling of the homeologous genes, sequencing of 6.4 kb of introns and exons of different nuclear genes from *Coffea eugenioides* and *C. canephora* (*C. arabica* ancestors) was done to evaluate the divergence between these species. Based on the results of this analysis, divergence between these sequences ranged from 0 to 2.47 polymorphic sites per 100 bp (i.e. 97.5% minimum similarity with an average of 1.3 polymorphic sites per 100 bp), confirming that the minimum similarity threshold used (95%) satisfied all the exigencies of the assembly.

After the assembly, bacterial sequence contaminations were analyzed using BLASTN with all the contigs against the NT database; the contigs with BLAST hits with *e*-values lower than $1e-5$ were removed. The pipeline used in this work is described in Figure 2.

SNP Discovery

QualitySNP was used as the core of SNP discovery with the default parameters. This software uses three filters for the identification of reliable SNPs. The first filter screens for all potential SNPs. False SNPs caused by sequencing errors are identified by the chromatogram quality given by Phred. Filter 2 is the core filter; it uses a haplotype-based strategy to detect reliable SNPs. In addition, the clusters with potential paralogs are identified using the differences in SNP number between potential haplotypes of the same contig.

Briefly, the SD of the normalized number of potential SNPs among potential haplotypes (D value) in one contig is calculated and used to identify haplotypes likely to be caused by paralogous sequences. The cutoff value of 0.6 was empirically observed by the authors of QualitySNP as adequate for the identification of paralogous genes in the assembly. Therefore, we considered that if D value is lower than 0.6, the contig is free of paralogs. All potential haplotypes consisting of only one sequence are removed, and singleton SNPs that are not linked to other polymorphism are not considered. This could lead to an underestimation of nucleotide diversity but guarantees that the false positives will be discarded. The last filter screens SNPs by calculating a confidence score, based upon sequence redundancy and base quality. All the information generated in QualitySNP with respect to contig, EST, and SNP (including haplotypes, SNP positions, etc.) was stored in a mysql database, which contains information about automatic (with BLAST against GenBank) and manual annotation. The scripts used to mine these data were developed in PERL (database available at <http://lge.ibi.unicamp.br/cafe/>).

Haplotype Identification, Assignment of *C. arabica* Haplotypes to Its Ancestral Genomes, and Diversity Analyses

The analysis performed on 6.4 kb in genes from *C. canephora* and *C. eugenioides* (data not shown) revealed divergences ranging between 0 and 2.47 polymorphisms per 100 bp. Given that *C. arabica* is a recent allotetraploid between these two species and assuming that the divergence between the two subgenomes stayed almost at the same level since their hybridization, an average of 13 sgSNPs within 1-kb contig sgSNPs will be detected between the two subgenomes. Therefore, assignment of the different haplotypes detected in *C. arabica* to the ancestral genomes was performed, taking into account that *C. arabica* subgenomes diverged at a low rate from their progenitor genome.

In QualitySNP, for a given contig, 80% of identities at all the polymorphic nucleotides are necessary to be assigned to the same haplotype. If different combinations of SNP alleles have at least 80% identity between them, QualitySNP allocates them in the same QualitySNP haplotype.

An identity higher than this threshold (greater than 80%) was expected between (1) the alleles of each homeolog derived from the CaCc and CaCe subgenomes (this homogenization is expected due to many generations of selfing) and (2) the homeologous genes from *C. canephora* and CaCc. As expected in example 2, comparison between *C. arabica* and *C. canephora* haplotypes revealed that some of the *C. arabica* haplotypes were highly similar to the *C. canephora* haplotypes (above the 80% threshold). Then, these haplotypes were clustered in the same QualitySNP haplotype by QualitySNP. The *C. arabica* haplotypes that were more divergent from *C. canephora* haplotypes were assigned as a different haplotype. The ESTs from *C. arabica* that clustered with *C. canephora* ESTs were considered as derived from the *C. canephora* ancestor (CaCc). By subtraction, all reads that were distant from the *C. canephora* haplotypes were considered as probably derived from the *C. eugenioides* ancestor (CaCe). To validate this strategy, some *C. eugenioides* ESTs were sequenced and mapped in this assembly.

As almost all polymorphisms within *C. arabica* must be derived from the divergence between the two subgenomes, homeologous genes are expected to be correctly identified in all cases using this approach, except when (1) the divergence between the gene of *C. canephora* and CaCc is higher than 80% (caused by a different evolution between the subgenome into *C. arabica* and the species *C. canephora*); (2) the divergence between CaCc and CaCe is very low (cases with no sgSNPs between the two subgenomes are possible); (3) some recombination occurred along the gene; or (4) there is no sequence from *C. canephora*. Only contigs with four or more ESTs from *C. arabica* and two or more ESTs from *C. canephora* were considered.

Homeologous Gene Frequencies

The differential expression of homeologous genes was calculated using Audic-Claverie statistics (Audic and Claverie, 1997). Contigs containing at least four ESTs, more than twice the number of reads from the same subgenome in comparison with the other, and with a *P* value less than 0.005 were considered as differentially expressed by the two subgenomes of *C. arabica*. A similar analysis was done using the cultivar information available from the *C. arabica* database with the exception that, in this case, we filtered contigs with at least two reads from each cultivar and at least two reads from each subgenome.

Differential Expression of Homeologous Genes by qPCR Analysis

Leaves from *C. arabica* cv Mundo Novo 376-4 were harvested in the Pólo Regional Nordeste Paulista from the Instituto Agronômico de Campinas, located in Mococa, São Paulo, Brazil (21°27'54''S/ 47°00'21''W, 640 m), and immediately frozen in liquid nitrogen. RNA was extracted using a method based on Azevedo et al. (2003) with modifications (protocol developed by Joan G. Barau, unpublished data). Samples of 785 ng of RNA were used for reverse transcription with random hexamer primers for first-strand synthesis and SuperScript III RNase reverse transcriptase (Invitrogen).

For the validation of in silico homeolog differential gene expression, an approach based on the real-time qPCR TaqMAMA method (Li et al., 2004) was applied. Six genes were chosen (Contig21552, Contig11105, Contig10821, Contig10284, Contig17875, Contig18072), observing the alignment of the reads in the contig. Forward primers contained the sgSNP in the last 3' nucleotide and a mismatch before it to increase the allele (homeolog) discrimination. Thus, two mismatches occur between a primer and the allele to be discriminated against, whereas only a single mismatch occurs with the allele of interest (the last nucleotide at the 3' region). The additional mismatches were selected based on the combination suggested by Li et al. (2004). Therefore, two primers were designed for each polymorphic site, one that preferentially amplifies allele 1 (homeolog 1) and one that preferentially amplifies allele 2 (homeolog 2). The reverse primers were designed to amplify a fragment of 100 bp. In addition, primers without sgSNPs and additional mismatches were designed (Supplemental Table S5).

qPCR was performed on the StepOne System (Applied Biosystems) with SYBR Green qPCR kits (Sigma). Reactions comprised 1× SYBR Green mix, 625 nM primer pairs, and 1 μL of template. The following cycling conditions were employed: initial denaturation at 94°C for 2 min, followed by 40 cycles of 94°C for 15 s, 30 s of annealing with the primer's temperature, 60°C per minute to amplification and melting curve of 95°C for 15 s, 60°C per minute, then 95°C for 15 s.

The data were analyzed by variation between logs of reaction efficiency at a given cycle threshold (Ct):

$$\log \frac{\text{expression allele 2}}{\text{expression allele 1}} = (\log E_{\text{allele2}} \times Ct_{\text{allele2}}) - (\log E_{\text{allele1}} \times Ct_{\text{allele1}})$$

according to Roberts et al. (2008). The efficiency were calculated by $E = 10^{(-1/b) - 1}$ (Rutledge and Côté, 2003), where b is the slope of the linear regression.

GO Analysis

A multilevel analysis for biological processes from the GO database (Ashburner et al., 2000) was performed using the BLAST2GO program (Conesa and Götz, 2008) within the contigs with at least one GO attributed in level 3 or higher. Hypergeometric distribution statistical analysis described in GOToolBox (Martin et al., 2004) was applied to select the GO terms with P values lower than 0.05, comparing the classes with differential expression between the *C. arabica* subgenomes and the total transcriptome.

All the Brazilian Coffee Genome Project ESTs were submitted to GenBank with the accession numbers GT669291 to GT734396 and GT640310 to GT640366 (*C. arabica*), GT645618 to GT658452 (*C. canephora*), and HO059040 to HO059057 (*C. eugenioides*).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Amplification plot and melting curve of sgSNPs by qPCR.

Supplemental Figure S2. Differential homeologous gene expression variation between paralogs in specific tissues.

Supplemental Table S1. Description of the EST libraries used in this work.

Supplemental Table S2. Top 50 contigs with more ESTs derived from one *C. arabica* subgenome than the other.

Supplemental Table S3. Correlation of EST coverage of contigs and differential expression of homeologous genes.

Supplemental Table S4. Manual annotation of contigs from each GO term described in Table II.

Supplemental Table S5. Sequences of primers used in allelic (homeologous) discrimination and differential expression of homeologous gene analysis by qPCR.

ACKNOWLEDGMENTS

We are grateful to Juan Lucas Argueso (Department of Molecular Genetics and Microbiology, Duke University Medical Center) for helpful discussion and suggestions about the manuscript, Johana Rincones (Braskem) for manuscript revision, and Xavier Argout and Pierre Charmetant (Centre de Coopération Internationale en Recherche Agronomique pour le Développement) for useful discussions about QualitySNP implementation to *Coffea* sequences and *Coffea* genetics. We appreciate the thoughtful criticisms and suggestions of the three colleagues who reviewed the manuscript. We also especially thank the researchers and technicians involved in the Brazilian coffee EST sequencing initiative.

Received July 9, 2010; accepted September 22, 2010; published September 23, 2010.

LITERATURE CITED

- Adams KL (2007) Evolution of duplicate gene expression in polyploid and hybrid plants. *J Hered* **98**: 136–141
- Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci USA* **100**: 4649–4654
- Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr Opin Plant Biol* **8**: 135–141
- Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor Appl Genet* **114**: 359–372
- Anthony F, Bertrand B, Quiros O, Wilches A, Lashermes P, Berthaud J, Charrier A (2001) Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. *Euphytica* **118**: 53–65
- Anthony F, Combes C, Astorga C, Bertrand B, Graziosi G, Lashermes P (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theor Appl Genet* **104**: 894–900
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29
- Audic S, Claverie JM (1997) The significance of digital gene expression profiles. *Genome Res* **7**: 986–995
- Azevedo H, Lino-Neto T, Tavares RM (2003) An improved method for high-quality RNA isolation from needles of adult maritime pine trees. *Plant Mol Biol Rep* **21**: 333–338
- Baudet C, Dias Z (2007) New EST trimming procedure applied to SUCEST sequences. In *Proceedings of the Second Brazilian Conference on Advances in Bioinformatics and Computational Biology*. Springer-Verlag, Berlin, pp 57–68
- Carvalho CHS, editor (2008) Cultivares de Café: Origem, Características e Recomendações. Embrapa, Brasília-DF, Brazil
- Cha RS, Zarbl H, Keohavong P, Thilly WG (1992) Mismatch amplification mutation assay (MAMA): application to the c-H-ras gene. *PCR Methods Appl* **2**: 14–20
- Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS, et al (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* **176**: 685–696
- Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**: 619832
- Cros J, Combes MC, Trouslot P, Anthony F, Hamon S, Charrier A, Lashermes P (1998) Phylogenetic analysis of chloroplast DNA variation in *Coffea* L. *Mol Phylogenet Evol* **9**: 109–117
- Cubry P, Musoli P, Legnaté H, Pot D, de Bellis F, Poncet V, Anthony F,

- Dufour M, Leroy T (2008) Diversity in coffee assessed with SSR markers: structure of the genus *Coffea* and perspectives for breeding. *Genome* **51**: 50–63
- 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
- Du CE, Liu HM, Li RZ, Li PB, Ren ZQ (2003) [Application of single nucleotide polymorphism in crop genetics and improvement]. *Yi Chuan* **25**: 735–739
- Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D (2009) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res* **37**: D951–D953
- Hendre PS, Phanindranath R, Annapurna V, Lalremruata A, Aggarwal RK (2008) Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex A. Froehner) showing broad cross-species transferability and utility in genetic studies. *BMC Plant Biol* **8**: 51
- Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF (2008a) Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. *Genetics* **179**: 1725–1733
- Hovav R, Udall JA, Chaudhary B, Rapp R, Flagel L, Wendel JF (2008b) Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proc Natl Acad Sci USA* **105**: 6191–6195
- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* **9**: 868–877
- 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
- Lashermes P, Paczek V, Trouslot P, Combes MC, Couturon E, Charrier A (2000) Single-locus inheritance in the allotetraploid *Coffea arabica* L. and interspecific hybrid *C. arabica* × *C. canephora*. *J Hered* **91**: 81–85
- Li B, Kadura I, Fu DJ, Watson DE (2004) Genotyping with TaqMAMA. *Genomics* **83**: 311–320
- Lin C, Mueller LA, McCarthy J, Crouzillat D, Pétiard 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
- Liu Z, Adams KL (2007) Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. *Curr Biol* **17**: 1669–1674
- Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B (2004) GOTool-Box: functional analysis of gene datasets based on Gene Ontology. *Genome Biol* **5**: R101
- Maurin O, Davis AP, Chester M, Mvungi EF, Jaufeerally-Fakim Y, Fay MF (2007) Towards a phylogeny for *Coffea* (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. *Ann Bot (Lond)* **100**: 1565–1583
- Mazzafera P, Carvalho A (1991) Breeding for low seed caffeine content of coffee (*Coffea* L.) by interspecific hybridization. *Euphytica* **59**: 55–60
- Mochida K, Yamazaki Y, Ogiwara 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
- Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312
- Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ, Lee HS, Comai L, Madlung A, Doerge RW, Colot V, et al (2003) Understanding mechanisms of novel gene expression in polyploids. *Trends Genet* **19**: 141–147
- Pindo M, Vezzulli S, Coppola G, Cartwright DA, Zharkikh A, Velasco R, Troglio M (2008) SNP high-throughput screening in grapevine using the SNPlex genotyping system. *BMC Plant Biol* **8**: 12
- Pinto-Maglio CAE, Cruz ND (1998) Pachytene chromosome morphology in *Coffea* L. II. *C. arabica* L. complement. *Caryologia* **51**: 19–35
- Roberts I, Ng G, Foster N, Stanley M, Herdman MT, Pett MR, Teschendorff A, Coleman N (2008) Critical evaluation of HPV16 gene copy number quantification by SYBR Green PCR. *BMC Biotechnol* **8**: 57
- Rutledge RG, Côté C (2003) Mathematics of quantitative kinetic PCR and the application of standard curves. *Nucleic Acids Res* **31**: e93
- Staples RC (2000) Research on the rust fungi during the twentieth century. *Annu Rev Phytopathol* **38**: 49–69
- Steiger L, Nagai C, Moore H, Morden W, Osgood V, Ming R (2002) AFLP analysis of genetic diversity within and among *Coffea arabica* cultivars. *Theor Appl Genet* **105**: 209–215
- Sylvain PG (1955) Some observations on *Coffea arabica* L. in Ethiopia. *Turrialba* **6**: 37–53
- Tang J, Leunissen JA, Voorrips RE, van der Linden CG, Vosman B (2008) HaploSNPer: a Web-based allele and SNP detection tool. *BMC Genet* **9**: 23
- Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JAM (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* **7**: 438
- Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB, et al (2006) A global assembly of cotton ESTs. *Genome Res* **16**: 441–450
- Vieira LGE, Andrade AC, Colombo CA, Moraes AHA, Metha A, Oliveira AC, Labate CA, Marino CL, Monteiro-Vitorello CB, Monte DC, et al (2006) Brazilian Coffee Genome Project: an EST-based genomic resource. *Braz J Plant Physiol* **18**: 95–108
- Wang JPZ, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC, dePamphilis CW (2004) EST clustering error evaluation and correction. *Bioinformatics* **20**: 2973–2984

DISCUSSÕES

O café é uma importante commodity agrícola, e tem grande importância econômica nos países produtores e consumidores. Embora *C. arabica* seja a principal espécie cultivada (~70%), ela tem uma base genética muito estreita. Esse baixo nível de diversidade é provavelmente um dos fatores que contribuem para a alta suscetibilidade a patógenos e pragas. *C. canephora* é uma das principais fontes de genes de resistência a doenças para programas de melhoramento de *C. arabica*, mas produz uma qualidade de bebida inferior. Assim, as características da bebida de híbridos resistentes a doenças, entre *C. canephora* e *C. arabica* podem ser inferiores ao de *C. arabica* parental. Esta limitação ressalta a necessidade para a compreensão da genética de mecanismos subjacentes à variabilidade fenotípica entre essas duas espécies, que podem apoiar estratégias alternativas de criação e de seleção orientada. Portanto, as descobertas feitas nesse trabalho são particularmente interessantes para espécies com baixa diversidade como *Coffea arabica*.

