the plant journal

SEB Society for Experimental Biology

The Plant Journal (2011) **67**, 305–317 doi: 10.1111/j.1365-313X.2011.04590.x

Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (*Coffea*)

Qingyi Yu[1,2,*,†], Romain Guyot[3,†], Alexandre de Kochko[3], Anne Byers[1], Rafael Navajas-Pérez[4], Brennick J. Langston[2], Christine Dubreuil-Tranchant[3], Andrew H. Paterson[4], Valérie Poncet[3], Chifumi Nagai[1] and Ray Ming[1,5]

[1]*Hawaii Agriculture Research Center, Waipahu, HI 96797, USA,*
[2]*Texas A&M University, AgriLife Research Center, Department of Plant Pathology and Microbiology, Weslaco, TX 78596, USA,*
[3]*Unité Mixte de Recherche Diversité, Adaptation et Développement (UMR DIADE), Evolution et Dynamique des Génomes (EVODYN), BP 64501, 34394 Montpellier Cedex 5, France,*
[4]*Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA, and*
[5]*Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA*

Received 7 January 2011; revised 21 March 2011; accepted 29 March 2011; published online 12 May 2011.

*For correspondence (fax 956 969 5620; e-mail qyu@ag.tamu.edu).
†These authors contributed equally to this work.

SUMMARY

Arabica coffee (*Coffea arabica* L.) is a self-compatible perennial allotetraploid species (2*n* = 4*x* = 44), whereas Robusta coffee (*C. canephora* L.) is a self-incompatible perennial diploid species (2*n* = 2*x* = 22). *C. arabica* ($C^aC^aE^aE^a$) is derived from a spontaneous hybridization between two closely related diploid coffee species, *C. canephora* (CC) and *C. eugenioides* (EE). To investigate the patterns and degree of DNA sequence divergence between the Arabica and Robusta coffee genomes, we identified orthologous bacterial artificial chromosomes (BACs) from *C. arabica* and *C. canephora*, and compared their sequences to trace their evolutionary history. Although a high level of sequence similarity was found between BACs from *C. arabica* and *C. canephora*, numerous chromosomal rearrangements were detected, including inversions, deletions and insertions. DNA sequence identity between *C. arabica* and *C. canephora* orthologous BACs ranged from 93.4% (between E^a and C^a) to 94.6% (between C^a and C). Analysis of eight orthologous gene pairs resulted in estimated ages of divergence between 0.046 and 0.665 million years, indicating a recent origin of the allotetraploid species *C. arabica.* Analysis of transposable elements revealed differential insertion events that contributed to the size increase in the C^a sub-genome compared to its diploid relative. In particular, we showed that insertion of a *Ty1*-copia LTR retrotransposon occurred specifically in *C. arabica*, probably shortly after allopolyploid formation. The two sub-genomes of *C. arabica*, C^a and E^a, showed sufficient sequence differences, and a whole-genome shotgun approach could be suitable for sequencing the allotetraploid genome of *C. arabica*.

Keywords: ***Coffea,* evolution, comparative genomics, allotetraploid, genetic divergence, transposable elements.**

INTRODUCTION

Coffee is one of the most important international traded commodities, and is ranked as the second most valuable primary commodity exported by developing countries (Pendergrast, 2009). It is a popular beverage, and is mostly consumed in industrialized countries, whereas over 90% of coffee production takes place in developing countries. More than 75 million farming families worldwide rely on coffee for their livelihood (Pendergrast, 2009).

Coffee belongs to the 4th largest flowering plant family, the Rubiaceae, which consists of more than 11 000 species in 660 genera (Robbrecht and Manen, 2006). Although there are more than 100 species in the genus *Coffea* L. (Maurin *et al.*, 2007), only two species are widely used in commercial production. *C. arabica* L., also known as Arabica coffee, accounts for 75–80% of the world's production. *C. canephora* Pierre ex Froehner (Robusta coffee) represents approximately 20% of the world production. *C. arabica* produces high-quality coffee, but it is susceptible to many diseases. Inter-specific hybridization between *C. arabica* and other coffee species to obtain desired traits, such as disease resistance, has been the major strategy in coffee breeding programs (Van der Vossen, 2001).

Coffea arabica is the only tetraploid ($2n = 4x = 44$) and self-fertile species in the *Coffea* genus, while all the other *Coffea* species are diploid, and most of them are self-sterile (Charrier and Berthaud, 1985; Davis *et al.*, 2006). *C. arabica* originated in southwestern Ethiopia, and its center of genetic diversity remained there (Aga *et al.*, 2005; Silvestrini *et al.*, 2007), while *C. canephora* originated in Central and Western Africa (Ferwerda, 1976; Gomez *et al.*, 2009; Musoli *et al.*, 2009). *C. arabica* was first cultivated in Ethiopia approximately 1500 years ago. From Ethiopia, *C. arabica* was introduced to other countries and continents in the 18th century, and formed the genetic base of the two major modern Arabica coffee cultivars Typica and Bourbon (Ferwerda, 1976). It has been shown that Typica and Bourbon were derived from one or very few plants, resulting in a very narrow genetic base for Arabica coffee cultivated worldwide (Ferwerda, 1976).

Although *C. arabica* is a tetraploid species, it shows diploid-like meiotic behavior. Chloroplast DNA variation studies suggested that *C. arabica* originated from two diploid *coffea* species (Berthou, 1983; Lashermes *et al.*, 1993, 1996). Genome *in situ* hybridization revealed that *C. arabica* ($C^aC^aE^aE^a$, $2n = 4x = 44$) has allopolyploid origin, and was derived from hybridization between ancestors of two closely related diploid coffee species, *C. eugenioides* (EE, $2n = 2x = 22$) and *C. canephora* (CC, $2n = 2x = 22$) (Lashermes *et al.*, 1999).

Several BAC libraries have been constructed for both *C. arabica* and *C. canephora* as a foundation for the study of coffee genome structure, comparative genomics, developmental and evolutionary biology (De Kochko *et al.*, 2010). A BAC library of *C. canephora* with nine genome equivalents was constructed from *C. canephora* genotype 126 (Leroy *et al.*, 2005). The first BAC library of *C. arabica* was constructed from the cultivar IAPAR 59, an introgressed variety derived from the Timor Hybrid (Noir *et al.*, 2004). The second BAC library of *C. arabica* was constructed from the high-cupping quality variety Tall Mokka (R. Ming, C. Nagai and Q. Yu, unpublished data). The availability of these BAC libraries allows us to perform a comparative study between tetraploid *C. arabica* and diploid *C. canephora*, and provides an opportunity to study the evolutionary history of these two genomes.

A BAC clone from *C. canephora* was sequenced and annotated to study the sequence organization surrounding the *CcEIN4* gene (Guyot *et al.*, 2009). *CcEIN4* encodes an ethylene receptor, a key element in transduction of the signal induced by the presence of the phytohormone ethylene (Bustamante-Porras *et al.*, 2007). In this study, we identified two homeologous BACs from an Arabica coffee BAC library of Tall Mokka, and compared the sequences with the ortholog in *C. canephora* to investigate the patterns and degree of DNA sequence divergence between Arabica and Robusta coffee genomes.

RESULTS

BAC library construction

The Arabica coffee variety Tall Mokka (MA2-7) was chosen to construct a BAC library because of its superior agronomic traits and because it is the parent of a mapping population used for genetic and QTL mapping (Pearl *et al.*, 2004; R. Ming, C. Nagai and Q. Yu, unpublished data). This BAC library consists of 79 872 clones in 208 384-well plates. From this library, 461 BACs were randomly selected to estimating the average insert size using *Not*I digestion and clamped homogenous electric fields (CHEF) gel electrophoresis. The BAC inserts ranged from 10 to 270 kb with an average insert size of 93.5 kb. As the genome size of *C. arabica* is 1.28 Gb (Clarindo and Carvalho, 2010), this *C. arabica* BAC library represents approximately 5.8 genome equivalents.

