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 Vidal[2], Jorge Maurício Costa Mondego[2], David Pot[2], 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, Brasilia-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; Sylvain, 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

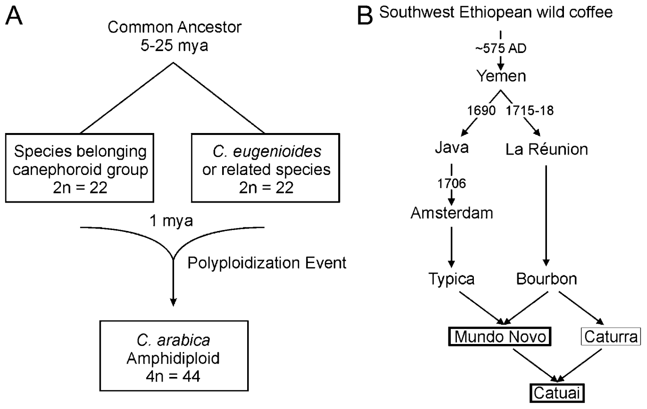

Figure 1. Evolutionary history of allotetraploid *C. arabica*. A, Origin of *C. arabica*. The progenitor genomes are represented by diploid *C. eugenioides* and *C. canephora*. *C. arabica* arose 1 to 2 million years ago (mya) from the fusion of *C. canephora* (or related species) and *C. eugenioides*. 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) assessment 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 sequences 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