As sequências de cDNA foram derivadas de duas iniciativas de seqüenciamento de transcriptoma via método Sanger (Lin et al, 2005; Vieira et al, 2006) e foram fontes de dados para a identificação de vários fenômenos neste trabalho que foi o primeiro estudo do perfil transcricional do genoma global de *Coffea arabica* e *Coffea canephora*. Através desse trabalho foi possível minerar profundamente todos os dados disponíveis dessas duas espécies e levantar importantes hipóteses.

A partir da montagem de cerca de 200 mil ESTs de *C. arabica* e *C. canephora* foi aplicado um conjunto diversificado de ferramentas de bioinformática para extrair informações sobre o conteúdo genético, diferenças de transcriptoma e novos genes e famílias de genes, além disso aprofundamos as análises em *C. arabica* identificando seus sugenomas ancestrais e novos padrões de expressão gênica dessa alotetraploide. As bibliotecas foram construídas a partir de diversos tecidos e órgãos. Porém, a maioria das bibliotecas de cDNA foram extraídas de sementes, folhas e flores.

Foram feitas várias análises comparativas entre as duas espécies e os resultados indicam uma prevalência de proteínas relacionadas ao metabolismo do açúcar em *C. arabica* e de

transdução de sinal em *C. canephora* podendo ser correlacionados com características agronômicas de cada espécie, devido à melhor qualidade da bebida de *C. arabica* e alta tolerância a estresses específicos em plantas de *C. canephora*.

A montagem híbrida entre *C. arabica* e *C. canephora*, juntamente com uma estratégia de identificação de haplótipos baseado nos padrões de SNPs nos permitiu analisar os dois subgenomas. Os reads de *C. arabica* que apresentam o mesmo padrão de SNPs de *C. canephora* foram identificados como derivados de *C. canephora* (CaCc), enquanto que os reads que não correspondem a esse padrão foram considerados como provenientes da segunda espécie ancestral, *C. eugenoides* (CaCe). Embora *C. arabica* seja autógama, *C. canephora* é alógama e, portanto, era esperado que apresentassem níveis mais elevados de diversidade de nucleotídeos. No entanto, a análise de polimorfismos mostrou que *C. arabica* apresenta maior frequência de polimorfismo (0,393 xSNP/100 pb) do que *C. canephora* (0,169 bp SNP/100, um resultado consistente com a presença de dois subgenomas em *C. arábica*. O nosso trabalho confirmou esta hipótese observando que a frequência de polimorfismos em *C. arábica* (0,393 xSNP/100 pb) é semelhante a frequência encontrada entre os subgenomas (0,359 bp sgSNP/100), indicando que a diversidade de sequência entre os subgenomas é a principal fonte de variabilidade genética em *Coffea arábica*. Descobrimos alguns poucos casos de polimorfismos dentro de subgenomas o que sugere que existe uma manutenção da heterozigozidade ancestral.

Apesar do transcriptoma de *C. arabica* ter metade dos transcritos de cada subgenoma, foi encontrado percentual de genes diferencialmente expressos similares aos encontrados em algodão. É possível que exista expressão diferencial entre os tecidos no café, mas nosso conjunto de dados não foi extenso o suficiente para detectar e provavelmente os genes que possuam essa expressão diferencial entre os tecidos estão subestimados em nossa análise, uma vez q a soma do pool de tecidos pode ocultar esses casos. Porém, apesar das limitações de análise, o fenômeno da expressão diferencial de homeologos em *C. arabica* é consistente como demonstrado pela validação experimental.

Nossos resultados indicam que o transcriptoma CaCc parece responsável por genes de ajuste fino em *C. arabica*, expressando mais proteínas relacionadas a regulação e transdução de sinal. Por outro lado a expressão do subgenoma CaCe parece estar mais estreitamente associado aos processos basais. *C. eugenoides* já foi utilizado em programas de melhoramento para reduzir

os níveis de cafeína (Mazzafera e Carvalho, 1991) e para melhorar a qualidade da bebida (Carvalho, 2008). Acreditamos que o resultado indica que o subgenoma de *C. eugenioides* contribui para importantes processos biológicos de *C. arabica* podendo fornecer novas estratégias para programas de melhoramento.

A diversidade genética observada entre os dois cultivares de *C. arabica* analisados neste estudo (Mundo Novo e Catuaí) é muito pequena, praticamente não foram encontrados SNPs entre os cultivares, apenas nove SNPs, mas que eram únicos e foram eliminados no filtro de qualidade, suspeitamos que exista uma diferença de expressão de homeologos entre os cultivares que talvez possa dar origem aos diferentes fenótipos, e encontramos algumas evidencias desse fenômeno, porém, inconclusivos devido aos diferentes tipos de tecidos analisados em ambos cultivares.

Os métodos utilizados se mostraram bastante confiáveis quando colocados a prova através do seqüenciamento de *C. eugenioides*, para validar o método de identificação dos subgenomas por subtração, e por um método baseado no TaqMAMA e qPCR para validar as tendências de expressão diferencial e, pelo que conhecemos, essa foi a primeira vez onde esse método foi empregado para análise de expressão diferencial entre homeologos.

O pipeline para descoberta de SNPs e a estratégia desenvolvida para identificação de genes homeologos são ferramentas eficientes para o estudo da diversidade e evolução de alopoliplóides recentes. Além disso, os dados incluem *C. arabica* como uma espécie poliploide com evidencias de expressão diferencial de genes homeologos, indicando que esse fenômeno é realmente difundido em poliplóides. Tais fenômenos são relevantes para a regulação do transcriptoma e podem ser um fator chave para compreensão do perfil de expressão de genes de uma espécie perene como *C. arabica*. Esses resultados indicam que é possível que genes úteis para o melhoramento genético de *C. arabica* podem já estar presentes no seu genoma, mas inativos devido à expressão particionada dos homeologos. Métodos que causam rearranjos no genoma podem ser alternativos para a hibridação convencional podendo ativar genes silenciados e, portanto, gerar novos fenótipos que possam fornecer características a serem selecionadas.

Acreditamos que esses dados são valiosos para a interpretação do desenvolvimento do café, fornecendo informações que possam ajudar os programas de melhoramento de café e indicar alvos potenciais para análise funcional e produtos de biotecnologia dessa espécie.

CONCLUSÕES

- Comparando *Coffea arabica* e *Coffea canephora*, existe uma prevalência de proteínas relacionadas ao metabolismo do açúcar em *C. arabica* e transdução de sinal em *C. canephora*;
- A prevalência de classes de genes diferenciando as duas espécies pode ser correlacionada com características agrônomicas de cada uma delas, devido à melhor qualidade da bebida de *C. arabica* e alta tolerância a estresses específicos em plantas de *C. canephora*;
- Foram identificados os transcriptomas dos subgenomas ancestrais de *C. arabica*;
- O transcriptoma de *C. arabica* é transcrito praticamente meio a meio por cada um dos ancestrais;
- 22% dos genes apresentam expressão diferencial de homeólogos.
- Em *Coffea arabica*, o subgenoma *C. eugenioides* (CaCe) parece ser responsável por processos biológicos basais;
- Em *Coffea arabica*, o subgenoma *C. canephora* (CaCc) parece ser responsável por mecanismos de transdução de sinais em *C. arabica*;
- Foi identificado um novo mecanismo de regulação baseado na expressão diferencial de subgenomas em parálogos;
- É possível que genes úteis para o melhoramento genético de *C. arabica* podem já estar presentes no seu genoma, mas inativados;
- Acreditamos que esses dados colaboram para o entendimento da fisiologia do café, fornecendo informações que poderão ajudar programas de melhoramento e indicando alvos potenciais para análise funcional e desenvolvimento de produtos biotecnológicos.

REFERÊNCIAS

- Adams KL (2007) Evolution of duplicate gene expression in polyploid and hybrid plants. *J. Hered* 98: 136-141
- Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U S A* 100: 4649-4654
- Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8: 135-141
- Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor. Appl. Genet.* 114: 359-372
- Anthony F, Bertrand B, Quiros O, Wilches A, Lashermes P, Berthaud J, Charrier A (2001) Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. *Euphytica* 118: 53-65
- Anthony F, Combes C, Astorga C, Bertrand B, Graziosi G, Lashermes P (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theor. Appl. Genet.* 104: 894-900
- Bridson DM, Verdcourt B. (1988) *Flora of tropical East Africa: Rubiaceae (Part 2)*. Cape Town: Iziko Museums of Cape Town, pp. 415-747
- Carvalho CHS, ed (2008) *Cultivares de Café: Origem, Características e Recomendações*, Ed 1. Embrapa
- Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS, Hwang EY, Yi SI, Young ND, Shoemaker RC, Van Tassell CP, Specht JE, Cregan PB (2007) A soybean transcript map: Gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* 176: 685-696

- Crepet, W.L., Nixon, K.C., and Gandolfo, M.A. (2004). Fossil evidence and phylogeny: the age of major angiosperm clades based on mesofossil and macrofossil evidence from Cretaceous deposits. *Am J Bot* 91(10): 1666-1682.
- Crane E. and Walker P. (1983) The Impact of Pest Management on Bees and Pollination. Tropical Development and Research Institute, College House, Wrights Lane, London, UK.
- Cubry P, Musoli P, Legnate H, Pot D, de Bellis F, Poncet V, Anthony F, Dufour M, Leroy T (2008) Diversity in coffee assessed with SSR markers: structure of the genus *Coffea* and perspectives for breeding Genome 51: 50-63
- Doyle, J.J. et al., (2008). Evolutionary Genetics of Genome Merger and Doubling in Plants. *Annual Review of Genetics*, 42(1), 443.
- Du CF, Liu HM, Li RZ, Li PB (2002) Application of single nucleotide polymorphism in crop genetics and improvement. *Hereditas* 25: 735-744
- Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D(2009) AutoSNPdb: An annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res.* 37: 951-953 Syvanen, A., 2005. Toward genome-wide SNP genotyping. *Nature Genetics*.
- Emahazion T., Feuk L., Jobs M., Sawyer SL., Fredman D., St Clair D., Prince JA., Brookes AJ. (2001) SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends in Genetics*. 17: 407-413(7)
- Free JB.(1993) Insect Pollination of Crops. Academic Press, London.
- Fazuoli, L.C. Genética e melhoramento do cafeeiro. In: Rena,A.B.; Malavolta,E.; Rocha,M.; Yamada,T. (eds).(1986) *Cultura do cafeeiro* – fatores que afetam a produtividade. Piracicaba, Associação Brasileira para Pesquisa da potassa e do fosfato. 87-113.
- Gandolfo, M.A., Nixon, K.C., Crepet, W.L., Stevenson, D.W., and Friis, E.M. (1998). Oldest known fossils of monocotyledons. *Nature* 394(6693): 532-533.
- Hendre PS, Phanindranath R, Annapurna V, Lalremruata A, Aggarwal RK (2008) Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex A. Froehner) showing broad cross-species transferability and utility in genetic studies. *BMC Plant Biol.* 8: 51

- Hoeven RV, Ronning C, Giovannoni J, Martin G, and Tanksley S (2002). Deductions about the Number, Organization, and Evolution of Genes in the Tomato Genome Based on Analysis of a Large Expressed Sequence Tag Collection and Selective Genomic Sequencing Plant Cell 14: 1441-1456.
- Hovav, R., Chaudhary, B. et al., (2008). Parallel Domestication, Convergent Evolution and Duplicated Gene Recruitment in Allopolyploid Cotton. *Genetics*, 179(3), 1725-1733.
- Hovav, R., Udall, J.A. et al., (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. *Proceedings of the National Academy of Sciences*, 105(16), 6191-6195.
- 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
- Leitch, I.J. & Bennett, M.D., (1997). Polyploidy in angiosperms. *Trends in Plant Science*.
- 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
- Liu Z, Adams KL (2007) Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. Curr. Biol. 17: 1669-1674
- Mazzafera P, Carvalho A (1991) Breeding for low seed caffeine content of coffee (*Coffea* L.) by interspecific hybridization. Euphytica 59: 55-60
- Masterson, J., (1994). Stomatal Size in Fossil Plants: Evidence for Polyploidy in Majority of Angiosperms. *Science*.
- Maurin O, Davis AP, Chester M, Mvungi EF, Jaufeerally-Fakim Y, Fay MF (2007) Towards a phylogeny for Coffea (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. Ann Bot (Lond) 100: 1565–1583
- Mochida, K., Yamazaki, Y. & Ogihara, Y., (2004). Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Molecular Genetics and Genomics*, 270(5), 371-377.
- Novaes E, Drost DR, Farmerie WG, Pappas Jr GJ, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an

- uncharacterized genome. *BMC Genomics* 9: 312
- Osborn, T.C. et al., (2003). Understanding mechanisms of novel gene expression in polyploids. *Trends in Genetics*.
- Otto, S.P. & Whitton, J., (2000). POLYPLOID INCIDENCE AND EVOLUTION. *Annual Review of Genetics*, 34(1), 401.
- Pindo M, Vezzulli S, Coppola G, Cartwright DA, Zharkikh A, Velasco R, Troglio M (2008) SNP high-throughput screening in grapevine using the SNPlex™ genotyping system. *BMC Plant Biol.* 8: 12
- Pinto-Maglio CAF, Cruz ND (1998) Pachytene chromosome morphology in *Coffea* L. II. *C. arabica* L. complement. *Caryologia* 51:19-35
- Purseglove JW. (1968) Tropical Crops. Dicotyledons I and II. Longmans, London, UK.
- Rick, C. (1971). Some cytogenetic features of the genome in diploid species. *Stadler Sym* 1: 153-174.
- Sherry ST, Ward MH , Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.* 29: 308-311.
- Syvänen AC. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2, 930-942
- Steiger L, Nagai C, Moore H, Morden W, Osgood V, Ming R (2002) AFLP analysis of genetic diversity within and among *Coffea arabica* cultivars. *Theor. Appl. Genet.* 105: 209-215;
- Sylvain PG (1955) Some observations on *Coffea arabica* L. in Ethiopia. *Turrialba* 6: 37-53
- Useche FJ, Gao G, Harafey M, Rafalski A. (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Inform.* 12:194-203.
- Vieira LGE, Andrade AC, Colombo CA, Moraes AHA, Metha A, Oliveira AC, Labate CA, Marino CL, Monteiro-Vitorello CB, Monte DC *et al.*: (2006) Brazilian coffee genome project: an EST-based genomic resource. *Brazil J Plant Physiol*, 18:95-108.