Identification and sequencing of *CcEIN4* orthologous loci from the *C. arabica* Tall Mokka BAC library

A *C. canephora* BAC clone, 46C02, containing an ethylene receptor gene, was fully sequenced (Guyot *et al.*, 2009). To identify orthologous BAC clones from the *C. arabica* BAC library, DNA fragments of the *CcEIN4* gene were used as probes to screen the *C. arabica* BAC library. Twenty-one positive BAC clones were identified. The positive BAC clones were then digested with *Hin*dIII and re-probed using the same probes. Southern analysis revealed two groups of positive BAC clones that showed different patterns when identical stringency was applied. One group of positive BACs showed a strong signal and the other showed a weaker signal. Two BACs, MA29G21 (from the strong signal group) and MA38M04 (from the weak signal group), were selected for shotgun sequencing. The sequencing result revealed extensive sequence collinearity between MA29G21 and 46C02. However, MA38M04 did not share sequence similarity with 46C02 except for one small region. All the positive BACs were then end-sequenced. The BAC end sequences were compared with the sequence of *C. canephora* BAC clone 46C02. Two more rounds of primer walking were performed to identify orthologous BAC clones of 46C02 from the *C. arabica* BAC library. The sequencing result revealed that nine BAC clones with weaker signal were not orthologous BACs of 46C02, and were detected due to containing an ethylene receptor paralogous gene. The other group of positive BAC clones with strong signals was differentiated into two sub-groups based on the BAC end sequences, and each contains a copy of an orthologous sequence of 46C02.

The insert sizes of these BAC clones were estimated by CHEF gel electrophoresis. Two BAC clones, MA29G21 and MA17P03, representing two orthologous sequences of 46C02 in allotetraploid *C. arabica*, were selected for complete sequencing.

The BAC clones MA29G21 and MA17P03 were fully sequenced using a shotgun sequencing approach. The GenBank accession numbers for MA29G21 and MA17P03 are HQ834787 and HQ832564, respectively. The insert sizes of MA29G21 and MA17P03 were estimated at 145 and 140 kb, respectively. Sanger reads were generated and provided approximately tenfold coverage for each BAC clone. The initial assembly was performed using PHRED/PHRAP/CONSED software. Gap filling was performed by primer walking. The total length of MA29G21 is 145 817 bp, with a gap in a GC-rich region. The complete sequence of MA17P03 is 140 269 bp. The overall GC content of both clones is 37%.

Gene content and repetitive sequences

We annotated the sequences of BACs MA29G21 and MA17P03 by BLAST similarity searches in GenBank. A total of 16 and 14 genes were predicted in BACs MA29G21 and MA17P03, respectively. All the predicted genes were validated by RT-PCR, and the start codon, intron junctions and stop codon of each gene were verified by RT-PCR. Each of the predicted genes was then subjected to BLASTP search of the National Center for Biotechnology Information non-redundant protein sequence database to assign putative function. Detailed annotations of each predicted gene are shown in Table 1.

The average gene density of BAC MA29G21 is one gene every 9.1 kb, and that for BAC MA17P03 is one gene every 10 kb. The coding regions of predicted genes range in size from 330 to 3471 bp (annotated genes that are partially covered by the BACs MA29G21 and MA17P03 were excluded). The coding regions of predicted genes in BAC MA29G21 cover 20 355 bp and account for 14.0% of the BAC. The coding regions of predicted genes in BAC MA17P03 cover 17 871 bp, and account for 12.7% of the BAC.

We searched for repetitive sequences including retrotransposons, DNA transposons, simple repeats and low-complexity repeats in BACs MA29G21 and MA17P03. The percentage of repetitive sequences in BAC MA29G21 is 7.9% and that in BAC MA17P03 is 16.0%. In total, 32 and 29 transposable elements (TEs) were annotated in MA17P03 and MA29G21, respectively. In particular, annotation revealed the presence of a full-length LTR retrotransposon, named CART no. 109 (*Coffea arabica* retrotransposon no. 109), within BAC MA17P03. CART no. 109, which belongs to the *Ty1*-copia group, had a complete size of 3939 bp. The presence of stop codons in the open reading frames coding for gag and pol suggests that this is a functionally defective TE. Overall, these two BACs cover gene-rich regions with a low percentage of repetitive sequences.

Sequence comparison between homeologous BACs MA17P03 and MA29G21 within the *C. arabica* genome

Comparative analyses were first performed between *C. arabica* homeologous BACs MA17P03 and MA29G21. A total of 140 269 bp of MA17P03 were aligned with 124 869 bp of MA29G21, indicating a 15 400 bp (12.3%) expansion in MA17P03. These two BACs have an average identity of 93.42%. Pairwise sequence comparisons between MA29G21 and MA17P03 revealed an overall similar organization comprising stretches of highly conserved segments interrupted by limited number of regions of difference.

Table 1 List of genes confirmed by RT-PCR on two *Coffea arabica* homeologous BACs MA29G21 and MA17P03 compared to *C. canephora* orthologous BAC 46C02

Gene ID	Location on MA29G21	Location on MA17P03	Location on 46C02	Gene ontology or function of putative ortholog
Gene 1	1–10 110 (partial)	N/A	1–1684 (partial)	Myosin-like protein XIE; motor/protein binding
Gene 2	13 154–15 653	1–1916 (partial)	4504–7023	DNA binding; transcription factor
Gene 3	19 336–27 703	8265–16 473	10 420–18 626	Sequence-specific DNA binding; transcription factor
Gene 4	31 694–33 024	20 495–21 815	21 416–23 546	putative c-Myc binding protein
Gene 5	35 019–41 373	24 106–30 388	25 585–31 885	DNA-binding family protein
Gene 6	45 630–49 957	34 870–39 205	36 147–40 482	Ethylene receptor
Gene 7	51 712–55 515	41 162–44 978	42 436–46 249	Methyltransferase
Gene 8	63 507–64 376	52 593–53 462	53 763–54 632	DNA-binding protein
Gene 9	70 195–72 700	58 911–61 430	59 858–62 377	Unnamed protein product
Gene 10	74 578–75 102	63 330–63 854	64 266–64 790	Cyclophilin
Gene 11	76 484–77 839	64 965–66 320	65 900–67 255	Unnamed protein product
Gene 12	85 554–86 258	74 857–75 561	75 661–76 365	Ethylene-responsive transcription factor
Gene 13	102 953–105 244	90 164–92 455	91 454–93 745	Serine protease
Gene 14	120 218–124 077	119 384–123 244	119 888–123 745	Serine/threonine protein phosphatase
Gene 15	133 920–135 194	132 633–133 919	132 647–134 033	Phosphoenolpyruvate carboxylase kinase
Gene 16	140 673–145 817	N/A	145 486–139 804	Plant glycogenin-like starch initiation protein

In addition, all coding regions were found to be strictly conserved in the same order and orientation between homeologous BAC sequences (Figures S1 and S2).

Sequence variations between homeologous BACs were carefully analyzed. Chromosomal rearrangements were observed between MA17P03 and MA29G21, including insertions, deletions and inversions. Most of the sequence insertions and deletions observed are limited to inter-genic regions. Close inspection of four large extra segments in MA17P03, accounting for a cumulative size of 12 kb, revealed insertion of three non-autonomous transposons (blue arrows, Figures S1 and S2) and one LTR retrotransposon (red arrow, Figures S1 and S2). In addition to insertions and deletions, sequence comparison clearly showed the presence of inversions of approximately 3430 bp (positions 84 779–88 209) in MA17P03 and 4352 bp (positions 96 249–100 601) in MA29G21 (black arrow, Figures S1 and S2). This inversion occurred in an inter-genic space (between genes 12 and 13), and did not involve any coding regions.

Dot-plot alignment of the MA17P03 and MA29G21 sequences (Figure S2) revealed a region located between the inversion and the insertion of the LTR retrotransposon that has undergone numerous small-scale rearrangements (93 059–116 195 bp in MA17P03; 110 281–117 032 bp in MA29G21). The conservation of this region is altered by a complex pattern of rearrangements, involving duplications, insertions and deletions of sequences.

Sequence comparison between orthologous BACs from diploid and allotetraploid genomes

Comparative analyses were then performed between Arabica and Robusta orthologous BACs. We aligned each of the two orthologous BAC sequences from *C. arabica* (C^a and E^a sub-genomes) with their orthologous BAC sequence from *C. canephora* (C genome). The phylogenetic tree generated from the sequence comparison showed clearly that MA17P03 and 46C02 are more closely related than MA29G21 and 46C02 (Figure S3). Thus, MA17P03 was assigned as a C^a BAC and MA29G21 was assigned as an E^a BAC.

