www.sciencemag.org/content/345/6201/1181/suppl/DC1

Supplementary Materials for

The coffee genome provides insight into the convergent evolution of caffeine biosynthesis

France Denoeud, Lorenzo Carretero-Paulet, Alexis Dereeper, Gaëtan Droc, Romain Guyot, Marco Pietrella, Chunfang Zheng, Adriana Alberti, François Anthony, Giuseppe Aprea, Jean-Marc Aury, Pascal Bento, Maria Bernard, Stéphanie Bocs, Claudine Campa, Alberto Cenci, Marie-Christine Combes, Dominique Crouzillat, Corinne Da Silva, Loretta Daddiego, Fabien De Bellis, Stéphane Dussert, Olivier Garsmeur, Thomas Gayraud, Valentin Guignon, Katharina Jahn, Véronique Jamilloux, Thierry Joët, Karine Labadie, Tianying Lan, Julie Leclercq, Maud Lepelley, Thierry Leroy, Lei-Ting Li, Pablo Librado, Loredana Lopez, Adriana Muñoz, Benjamin Noel, Alberto Pallavicini, Gaetano Perrotta, Valérie Poncet, David Pot, Priyono, Michel Rigoreau, Mathieu Rouard, Julio Rozas, Christine Tranchant-Dubreuil, Robert VanBuren, Qiong Zhang, Alan C. Andrade, Xavier Argout, Benoît Bertrand, Alexandre de Kochko, Giorgio Graziosi, Robert J Henry, Jayarama, Ray Ming, Chifumi Nagai, Steve Rounsley, David Sankoff, Giovanni Giuliano, Victor A. Albert,* Patrick Wincker* Philippe Lashermes*

*Corresponding author. E-mail: vaalbert@buffalo.edu (V.A.A.); pwincker@genoscope.cns.fr (P.W.); philippe.lashermes@ird.fr (P.L.)

Published 5 September 2014, *Science* **345**, 1181 (2014)
DOI: 10.1126/science.1255274

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/345/6201/1181/suppl/DC1)

The Coffee Genome Provides Insight into the Convergent Evolution of Caffeine Biosynthesis

Material and Methods

Plant material and DNA preparation

Although the *Coffea* genus includes more than 124 species (*1*), commercial coffee production relies exclusively on two related species, *C. arabica* and *C. canephora*, which account for 65% and 35% of global production, respectively (http://www.ico.org). *C. canephora* (2n=2x=22) is an out-crossing, highly heterozygous diploid, while *C. arabica* is a recent allotetraploid (2n=4x=44) derived from hybridization between *C. canephora* and *C. eugenioides*(*2, 28*). The *C. canephora* accession DH 200-94 was selected for sequencing. This accession is a doubled haploid plant produced from the clone IF200 based on haploid plants occurring spontaneously in association with polyembryony (*29*). Young expanding leaves from two greenhouse-grown plants at IRD-Montpellier (France) were harvested and stored at -80°C prior to DNA extraction. A large quantity of genomic DNA was extracted by means of a nuclei isolation step as described in (*30, 31*). For BAC library construction, high molecular weight DNA was isolated from 20g of young leaf tissue as reported in (*32*).

Genome sequencing

The genome was sequenced using a whole genome shotgun strategy. All data were generated using next generation sequencers (Roche/454 GSFLX and Illumina GAIIx), except for sequences of BAC ends that were produced by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers (**Table S1**).

Transcriptome sequencing

Vegetative and reproductive tissues (root, leaves, pistil, and stamen) were taken from greenhouse-grown *C. canephora* plants for RNA isolation. To investigate the effect of the growing temperature on the transcriptome, *C. canephora* plants were cultivated for two months in four sets of contrasted growing conditions with different diurnal/nocturnal temperatures: 18/14, 23/19, 28/24 and 33/29 °C. In each climatic chamber, the photoperiod was set at 12 h per day, humidity at 80-90% and luminosity at 600 µmol photon m^{-2} s^{-1}. At the same time, 3 h after the beginning of the diurnal period, young, fully expanded leaves were harvested on two plants (biological replicates) in each growing condition. Finally, perisperm and endosperm samples were prepared from fruits collected from *C. canephora* plants at different developmental stages: 120, 150 and 180 days after pollination (DAP) for perisperm, and 180, 260 and 320 DAP for endosperm. After collection, the samples were immediately flash frozen in liquid nitrogen and stored at -80 °C until RNA extraction.