OUTROS TRABALHOS PUBLICADOS PELO AUTOR

1. Maciel, B. ; Santos, ACF. ; Dias, JCT ; Vidal, RO. ; Cascardo, JCM. ; Rezende, RP. **Simple DNA extraction protocol for a 16S rDNA study of bacterial diversity in tropical landfarm soil used for biore-mediation of oil waste.** *Genetics and Molecular Research*, v. 8, p. 375-388, 2009.
2. Carels, N. ; Vidal, RO. ; Frias, DG . **Universal Features for the Classification of Coding and Non-Coding DNA Sequences.** *Bioinformatics and Biology Insights*, v. 3, p. 37-49, 2009
3. Mondego, Jorge MC ; Carazzolle, Marcelo F ; Costa, Gustavo GL ; Formighieri, Eduardo F ; Parizzi, Lucas P ; Rincones, Johana ; Cotomacci, Carolina ; Carraro, Dirce M ; Cunha, Anderson F ; Carrer, Helaine ; VIDAL, R. O. ; Estrela, Raissa C ; Garcia, Odalys ; Thomazella, Daniela PT ; de Oliveira, Bruno V ; Pires, Acassia BL ; Rio, Maria Carolina S ; Araujo, Marcos Renato R ; de Moraes, Marcos ; Castro, Luis AB ; Gramacho, Karina P ; Goncalves, Marilda S ; Moura Neto, Jose P ; Goes Neto, Aristoteles ; Barbosa, Luciana V ; Guiltinan, Mark J ; Bailey, Bryan A ; Meinhardt, Lyndel W ; Cascardo, Julio CM ; Pereira, Goncalo AG . **A genome survey of *Moniliophthora perniciosa* gives new insights into Witches Broom Disease of cacao.** *BMC Genomics*, v. 9, p. 548, 2008.
4. Argueso, J. L. ; Carazzolle, M. F. ; Mieczkowski, P. A. ; Duarte, F. M. ; Netto, O. V.C. ; Missawa, S. ; Galzerani, F. ; Costa, G. G.L. ; VIDAL, R. O. ; Noronha, M. F. ; Dominska, M. ; Andrietta, M. G.S. ; Andrietta, S. R. ; Cunha, A. F. ; Gomes, L. H. ; Tavares, F. C.A. ; Alcarde, A. R. ; Dietrich, F.; McCusker, J. H. ; Petes, T. D. ; Pereira, G. A.G. . **Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production.** *Genome Research*, v. 19, p. 2258-2270, 2009.
92. Cardoso, Kiara C ; Da Silva, Marcio J ; Costa, Gustavo GL ; Torres, Tatiana T ; Del Bem, Luiz Eduardo; Vidal, Ramon O ; Menossi, Marcelo ; Hyslop, Stephen . **A transcriptomic analysis of gene expression in the venom gland of the snake *Bothrops alternatus* (urutu).** *BMC Genomics*, 2010.

ANEXO 1 – Material suplementar do Capítulo I

Additional File I: Description of Brazilian coffee ESTs and description of libraries

A) Confection of Brazilian initiative cDNA libraries

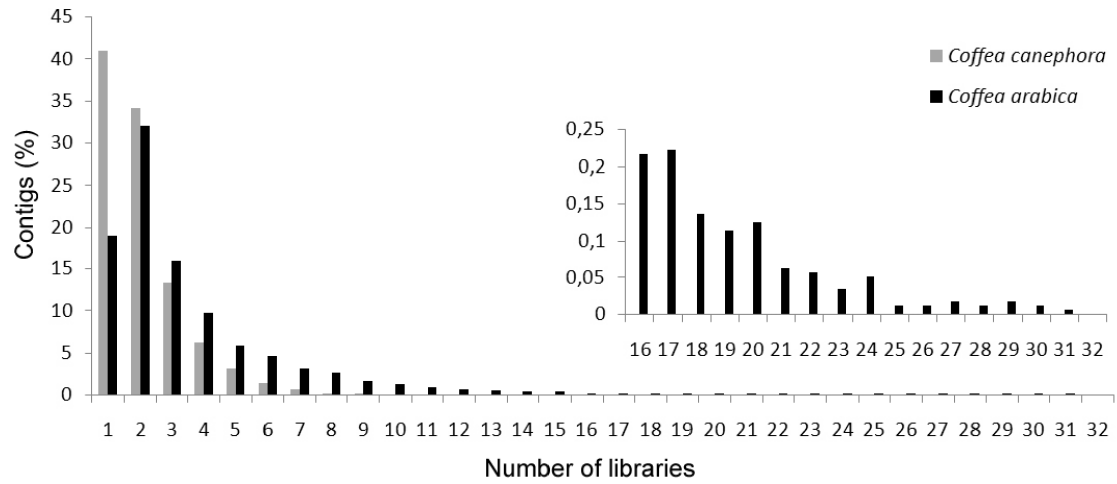
RNA from coffee tissues was extracted from different developmental stages and from plant suffering different stress conditions. Poly(A)+ RNA was purified from total RNA using the Oligotex Kit (Qiagen, USA). cDNA libraries were constructed using the SuperScript Plasmid System and Plasmid Cloning Kit (Invitrogen, USA) with about 1-2 µg poly(A)+ RNA. The efficiency of cDNA synthesis was monitored with radioactive nucleotides. cDNA were size fractionated on a Sepharose CL-2B column. Aliquots of each fraction were electrophoresed in agarose gel to determine the size range of cDNAs. Fractions containing cDNA larger than 500 bp were ligated into pSPORT1 and pSPORT6 vectors (Invitrogen) at the Sall-NotI site. The resulting plasmids were transformed in *E. coli* DH10B or DH5α cells (Invitrogen) by electroporation. Plasmid DNA was purified using a modified alkaline lysis method (Sambrook et al., 1989). Sequencing reactions were conducted using the ABI BigDye Terminator Sequencing kit (Applied Biosystems). cDNA inserts were sequenced from the 5' end with T7 promoter primer (5'-TAATACGACTCACTATAGGG-3') or M13 Rev in the pSPORT1 vector with SP6 primer (5'-ATTTAGGTGACACTATAG-3') in the pSPORT6. Sequencing reaction products were analyzed on ABI 3700 sequencers (Applied Biosystems).

B) Description of the coffee ESTs libraries

<i>Coffea arabica</i>	Library	Description	Cultivar	Source
	AR1	Leaves treated with araquidonic acid	Mundo novo	Brazil
	LP1	Plantlets treated with araquidonic acid	Mundo novo + Catuai	Brazil
	CB1	Suspension cells treated with benzothiadiazole and brassinoesteroids	Catuai	Brazil
	CL2	Hypocotyls treated with benzothiadiazole	Mundo novo + Catuai	Brazil
	EA1, IA1, IA2	Embryogenic calli	Catuai	Brazil
	EB1	Zygotic embryo	Mundo novo + Catuai	Brazil
	EM1, SI3	Germinating seeds (whole seeds and zygotic embryos)	Catuai	Brazil
	FB1, FB2, FB4	Flower buds in different developmental stages	Mundo novo	Brazil
	FR1, FR2	Flower buds + pinhead fruits + fruits at different stages	Mundo novo	Brazil
	CA1	Non embryogenic calli	Mundo novo + Catuai	Brazil
	IC1	Non embryogenic calli	Catuai	Brazil
	PC1	Non embryogenic calli + 2,4-D	Mundo novo + Catuai	Brazil
	LV4, LV5	Young leaves from orthotropic branch	Mundo novo	Brazil
	LV8, LV9	Mature leaves from plagiotropic branches	Mundo novo	Brazil
	NS1	Roots infected with nematodes	Mundo novo + Catuai	Brazil
	PA1	Primary embryogenic calli	Mundo novo + Catuai	Brazil
	RM1	Leaves infected with leaf miner and coffee leaf rust	Mundo novo	Brazil
	RT3	Roots	Mundo novo	Brazil
	RT5	Roots with benzothiadiazole	Mundo novo	Brazil
	RT8	Suspension cells with stressed with aluminum	Catuai	Brazil
	RX1	Stems infected with <i>Xylella</i> spp	Catuai	Brazil
	SH2	Water deficit stresses field plants (pool of tissues)	Catuai	Brazil
	SS1	Well-watered field plants (pool of tissues)	Catuai	Brazil
	CS1	Suspension cells with mannose NaCl and KCL	Catuai	Brazil
	BP1	Suspension cells treated with acibenzolar-S-methyl	Catuai	Brazil
	PL1	?	?	Brazil
	SI1	Germinating seeds	Rubi	Brazil
	SI2	Germinating seeds	Rubi	Brazil
	CD1	Suspension cells	Catuai	Brazil
	CL1	Suspension cells	Catuai	Brazil
	CM1	?	Mundo novo + Catuai	Brazil
	LM3	?	Mundo novo + Catuai	Brazil
	RT7	Root	Mundo novo + Catuai	Brazil
	FB3	Flower buds	Mundo novo	Brazil
	FP2	?	Mundo novo + Catuai	Brazil
<i>Coffea canephora</i>	Library	Description	Cultivar/ Varieties	
	LF1	Young leaves,	BP409	Nestlé
	PP1	Pericarp, all developmental stages	BP358,BP409,BP42, BP961,Q121	Nestlé
	SE1	Whole cherries,18 and 22 week after pollination	BP358,BP409, BP42,Q121	Nestlé

SE2	Whole cherries, 18 and 22 week after pollination	BP358, BP409, BP42, Q121	Nestlé
SE3	Endosperm and perisperm, 30 week after pollination	BP409, BP961, Q121	Nestlé
SE4	Endosperm and perisperm, 42 and 46 weeks after pollination	BP358, BP409, BP42, BP961, Q121	Nestlé
EC1	Embriogenic calli	Conilon	Brazil
SH1	Leaves from water deficit stressed plants	Conilon	Brazil
SH3	Leaves from water deficit stressed plants (drought resistant clone)	Conilon	Brazil

Additional File 2: Number of contigs composed from sequence originated from one or more libraries. The inset details contigs present in more than 16 libraries.



<i>Coffea arabica</i>		CT	AG	AT	AC	TG	CG
Total Contigs with SNPs	4,535						
True SNPs	18,390						
Transitions	10,142	5,234	4,908	-	-	-	-
Transversions	6,078	-	-	1,573	1,325	1,493	1,687
Indels	2,126						
Frequency Transitions	55.28%	51.61%	48.39%				
Frequency Transversions	33.13%			25.88%	21.80%	24.56%	27.76%
Frequency Indels	11.59%						
Tri-allelic Pol	44						
Tetra-allelic Pol	0						
Total length	5,121,760						
Frequency of SNPs+Indels	0.35906						
Frequency SNPs	0.31669						
Frequency INDELS	0.04151						
SNPs+Indels per Contig	4.05513						
SNPs per Contig	3.57663						
INDELS per Contig	0.46880						
# Haplotypes	11,288						
Haplotypes/Contig	2.489085						
<i>Coffea canephora</i>		CT	AG	AT	AC	TG	CG
Total Contigs with SNPs	2,000						
True SNPs	4,724						
# Transitions	2,384	1,211	1,173	-	-	-	-
# Transversions	1,588	-	-	358	356	468	406
# Indels	727						
% Transitions	50.73%	50.80%	49.20%				
% Transversions	33.79%			22.54%	22.42%	29.47%	25.57%
% Indels	15.47%						
Tri-allelic Pol	25						
Tetra-allelic Pol	0						
Total length	2,077,254						
Frequency SNPs+Indels	0.22742						
Frequency SNPs	0.19121						
Frequency INDELS	0.03500						
SNPs+Indels per Contig	2.36200						
SNPs per Contig	1.98600						
INDELS per Contig	0.36350						
# Haplotypes	5,360						
Haplotypes/Contig	2.68						

Additional File 4: Annotation of KA/KS ratio in *Coffea* spp. contigs.a): Annotation of Top 20 *C. arabica* contigs with highest and lowest KA/KS ratio