The multiple sequence alignments of *C. arabica* BACs MA29G21 and MA17P03 with *C. canephora* BAC 46C02 are shown in Figure 1. Sequence comparison between *C. canephora* BAC 46C02 and *C. arabica* BAC MA29G21 revealed similar pattern of conservation and divergence as observed between MA17P03 and MA29G21 within *C. arabica*. A chromosomal inversion was detected between homeologous BACs MA17P03 and MA29G21 and between 46C02 and MA29G21, but not between 46C02 and MA17P03. Overall, MA29G21 shares 93.75% identity with 46C02 based on the gapless alignment of 120 321 bp (Table 2). A total of 137 388 bp of *C. arabica* BAC MA29G21 aligned with 144 998 bp of *C. canephora* BAC 46C02, showing 7611 bp (5.5%) expansion in the *C. canephora* BAC 46C02 (Table 2).

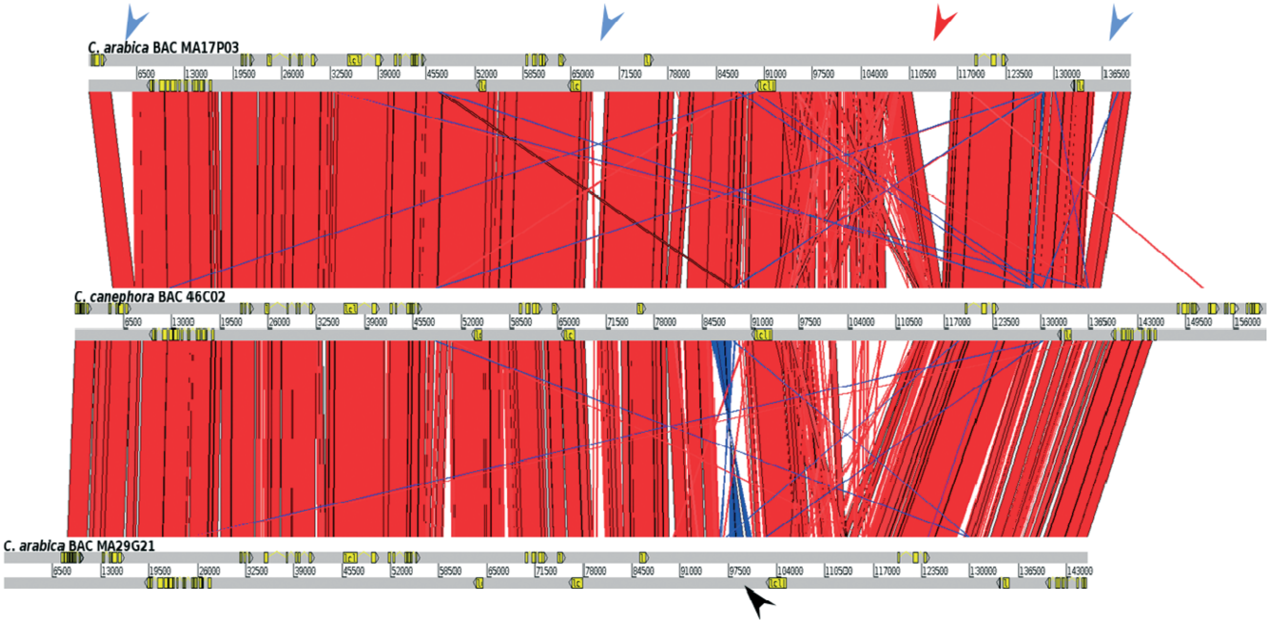

Figure 1. Sequence comparison between *Coffea canephora* (46C02) and *C. arabica* (MA29G21 and MA17P03).
The red lines link conserved regions and blank regions symbolize disrupted conservation. The blue and red arrows represent insertion of transposons and retrotransposons, respectively, in the MA17P03 BAC sequence. The black arrow indicates a chromosomal inversion. Comparisons were performed using the BLASTZ algorithm (Schwartz *et al.*, 2003) and visualized using the Artemis Comparison Tool (Carver *et al.*, 2005).

Table 2 Summary of gapless sequence comparison between *C. arabica* (MA17P03 and MA29G21) and *C. canephora* (46C02) BAC sequences

	MA17P03/46C02	MA29G21/46C02	MA17P03/MA29G21
Length of BAC (bp)	140 269/160 404	145 817/160 404	140 269/145 817
Aligned sequence (bp)	125 643	120 321	110 439
Span of aligned BAC (bp)	126 439/126 535	123 522/123 810	112 641/113 064
% alignment coverage	99.37/99.29	97.41/97.18	98.00/97.67
Average percentage identity	94.63	93.75	93.42

Comparisons between *C. canephora* BAC 46C02 and *C. arabica* BAC MA17P03 revealed a high degree of microcollinearity, with an overall sequence identity of approximately 94.6% in the aligned region covering 125 643 bp (Table 2). Sequence alignments also revealed several insertions and deletions (Figure 1 and S4). Among these insertions and deletions, four are caused by insertions of large TEs in MA17P03 (blue and red arrows, Figure S4). A total of 133 001 bp of *C. canephora* BAC 46C02 aligned with 140 269 bp of *C. arabica* BAC MA17P03, showing 7269 bp (5.5%) expansion in the *C. arabica* BAC MA17P03, probably due to the insertion of TEs (Table 2).

Altogether, based on the conservation of sequence identity and chromosomal rearrangement, the MA17P03 BAC in *C. arabica* is orthologous to BAC 46C02 from one of its diploid progenitors, *C. canephora*.

Comparative analysis of transposable elements

To verify the orthologous relationship between 46C02 and MA17P03, we compared TE accumulation in each of the three BAC clones. A total of 32, 29 and 29 TEs were annotated for C^a BAC MA17P03, C BAC 46C02 and E^a BAC MA29G21, respectively. In total, 15 TEs were conserved in all three BACs, and may represent ancient insertions prior to divergence of the ancestor of *C. canephora* and *C. eugenioides*, and thus prior to the speciation event of *C. arabica* (gray boxes, Figure 2). Most of these elements are classified as miniature-inverted transposable elements (MITE) or non-autonomous transposons. In addition, all three segments shared a remnant of a long interspersed element (LINE) upstream of the start codon of gene 4.

The *C. canephora* BAC 46C02 and C^a BAC MA17P03 sequences share eight TEs, all belonging to the MITE and non-autonomous transposon groups, whereas 46C02 and E^a BAC MA29G21 do not share any TEs that are not also conserved between 46C02 and C^a BAC MA17P03. These analyses confirm that the *C. canephora* BAC 46C02 and the *C. arabica* BAC MA17P03 are orthologous.

Seven, twelve and six insertions of transposable elements were found to be unique to MA17P03, MA29G21and 46C02, respectively. After a TE moves, it usually leaves behind footprints in the form of a short duplicated segment or a small section of the element, due to imprecise excision from the original insertion site. To test whether these unique elements were inserted in one ortholog or deleted from the others, we investigated the presence of such footprints at the orthologous insertion sites. We did not find any trace of footprints in segments where TEs were missing, suggesting that these elements that differentiate the three sequences were probably inserted after the divergence of the diploid progenitors *C. canephora* and *C. eugenoides*, and after the formation of the *C. arabica* genome.

Most of the unique transposable elements found in the three BACs fall into the MITE and non-autonomous transposon groups, with the exception of a complete LTR retrotransposon (CART no. 109) inserted in C^a BAC MA17P03. The 530 bp long-terminal repeats of CART no. 109 are approximately 99% identical, suggesting that the retrotransposon was probably recently inserted in *C. arabica* at this locus. In order to test this hypothesis, we used a retrotransposon-based polymorphism assay (Flavell *et al.*, 1998) to test for the presence of the CART no. 109 LTR retrotransposon in a collection of *C. arabica*, *C. eugenioides* and *C. canephora* genotypes.

Of the 34 genotypes analyzed, all four *C. arabica* genotypes showed the presence of the CART no. 109 insertion (Figure S5), and the remaining genotypes, which are either *C. eugenioides* or *C. canephora*, with all diversity groups as defined by Gomez *et al.* (2009) being represented, showed the absence of CART no.109. In addition to PCR, we estimated the date of the insertion of CART no. 109 using divergence of both long-terminal repeat sequences (San-Miguel *et al.*, 1998). Using a molecular clock of 7.1×10^{-9} synonymous substitutions per site per year (Ossowski *et al.*, 2010), we estimated that CART no. 109 was inserted in MA17P03 approximately 528 000 years ago.

Our analysis based on the TE content confirmed that the MA17P03 BAC in *C. arabica* is orthologous to the *C. canephora* BAC 46C02, and thus belongs to the C^a sub-genome of *C. arabica*. In addition, our results indicate that the TE CART no. 109 was inserted in the C^a sub-genome after the speciation event of *C. arabica*.