Total RNA was isolated from 1 g of material from each plant. Harvested samples were ground in liquid nitrogen and the powders suspended in 20 ml of an extraction buffer (5M guanidinium isothiocyanate, 31 mM sodium acetate (pH 8), 1% β-mercaptoethanol, 0.88% (w/v) N-lauroyl sarcosine and 1% (w/v) polyvinylpyrrolidone -PVP40-). The solutions were centrifuged at 15,000 g for 20 min at 4 °C. RNA was purified on a 5.7 M cesium chloride layer (18,000 g for 20 h at 20 °C). The RNA pellets were rinsed twice with 70% (v/v) ethanol, and dissolved in 100 µl of RNAse-free water. The quality and the concentration of extracted RNA samples were determined using the Agilent DNA1000 (Agilent, Santa Clara, CA, USA).

Using Illumina GAII technology, either 72-nt or 100-nt reads were generated from libraries derived from a subset of the RNA samples The reads were aligned using BWA

(version 0.7.2, BWA-MEM algorithm) (*33*) against the *C. canephora* 25,574 protein-coding gene models used as reference transcriptome. Uniquely mapped sequence counts were processed using the "DESeq" package (*34*) to estimate the transcript level. The count data were normalized on the total number of counts, taking the variance and the mean of the biological replicates into account. For each sample, the normalized expression level for each gene of the reference transcriptome was expressed as reads per kilobase of transcript sequence per million mapped reads (RPKM). For each gene, the fold change of the gene expression between conditions was analyzed and the statistical significance estimated using an adjusted pvalue for multiple testing (Benjamini-Hochberg method).

Genome assembly and automatic error corrections with Solexa/Illumina reads

454 reads and Sanger BAC ends were assembled using Newbler version MapAsmResearch-04/09/2010-patch-18/17/2010. From the initial 54,415,922 reads, about 86.31% were assembled. Following removal of contaminants, 91,439 contigs were assembled and linked into 13,345 scaffolds. The contig N50 was 14.8 kb, and the scaffold N50 was 1.3Mb (**Table S2**). The cumulative scaffold size was 569.4 Mb, about 20% smaller than the estimated genome size of 710 Mb. The *C. canephora* cDNA unigene resources were aligned with the assembly using the Blat (*35*) algorithm with default parameters, and only the best match was kept for each unigene. The assembly contains 94% of the 56,216 *Coffea* unigenes, confirming the high coverage of the genome.

One way to improve the 454 assembly is to complement it with another type of data with a different error type bias, as described previously (*36*). Short-read sequences (around 55X of paired-end reads and 10X of single-end reads) were aligned on the *C. canephora* genome assembly using BWA (*33*). Only uniquely mapped reads were retained. Each difference was then considered and kept only if it met the following three criteria: (1) an error was not located in the first 5 bp or the last 5 bp of the read, (2) in terms of the quality of the bases being considered, the previous and the next base were above 20, and (3) the remaining sequences around the error (before and after) were not homopolymers (to avoid misalignment at boundaries). In the next stage, pile-up errors located at the same position were identified, particularly errors that occurred within homopolymers (since two reads that tag the same error can report different positions). Finally, each detected error was corrected if at least three reads detected the given error (in both orientations) and 70% of the reads located at that position agreed. Since we only allowed uniquely mapped reads, several regions were devoid of Illumina reads. In a first step, one or several errors were corrected, and during subsequent iterations of the strategy, regions devoid of Illumina reads were also covered. During the first step, 119,010 nucleotides were corrected, and during the second step, reads were mapped on the corrected scaffolds from the first step and 32,380 errors were corrected. We decided to stop the iteration after eight cycles.