High KA/KS						
Sequence	KS	KA	KA/KS	Firts Hit (BlastX-NR)	E-value	Annotation
Contig9578	0.004	0.0094	2.0952	emb CAO71103.1 unnamed protein product [<i>Vitis vinifera</i>]	2.00E-62	Major intrinsic protein (MIP) superfamily
Contig4156	0.003	0.0064	2.0936	ref NP_568215.2 SNG2 (Sinapylglucose accumulator 2)[<i>Arabidopsis thaliana</i>]	1.00E-133	Serine carboxypeptidase-like
Contig1735	0.006	0.0127	2.0609	gb ABK94488.1 unknown [<i>Populus trichocarpa</i>]	5.00E-81	Glutathione peroxidase
Contig12903	0.006	0.0113	1.8721	emb CAA46808.1 Rieske FeS [<i>Nicotiana tabacum</i>]	1.00E-103	Cytochrome b6-f complex iron-sulfur
Contig15568	0.002	0.0037	1.8685	emb CAO45533.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Leucine-rich repeat transmembrane protein kinase
Contig6193	0.018	0.0341	1.8471	gb ABK91930.1 Mal d 1 isoallergen [<i>Malus x domestica</i>]	7.00E-45	Major allergen Mal d/ PR10-like proteins
Contig9214	0.006	0.0102	1.8168	emb CAO65210.1 unnamed protein product [<i>Vitis vinifera</i>]	5.00E-57	Jasmonate ZIM-domain protein 1
Contig5255	0.009	0.0158	1.7112	gb ABK91930.1 Mal d 1 isoallergen [<i>Malus x domestica</i>]	1.00E-44	Major allergen Mal d/PR10-like proteins
Contig10695	0.022	0.0349	1.5664	gb AAX49391.1 OLE-3 [<i>Coffea canephora</i>]	4.00E-61	Oleosin
Contig17112	0.003	0.0052	1.5054	emb CAO40012.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-152	Beta-glucosidase
Contig2205	0.004	0.0064	1.4983	gb AAP42136.1 erg-1 [<i>Solanum tuberosum</i>]	1.00E-121	Phosphate-responsive protein (phi-1)
Contig1918	0.008	0.012	1.4876	emb CAO45507.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-69	Protein phosphatase 2C-like protein
Contig82	0.008	0.0116	1.4828	emb CAO17977.1 unnamed protein product [<i>Vitis vinifera</i>]	8.00E-71	Vegetative storage protein
Contig7582	0.013	0.0189	1.4709	emb CAO40168.1 unnamed protein product [<i>Vitis vinifera</i>]	8.00E-52	PII protein
Contig2035	0.004	0.0059	1.4164	gb AAP40022.1 callus-expressing factor [<i>Nicotiana tabacum</i>]	1.00E-97	Ethylene-responsive element binding protein ERF2
Contig2469	0.007	0.0092	1.3556	dbj BAB09523.1 unnamed protein product [<i>Arabidopsis thaliana</i>]	5.00E-58	Cytochrome b5 domain-containing protein
Contig2994	0.008	0.0102	1.3395	gb ABK92934.1 unknown [<i>Populus trichocarpa</i>]	4.00E-82	Coated vesicle membrane protein
Contig2672	0.006	0.0076	1.3247	gb ABK92454.1 unknown [<i>Populus trichocarpa</i>]	3.00E-98	5'-Methylthioadenosine Nucleosidase
Contig8174	0.004	0.0055	1.3046	dbj BAA03526.1 F1-ATPase gamma subunit [<i>Ipomoea batatas</i>]	1.00E-137	F1-ATPase gamma subunit
Contig16950	0.02	0.0263	1.2831	gb ABG73415.1 chloroplast pigment-binding protein CP29 [<i>Nicotiana tabacum</i>]	1.00E-92	Chlorophyll A-B binding protein CP29
Low KA/KS						
Sequence	KS	KA	KA/KS	Firts Hit (BlastX-NR)	E-value	Annotation
Contig6524	0.057	0.0017	0.0298	dbj BAD10939.1 14-3-3 protein [<i>Nicotiana tabacum</i>]	9.00E-133	14-4-3 protein
Contig2240	0.027	0.001	0.0354	gb EAZ34301.1 hypothetical protein OsJ_017784 [<i>Oryza sativa</i> (japonica cultivar-group)]	0	Tubulin beta-2 chain
Contig15974	0.028	0.001	0.036	emb CAO23450.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-172	Male sterility protein 2/ acyl CoA reductase
Contig5581	0.025	0.0009	0.0363	emb CAO15686.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Rubisco activase
Contig15938	0.028	0.0011	0.038	emb CAA81527.1 S-adenosyl-L-homocysteine hydrolase [<i>Catharanthus roseus</i>]	0	S-adenosyl-L-homocysteine hydrolase
Contig4350	0.038	0.0015	0.0406	emb CAO44494.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-134	Light-harvesting complex II protein 5
Contig13838	0.021	0.0009	0.0411	emb CAO66235.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Catalase
Contig5884	0.047	0.002	0.0415	gb AAD56018.1 60S ribosomal protein L10 [<i>Vitis riparia</i>]	1.00E-123	60S ribosomal protein L10
Contig2627	0.065	0.0027	0.0416	gb AAS48586.1 eukaryotic initiation factor 5A2 [<i>Capsicum annuum</i>]	2.00E-72	Eukaryotic translation initiation factor 5
Contig11187	0.016	0.0006	0.0417	emb CAA42660.1 luminal binding protein (BiP) [<i>Nicotiana tabacum</i>]	0	Luminal-binding protein
Contig3104	0.033	0.0014	0.0433	gb EAZ04358.1 hypothetical protein OsI_025590 [<i>Oryza sativa</i> (indica cultivar-group)]	1.00E-107	Putative secretory carrier-associated membrane protein 1
Contig6370	0.039	0.0017	0.0441	emb CAN79984.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-124	LHCA4 (Photosystem I light harvesting complex gene 4)

Contig12505	0.023	0.001	0.045	gb ABV80356.1 phosphoenolpyruvate carboxylase [<i>Gossypium hirsutum</i>]	0	Phosphoenolpyruvate carboxylase
Contig6753	0.057	0.0025	0.0451	emb CAO66090.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-64	Iron-sulfur cluster assembly complex protein
Contig1269	0.018	0.0008	0.0456	gb ABP98813.1 chloroplast biotin carboxylase [<i>Gossypium hirsutum</i>]	0	Biotin carboxylase
Contig8981	0.023	0.001	0.0458	emb CAO22101.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-173	60S ribosomal protein L4/L1
Contig4659	0.024	0.0011	0.0465	emb CAN80621.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-139	Beta-1.3-glucanase
Contig9099	0.031	0.0015	0.0475	ref NP_177596.1 NRP1 (NAP1-RELATED PROTEIN 1) [<i>Arabidopsis thaliana</i>]	1.00E-90	NRP1 (Nap1-related protein 1)
Contig4086	0.014	0.0007	0.0476	emb CAO61278.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Leucine Rich Repeat family protein
Contig3271	0.041	0.002	0.0484	emb CAO71073.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-81	Glutamate binding protein

b) Annotation of Top 20 *C. canephora* contigs with highest and lowest KA/KS ratio

103

High KA/KS						
Sequence	KS	KA	KA/KS	Firts Hit (BlastX-NR)	E-value	Annotation
Contig2864	0.0069	0.00905	1.30796	gb AAT40548.1 Putative vicilin, identical [<i>Solanum demissum</i>]	1.00E-151	Vicilin/ globulin
Contig386	0.0047	0.0061	1.2914	gb AAL35365.1 ascorbate peroxidase [<i>Capsicum annuum</i>]	1.00E-134	Ascorbate peroxidase
Contig3937	0.0061	0.00749	1.23655	gb ABK93197.1 unknown [<i>Populus trichocarpa</i>]	3.00E-54	Membrane steroid-binding protein
Contig2694	0.0163	0.01959	1.20048	gb AAF31403.1 putative glycine-rich RNA binding protein 3 [<i>Catharanthus roseus</i>]	3.00E-38	Glycine-rich RNA binding protein-like
Contig1112	0.0038	0.00446	1.18711	gb ABK95575.1 unknown [<i>Populus trichocarpa</i>]	1.00E-170	Aminopeptidase N
Contig3653	0.0039	0.00446	1.13298	gb AAQ94896.1 putative N-methyltransferase [<i>Coffea canephora</i>]	0	<u>Dimethylxanthine Methyltransferase</u>
Contig6678	0.0032	0.00362	1.13002	gb AAL37719.1 AF413204_1 beta-mannosidase [<i>Solanum lycopersicum</i>]	0	Beta-mannosidase enzyme
Contig175	0.0029	0.00311	1.07339	emb CAO24398.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-123	Fasciclin-like arabinogalactan protein
Contig5988	0.0099	0.01001	1.00908	—		No Hits Found
Contig2645	0.003	0.00292	0.97927	dbj BAA22813.1 CND41, chloroplast nucleoid DNA binding protein [<i>Nicotiana tabacum</i>]	1.00E-166	Nucleoid DNA-binding protein cnd41-like protein
Contig3544	0.0033	0.00308	0.94748	emb CAN68737.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-107	7S globulin 2 precursor small subunit
Contig4994	0.007	0.00628	0.89978	emb CAO65935.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-94	Beta-adaptin
Contig1021	0.0089	0.00779	0.87871	gb ABB13620.1 USP-like protein [<i>Astragalus sinicus</i>]	3.00E-58	Universal stress protein family protein
Contig1581	0.0184	0.01616	0.87812	emb CAN65185.1 hypothetical protein [<i>Vitis vinifera</i>]	3.00E-57	Small heat-shock protein
Contig409	0.01	0.00857	0.8546	emb CAO40936.1 unnamed protein product [<i>Vitis vinifera</i>]	7.00E-48	Bet v I allergen family protein/ PR10-like proteins
Contig1866	0.0102	0.00853	0.83941	emb CAO40936.1 unnamed protein product [<i>Vitis vinifera</i>]	2.00E-49	Bet v I allergen family protein/ PR10-like proteins
Contig8158	0.0034	0.00282	0.82975	gb ABK94910.1 unknown [<i>Populus trichocarpa</i>]	1.00E-111	E3 ubiquitin-protein ligase PRT1
Contig1010	0.0049	0.00405	0.8204	gb AAP03998.1 EIL2 [<i>Nicotiana tabacum</i>]	0	Ethylene-insensitive3-like1
Contig3075	0.007	0.00475	0.67981	emb CAO15071.1 unnamed protein product [<i>Vitis vinifera</i>]	8.00E-38	Zinc finger (AN1-like) family protein
Contig3473	0.0082	0.0055	0.67422	gb AAM63420.1 unknown [<i>Arabidopsis thaliana</i>]	4.00E-43	MD-2-related lipid recognition domain-containing protein
Low KA/KS						
Sequence	KS	KA	KA/KS	Firts Hit (BlastX-NR)	E-value	Annotation
Contig5300	0.0428	0.00179	0.04179	gb ABK96261.1 unknown [<i>Populus trichocarpa</i> x <i>Populus deltoides</i>]	1.00E-125	Peroxisomal membrane protein-related
Contig3566	0.0238	0.00118	0.04963	emb CAO24361.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Methionine aminopeptidase
Contig8165	0.017	0.00086	0.05052	gb ABB87123.1 aspartic protease precursor-like [<i>Solanum tuberosum</i>]	9.00E-133	Aspartic proteinase
Contig4689	0.0509	0.00277	0.05441	gb ABQ11264.1 mago nashi-like protein 1 [<i>Physalis pubescens</i>]	2.00E-76	Mago Nashi like protein
Contig8253	0.035	0.00205	0.05867	emb CAO40052.1 unnamed protein product [<i>Vitis vinifera</i>]	6.00E-90	60S ribosomal protein L19
Contig6691	0.0181	0.00112	0.06173	dbj BAA05641.1 chalcone synthase [<i>Camellia sinensis</i>]	0	Chalcone synthase
Contig7328	0.024	0.00165	0.06904	emb CAO61870.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-62	Chloroplast photosystem II 22 kDa
Contig3171	0.0154	0.00114	0.07425	emb CAI47559.1 alpha galactosidase [<i>Coffea arabica</i>]	0	Alpha galactosidase
Contig8251	0.0125	0.00095	0.07596	gb AAL99198.1 UTP:alpha-D-glucose-1-phosphate uridylyltransferase [<i>Solanum tuberosum</i>]	0	UTP:alpha-D-glucose-1-phosphate uridylyltransferase
Contig3901	0.015	0.00115	0.07662	gb ABF61806.1 alcohol dehydrogenase [<i>Dimocarpus longan</i>]	0	Alcohol dehydrogenase
Contig944	0.0187	0.00144	0.07722	emb CAO70082.1 unnamed protein product [<i>Vitis vinifera</i>]	8.00E-72	Nodulin MtN3 family protein
Contig8189	0.0477	0.00387	0.08126	dbj BAA34348.1 elongation factor-1 alpha [<i>Nicotiana paniculata</i>]	0	Elongation factor-1 alpha
Contig3388	0.0215	0.0018	0.084	emb CAO70406.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-37	Putative AP2/EREBP transcription factor

Contig4389	0.0238	0.00205	0.08607	gb AAO85557.1 photosystem I subunit XI [<i>Nicotiana attenuata</i>]	4.00E-84	Photosystem I subunit XI precursor
Contig1320	0.0157	0.00147	0.09373	emb CAO65178.1 unnamed protein product [<i>Vitis vinifera</i>]	2.00E-47	Zinc finger (C3HC4-type RING finger) family protein
Contig3416	0.024	0.00226	0.09405	gb AAM18501.1 N-methyltransferase [<i>Coffea arabica</i>]	0	<u>3,7-dimethylxanthine N-methyltransferase</u>
Contig6826	0.0122	0.00117	0.09613	dbj BAD34459.1 flavanone 3-hydroxylase [<i>Eustoma grandiflorum</i>]	0	Flavanone 3-hydroxylase
Contig2822	0.0087	0.00085	0.09744	gb AAC61844.1 tyrosine/dopa decarboxylase [<i>Papaver somniferum</i>]	0	Tyrosine decarboxylase
Contig1066	0.0089	0.00089	0.09935	emb CAN70603.1 hypothetical protein [<i>Vitis vinifera</i>]	0	Vacuolar-processing enzyme precursor (VPE)
Contig4399	0.017	0.00176	0.10343	emb CAN79985.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-113	Alpha-expansin precursor

<i>Coffea arabica</i> ^a	Hits ^b	% Hits ^c	<i>Coffea canephora</i> ^a	Hits ^b	% Hits ^c
Pfam00069, Protein Serine/Threonine kinase	235	2.38%	Pfam00069, Protein Serine/Threonine kinase	127	2.32%
Pfam00067, Cytochrome P450	182	1.84%	Pfam00067, Cytochrome P450	72	1.31%
Pfam07714, Protein tyrosine kinase	151	1.53%	Pfam07714, Protein tyrosine kinase	68	1.24%
Pfam00076, RNA recognition motif	102	1.03%	Pfam00076, RNA recognition motif	60	1.10%
Pfam04554, Extensin-like region	99	1.00%	Pfam03171, 2OG-Fe(II) oxygenase superfamily	43	0.78%
Pfam07172, Glycine rich protein family	82	0.83%	Pfam00071, Ras family	39	0.71%
Pfam00106, Short chain dehydrogenase	63	0.64%	Pfam00106, Short chain dehydrogenase	38	0.69%
Pfam00153, Mitochondrial carrier protein	59	0.60%	Pfam07172, Glycine rich protein family	34	0.62%
Pfam00083, Sugar transporter	58	0.59%	Pfam00005, ABC transporter	33	0.60%
Pfam00179, Ubiquitin-conjugating enzyme	58	0.59%	Pfam00153, Mitochondrial carrier protein	31	0.57%
Pfam00201, UDP-glucuronosyl and UDP-glucosyl transferase	53	0.54%	Pfam00450, Serine carboxypeptidase	30	0.55%
Pfam00005, ABC transporter	51	0.52%	Pfam00248, Aldo/keto reductase family	28	0.51%
Pfam00226, DnaJ domain	48	0.49%	Pfam04554, Extensin-like region	28	0.51%
Pfam01370, NAD dependent epimerase/dehydratase	48	0.49%	Pfam00004, AAA, atpase family	28	0.51%
Pfam00450, Serine carboxypeptidase	47	0.48%	Pfam00083, Sugar transporter	25	0.46%
Pfam00931, NB-ARC domain	45	0.46%	Pfam00481, Protein phosphatase 2C	23	0.42%
Pfam00071, Ras family	45	0.46%	Pfam00240, Ubiquitin family	22	0.40%
Pfam03171, 2OG-Fe(II) oxygenase superfamily	45	0.46%	Pfam00226, dnaJ domain	22	0.40%
Pfam00854, proton-dependent oligopeptide transport, POT	44	0.45%	Pfam01490, Transmembrane amino acid transporter	22	0.40%
Pfam02458, Transferase	44	0.45%	Pfam00847, AP2 domain	20	0.37%
Pfam00240, Ubiquitin family	42	0.42%	Pfam00270, DEAD/DEAH box helicase	20	0.37%
Pfam03552, Cellulose synthase	42	0.42%	Pfam00190, Cupin	20	0.37%
Pfam00004, AAA, ATPase family	41	0.41%	Pfam00141, Peroxidase	20	0.37%
Pfam00141, Peroxidase	41	0.41%	Pfam00011, Hsp20/alpha crystallin family	20	0.37%
Pfam00481, Protein phosphatase 2C	40	0.40%	Pfam00504, Chlorophyll A-B binding protein	20	0.37%
Pfam00248, Aldo/keto reductase family	39	0.39%	Pfam02458, transferase	19	0.35%

a – PFAM family identity

b – Number of ESTS present in each family

c – Percentatge of ESTs present in each family

Additional File 6: Annotation of 20 genes with the widest distribution among *Coffea* spp.
cDNA libraries