Divergence in coding regions and tentative timing of allopolyploid formation of *C. arabica*

We calculated the synonymous (K_s) and non-synonymous (K_a) substitution rates between the *C. arabica* and *C. canephora* orthologous coding regions (Table 3). Based on the annotated genes, we identified 14 complete gene pairs between E^a BAC MA29G21 and C BAC 46C02, and 13

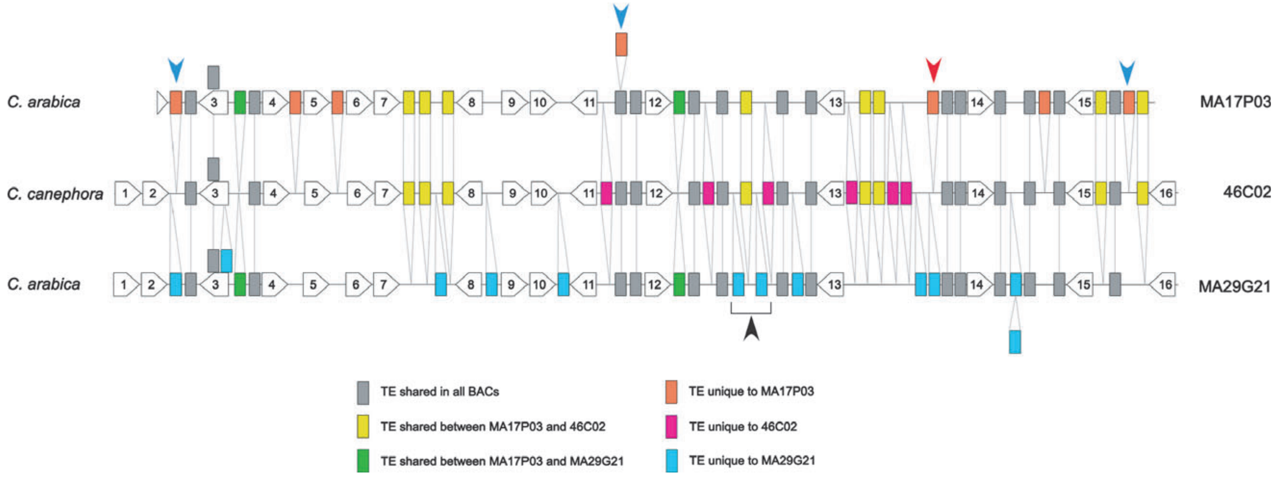

Figure 2. Schematic representation of the conservation of transposable elements between orthologous and homeologous BACs in *Coffea canephora* and *C. arabica*.

Coding regions are represented as white arrowheads, with the direction of the arrow indicating the gene orientation. Colored boxes symbolize annotated transposable elements in BAC sequences. Specific colors are used to indicate transposable elements that are unique to a specific genome or sub-genome. The blue and red arrows represent insertion of transposons and retrotransposons, respectively, in the MA17P03 BAC sequence. The black arrow indicates a chromosomal inversion.

Table 3 Estimates of synonymous and non-synonymous nucleotide divergence between *Coffea arabica* BAC MA17P03 and *C. canephora* BAC 46C02, and *C. arabica* BACs MA17P03 and MA29G21

	Number of sites (bp)				Number of mutations		Sequence divergence		
Gene ID	Total sites	Total coding sites	Synonymous sites	Non-synonymous sites	Synonymous mutations	Non-synonymous mutations	Synonymous sites (K_s)	Non-synonymous sites (K_a)	K_a/K_s
C. arabica BAC MA17P03 versus *C. canephora* BAC 46C02									
3	8207	3348	750.67	2597.33	2	2	0.00267	0.00077	0.29
4	2131	327	71.17	255.83	0	0	N/A	N/A	N/A
5	6278	1053	274.83	778.17	0	4	N/A	0.00516	N/A
6	4336	2295	521.00	1774.00	2	0	0.00385	N/A	N/A
7	3814	951	216.17	734.83	0	2	N/A	0.00273	N/A
9	2520	1218	280.17	937.83	2	1	0.00717	0.00107	0.15
14	3858	1017	229.17	787.83	0	0	N/A	N/A	N/A
15	1287	843	194.67	648.33	2	2	0.01034	0.00309	0.30
C. arabica BACs MA17P03 versus MA29G21									
3	8203	3348	751.00	2597.00	9	11	0.01208	0.00425	0.35
4	2115	327	71.17	255.83	1	2	0.01418	0.00786	0.55
5	6237	1053	274.50	778.50	3	3	0.01101	0.00386	0.35
6	4327	2295	519.92	1775.08	12	14	0.02344	0.00793	0.34
7	3791	951	215.33	735.67	3	5	0.01406	0.00683	0.49
9	2497	1218	280.92	937.08	16	9	0.05924	0.00967	0.16
14	3840	1017	228.75	788.25	2	6	0.00879	0.00765	0.87
15	1272	843	195.17	647.83	3	3	0.01553	0.00465	0.30

N/A, not applicable.

complete gene pairs between C^a BAC MA17P03 and C BAC 46C02. Except for the single-exon genes 8, 10, 11, 12 and 13, all other full-length genes (genes 2–7, 9, 14 and 15) were subjected to synonymous and non-synonymous divergence analysis. Due to high similarity, several gene pairs (genes 4–7 and 14) between C^a BAC MA17P03 and C BAC 46C02 showed either no synonymous or non-synonymous mutations or both, thus there is no estimate of the K$_a$/K$_s$ ratio for these gene pairs. The K$_a$/K$_s$ ratios of all other gene pairs are <1, suggesting their sequence divergence has been functionally constrained (i.e. purifying selection).

We estimated the time of divergence between *C. arabica* and *C. canephora* gene pairs by calculating silent-site nucleotide divergence (K_{sil}). The degree of silent-site divergence was low between gene pairs in C^a BAC MA17P03 and C BAC 46C02 (ranging from 0.00065 to 0.00945), and higher between the C^a BAC MA17P03 and the E^a BAC MA29G21 (ranging from 0.01265 to 0.06002) (Table 4).

We used the upper limit of the molecular divergence to set the minimum divergence time between species (Table 4). Assuming a mean substitution rate of 7.1×10^{-9} substitutions per site per year, which is the mean mutation rate in *Arabidopsis thaliana* based on mutation accumulation experiments (Ossowski *et al.*, 2010), we estimated the time of divergence as approximately 0.665 million years (0.046–0.665) between C^a BAC MA17P03 and C BAC 46C02, and approximately 4.2 million years (0.891–4.227) between C^a BAC MA17P03 and E^a BAC MA29G21.

Our results suggest that *C. eugenoides* and *C. canephora* shared a common ancestor living approximately 4.2 million years ago, and recent formation of the allotetraploid *C. arabica* approximately 665 000 years ago. Based on these results, a schematic representation of the evolutionary history of *C. arabica* is proposed in Figure 3.

Comparative analysis of Coffea and other reference dicotyledonous homeologous sequences

The availability of several plant genome sequences makes it possible to identify conserved microsynteny and perform comparative genomics analysis to assess genome evolution. We BLAST-searched *C. arabica* BAC sequences in genome sequence databases of two fruit crops, tomato (*Solanum lycopersicum*) and grape (*Vitis vinifera*). Two scaffolds from the tomato genome sequence database (version SL 1.03), scaffolds 01386 and 01260, were identified as sharing conserved microsynteny with *C. arabica* BAC sequences (Figure 4a). We aligned the tomato scaffolds 01386 and 01260 with *C. arabica* BAC MA29G21, which contains 16 genes (Figure 4a). Among the 16 annotated genes, we identified conserved sequences for genes 1, 3, 5–8 and 10–12 in tomato scaffold 01386, and for genes 15 and 16 in tomato scaffold 01260. The gene order is conserved among tomato, *C. arabica* and *C. canephora.* However, approximately 106 kb sequence of tomato aligned to approximately 86 kb of *C. arabica* sequqnce (MA29G21), suggesting a 20 kb (23%) expansion in tomato. A similar result was obtained when *C. arabica* BAC sequences were compared with their homeologous sequences from grape. We identified two scaffolds from the grape genome sequence database (*Vitis vinifera* 8X), scaffolds 127 and 128, that share conserved microsynteny with *C. arabica* BAC MA29G21 (Figure 4b). Genes 5–8 and 10–12 were found in grape scaffold 127 and genes 13 and 14 were found in grape scaffold 128. The relative location of the genes remained the same among grape, *C. arabica* and *C. canephora.* However, a total of 257 kb grape sequence aligned to 51 kb of the *C. arabica* (MA29G21) sequence, indicating a 206 kb (404%) expansion in this region in grape.