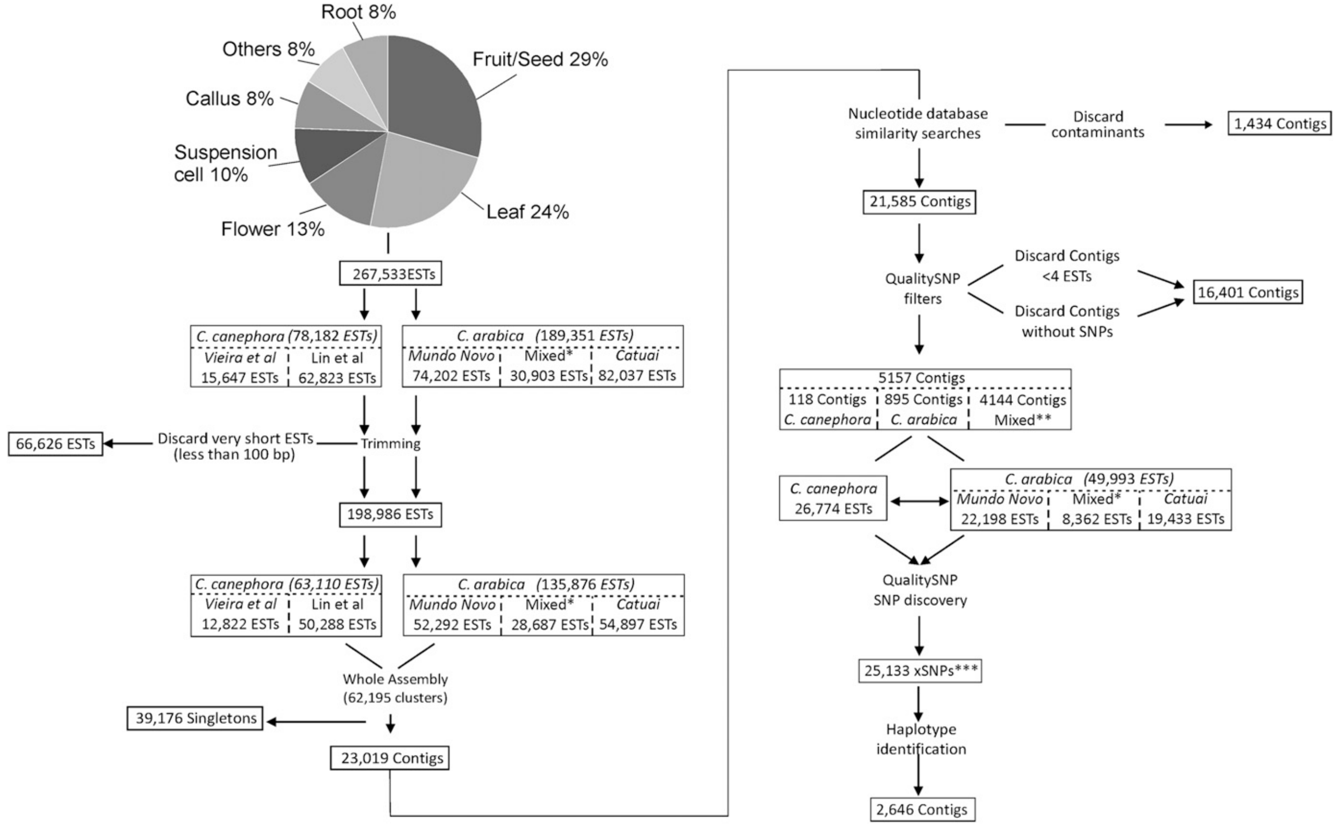

Figure 2. Flow diagram of the pipeline for data cleaning, EST assembly, SNP discovery, and analysis in *Coffea* species. * It is not possible to know the origin of these sequences because the libraries were constructed from cv Catuai and cv Mundo Novo. ** Contigs with ESTs from both *C. arabica* and *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. eugenioides* 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. eugenioides* 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. eugenioides*, and were labeled as CaCe (subgenome of the *C. eugenioides*-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

[a] Depending 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.

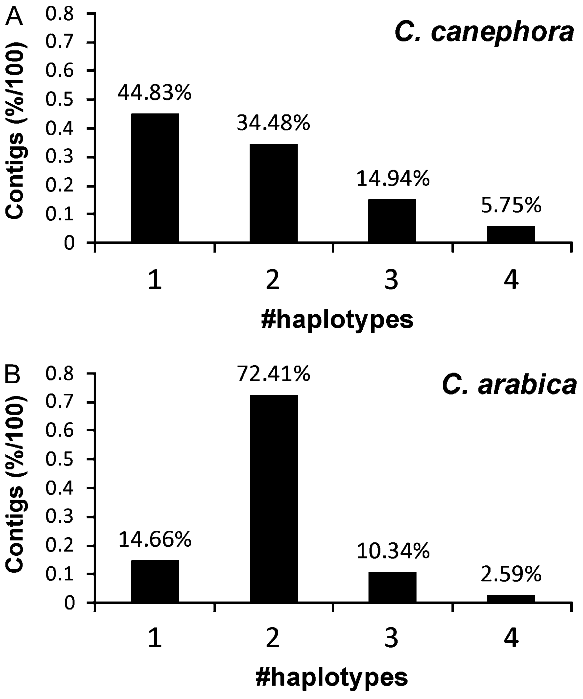

Figure 3. Variability of the number of haplotypes per contig in *C. arabica* (A; only the contigs with at least eight reads were considered) and *C. canephora* (B; only contigs with at least four reads were considered).

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. eugenioides* ESTs and mapped them in the assembly. It was possible to map 18 *C. eugenioides* ESTs in 16 of those 2,646 contigs. *C. eugenioides* 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. eugenioides* 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. eugenioides*. For all the genes analyzed, the *C. eugenioides* 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