<i>Coffea arabica</i>					
Contig	#libraries	#ESTs	First Hit (BlastX-NR)	E-value	Annotation
Contig1217	30	207	gb EAZ38040.1 hypothetical protein OsJ_021523 [<i>Oryza sativa</i> (japonica cultivar-group)]	0	Polyubiquitin
Contig9379	30	234	gb ABK92924.1 unknown [<i>Populus trichocarpa</i>]	1.00E-165	Cysteine proteinase
Contig16478	29	162	gb ABK93203.1 unknown [<i>Populus trichocarpa</i>]	1.00E-163	Glyceraldehyde 3-phosphate dehydrogenase
Contig16878	29	245	gb AAY26520.1 secretory peroxidase [<i>Catharanthus roseus</i>]	1.00E-166	Peroxidase
Contig3635	28	148	emb CAO63006.1 unnamed protein product [<i>Vitis vinifera</i>]	8.00E-99	Aquaporin 1
Contig3702	28	147	emb CAA66667.1 polyubiquitin [<i>Pinus sylvestris</i>]	0	Polyubiquitin
Contig1691	27	203	No hits Found		
Contig3648	27	217	sp P43396 MT1_COFA Metallothionein-like protein 1 (MT-1)	3.00E-07	Metallothionein
Contig1691	27	203	No hits found		
Contig4777	26	108	gb ABK94573.1 unknown [<i>Populus trichocarpa</i>]	0	eIF4-gamma/eIF5/eIF2-epsilon domain-containing protein
Contig3524	26	301	emb CAA85426.1 catalase [<i>Nicotiana glauca</i>]	0	Catalase
Contig13370	25	194	emb CAI56307.1 sucrose synthase [<i>Coffea canephora</i>]	0	Sucrose synthase
Contig9414	25	81	dbj BAA34348.1 elongation factor-1 alpha [<i>Nicotiana glauca</i>]	0	Elongation Factor 1
Contig6243	24	123	emb CAN62488.1 hypothetical protein [<i>Vitis vinifera</i>]	0	Heat shock protein 90
Contig9342	24	79	emb CAN69723.1 hypothetical protein [<i>Vitis vinifera</i>]	6.00E-85	Eukaryotic translation initiation factor 5A
Contig16384	24	77	gb ABF47216.1 cathepsin B [<i>Nicotiana glauca</i>]	1.00E-142	Cathepsin B-like cysteine proteinase
Contig11332	24	107	emb CAN81694.1 hypothetical protein [<i>Vitis vinifera</i>]	0	Heat shock protein 70
Contig2078	24	58	gb AAQ63462.1 calmodulin 8 [<i>Daucus carota</i>]	4.00E-79	Calmodulin
Contig1870	24	119	emb CAN72774.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-153	YT521-B-like protein
Contig16384	24	77	gb ABF47216.1 cathepsin B [<i>Nicotiana glauca</i>]	1.00E-142	Cathepsin B-like cysteine proteinase
Contig9342	24	79	emb CAO64503.1 unnamed protein product [<i>Vitis vinifera</i>]	6.00E-85	Eukaryotic translation initiation factor 5A
Contig10847	24	111	emb CAA58474.1 methionine synthase [<i>Catharanthus roseus</i>]	0	Methionine synthase
Contig11332	24	107	emb CAO21681.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Heat shock protein 70
<i>Coffea canephora</i>					
Contig	#libraries	#ESTs	First Hit (BlastX-NR)	E-value	Annotation
Contig7932	9	22	gb ABP65665.1 VTC2-like protein [<i>Actinidia chinensis</i>]	0	GDP-L-galactose: hexose 1-phosphate guanylyltransferase
Contig559	8	87	emb CAN72774.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-148	YTH2 protein; Pseudouridine synthase
Contig2001	8	66	gb AAD03341.1 ubiquitin [<i>Pisum sativum</i>]	0	Ubiquitin
Contig2882	8	31	emb CAN74796.1 hypothetical protein [<i>Vitis vinifera</i>]	0	Chaperonin
Contig3120	8	39	gb AAC33305.1 fiber annexin [<i>Gossypium hirsutum</i>]	1.00E-128	Annexin
Contig6320	8	63	emb CAN73572.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-139	Single-stranded nucleic acid binding R3H
Contig6424	8	96	gb ABK92924.1 unknown [<i>Populus trichocarpa</i>]	1.00E-166	Papain-like cysteine proteinase
Contig6667	8	30	gb AAB39248.1 NADP-isocitrate dehydrogenase [<i>Eucalyptus globulus</i>]	0	NADP-isocitrate dehydrogenase
Contig7234	8	25	emb CAN68309.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-94	Tetratricopeptide domain-containing Thioredoxin
Contig8231	8	77	emb CAI56307.1 sucrose synthase [<i>Coffea canephora</i>]	0	Sucrose synthase
Contig5136	7	30	emb CAO68932.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-180	Adenosine kinase isoform 2S
Contig3668	7	23	gb AAA33697.1 1-aminocyclopropane-1-carboxylate oxidase [<i>Petunia x hybrida</i>]	1.00E-140	ACC oxidase
Contig1417	7	54	emb CAC80550.1 cyclophilin [<i>Ricinus communis</i>]	3.00E-78	Cyclophilin
Contig5950	7	23	emb CAO17373.1 unnamed protein product [<i>Vitis vinifera</i>]	0	Shaggy-related protein kinase alpha
Contig5037	7	51	gb ABK95178.1 unknown [<i>Populus trichocarpa</i>]	1.00E-83	Ubiquitin conjugating enzyme
Contig3525	7	30	gb ABG33750.1 cysteine protease [<i>Hevea brasiliensis</i>]	0	Cysteine proteinase
Contig811	7	47	emb CAO69769.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-154	ADP, ATP carrier-like protein
Contig3198	7	42	gb ABK94655.1 unknown [<i>Populus trichocarpa</i>]	1.00E-169	Elongation factor 1 gamma-like protein

Contig6702	7	17	dbj BAB68527.1 14-3-3 protein [<i>Nicotiana tabacum</i>]	1.00E-128	14-3-3 protein
Contig4522	7	41	gb EAZ23241.1 hypothetical protein [<i>Oryza sativa</i>]	0	Translation elongation factor 2
Contig1340	7	17	emb CAO39543.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-100	pre-mRNA cleavage factor im, 25kD subunit
Contig978	7	26	emb CAK22271.1 40S ribosomal protein S11 [<i>Chenopodium rubrum</i>]	6.00E-74	Ribosomal protein S11

Additional File 7: Annotation of 20 genes with the highest expression among *Coffea* spp. cDNA libraries.

<i>Coffea arabica</i>					
Contig	#libraries	#ESTs	First Hit (BlastX-NR)	E-value	Annotation
Contig16809	22	493	emb CAD11991.1 rubisco small subunit [<i>Coffea arabica</i>]	5.00E-93	Rubisco small subunit
Contig5072	20	388	emb CAO17297.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-89	Sterol desaturase
Contig13384	15	383	No hits found		
Contig4415	17	376	emb CAJ43737.1 class III chitinase [<i>Coffea arabica</i>]	7.00E-63	Class III chitinase
Contig1271	4	367	gb AAK15088.1 AF240005_1 2S albumin [<i>Sesamum indicum</i>]	3.00E-09	2S albumin
Contig13751	18	333	dbj BAA14339.1 cyc02 [<i>Catharanthus roseus</i>]	8.00E-14	Antimicrobial peptides precursor
Contig660	11	322	emb CAJ43737.1 class III chitinase [<i>Coffea arabica</i>]	1.00E-120	Class III chitinase
Contig3524	26	301	emb CAA85426.1 catalase [<i>Nicotiana plumbaginifolia</i>]	0	Catalase
Contig14309	23	273	gb AAV44205.1 unknow protein [<i>Oryza sativa</i> (japonica cultivar-group)]	2.00E-25	Sucrose synthase
Contig2929	13	272	emb CAA36249.1 metallothionein [<i>Mimulus guttatus</i>]	1.00E-17	Metallothionein
Contig16878	29	245	gb AAY26520.1 secretory peroxidase [<i>Catharanthus roseus</i>]	1.00E-166	Peroxidase
Contig9379	30	234	gb ABK92924.1 unknown [<i>Populus trichocarpa</i>]	1.00E-165	Hypothetical protein
Contig3648	27	217	sp P43396 MT1_COFA Metallothionein-like protein 1 (MT-1)	3.00E-07	Metallothionein
Contig1217	30	207	gb AAD03341.1 ubiquitin [<i>Pisum sativum</i>]	0	Polyubiquitin
Contig16715	17	207	emb CAN79558.1 hypothetical protein [<i>Vitis vinifera</i>]	1.00E-41	Major allergen Mal
Contig1691	27	203	gb EES12155.1 hypothetical protein SORBIDRAFT_06g016540 [<i>Sorghum bicolor</i>]	1.00E-06	ABA/WDS induced protein
Contig16012	12	203	emb CAA36249.1 metallothionein [<i>Mimulus guttatus</i>]	2.00E-18	Metallothionein
Contig12496	15	195	emb CAA41188.1 chlorophyll a/b binding protein [<i>Nicotiana tabacum</i>]	1.00E-139	Chlorophyll a/b binding protein
Contig13370	25	194	emb CAI56307.1 sucrose synthase [<i>Coffea canephora</i>]	0	Sucrose synthase
Contig15294	7	192	No hits found		
<i>Coffea canephora</i>					
Contig	#libraries	#ESTs	First Hit (BlastX-NR)	E-value	Annotation
Contig5887	6	1395	gb AAK15088.1 AF240005_1 2S albumin [<i>Sesamum indicum</i>]	3.00E-08	2S albumin
Contig4069	5	725	gb AAC61881.1 11S storage globulin [<i>Coffea arabica</i>]	0	11S albumin
Contig2553	4	308	emb CAO69959.1 unnamed protein product [<i>Vitis vinifera</i>]	2.00E-23	Lipid transfer protein
Contig2650	5	256	emb CAJ40777.1 alpha galactosidase precursor [<i>Coffea arabica</i>]	1.00E-179	Alpha Galactosidase
Contig1953	6	216	No Hits Found		No Hits Found
Contig6917	5	212	ref NP_190972.1 photoassimilate-responsive protein-related [<i>Arabidopsis thaliana</i>]	3.00E-34	PAR-1 protein
Contig3726	1	190	No Hits Found		No Hits Found
Contig2403	6	188	dbj BAB90396.1 ADP-ribosylation factor [<i>Oryza sativa</i> (japonica cultivar-group)]	6.00E-99	ADP ribosylation factor
Contig2495	6	176	gb AAY26520.1 secretory peroxidase [<i>Catharanthus roseus</i>]	1.00E-166	Peroxidase
Contig7356	6	173	emb CAD11990.1 rubisco small subunit [<i>Coffea arabica</i>]	7.00E-85	Rubisco small subunit
Contig890	2	168	ref NP_179721.1 mannose 6-phosphate reductase [<i>Arabidopsis thaliana</i>]	1.00E-118	Mannose 6-phosphate reductase
Contig2549	6	151	sp P43396 MT1_COFA Metallothionein-like protein 1 (MT-1)	2.00E-07	Metallothionein
Contig1103	5	150	emb CAA95858.1 S-adenosyl-L-methionine synthetase 3 [<i>Catharanthus roseus</i>]	0	SAM synthase
Contig3776	4	150	ref NP_566847.1 unknown protein [<i>Arabidopsis thaliana</i>]	2.00E-86	Hypothetical protein
Contig3742	4	132	emb CAJ43737.1 class III chitinase [<i>Coffea arabica</i>]	5.00E-64	Class III chitinase
Contig4591	2	128	emb CAO49414.1 unnamed protein product [<i>Vitis vinifera</i>]	1.00E-106	WRKY family transcription factor
Contig3494	5	127	emb CAA66109.3 specific tissue protein 2 [<i>Cicer arietinum</i>]	7.00E-05	Hypothetical protein
Contig6863	7	126	gb ABB29942.1 S-adenosyl methionine synthase-like [<i>Solanum tuberosum</i>]	0	SAM synthase
Contig7290	6	124	emb CAA85426.1 catalase [<i>Nicotiana plumbaginifolia</i>]	0.00E+00	Catalase
Contig6466	4	118	gb ABK92757.1 unknown [<i>Populus trichocarpa</i>]	1.00E-117	Mob1-like protein

Additional File 8: Annotation of selected differentially expressed genes in coffee EST libraries according to hierarchical clustering analysis. Worksheet CA: *C. arabica* contigs; Worksheet CC: *C. canephora* contigs. Libraries: Tissues and organs used in the libraries construction; Nomenclature: code of the library; Contig ID: Contig number; Annotation: automatic annotation based in AutoFACT results.

Arquivo disponível no endereço: www.lge.ibi.unicamp.br/~vidal/S8.xls

Additional File 9: Results concerning some genes related with drought abiotic stress (Dehydrins, LEAs, Metallothioneins).

Dehydrins are extensively characterized as proteins expressed during drought stress [1]. However, Hinniger et al. [2] isolated and characterized dehydrins expressed during *C. canephora* and *C. arabica* fruit development. For both species they found that dehydrin CcDH2 and CcDH1 are expressed during the final stages of grain development, but CcDH1 are also detected in the pericarp, leaves and flowers. CcContig7329 corresponds to CcDH1a isoform and, according to expression clustering analysis, it seems to be preferentially expressed in leaves (Additional File 7). CcContig1448, which corresponds to CcDH2, is mostly expressed in SE3 library (Middle stage seeds; Additional File 7), coinciding with previous data [2]. Other desiccation tolerance-related gene characterized by those authors was a *LEA* (Late Embryogenesis Abundant) [3] detected during a brief period of mid-stage development. CcContigs 1491 and 7919 were also preferentially expressed in SE3 library (Additional File 7; Figure 6A).

Metallothioneins (MTs) are small Cys-rich proteins that bind essential and non-essential heavy metals. MTs are involved in zinc (Zn) homeostasis and have antioxidant function [4,5]. There is evidence that MTs scavenge oxygen free radicals and avoid DNA damage and lipid peroxidation [4]. In *C. arabica* 6 MTs were found to be preferentially expressed in libraries from plants treated with arachidonic acid (AA) (Additional File 7). AA is a polyunsaturated fatty acid (PUFA) present in pathogens, such as oomycete *Phytophthora* spp. AA has toxic effects, which are associated with mitochondrial damage and lipid peroxidation that can induce program cell death in plants [6]. It was suggested that zinc has a protective role against AA toxicity by inducing MT that could scavenge ROS, alleviating the stress [7]. In this scenario, the amount of MTs

expressed in plants treated with AA can be a consequence of a protective signaling cascade against damaging effects of such substance.