Table 4 Estimated time of divergence between *Coffea arabica* BAC MA17P03 and *C. canephora* BAC 46C02, and *C. arabica* BACs MA17P03 and MA29G21

Gene ID	Silent sites	Silent mutations	Silent-site divergence (K_{sil})	Estimated age (million years)
C. arabica BAC MA17P03 and *C. canephor*a BAC 46C02				
3	5609.67	8	0.00143	0.101
4	1875.17	14	0.00750	0.528
5	5499.83	25	0.00456	0.321
6	2562.00	7	0.00274	0.193
7	3079.17	12	0.00391	0.275
9	1582.17	4	0.00253	0.178
14	3070.17	2	0.00065	0.046
15	638.67	6	0.00945	0.665
C. arabica BACs MA17P03 and MA29G21				
3	5606.00	80	0.01441	1.015
4	1859.17	50	0.02739	1.929
5	5458.50	132	0.02458	1.731
6	2551.92	32	0.01265	0.891
7	3055.33	84	0.02801	1.973
9	1559.92	54	0.03544	2.496
14	3051.75	44	0.01456	1.025
15	624.17	36	0.06002	4.227

DISCUSSION

Polyploidy is very common in flowering plants, as genomic studies indicated that most angiosperm species are either recent or ancient polyploids. Allopolyploids arise from hybridization of different but closely related species. During evolution, this process has shaped numerous crop genomes such as tobacco (*Nicotiana tabacum*), rapeseed (*Brassica napus*), cotton (*Gossypium hirsutum*) and bread wheat (*Triticum aestivum*), which are now considered model systems to study the genetic consequences of allopolyploidy. In these allopolyploid genomes, hybridizations were estimated to have occurred between 200 000 years ago for tobacco and 1.5 million years ago for cotton (Senchina *et al.*, 2003). This process is still occurring, and several examples of natural allopolyploids have been identified in the past 150 years (Ainouche *et al.*, 2009).

A central question when studying evolution of allopolyploid genomes is not only to determine when parental species hybridized to create the new polyploid species, but also to identify the mechanisms of genome divergence of allopolyploid genomes. Both are required to interpret the consequences of polyploidy events in an evolutionary framework.

The cultivated Arabica coffee is an allotetraploid plant with two different homeologous sub-genomes (named C^a and E^a), that are most likely derived from two diploid *Coffea* species, *C. canephora* (CC) and *C. eugenioides* (EE) (Lashermes *et al.*, 1999). In this study, we aimed to estimate the age

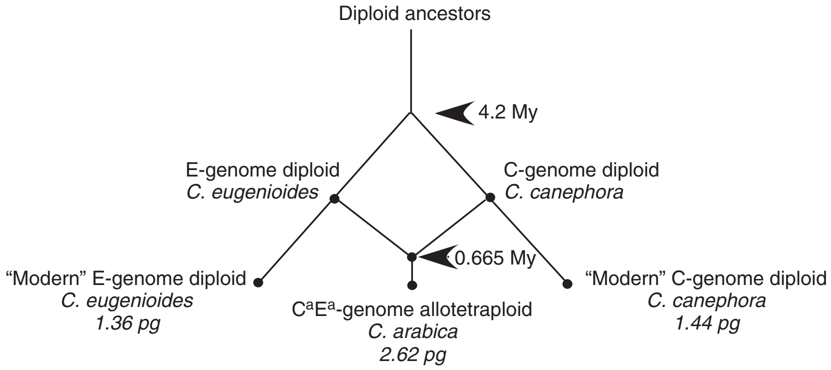

Figure 3. Schematic representation of the evolutionary history of *Coffea arabica* as inferred from gene divergence analysis.
The allotetraploid species *C. arabica* originated 0.665 million years ago, following divergence of its diploid progenitors 4.2 million years ago. *C. arabica* has a slightly smaller genome size than the combined genome size of its diploid progenitors.

of the *C. arabica* species by analyzing orthologous coding sequence divergence.

We constructed a BAC library from the *C. arabica* cultivar Tall Mokka. Two BAC clones, MA17P03 and MA29G21, were isolated from the two sub-genomes, corresponding to an orthologous *C. canephora* BAC containing an ethylene receptor gene (BAC 46C02, *CcEIN4* region) (Guyot *et al.*, 2009). Based on a greater overall nucleotide sequence identity, and better conservation of microsynteny and transposable elements, we identified MA17P03 and MA29G21 as homeologous BACs to the C^a (*C. canephora*) and E^a (*C. eugenioides*) sub-genomes of *C. arabica*, respectively.

Both homeologous sequences showed specific TE insertion events involving both classes of TEs (I and II). No conclusion can be drawn from our results concerning those specific to MA29G21 derived from the E^a sub-genome that may have been present in the original *C. eugenioides* genome, as we do not have the corresponding sequence from *C. eugenioides*. The comparison of *C. canephora* and the C^a orthologous sequence indicated that some transposable elements may be specific to the C^a sub-genome, such as the CART no. 109 LTR retrotransposon, whose presence was checked in a representative set of *C. canephora* genotypes. Our results show that the CART no. 109 LTR retrotransposon was specifically and recently inserted in the C^a sub-genome, most likely after the formation of *C. Arabica*, because it was only identified in *C. arabica* genotypes and not in *C. canephora* and *C. eugenoides* genotypes. Sequence comparison between the two long-terminal repeats of this element allowed us to date the insertion of this element no later than 528 000 years ago, shortly after the speciation of *C. arabica* 665 000 years ago.

In an earlier analysis, the *CcEIN4* region was isolated from the diploid genome of *C. canephora*. It was found that it is a gene-rich and TE-poor region that is remarkably conserved at the genic level over distant dicotyledonous genomes (Guyot *et al.*, 2009). The availability of the BAC library for the *C. arabica* cultivar Tall Mokka allowed us to study genome evolution in diploid and allopolyploid coffee over two time scales: the 665 000 years since formation of the allopolyploid genome and the 4.2 million years since the probable divergence of the two diploid progenitors. In addition, such analysis provides the opportunity to study the genomic effects of allopolyploidy in *C. arabica*.

Based on eight orthologous gene pairs, we estimated the age of the allotetraploid *C. arabica* as approximately 665 000 years. Due to the limited sequence information available, this estimate may not reflect its actual age. Furthermore, the divergence analysis was based on the 'modern' *C. canephora* species and the C^a sub-genome of *C. arabica*, which may include divergences between the *C. canephora* individual involved in the polyploidization event and the direct ancestor of the 'modern' *C. canephora* before polyploidization in addition to divergences after polyploidization. Accurate estimation of the divergence time requires genome-wide sequence data and good fossil data support.

The comparative analysis demonstrated high degree of sequence conservation in coding regions and no rapid divergence or dramatic re-organization in either diploid or allotetraploid *EIN4* regions. These observations are reinforced by overall low coding region variability as indicated by the K_a/K_s ratio, indicating that purifying selection acted on the *EIN4* region of the studied *Coffea* species. However, studies from model polyploidy species have revealed rapid and dynamic changes in genome structure immediately after allopolyploidization (reviewed by Adams and Wendel, 2005; Chen and Ni, 2006; Parisod *et al.*, 2009), called the 'revolutionary phase' (Feldman and Levy, 2009), followed by slow changes in DNA sequences called the 'evolutionary phase' in later generations. In contrast to the expected major changes, our results demonstrate that the genome microstructure in the *C. arabica EIN4* region was very similar to that in its diploid relatives. Hence, in the segments analyzed, the *C. arabica* genome microstructure was not affected by the polyploidization event.