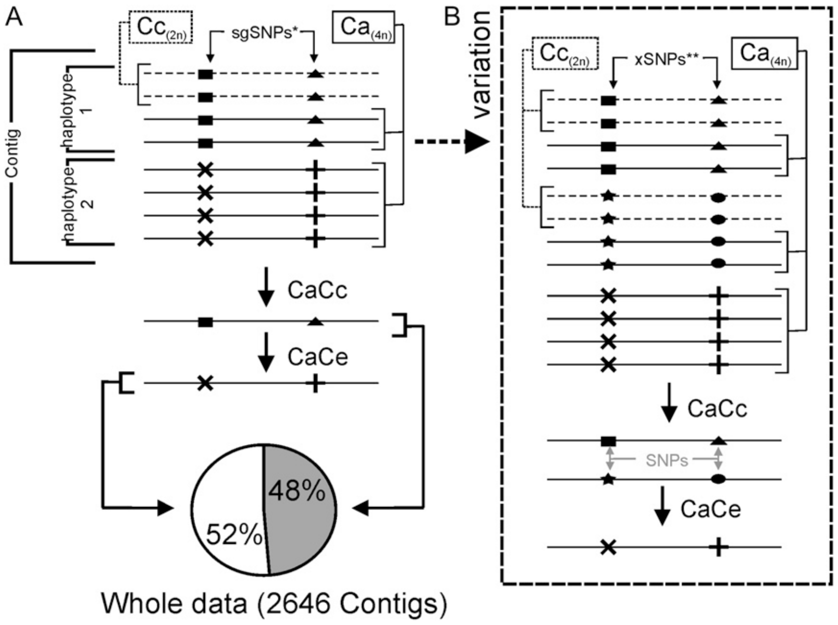

Figure 4. Identification of homeologous genes. A, Scheme showing the assembly of *C. canephora* ESTs (Cc) with *C. arabica* ESTs (Ca) into the same haplotype in the same contig. ESTs from *C. arabica* presenting the same pattern as *C. canephora* were labeled as derived from the CaCc subgenome, and the 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 transcripts 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 proteosome 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

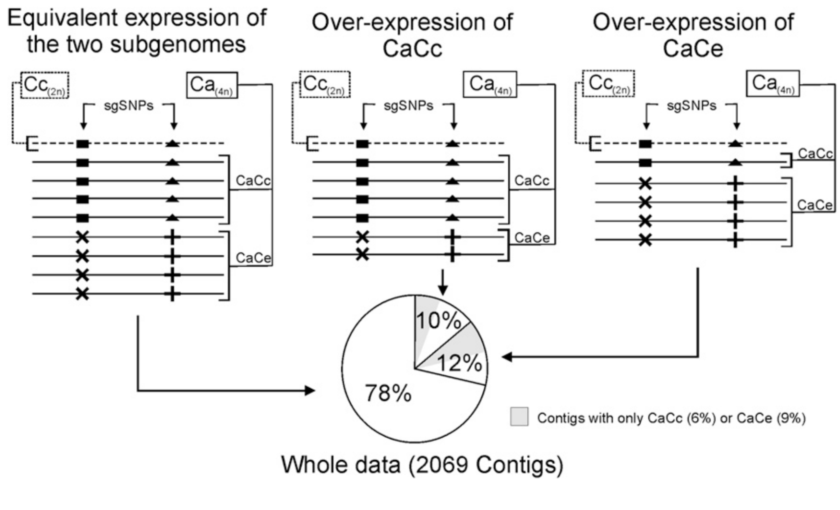
Equivalent expression of the two subgenomes
Over-expression of CaCc
Over-expression of CaCe
Cc(2n)
Ca(4n)
sgSNPs
CaCc
CaCe
10%
12%
78%
Contigs with only CaCc (6%) or CaCe (9%)
Whole data (2069 Contigs)

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.

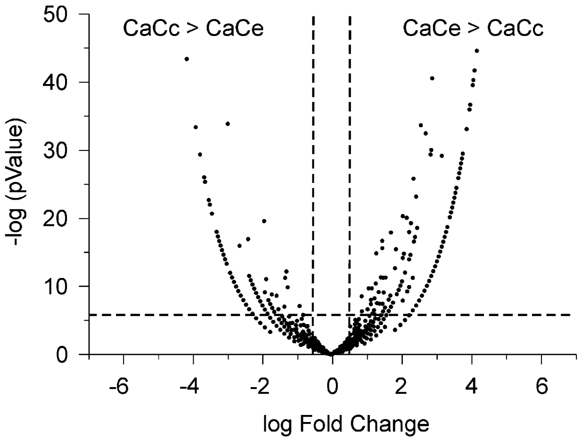

Figure 6. Volcano plot with the 2,645 contigs with CaCc and CaCe ESTs identified. The *x* axis corresponds to the fold change values calculated according to the following formulas: if the number of CaCe ESTs (#CaCe) is higher than the number of CaCc ESTs (#CaCc), the fold change is (#CaCe + 1)/(#CaCc + 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 in 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

[a]Normalized 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

[a] Protein 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, EF-1γ, 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. eugenioides* 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. eugenioides* 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. eugenioides* 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 THI1 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. eugenioides* 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):

$$\log \frac{\text{expression allele 2}}{\text{expression allele 1}} = \left(\log E_{\text{allele2}} \times \text{Ct}_{\text{allele2}}\right) - \left(\log E_{\text{allele1}} \times \text{Ct}_{\text{allele1}}\right)$$

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

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, Troggio 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

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