REFERENCES

1. Allagulova Ch R, Gimalov FR, Shakirova FM, Vakhitov VA: **The plant dehydrins: structure and putative functions.** *Biochemistry (Mosc)* 2003, **68**(9):945-951.
2. Hinniger C, Caillet V, Michoux F, Ben Amor M, Tanksley S, Lin C, McCarthy J: **Isolation and characterization of cDNA encoding three dehydrins expressed during *Coffea canephora* (Robusta) grain development.** *Ann Bot* 2006, **97**(5):755-765.
3. Wise MJ, Tunnacliffe A: **POPP the question: what do LEA proteins do?** *Trends Plant Sci* 2004, **9**(1):13-17.
4. Bourdineaud JP, Baudrimont M, Gonzalez P, Moreau JL: **Challenging the model for induction of metallothionein gene expression.** *Biochimie* 2006, **88**(11):1787-1792.
5. Freisinger E: **Plant MTs-long neglected members of the metallothionein superfamily.** *Dalton Trans* 2008(47):6663-6675.
6. Knight VI, Wang H, Lincoln JE, Lulai EC, Gilchrist DG, Bostock RM: **Hydroperoxides of fatty acids induce programmed cell death in tomato protoplasts.** *Physiol Mol Plant Pathol* 2001, **59**(6):277-286.
7. Perez MJ, Cederbaum AI: **Metallothionein 2A induction by zinc protects HEPG2 cells against CYP2E1-dependent toxicity.** *Free Radic Biol Med* 2003, **34**(4):443-455.

Additional File 10: RALF and RALF-like peptides in *Coffea* spp. In magenta: dibasic sites; In Yellow: cysteine residues

A) RALF Peptides

Coffea arabica

CaContig13668

MGWMMMPGMARSGLVGEDDGVEFELDSESNRRILATTRYISYGALQKNSVPCSSRGQSYYNCRPGAPANPYSRGCSAITRCRS

CaContig4015

MLVSFWAVGDAASGSHELSEYFFPAVTTSTASFENGGSIESCIMSEQEELEMDSETNRRILYWRRRYISYSALTRDRVPCSSRGQSYYNCRPGRPVNPYNRGCAITRCRR

Coffea canephora

Contig3558

MANSSSLSTLLFALSLLTALVLSSTTVVSASGGDHYDAAQMGWMMMPGMARSGLVGEDDGVEFELDSESNRRILATTRYISYGALQKNSVPCSSRGQSYYNCRPGAPANPYTRGCSAITRCRS

Contig742

MVKPSAGLFLISATLFAATMLVSFWAVGVAASGSHELSEYFFPAVTTSTASFENGGSIESCIMSEQEEEGD
DDDQEELEMDSETNRRILYWRRRYISYGALTRDRVPCSSRGQSYYNCRPGRPVNPYNRGCAITRCRR

Contig4772

MAKSGLAGKNNGGEFKLDSKSNNGILATTRYIS*GALQNRAPCSRRGKFYYNCRPGAPVNPYTRGCRAITRCRSKNFRTSIHLAKSFGFPPFPLGGK

CC00-XX-PP1-077-C04-TL.F

MSVLDLNSMKNGLDAMVKRACAGKMSDCPTVSLEEEEEEMDSESHRMLLMRRFISYDTLRRDFAPCN
RPGSSYYNCKGAGPVNTYNRGCEIITRCDRGD

Contig3823

MMSLYDAADDVVVDNDDMEMDDDAVSSRRSLFWRRVRYIISYAALSANRIPCPRSGRSYYTHHCYFASGPVHPYNRGCSAITRCRR

B) RALF-like Peptides

Coffea arabica

CaContig203

MEKSASKSLCIFPVVAWLQLTTTVLLSATSISQSVNASVSWDWDGSDTVGSTVVADDQEFLLMDSQFGNVLAS

GGSNVYRALGRKPI^CNNARYAN^CLGAGANGRP^CDYTNRC^CAKH

CA00-XX-RM1-050-F02-AC.F

MEKSASKSLCIFPVVAWLLLLITTVLLSATSISQSVNASVSWDWDNSTVGSTVADDQEFLLVDSQFGNVLAVP
QORYVT^{RR}VLQPPP^I^CDRTRYAN^CIQPGANQRP^CDLHNRC^CARHI*Coffea canephora*

CcContig7693

WSPSKSKSLCIFPAVAWLLLLTTIVLLSATSISQSVNGSISWDWGNSTIGSTTAGDQEFLLMDSQFGNVLAS
^{RR}GVAYRVLGRKPI^CNNPRYAN^CIGAGANGRN^CGYDNR^CLRHS

CcContig15

MEKSASKSLCIFPVVAWLLLLTTTLLLSATSISQSVNGSVSGWDWDGSDTVGSTVVADDQEFLLMDSQFGNVLAS
GVSTSKVPLQKGPFC^SRSLYYNH^CIQRFGRPDKDREC^CDYTNH^CGRQSPH

CcContig5266

KSKSKSKSLCIFPVVAWLLLLTTTLLLSATSISQSVNGSVSWDWDGPTVGSTVVVDDQEFLLVDSQFGNVL
AVPPRGKLSYRGLEQPAI^CGLAVYYH^CIQRFGRPDKDREC^CLYREL^CRH

Contig2070

MEKSAPKSLCIFPVVAWLLLLTTTVLLSAASISQSVNGSVSWDWDNSTVGSTVADDQEFLLMDSQFGNVLASG
GSNVYRALQRKPF^CDNARYAN^CIGAGAKANGSP^CRFSDHC^CRHNVG

CC00-XX-LF1-040-H01-TL.F

MEKSAPKSLCIFPVVAWLLLLTTTVLLSAASISQSVNGSVSWDWDNSTVGSTVADDQEFLLMDSQFGNVLASG
GSNVYRALQRKPF^CDNARYAN^CIGAGAKANGSP^CRFSDHC^CRHNVG

Additional File 11: *Coffea* spp. OrthoMCL families of Glycine Rich Proteins (GRP). In Yellow: cysteine residues; Underlined: signal peptide for secretion

A) Family1231 – Class I-like GRPs

Coffea arabica

CA00-XX-CB1-036-E04-MC.F

MAVVLMITSEVAAKSVDNSKTIVETNEEGEAKYHGGGYGGGHGGGYGGGHGGHGGGYGGGHGGHGGHGG
GYGGHPGEGNGDGHGGYGGGGHGGYGHGGGSHGGYGHGGHGGGGHGGHPGEAADAEPQN

CaContig13520

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETNEEGEAKYHGGGHGGGHGGGYGGGHGGHGGGYGG
GHGGGHGGHGGGYGGHPGEGNGDGHGGYGGGGHGGYGHGGGGHGGYGHGGHGGGGHGGHPGEAADAEP L
N

CA00-XX-CS1-066-A03-CC.F

ISLHFHGFQDTSFLFHFPGSSNDHLRGGYGGGHGGGHGGGYGGGHGGGHGGHGGGYGGHPGGGNGDGH
GGYGGGGHGGYGHGGGGHGGYGHGGHGGGGHGGHPGEAADAEPQN

CaContig11073

IRYHTQFTSISHHHGGGHGGGYGGGHGGHGGGYGGGHGGGHGGHGGGYGGHPGEGNGDGHGGYGGGGH
GYHGGGGHGGYGHGGHGGGGHGGHPGEAADAEP LN

CaContig13384

MGSKTLLFFFISLAVVLMITSEVAAKSVDNSKTIVETNKEGEAKYHGGGYGGGHGGGHGGGYGGGHGGH
GHHGGGYGGHPGGGNGDGHGGYGGGGHGGYGHGGGGHGGYGHGGHGGGGHGGHPGEAADAEPQN

CaContig14011

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETKEEGEAKYHGGGYGGHGGGYGGGHGGYGGGGH
GHPGEAADAEPQ

CaContig16626

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETNEEGEAKYHGGGYGGGHGGGHGGGYGGGHGGH
GHHGGGYGGHPGGGYGGGHGGYGGGGHGGYGHGGHGGGGHGGHPGEAADAEPQN

CaContig5866 (?)

MSSMKILFFCISLALVLMITSQVAARELVEITSNSVDNSKTDEANGLKEAKYPGGYGGYPGGGYGGYPGG
GYGGYPGGYEGYPGGGYGGYPGGRYGGYPGGRYGGYPGGGRGGYGGNCRFGCCGRNYGGGCRCCYYPG
QAVDAEPQN

CaContig14011

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETKEEGEAKYHGGGYGGHGGGYGGGHGGYGGGGH
GHPGEAADAEPQN

CaContig16172

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETNEEGEAKYHGGGYGGGHGGGHGGGYGGGHGGY
GHHGGGYGGGHGGYGGGGHGGYGHGGHGGGGHGGHPQAAGAEPQN

CaContig8936

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETNEEGEAKYHGGGYGGGHGGGYGGGHGGYGGHH
GGYGGGHGGYGGGGHGGYGHGGHGGGGHGGHPGEAADAEPQN

CA00_XX_CB1_108_B01_RF_F

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTIVETNEEGEAKYLTVTTHLTSRHTMFP IHQSC HQFTRS
HHLMTTTHQFTRSHQFTRSHHLMVTTHLTSRHIMFPSHQSC HQYTRSHHLMVTTHLTSRHGGHGGGYGGG
HGGGHGGHGGGYGGHPGEGNGDGHGGYGGGGHGGYGHGGGSHGGYGHGGHGGGGHGGHPGEAADAEPQN

CA00_XX_PA1_013_C02_EC_F

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTVETNEEGEAKYHGGGGHGGYGHGGHGGGGHGGHPGEA
ADAKPLN

Coffea canephora

CC00-XX-SH3-001-A09-EM.F

SHVWYGGGGHGGYGHGGGGHGGYGPADYAHGRHGGGGHGGHPGEAADAHPQT

CcContig2017

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTVETNEEGEAKYHGGGYGGGHGGGYGGHGGGYGGGHGG
HHGGGYGGHPGEGNGDRHGGYGGGGHGGYGHGGHGGGGHGGHPGEAANAKPQN

CcContig376

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTVETNEEGEAKYHGGGYGGGHGGGYGGHGGGYGGGHGG
HHGGGYGGHPGEGNGDRHGGYGGGGHGGYGHGGGGHGGYGHGGHGGGGHGGHPGEAADAKPQN

CcContig1840

MGSKTLLFFFISMAVVLMITSEVAAKSVDNSKTVETNEEGEAKYHGGGYGGGHGGGYGGHGGGYGGGHGG
HHGGGYGGHPGEGNGDRHGGYGGGGHGGYGHGGGGHGGYGHGGGGHGGYGHGGHGGGGHGGHPGEAADAK
PQN

CcContig6984

MGSKTLLFFCISLAIVLTIASQVAARELAETTTSAENSKTDETTGVEEGKYGGGYGGYGGYPGYGGYGGR
GGYGGYGGRGYGGYGGRGYGGYGGRGYGGYGGRGYGGYPGGGYGGRGYGGYPGGGYGHGGYPGQA
VDAEPQN

B) Family4011 - Class II-like GRPs

Coffea arabica

CA00-XX-CA1-004-H01-EZ.F

MGSKAII¹LLCL²LA³AVLMI⁴ASEVTARD⁵LA⁶ENTNAA⁷EKST⁸EGL⁹EESK¹⁰Y¹¹GGG¹²GR¹³YGGGG¹⁴GH¹⁵YGGGG¹⁶GH¹⁷YGGG¹⁸
GGHYGGGG¹⁹GH²⁰YGGGG²¹GG²²YHG²³CC²⁴GGGG²⁵YGG²⁶CR²⁷CCT²⁸YAGEPKDAGYTEPETK²⁹PQ

CaContig5329

MGSKAII¹LLC²LLA³AVLMIASEVTARDLAENTNAAEKSTEGLEESKYGGGGGRYGGGGGRYGGGGH⁴C⁵YGGH⁶
CGGGGGGGGHYGGGGGG⁷C⁸NHG⁹CC¹⁰GGGGYGG¹¹CR¹²CTYAGEPKDAGYTEPETKPQN

CaContig6646

MSSKAI¹LLLC²LLAAVLMIASEVTARDLAENTNAAEKSTEGLEESKYGGGGGCHGYGCGGGGGGCHGYGCH
GGGGGGGGGGHCYHGCGGHGYGGCRCC³TYAGEPKDAGYTEPETTKPQN

CaContig2625

MGSKAII LLLCLLA AVLMIASEVTARDLA ENTNAAEKSTEGLEESKYGRGGG CYGRG CGGGGGGGGR CYHGC
CGGGYGGRCRCC TYAGEPKDAGYTEPETKPN

CA00-XX-LP1-021-G09-EB.F

MGSKAII^{LL}CLLA^{AV}LMIA^{SE}VTARD^{LA}ENTNAA^{EK}STEG^{LE}ESKYGGGG^{CH}GYG^{CG}GGGGG^{GH}CYGG^{HC}
GGGGGGG^{GH}CYHG^{CC}GHGYGG^{CR}CC^{CC}TYAGEPKDAGYTEPETK^{PQ}

CaContig16496

MGSKAII~~LL~~CLLA~~AV~~LMIASEVTARDLAENTNAAEKSTEGLEESKY~~Y~~GGGGGGGGRG~~C~~YGRG~~C~~GGGGGG~~C~~Y
HG~~CC~~GGGGGYGYGHGG~~CR~~~~CC~~TYAGEPKDAGYTEPETKPQN

CaContig1435

MGSKAII^{LL}CL^{LA}AV^{LM}IA^{SE}VT^{AR}DL^{AE}NT^{NA}AE^{KS}TE^{GL}ES^{SK}YGR^{GGG}CYGR^GCGGGGGGG^GCHYG^C
GGGGGGGGH^GCYH^GCCGGGYGG^{CR}CC^{CC}TYAGEPKDAGYTEPE^TETEP^QN

CaContig3765

MGSKAILLLCLLAAVLMIASEVTARDLAENTNAAEKSTEGLEESKY^{YGGGGHYGGGGGHYGGGGGHYGGG}
GGHYGGGGGHYGGGGGHYGGGGGCYNGCCHGYGGCRCTYAGEPKDAGYTEPETKPQN

CA00-XX-IA2-030-G11-EC.F

LAENTNAGEKSNEGLEESKYGGGGGCHGYGCGGGGGGHHCYHGCGGHGYGGCRCCTYAGEPKDAGYTEPE
TKPQN

CA00-XX-LP1-002-E03-EB.F

MGSKAII¹LLLC²LLAAV³LMIASEVTARD⁴LAENTNAAEKSTEGLEESKYGRGGG⁵CHGYG⁶CGGGGGGWWPWLLP⁷
WLLWRRLWRLQMLHICW⁸

C) GRPs differentially expressed in *C. arabica* plantlets treated with AA containing 12 Cys

CaContig10126

MGSKAILLCLLA AVLMIASEVTARDLAENTNAAEKSTEGLEESKYGRGGG CYGRG CGGGGGG CYGGH CG
GGGGGGH CYGGH CGGGGGGGH GYHG CC GGGYGG CRCC TYAGEPKDAGYTEPETKPQN

CaContig1089

MGSKAILLCLLA AVLMIASEVTARDLAENTNAAEKSTEGLEESKYGGGGG CHYG CGGGGGG CHYG CG
GGGGGGH CYGGH CGGGGGGGH GYHG CC GHGYGG CRCC TYAGEPKDAGYTEPETKPQ