The major chromosomal rearrangements were observed in the inter-genic regions of these BACs. The most striking difference is a small paracentric inversion between homeologous regions within *C. arabica.* Chromosomal rearrange-

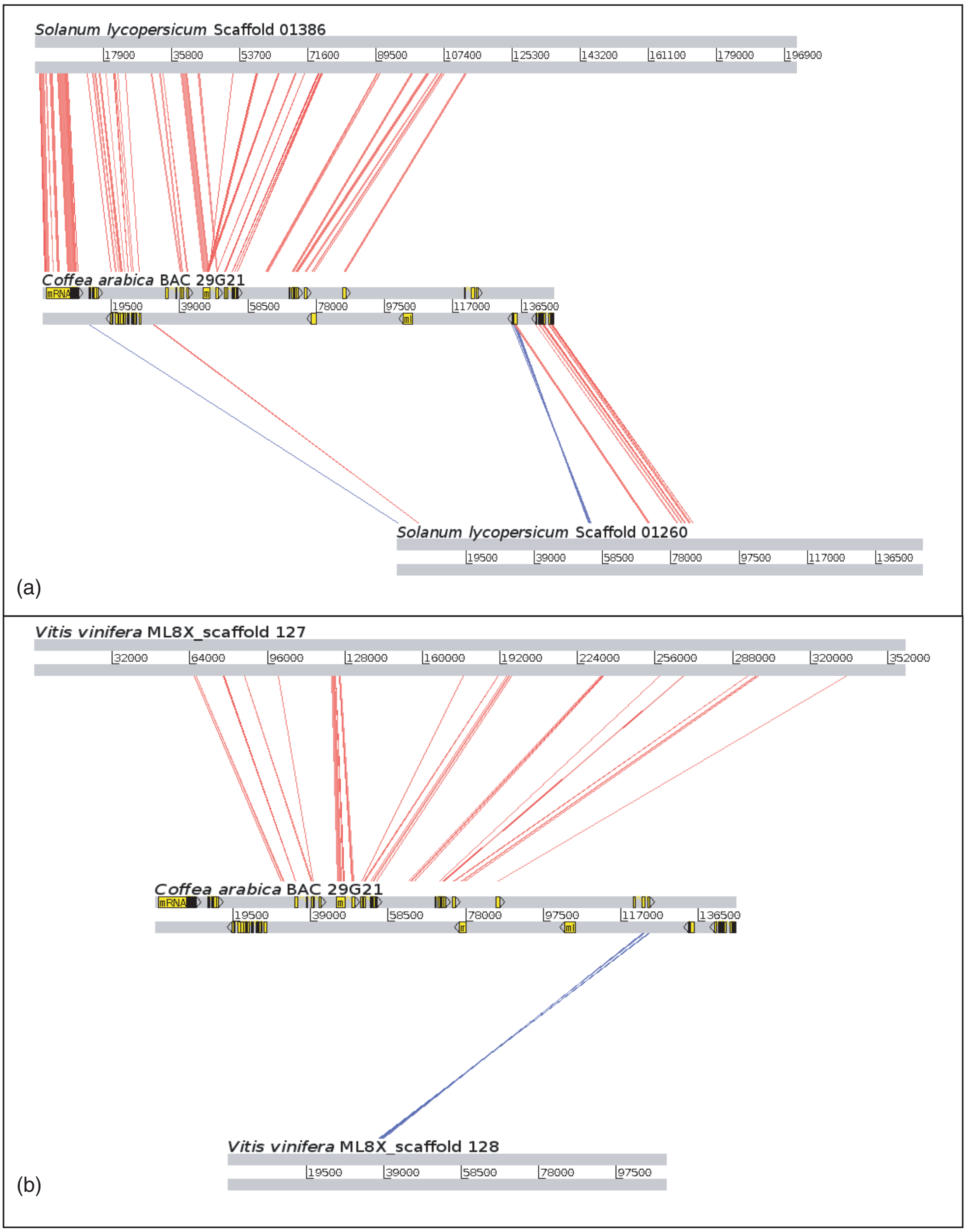

Figure 4. Pairwise comparison of homologous sequences from *Coffea arabica*, *Solanum lycopersicum*, and *Vitis vinifera*. (a) Sequence comparison between *Coffea arabica* BAC MA29G21 and tomato scaffolds 01386 and 01260.
(b) Sequence comparison between *C. arabica* BAC MA29G21 and grape scaffolds 127 and 128.
The red lines link conserved regions and blank regions symbolize disrupted conservation. Comparisons were performed using the BLASTZ algorithm (Schwartz *et al*., 2003) and visualized using the Artemis Comparison Tool (Carver *et al*., 2005).

ments such as local or large inversions appear to be very common in eukaryotes, even between closely related species (Coghlan *et al*., 2005). The chromosomal inversion may have occurred in either *C. arabica* or one of the two diploid progenitors, and may not be a consequence of the polyploidization of the *C. arabica* genome. The orthologous sequences of *C. eugenioides* (EE) are required to resolve the origin of this rearrangement.

In addition to rearrangements, we found lineage-specific TEs in *C. arabica* and its diploid progenitor. These TE insertions allowed unambiguous confirmation of orthologous relationships between *C. canephora* and the C^a sub-genome of *C. arabica*. Sequence expansion in the C^a homeologous regions relative to their diploid progenitors is probably caused by insertions of TEs after polyploidization. However, there is the possibility that the insertions occurred in the progenitor of the C^a sub-genome before polyploidization, as the diploid genome we tested is the modern *C. canephora* genome, and may not fully represent the real progenitor of the C^a sub-genome. We have shown that insertion of a *Ty1*-copia retrotransposons (CART no. 109) was specific to *C. arabica*. It has been shown that allopolyploidization may trigger the activity of certain TE family (Chen and Ni, 2006), and may induce structural changes in the TE genome fraction (Parisod *et al.*, 2010). Differential proliferation of active elements was observed in *Nicotiana* allopolyploid species, but loss of fragments, as observed by sequence specific amplified polymorphism (SSAP) experiments, appears to be more frequent (Parisod *et al.*, 2010). Analyses of the homeologous segments in the allohexaploid wheat genome revealed extensive deletion of TE content, suggesting that TEs may promote loss of DNA through unequal or illegitimate recombination in response to polyploidization (Chantret *et al.*, 2005).

Genome size reduction or massive loss of DNA after polyploidization appears to be a widespread consequence in numerous polyploid species (Leitch and Bennett, 2004). The 2C genome sizes of *C. canephora*, C. *eugenioides* and *C. arabica* were estimated to represent 1.44, 1.364 and 2.622 pg of DNA, respectively (Noirot *et al.*, 2003; Clarindo and Carvalho, 2010). The genome size of *C. arabica* is slightly smaller than the combined genome size of the two progenitors (2.804 pg), indicating that the allopolyploid *C. arabica* genome was not much affected by the allopolyploidization event.

Coffee belongs to the Rubiaceae family, which shares a common ancestor with the closely related Solanaceae family 83–89 million years ago (Wikström *et al.*, 2001). Both the Rubiaceae and Solanaceae family are in the Asterids clade. *V. vinifera* is a member of the *Vitaceae* family in the Rosids clade. Asterids and Rosids diverged from a common ancestor approximately 114–125 million years ago (Wikström *et al.*, 2001). In the study, we compared homeologous regions among coffee, tomato and grape. It is interesting that the homeologous regions are highly collinear among these species. Of 16 annotated genes, nine showed conserved relative location between coffee and tomato, and seven were in consistent positions between coffee and its distantly related species, *V. vinifera*. It is widely accepted that the gene arrangements on chromosomes are determined by a balance between chromosomal rearrangements and functional constrains. All genomes have undergone sequence changes and chromosomal rearrangements by the driving force of natural selection. However, some gene clusters remain linked over genome evolution because of functional constraints. The major histocompatibility complex (MHC) is a good example of such gene clusters, showing surprising conservation of synteny among distantly related vertebrates (Ohta *et al.*, 2000). Identification of these linked syntenic genes is important for understanding evolutionary processes that shape genome evolution. The syntenic gene cluster we identified contains two genes associated with fruit ripening, the ethylene receptor and an ethylene-responsive transcription factor. As the three species we compared are fruit crops, this may indicate that this gene cluster plays an important role in fruit development. Only a limited degree of conservation of this region was observed among the non-fruit crops *Arabidopsis thaliana*, *Medicago sativa* (alfalfa) and *Populus trichocarpa* (black cottonwood) (Guyot *et al.*, 2009), although all of them are members of the Rosids clade and thus relatively closely related to grape.

EXPERIMENTAL PROCEDURES

Plant material

Young leaf tissue from *Coffea arabica* cultivar Tall Mokka (MA2-7) was used for nuclei extraction. The leaf tissues, young flower buds, young seedlings and shoot apical meristems of the same cultivar were used for RT-PCR amplification of target cDNA sequences to annotate coding regions of BAC sequences.