CaContig3317

MGSKAIFLLCLLA AVLMIASEVTARDLAENTNAAEKSTEGLEESKYGRGGGG CYGGH CGGGGG CYGGH CG
GGGGGGH CYGGH CGGGGGGGH GYHG CC GGGYGG CRCC TYAGEPKDAGYTEPETKPQN

Additional File 12: *Coffea* spp. OrthoMCL families of cystatins: In green: variation of LARFAV motif; In yellow: new motif GG-X-YY; In blue: QVVAG motif

Family544

Coffea arabica

CA00_XX_AR1_001_B01_EB_F

MAAAKFAIGTGQTDISSLEPVKPADPHVIQIGKFVEQHHHGKLLSVAVVGGFTWSGDG
GNYIALIIENQDSDGATCLHKHKVLVLETPSETKLIWHKK

CaContig113

MSTVAARSATPAIGAGQKNMVGVPVPMASSTVKRTDPGVIGIANFAVEKYNERNETALA
VINVEFGFLWPHGGHYYYMLAIITQDDKGTHHDVAYVRDAGKSNAHAYEFMWYNHNNN

CA00_XX_LP1_007_B05_EB_F

MDQVPVNPEDVDRIGRFAVNEENRKRPNQLTFVHVYAFKGSAGEDKIYPLIIKIRDVN
DKPFKHKALVLEKTDGSLNLKGYWE

CA00_XX_RT8_064_F07_EQ_F

MAAAKSAIGTGKNDISSLEPVKPADRHVIQIGEFVVEQCHHGQLLFVAVVGGFTWSGDG
GYYIALIIENQDSEGATYLHKALVLETPNETKLIWHKK

CaContig8767

MAAAKSAIGTGKNDISELEPVKPADPHVQIGEFVVEQCHHGQLLFVAVVGGFTWSGDG
GYYIALIIENQDSDGATYLHKALVLQTPSETKLIWHKK

CaContig10690

MAAAKSAIDTGKNDISSLEPVKPADPHVIQVGQFVVEQCYHGQLLFVAVVGGFTWSGDG
GYYIALIIENQDNDGATYLHKALVLETPSETKLIWHKK

CA00_XX_RX1_054_H06_EB_F

MPGQIDVDGLVPVKPTDPPVIAIGKFAVEEYKQKQPIEIVAVVSGFTGSGDGGNYILLI
IETQDSNGAIFLHKALVFKDTNGGLKVKGWGF

CaContig13279

MAAVAANFPVAGVAKNPMQGLKPALVVGALNQLAGQKQGQGNAAVPDDWTPVNPLDRHI
QELGAFAVDEHNKQTKDQLVFVAVLSGIQKTEDDRSTYCLLISAKDSTGKLGRYYAVII
EYNTGCQQLLQFEPSP

CaContig11025

MAAAKSAIGTGQTDISSLEPVKPADPHVIQIGRFVVEQAHHGKLLFVAVVGGFTWSVIG
GNYIALIIENQDYEGATYLHKALVFETPDGVLELIWHKK

CaContig5403

MAAAKSAIGTGQTDISSLEPVKPADPHVIQIGQFVVEQAHHGKLLFVAVVGGFTWSVIG
GNYYALIIENQDYEGATYLHKALVFETPDGVLTLIWHKK

CaContig16944

MAEAKSATVTDQIDINSIQPVAPADPHVVGIGQFVVEKFHHGKLLFIAVLGGFTWKCEG
GKYYALIIQNQDYEGATFIHKALVVEAKGETKLLWHRN

CaContig4522

MAEAKSATVTDQIDITSIQPVAPADPHVVGIGQFVVEKFHHGKLHFIAVIGGFTWNCEG
GKYYALIIQNQDYEGATFIHKALVVEAKGETKLLWHRN

CaContig16895

MAAAKSAIGTGQTDISSLEPVKPADPRVIQIGQFVVEQAHHGKLLFVAVVGGFTWSVIG
GNYYALIIENQDYEGATYLHKALVFETPDGVLTLIWHKK

CaContig8410

MAEAKSATVTDQIDINSIQPVKPADPRVVEIGQFVVEKFHHGKLLFIAVLGGFTWKCEG
GKYYALVIENQDYEGATFIHKALVVEAPGETKLLWHKN

CaContig4566

MAAAKSAIGTGQTDISSLEPVKPADPHVIQIGQFVVEQAHHGKLLFVAVVGGFTWSVIG
GNYYALIIENQDYEGATYLHKALVFETPDGVLTLIWHKK

CaContig12045

MAKFSVDKYNEEAGTKLVFMKVIACALWNLGVVTVYALLIQTQDSKGTYIDKAVAVDVT
IIGKKLLWYKH

CA00-XX-RT8-047-A07-EP.F

MAAAKSAIGTGKNDISSLEPVKPADPHVIQIGQFVVEQCHHGQLLFVAVVGGFTWSGDG
GYYYALIIENQDSDGATYLHKALVLETPSETKLIWHMK

CaContig15921

MVGVPVPMAS TVKRTDPGVIGIANFAVEKYNERNETALAVINVEFGFLWPHGGHYYYM
LAIITXDDKGTHHDVAYVRDAGKTMLTLMNSCGTIITIIDLALLIS

CaContig8137

MAAAKSAIGAGKNDIDALEPVKPADPRVIEIGRFAVTEHGHALLFVG VGGFRWAIPGG
DHYALIIETQDDNGATYLHKALVVMVEVEGQPLRLIWKYK

CaContig17257

MAAAKSAIGTGKIDISSLEPVKPADPHVIQIGKFVEQQHHHGKLLCVAVVGGFTWSGDG
GNYYALIIENQDSDGATYLHKHKVLVLETPSEMCLIWHKK

Coffea canephora

CcContig2160

MAEAKSATVTDQIDINSIQPVAPADPHVVGIGQFVVEKFHHGKLLFIAVLGGFTWKCEG
GKYIALIIQNQDYEGATFIHKALVVEAKGETKLLWHRN

CcContig4504

MATVAAKSATAAIGAGQKNMVGGLSSTVPPRSSTVNPKDPHVIQIAQFAVANYNAKAG
TTVVWLNVEYGFWWIDDDTYYMLAIKTQDLTGTHCDVALVREISESNGTYSLKWYNHNN
K

CC00-XX-PP1-087-G12-TL.F

MDQVPVNPEDVDRIGRFVNEENRKRPNQLTFVHVYAFKGSAGEDKIYPLIIKIRDVN
DKPFKHKALVLEKTDGSLNLKGYW

CcContig3825

MSTVAARSATPAIGAGQKNMMGGGVSCIIPPATTVKVEDACVIEIAKFAVAQITGRVFI
KVEFGFWWKIEIGPNA GTYYMLAIITQDNNRTHCDVALVCDLETSNGHTLIWYNDKNN

CcContig7886

MAEAKSATVTDQIDINSIQPVAPADPRVAEIGQFVVEKFHHGKLLFIAVIGGFTWKCEG
GKYIALIIQNQDYEGATFIHKALVVEAPGETKLLWHRN

CcContig1043

MVGGLSSTVPPRSSTVNPKDPHVIQIAQFAVANYNAKAGTTVVWLNVEYGFWWIDDDT
YYMLAIKTQDLTGTHCDVALVREISESNGTYSLKWYNHNNK

Family2703

Coffea arabica

CaContig1058

MTEVIANYNISVNEFAANMAVEGFQSAEVEAIMKAVGENKTWNAIEGLSDTNANLRGLC
GTTTAQNVDKTVPPDVQE MAEFVAEYNR IAGTKLVLIKVLAYVKLVVVFGT FYGLHML
TQDDKGYKDQALTLKLNNGMKVLLWYKHN

CaContig4053

MAVTAKCQKTELANNYVKQFQSAEVDAILKQAGETKLIVHGGWTPVNPADPHIQE LGRF
AVDEHN KQTGDKLVFVAVVAGLKKPVELATLYWLI IEAKSDGNQNIYKALVQETDLEM
KKLLYFGEVPPVN

CaContig5345

MTEVIANYNINVNEFAANMAVEGFQSAEVEAIMKAVGENKTWNAIEGLSDTNANLQGLC
GTTTAQNVDKTVPPDVQE MAEFVAEYNR IAGTNLVLIKVLAYVKRVVVFGTLYRLHML
TQDDKGIHNDQALTLKLKNGKKVLLSYKHN

CaContig4160

MAAAKSGIGSGQKDEPIIPMASTVNPNDVVIQ KAKFAVD SYNGQAGTGLKFNSVEFGF
CWSVSDVTDYLLAINTHDDKGPYCDPALVSDTLKSNAHTYELIWYNHKKK

CaContig7667

MATVAAKSATAAIGAGQKNMVGGLSSTVPPRSSTVNPKDPHVIQ IAQFAV ANYNAKAG
TTVVWLNVEYGFWWIDDDTYMLAIKTRDLTGTHCDVALVREISESNGTYSLKWYNHNN

CaContig7242

DPHIQELGRFAVNEHNRQTRDKLVFVAVVAGLKKPVELATLYWLI IEAKDRNGNQNIYK
A

Coffea canephora

CcContig6730

MTEVVANYNINVNEFAANMAVEGFQSAEVEAIMKAVGENKTWNAIEGLSDTNANLQGLR
GTTTAQNVDKTVPPDVQE MAEFVAEYNR IAGTNLVLIKVLAYVKRVVVFGTLYGLHML
TQDDKGIHHDQALTLKFKNKKVLLWYKHNNH

CcContig4176

MTEVIANYNINVNEFAANMAVEGFQSTEVEAIMKAVGENKTWNAIEGLNDTNANLQGLC
GTTTAQNVDKAVPPDVQE MAEFVAEYNR RAGTKLVLIKVLRYVKRVVVFGT FYGLHML
TQDDKGYKDQALALKFKNGKKVLVWYKHNN

CC00-XX-SH3-075-F10-EM.F

MAAAKSGIGSGQKDQPIIPVASTVKPKDDKVIEAAQFAVVTYNKQAGTDLVCINVEFGF
WWSITGATYYMLAIKTQDAKGTychvalvadvLVSGGNHTYDliWYNHKN

Family942

Coffea arabica

CaContig2092

MAAVVANPHINITEITANMKAEGVQSPEIEAIVKALSDDTIWKTIEGFKGKDMSTQEKM
INNMVAGGHLPQVGVP LPTPVNPTDPHVISVAKFALAKYNDKHG TKLVFN RVNGGLQWK
IVIGTLYILVLATQDSKGT YTDYAVVFETFLGQKYLFWYKH

CaContig2323

MAAVAANFPVAGVAKNPMQGLKPALVVGALKQLAGQKQGQGNAAVPDDWTVVSTLDRHI
QALGAFAVDEHNKQTKNQLVFVAVLSGIKKTEDDRSTYCLLISAKNSTGKLGSYNAVII
EYNTGCQQLLQFEESP

CaContig 13328

MDLVPVNPAEPHVTAIGQFAVD EENKKRPTNKLN FVAVVGGYHGPVTGATRYPLILGTQ
DGKGHTFLHKALVHEKPDGSLELKGYW

CaContig5297

MASAFPHLLLLTTLAAICLFSDVPSAALGGRPKDALVGGWSKADPKDPEVVE NGKFAVD
EHNKEAKTKLEFKTVVEAQQQV VAGTNYKIVIKALDGTASNLYEAIWVKPWLKFKKLT
SFRKLP

CaContig15270

MASAFPHLLLLTTLAAICLFSDVPSAALGGRPKDALVGGWSKADPKDPEVVE NGKFAID
EHNKEAGTKLEFKTVVEAQEQV VAGTNYKIVIKALDGTASNLYEAIWVKPWLKFKKLT
SFRKLP

CaContig14147

MDMCDEDFVVTGGGKDTKLVGIAGVPLPKPVDKTSPHVIKIAQFAVKKHNEKAGTKLVF
IKVVGGVKWSAIAGT FYALQIETQDSKGT YRDKTLVVEAVTGHKKLIWYKH

CaContig3848

MTEVTVNYNFNITEVAANMAVEGFQSAEVEAIMKTAGDDMIWNAIEDTKDMDMCDEDF
VTGGGKDTKLVGIAGVPLPKPVDKTSPHVIKIAQFAVKKHNEKAGTKLVF IKVVGGVKW
SAIAGT FYALQIETQDSKGTHRD KTLVVEAITGHKKLIWYKH

CA00-XX-IA2-005-D11-EC.F

MAAVVANYNINISEITANMKAEGVQSPMEAILKATAEDAIWNTIERFKGMDMSNKKKM
 INNRMGSGGRAQLGIPLPEPVNPTDPHVIAIAKFAVEKHNEAGTSLVFIQVIGGLQWN
 LLIGALYMLIITTQDSKGTYYDKTVVFETCLGQKYLWYKH

Coffea canephora

CcContig1026

MASAFPHLLLLTTLAAICLFSDVPSAALGGRPKDALVGGWSKADPKDPEVLENGKFAID
 EHNKEAGTKLEFKTVVEAQEQVVAGTNYKIVIKALDGTASNLYEAIWVVKPWLKFKKLT
 SFRKL

CcContig6451

MTEVTVNYNFNITEVAANMAVEGFQSVEAEAIMKTAGDDMIWNAIEDTKDMDMCDEDF
 VTGGGKDTKLVGIAGVPLPKPVDKTSPhVIKIAQFAVKKHNEKAGTKLVFIKVVGGVKW
 SAIAGTFYALQIETQDSKGTHRDKTLVVEAITGHKKLIWYKH

CcContig7844

MAAVVANPHINISEITANMKAEGVQSPEIEAIVKALSDDTIWNTIEGFKGKDMSTQEKM
 INNMVAGGHLPPQGVPLPTPVNPTDPHVISVAKFAVAKYNDKHGTKLVFN RVNGGLQWK
 IVIGTLYILILATQDSKGTYYTDYAVVFETFLGQKYLFWYK

Additional File 13: *Coffea* spp. OrtoMCL families of PinII serine proteinase inhibitors

Family7241

Coffea arabica

CA00_XX_CL2_115_D10_JF_F
MAINKIGAMAILFCGMILLGANVEVTAVRPGPEQICPLYCIVGIEYVDCDGEKTYTDCT
NCCFENGCTLHFKDGTSYFCTWPAKQELGFGKGVYKI

CaContig12344
MAINKIGAMAILFCGMILLGANVEVKAVRPGPGPVCPQYCILGIEYVDCDGEKIYTDCT
NCCLSEGCTLHFDTGTEEYCEPVGKGVYKI

CaContig5418
MMAVNKIGAMAILFCGMILLGANVEVTAVRPGPDQICPLYCIVGIEYIVCDGEKIYTDCT
TNCCFANGCTLHFDTGTSYYCTWPAQQELGYGKGVYKI

CaContig13131
MAINKIGAMAILFCGMILLGANVEVTAVRPGPEQICPLYCIVGIEYVDCDGEKTYTDCT
NCCFENGCTLHFKDGTSYFCTWPAKQELGFGKGVYKI

CaContig7989
MAVNKIGAMAILFCGMILLGANVEVTAVRPVPQICPLYCILGIEYVVCDGEKTYKGCTN
CCFENGCTLHFFEDGTEKYCTWPTEQKLGLANIMLNNMPF

Coffea canephora

CC00_XX_PP1_063_C07_TL_F
MAVNKIGAMVILFCGMILLGANVEVTAVRPGPEQICPLYCIVGIEYVDCDGEKTYTDCT
NCCFENGCTLHFKDGTSYFCTWPAKHELGFGKGVYKI

Family10273*Coffea arabica*

CA00_XX_CA1_003_B05_EZ_F

MAINKIGAMAILFCGMILLGANVEVKAVRPGPVRPCPRNCIGGTLYQICNGTKTYTTCT
NCCVSDGCTLYFLDGSSLYCDWPDACY

CaContig6030

MGINKIGAMAILFCGMILLGANVEVKAVRPGLLQPCPRNCIGGTVFQICNGTKTYTTCT
NCCVSNNGCTLYFLDGSSLYCDWPDACY

CaContig2158

MAINKIGAMAILFCGMILLGANIEVKAVRQAPLRPCPRNCIGGTVYKVCNGTKTYTDCT
NCCVSDGCTLYFEDGSSLYCDWPYAKY

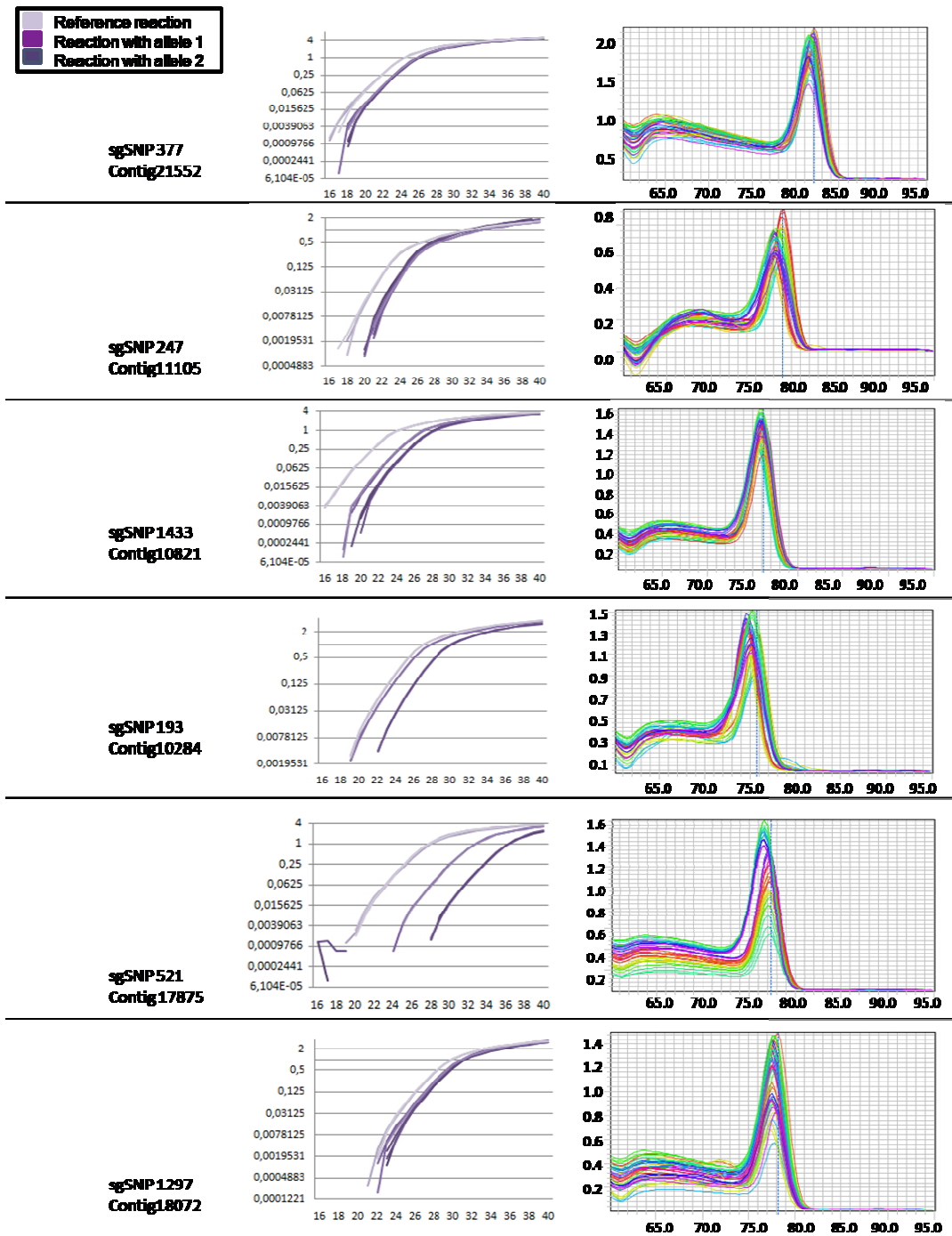
CaContig14018

MILLSSNVEVKVVEACPQYCLDVEYMTCGNSETKLPPRCNCCLAPKGCTLHLADGTSQY
CS*Coffea canephora*

CcContig3974

MAISKIGAMAILFCGMILLGANVEVMAVRPGPIRPCPLICLLTEYKICNGTKTYTNCTN
CCVDDGCTLYFEDGSSSIYCEWPWAKY

ANEXO II – Material Suplementar do Capítulo II



Supplemental Figure S1. Amplification plot and melting curve of sgSNPs by qPCR.

		Leaves	Bud Flowers	Seeds	Callus	Root	Suspension Cells
MLPs	Paralog 1	20 0	23 0				
	Paralog 2	0 9	0 11				
THI1	Paralog 1	33 4		0 0			
	Paralog 2	0 0		1 17			
Osmotin	Paralog 1	12 0		11 0	12 0		
	Paralog 2	0 19		0 0	0 0		
FLP	Paralog 1	4 1	3 0		3 0		
	Paralog 2	1 4	0 0		0 0		
SAMDC	Paralog 1		19 0	0 0		1 0	1 0
	Paralog 2		0 3	0 5		0 0	0 0

Supplemental Figure S2. Differential homeologous gene expression variation between paralogs in specific tissues.

Supplemental Table S1. Description of the EST libraries used in this work.

C) Confection of *Coffea arabica* cDNA libraries

Total RNA was extracted from coffee tissues at different developmental stages and also submitted to different stress conditions. Poly(A)⁺ RNA was purified from total RNA using the Oligotex Kit (Quiagen), following the manufacturer's directions. The mRNA purity and integrity were estimated by absorbance at 260/280 nm and agarose gel electrophoresis. cDNA libraries were constructed using the SuperScript Plasmid System and Plasmid Cloning Kit (Invitrogen) with about 1-2 µg poly(A)⁺ RNA. The efficiency of cDNA synthesis was monitored with radioactive nucleotides. cDNA were size fractionated on a Sepharose CL-2B column. Aliquots of each fraction were eletrophoresed in agarose gel to determine the size range of cDNAs. Fractions containing cDNA larger than 500 pb were ligated into pSPORT1 and pSPORT6 vectors (Invitrogen) at the Sall-NotI site. The resulting plasmids were transformed in *E. coli* DH10B or DH5α cells (Invitrogen) by electroporation. Plasmid DNA was purified using a modified alkaline lysis method (Sambrook et al., 1989). Sequencing reactions were conducted using the ABI BigDye Terminator Sequencing kit (Applied Biosystems). cDNA inserts were sequenced from the 5' end with T7 promoter primer (5'-TAATACGACTCACTATAGGG-3') or M13 Rev in the pSPORT1 vector with SP6 primer (5'- ATTTAGGTGACACTATAG-3') in the pSPORT6. Sequencing reaction products were analyzed on ABI 3700 sequencers (Applied Biosystems).

D) Description of the coffee ESTs libraries

<i>Coffea arabica</i>	Library	Description	Cultivar	Source
	AR1	Leaves treated with araquidonic acid	Mundo novo	Brazil
	LP1	Plantlets treated with araquidonic acid	Mundo novo + Catuai	Brazil
	CB1	Suspension cells treated with benzothiadiazole and brassinoesteroids	Catuai	Brazil
	CL2	Hypocotyls treated with benzothiadiazole	Mundo novo + Catuai	Brazil
	EA1, IA1, IA2	Embryogenic calli	Catuai	Brazil
	EB1	Zygotic embryo	Mundo novo + Catuai	Brazil
	EM1, SI3	Germinating seeds (whole seeds and zygotic embryos)	Catuai	Brazil
	FB1, FB2, FB4	Flower buds in different developmental stages	Mundo novo	Brazil
	FR1, FR2	Flower buds + pinhead fruits + fruits at different stages	Mundo novo	Brazil
	CA1	Non embryogenic calli	Mundo novo + Catuai	Brazil
	IC1	Non embryogenic calli	Catuai	Brazil
	PC1	Non embryogenic calli + 2,4-D	Mundo novo + Catuai	Brazil
	LV4, LV5	Young leaves from orthotropic branch	Mundo novo	Brazil
	LV8, LV9	Mature leaves from plagiotropic branches	Mundo novo	Brazil
	NS1	Roots infected with nematodes	Mundo novo + Catuai	Brazil
	PA1	Primary embryogenic calli	Mundo novo + Catuai	Brazil
	RM1	Leaves infected with leaf miner and coffee leaf rust	Mundo novo	Brazil
	RT3	Roots	Mundo novo	Brazil
	RT5	Roots with benzothiadiazole	Mundo novo	Brazil
	RT8	Suspension cells with stressed with aluminum	Catuai	Brazil
	RX1	Stems infected with <i>Xylella</i> spp	Catuai	Brazil
	SH2	Water deficit stresses field plants (pool of tissues)	Catuai	Brazil
	SS1	Well-watered field plants (pool of tissues)	Catuai	Brazil
	CS1	Suspension cells with mannose NaCl and KCL	Catuai	Brazil
	BP1	Suspension cells treated with acibenzolar-S-methyl	Catuai	Brazil
	PL1	?	?	Brazil
	SI1	Germinating seeds	?	Brazil
	SI2	Germinating seeds	?	Brazil
	CD1	Suspension cells	?	Brazil
	CL1	Suspension cells	?	Brazil
	CM1	?	Mundo novo + Catuai	Brazil
	LM3	?	Mundo novo + Catuai	Brazil
	RT7	Root	Mundo novo + Catuai	Brazil
	FB3	Flower buds	Mundo novo	Brazil

	FP2	?	Mundo novo + Catuai	Brazil
<i>Coffea canephora</i>	Library	Description	Cultivar/ Varieties	
	LF1	Young leaves,	BP409	Nestlé
	PP1	Pericarp, all developmental stages	BP358,BP409,BP42, BP961,Q121	Nestlé
	SE1	Whole cherries,18 and 22 week after pollination	BP358,BP409, BP42,Q121	Nestlé
	SE2	Whole cherries,18 and 22 week after pollination	BP358,BP409, BP42,Q121	Nestlé
	SE3	Endosperm and perisperm, 30 week after pollination	BP409,BP961,Q121	Nestlé
	SE4	Endosperm and perisperm, 42 and 46 weeks after pollination	BP358,BP409,BP42, BP961,Q121	Nestlé
	EC1	Embriogenic calli	Conilon	Brazil
	SH1	Leaves from water deficit stressed plants	Conilon	Brazil
	SH3	Leaves from water deficit stressed plants (drought resistant clone)	Conilon	Brazil

Supplemental Table S2. Top 50 contigs with more ESTs derived from one *C. arabica* subgenome than the other.

Arquivo disponível no endereço: www.lge.ibi.unicamp.br/~vidal/S2.xls

Supplemental Table S3. Correlation of EST coverage of contigs and differential expression of homeologous genes.

	>6 ESTs	>20 ESTs	>30 ESTs	>50 ESTs
Equal Expression in				
both subgenomes	71%	51%	47%	40%
More Expressed in				
CaCc	8%	17%	21%	25%
More Expressed in				
CaCc	9%	12%	11%	8%
Expressed only in CaCc	5%	12%	17%	23%
Expressed only in CaCe	7%	8%	6%	4%
Number of Contigs in				
each class	2069	992	494	172

Supplemental Table S4. Manual annotation of contigs from each GO term described in Table II.

Arquivo disponível no endereço: www.lge.ibi.unicamp.br/~vidal/S4.xls

Supplemental Table S5. Sequences of primers used in allelic (homeologous) discrimination and differential expression of homeologous gene analysis by qPCR

Contig	Primer	Primers Sequence	Primers Design
Contig18072	SNP_1297_F	5'-GGTGGAGATGACTTCAAGAG-3'	Reference Primer
	SNP_1297_1F	5'-GTGGAGATGACTTCAAGAC <u>T</u> -3'	Forward Allele1 Primer
	SNP_1297_2F	5'-GTGGAGATGACTTCAAGAC <u>C</u> -3'	Forward Allele2 Primer
	SNP_1297_R	5'-GTAGCTCACAATTGAAGTTG-3'	Reverse Primer
Contig17875	SNP_521_F	5'-CAAGTGACCCAGAGCTTATAT-3'	Reference Primer
	SNP_521_1F	5'-CAAGTGACCCAGAGCTTATAG <u>A</u> -3'	Forward Allele1 Primer
	SNP_521_2F	5'-CAAGTGACCCAGAGCTTATAG <u>T</u> -3'	Forward Allele2 Primer
	SNP_521_R	5'-CCAATACCCCATCACCTTC-3'	Reverse Primer
Contig10284	SNP_193_F	5'-AAGGAAAAGTTGGTGCATC-3'	Reference Primer
	SNP_193_1F	5'-AAGGAAAAGTTGGTGCAT <u>T</u> A-3'	Forward Allele1 Primer
	SNP_193_2F	5'-AAGGAAAAGTTGGTGCAT <u>A</u> G-3'	Forward Allele2 Primer
	SNP_193_R	5'-ATCTTCAATGAGGAGGAGCT-3'	Reverse Primer
Contig10821	SNP_1433_F	5'-ATCCTCAAACAGTGGGTTG-3'	Reference Primer
	SNP_1433_1F	5'-ATCCTCAAACAGTGGGTT <u>C</u> C-3'	Forward Allele1 Primer
	SNP_1433_2F	5'-ATCCTCAAACAGTGGGTT <u>C</u> T-3'	Forward Allele2 Primer
	SNP_1433_R	5'-GGATTTCCTGCCACCATGC-3'	Reverse Primer
Contig21552	SNP_377_F	5'-CCGTCCACGGCGTTACTA-3'	Reference Primer
	SNP_377_1F	5'-CCGTCCACGGCGTTACT <u>C</u> A-3'	Forward Allele1 Primer
	SNP_377_2F	5'-CCGTCCACGGCGTTACT <u>T</u> G-3'	Forward Allele2 Primer
	SNP_377_R	5'-CAGGGAGTCTGAGCCGTCG-3'	Reverse Primer
Contig11105	SNP_247_F	5'-TTGCTCTTCGTGAAATCCG-3'	Reference Primer
	SNP_247_1F	5'-TTGCTCTTCGTGAAATCC <u>C</u> T-3'	Forward Allele1 Primer
	SNP_247_2F	5'-TTGCTCTTCGTGAAATCC <u>C</u> C-3'	Forward Allele2 Primer
	SNP_247_R	5'-TTGAAGTCCTGAGCAATTCTC-3'	Reverse Primer