Construction of a *C. arabica* BAC library

Nuclei was isolated from the young leaf tissues of Arabica coffee cultivar Tall Mokka (MA2-7) as described by Ming *et al.* (2001). The high molecular weight DNA was released from nuclei and embedded in agarose and partially digested using *Hin*dIII. The fraction at approximately 100 kb was recovered and ligated into the pIndigo-BAC-5 vector (Epicentre, http://www.epibio.com). A total of 79 872 BAC clones were picked and archived in 384-well plates with freezing medium. BAC clones were gridded onto Performa II Nylon Filters (Genetix, http://www.genetix.com) using Q-Pix 2 (Genetix).

Screening the BAC library

High-density membranes of the *C. arabica* BAC library were prehybridized in 0.5 M Na_2HPO_4, 7% SDS, 1 mM EDTA, 100 μg ml^{-1} heat-denatured herring sperm DNA for at least 4 h. Probes were labeled using a random primer labeling system (Rediprime II DNA Labeling System, GE Healthcare, http://www.gelifescience.com). The hybridization was performed overnight at 55°C in 0.5 M Na_2HPO_4, 7% SDS, 1 mM EDTA, 100 μg ml^{-1} heat-denatured herring sperm DNA using ^{32}P-labeled probes. Hybridized membranes were washed twice in 0.5 × SSPE/0.5% SDS for 10 min each time.

Southern hybridization

BAC DNA was isolated by the alkaline lysis method (Sambrook *et al.*, 1989) and digested using *Hin*dIII. The digested DNA samples were electrophoresed through 0.8% agarose gel. After electrophoresis, the gel was blotted onto Hybond N+ membranes (GE Healthcare) using standard methods (Sambrook *et al.*, 1989). The membranes were pre-hybridized (6 × SSC, 10 × Denhardt's reagent, 1% SDS, 100 μg ml^{-1} heat-denatured herring sperm DNA) at 55°C for 2 h. Probes were labeled using a random primer labeling system (Rediprime II DNA Labeling System, GE Healthcare). The hybridization was performed overnight at 55°C in 6 × SSC, 5% dextran sulfate, 1% SDS, 100 μg ml^{-1} heat-denatured herring sperm DNA, using ^{32}P-labeled probes. The labeled membranes were washed three times for 30 min each at 55°C. The washing solutions were 2 × SSC/0.1% SDS, 1 × SSC/0.1% SDS and 0.3 × SSC/0.1% SDS, respectively. The washed membranes were exposed to X-ray films at –80°C for 1–6 days depending on the intensity of the signal.

Sequencing BAC clones and sequence assembly

The BAC clones were sequenced using the shotgun approach with at least tenfold coverage. BAC DNAs were isolated using a Qiagen large-construct kit (http://www.qiagen.com/) and randomly sheared using Hydroshear (Genomic Solutions, http://www.genomicsolutions.com) to generate approximately 3 kb insert fragments. The sheared DNAs were size-selected on an agarose gel, purified using a QIAquick gel extraction kit (Qiagen), end-repaired using a DNATerminator end repair kit (Lucigen, http://www.lucigen.com), and ligated into the pSMART-HCKan vector (Lucigen). The initial assembly of MA29G21 and MA17P03 BAC clones was performed using PHRED/PHRAP/CONSED software (University of Washington, http://www.phrap.org). Gap filling was performed by primer walking.

Sequence annotation

We annotated the sequences of BACs MA29G21 and MA17P03 by BLAST similarity searches in GenBank. Genes and TEs were identified as described by Guyot *et al.* (2009). Dating of the LTR retrotransposon insertion was performed as described by SanMiguel *et al.* (1998) with the EMBOSS package, using a molecular clock of 7.1 × 10^{-9} base substitutions per site per year (Ossowski *et al.*, 2010).

Plant material for retrotransposon-based polymorphism analysis

Except for the Ugandan genotypes, all the plants are maintained in tropical greenhouses at the Institut de Recherche pour le Développement (IRD) Center in Montpellier (France). Twenty-three *C. canephora* genotypes representing the various diversification groups of the species (Gomez *et al.*, 2009; Musoli *et al.*, 2009), eight *C. eugenioides* genotypes and four *C. arabica* genotypes were used for the analysis. DNA was extracted from young fresh leaves as described by Gomez *et al.* (2009). DNAs from Ugandan genotypes were kindly provided by Coffee Research Center (COREC, Kampala, Uganda) and Centre International de Recherche Agronomique pour le Développement (CIRAD, Montpellier, France).

Survey of the CART no. 109 LTR retrotransposon insertion at the EIN4 locus by retrotransposon-based polymorphism analysis

For retrotransposon-based polymorphism (RBIP) experiments, three types of primer were designed as follows: RBIP2, which is located inside the long-terminal repeat region of the retrotransposons (oriented towards the exterior of the element), and RBIP1 and RBIP3, which are located in the flanking regions outside the retrotransposons (RBIP1: 5′-CAGACAAGGGATCAACAGCA-3′; RBIP2: 5′-CTCCGATGTGGGATGAGAAG-3′; RBIP3: 5′-AAAATTGGAAGG-GAGGGTTG-3′). We tested for the presence/absence of the *CART no. 109* LTR retrotransposon in a large collection of *C. canephora* (23 individuals), *C. arabica* (four individuals) and *C. eugenioides* (eight individuals).

PCR was performed in 25 μl total volume with 50 ng template DNA, 0.1 μM of each primer, 0.2 mM dNTPs, 5 mM $MgCl_2$, 0.5 units of GoTaq® DNA polymerase (Promega, http://www.promega.com) and 1× GoTaq reaction buffer. The following program was used: 94°C for 2 min, five cycles of 94°C for 30 sec, 60°C for 30 sec with a decrease of 1°C per cycle, and 72°C for 30 sec, then 30 cycles of 94°C for 30 sec, 55°C for 30 sec and 72°C for 30 sec, followed by extension at 72°C for 8 min. The amplified products were visualized after migration on a 1% agarose gel and ethidium bromide staining. A retrotransposon was present when a PCR product of 236 bp was visualized. A PCR product of 371 bp indicates absence of the retrotransposon at the investigated site.

RT-PCR

At least one intron was covered by primers designed for RT-PCR experiments to control genomic DNA contamination. Total RNA was extracted from young flower buds, cherries, leaf tissues and young seedlings. Approximately 2 μg total RNA was treated with RNase-free DNase I (Promega) and reverse-transcribed using a RETROscript kit (Ambion, http://www.ambion.com). The synthesized cDNAs served as templates for RT-PCR.

Sequence divergence of gene pairs

Exon and intron regions of gene pairs were manually aligned using BioEdit (Hall, 1999). The numbers of synonymous substitutions per synonymous site (K_s), non-synonymous substitutions per non-synonymous site (K_a), and synonymous and non-coding (silent) substitutions per silent site (K_{sil}) were estimated according to the method developed by Nei and Gojobori (1986) and implemented in DnaSP 4.0 (Rozas *et al.*, 2003). Divergence times for the paired alleles were determined using K_{sil} and the methods described by Li (1997), using a mean substitution rate of 7.1 × 10^{-9} substitutions per site per year, which is the mean mutation rate in *A. thaliana* based on mutation accumulation experiments (Ossowski *et al.*, 2010).

BAC sequence comparison analysis

The BAC sequence comparison study was performed using the Artemis Comparison Tool (Carver *et al.*, 2005). Large-scale alignments between homeologous BACs were performed using BLASTZ (Schwartz *et al.*, 2003). The grapevine reference genome sequences were downloaded from the Grape Genome Browser (http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/) (version *Vitis vinifera* 8X, released on 30 August 2007). The tomato reference genome sequences were downloaded from the SOL Genomics Network website (http://solgenomics.net/) (version SL 1.03, released in January 2010).

ACKNOWLEDGEMENTS

We thank Meghan R. Jones (Hawaii Agriculture Research Center, Waipahu, Hawaii, USA), Rebecca Heinig (Hawaii Agriculture Research Center, Waipahu, Hawaii, USA) and Marie Couderc (Unité Mixte de Recherche Diversité, Adaptation et Développement, Evolution et Dynamique des Génomes, Montpellier, France) for technical assistance, and Dr Pascal Musoli (Coffee Research Center, Kampala, Uganda) and Dr Thierry Leroy (Centre International de Recherche Agronomique pour le Développement, Montpellier, France) for kindly providing the Ugandan *C. canephora* DNAs. This work was partially supported by a US Department of Agriculture, Agricultural Research Service Cooperative Agreement (CA 58-3020-134) with the Hawaii Agriculture Research Center.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Sequence comparisons between *C. arabica* homeologous BACs MA29G21 and MA17P03.

Figure S2. Dot-plot comparison between the homeologous *C. arabica* BACs MA29G21 and MA17P03.

Figure S3. Phylogenetic tree generated from the sequence comparison between MA17P03, 46C02 and MA29G21.

Figure S4. Dot-plot comparisons between the orthologous BACs MA17P03 and 46C02, and between MA29G21P03 and 46C02.

Figure S5. PCR detection of the polymorphic CART no. 109 retrotransposon insertion in *C. canephora*, *C. arabica* and *C. eugenioides*.

Please note: As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

REFERENCES

Adams, K.L. and Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* **8**, 135–141.

Aga, E., Bekele, E. and Bryngelsson, T. (2005) Inter-simple sequence repeat (ISSR) variation in forest coffee trees (*Coffea arabica* L.) populations from Ethiopia. *Genetica*, **124**, 213–221.

Ainouche, M.L., Fortune, P.M., Salmon, A., Parisod, C., Grandbastien, M.-A., Fukunaga, K., Ricou, M. and Misset, M.-T. (2009) Hybridization, polyploidy and invasion: lessons from Spartina (Poaceae). *Biol. Invasions*, **11**, 1159–1173.

Berthou, F. (1983) Chloroplast and mitochondrial DNA variation as indicator of phylogenetic relationships in the genus *Coffea* L. *Theor. Appl. Genet.* **65**, 77–84.

Bustamante-Porras, J., Poncet, V., Campa, C., Noirot, M., Hamon, S. and de Kochko, A. (2007) Characterization of three ethylene receptor genes in *Coffea canephora* Pierre. In *Advances in Plant Ethylene Research* (Ramina, A., Chang, C., Giovannoni, J., Klee, H., Perata, P. and Woltering, E., eds). Dordrecht, The Netherlands: Springer, pp. 53–56.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool. *Bioinformatics*, **21**, 3422–3423.

Chantret, N., Salse, J., Sabot, F. *et al.* (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell*, **17**, 1033–1045.

Charrier, A. and Berthaud, J. (1985) Botanical classification of coffee. In *Coffee: Botany, Biochemistry and Production of Beans and Beverage* (Clifford, M.N. and Willson, K.C., eds). Beckenham, UK: Croom Helm Ltd, pp. 13–47.

Chen, Z.J. and Ni, Z. (2006) Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. *Bioessays*, **28**, 240–252.

Clarindo, W.R. and Carvalho, C.R. (2010) Flow cytometric analysis using SYBR Green I for genome size estimation in coffee. *Acta Histochem.* **113**, 221–225.

Coghlan, A., Eichler, E.E., Oliver, S.G., Paterson, A.H. and Stein, L. (2005) Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet.* **21**, 673–682.

Davis, A.P., Govaerts, R., Bridson, D.M. and Stoffelen, P. (2006) An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). *Bot. J. Linn. Soc.* **152**, 465–512.

De Kochko, A., Akaffou, S., Andrade, A.C. *et al.* (2010) Advances in *Coffea* genomics. *Adv. Bot. Res.* **53**, 23–63.

Feldman, M. and Levy, A.A. (2009) Genome evolution in allopolyploid wheat – a revolutionary reprogramming followed by gradual changes. *J. Genet. Genomics*, **36**, 511–518.

Ferwerda, F.P. (1976) Coffee. In *Evolution of Crop Plants* (Simmonds, N.W., ed). London: Longman, pp. 257–260.

Flavell, A.J., Knox, M.R., Pearce, S.R. and Ellis, T.H. (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. *Plant J.* **16**, 643–650.

Gomez, C., Dussert, S., Hamon, P., Hamon, S., Kochko, A. and Poncet, V. (2009) Current genetic differentiation of *Coffea canephora* Pierre ex A. Froehn in the Guineo-Congolian African zone: cumulative impact of ancient climatic changes and recent human activities. *BMC Evol. Biol.* **9**, 167.

Guyot, R., de la Mare, M., Viader, V., Hamon, P., Coriton, O., Bustamante-Porras, J., Poncet, V., Campa, C., Hamon, S. and de Kochko, A. (2009) Microcollinearity in an ethylene receptor coding gene region of the *Coffea canephora* genome is extensively conserved with *Vitis vinifera* and other distant dicotyledonous sequenced genomes. *BMC Plant Biol.* **9**, 22.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98.

Lashermes, P., Cros, J., Marmey, P. and Charrier, A. (1993) Use of random amplified DNA markers to analyse genetic variability and relationships of *Coffea* species. *Genet. Resour. Crop Evol.* **40**, 91–99.

Lashermes, P., Cros, J., Combes, M.C., Trouslot, P., Anthony, F., Hamon, S. and Charrier, A. (1996) Inheritance and restriction fragment length polymorphism of chloroplast DNA in the genus *Coffea* L. *Theor. Appl. Genet.* **93**, 626–632.

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

Leitch, I.J. and Bennett, M.D. (2004) Genome downsizing in polyploid plants. *Biol. J. Linn. Soc.* **82**, 651–663.

Leroy, T., Marraccini, P., Dufour, M. *et al.* (2005) Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes. *Theor. Appl. Genet.* **111**, 1032–1041.

Li, W.H. (1997) *Molecular Evolution*. Sunderland, MA: Sinauer Associates Inc.

Maurin, O., Davis, A.P., Chester, M., Mvungi, E.F., Jaufeerally-Fakim, Y. and Fay, M.F. (2007) Towards a phylogeny for *Coffea* (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. *Ann. Bot.* **100**, 1565–1583.

Ming, R., Moore, P.H., Zee, F., Abbey, C.A., Ma, H. and Paterson, A.H. (2001) Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. *Theor. Appl. Genet.* **102**, 892–899.

Musoli, P., Cubry, P., Aluka, P., Billot, C., Dufour, M., De Bellis, F., Pot, D., Bieysse, D., Charrier, A. and Leroy, T. (2009) Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda. *Genome*, **52**, 634–646.

Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426.

Noir, S., Patheyron, S., Combes, M.-C., Lashermes, P. and Chalhoub, B. (2004) Construction and characterization of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). *Theor. Appl. Genet.* **109**, 225–230.

Noirot, M., Poncet, V., Barre, P., Hamon, P., Hamon, S. and de Kochko, A. (2003) Genome size variations in diploid African *Coffea* species. *Ann. Bot.* **92**, 709–714.

Ohta, Y., Okamura, K., McKinney, E.C., Bartl, S., Hashimoto, K. and Flajinik, M.F. (2000) Primitive synteny of vertebrate major histocompatibility complex class I and class II gene. *Proc. Natl Acad. Sci. USA*, **97**, 4712–4717.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D. and Lynch, M. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.A. and Ainouche, M. (2009) Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. *New Phytol.* **184**, 1003–1015.

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., Ainouche, M., Chalhoub, B. and Grandbastien, M.A. (2010) Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.* **186**, 37–45.

Pearl, H.M., Nagai, C., Moore, P.H., Steiger, D.L., Osgood, R.V. and Ming, R. (2004) Construction of a genetic map for arabica coffee. *Theor. Appl. Genet.* **108**, 829–835.

Pendergrast, M. (2009) Coffee second only to oil? Is coffee really the second largest commodity? Mark Pendergrast investigates and finds some startling results. *Tea Coffee Trade J.* **181**, 38–41.

Robbrecht, E. and Manen, J.-F. (2006) The major evolutionary lineages of the coffee family (Rubiaceae, angiosperms). Combined analysis (nDNA and cpDNA) to infer the position of Coptosapelta and Luculia, and supertree construction based on rbcL, rps16, trnL-trnF and atpB-rbcL data. A new classification in two subfamilies, Cinchonoideae and Rubioideae. *Syst. Geogr. Plants*, **76**, 85–146.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998) The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107.

Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A. and Wendel, J.F. (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20**, 633–643.

Silvestrini, M., Junqueira, M.G., Favarin, A.C., Guerreiro, O., Maluf, M.P., Silvarolla, M.B. and Colombo, C.A. (2007) Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. *Genet. Resour. Crop Evol.* **54**, 1367–1379.

Van der Vossen, H.A.M. (2001) Agronomy I: coffee breeding practices. In *Coffee: Recent Developments* (Clarke, R.J. and Vitzthum, O.G., eds). Oxford, UK: Blackwell Science, pp. 184–201.

Wikström, N., Savolainen, V. and Chase, M.W. (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. Lond. B Biol. Sci.* **268**, 2211–2220.