Gaps between consecutive contigs were filled using the GapCloser software (from SOAPdenovo package (*37*)) and Illumina paired-end reads. This last step greatly improved the contig contiguity as well as the completeness of the genome assembly; as a consequence the N50 went from 14.8kb to 50kb, and the proportion of N decreased from 24 to 17% (**Table S2**).

Validation of the structural integrity of the genome assembly

The 454 mate-pair data were used to assess the structural accuracy of the *de novo* genome assembly. Filtered 454 mate pair reads from all the library classes – short (3 kb), medium (8 kb) and long-span (13 kb) – were aligned to the scaffolds using BLASTN. In order to ensure unambiguous mapping, only sequences of at least 30 nucleotides that aligned to a unique location with a coverage of 90% or more and an identity of 95% or better were used.

In total, 26.95% of the mate-pairs could be aligned to a unique position on the scaffolds using these stringent criteria (**Figure S1A**). Pairs of sequences that aligned to a single scaffold with an incorrect orientation were considered indicators of potential misassemblies. In total, only 7,926 out of 6,385,814 filtered mate-pairs (0.12%) aligned inconsistently with the assembly, indicating its overall accuracy. These data are consistent with other high-quality genome assemblies using next generation sequence data, such as tomato (*10*). Next, the average span distance of mate-pairs mapping on the same scaffold was computed to provide a further estimate of the assembly accuracy. The histogram derived shows that approximately 95% of the uniquely and correctly mapped pairs have a span within ±2 kilobases (Kb) from the calculated average for each library class (**Figure S1B**).

Additionally, we searched for *Coffea canephora* bacterial artificial chromosome (BAC) sequences to validate the scaffolding. Seven BACs from the sequenced genotype (HD200-94) were found, among which only four could be used (for the three others, the contigs have been ordered on the BAC using the genome assembly), JX227993(*38*), HQ696507, HQ696512, HQ696513 (*39*), all corresponding to the same genomic region, SH3 (R-genes cluster). Two BACs from another clone (IF126) were also used, HM635075 (BAC 111018, ovate locus) (*40*) and EU164537 (ethylene receptor) (*41*). We aligned the assembled scaffolds and the BACs using MUMmer (nucmer) (*42*) and checked the colinearity on a dotplot (**Figure S2**). In all cases, BACs and scaffolds are colinear, which confirms the organization of contigs inside scaffolds (gaps are displayed in grey on the y axis). In one case, a small scaffold (scaffold1573) could have been included in the gap of a larger scaffold (scaffold3), and in another case the BAC sequence confirms the anchoring of scaffolds (scaffold8 (+) is followed by scaffold7 (-) on the chr7 pseudomolecule). In conclusion, although the BAC dataset is small, it provides independent support for the validity of the assembly.

Construction of a high-density genetic map

A consensus high-density genetic map of *C. canephora* was developed (*43, 44*) based on a F1 cross between two highly heterozygous genotypes, a Congolese group genotype (BP409) and a Congolese x Guinean hybrid parent (Q121). The segregating population was composed of 93 F1 individuals. Several types of markers were used in this coffee mapping population, including Restriction Associated DNA sequencing (RADseq). RFLP mapping was performed using eight restriction enzymes (DraI, EcoRV, HindIII, HaeIII, RsaI, ScaI, HincII and PvuII). Several sources of molecular markers were used from EST and genomic libraries (*45, 46, 32*). SSR primer pairs were designed to obtain a range of PCR amplicon lengths from 100 to 300 bp in genomic libraries and EST databases. Allele detection was obtained using an ABI Prism 3500 analyzer. SNPs identified in sequenced amplicons from the two parents were mapped using the MGB TaqMan or HRM technologies. Progeny genotyping was performed using allelic discrimination assays. Analysis of results was done using a LightCycler 480. The RAD libraries were made from digestion of DNA using two restriction enzymes, NsiI (6-base cutter) and MseI (4-base cutter). The fragments (150 - 500 bp) were selected to ligate to two adaptors including one with a tag for each progeny. Equal amounts of amplicons from each individual were pooled to build Illumina RNA-Seq libraries with individual tags for each library. Co-segregating markers within 50 Kb regions (< 1 cM) based on the aligned template scaffold were sorted as bins. One marker from each bin was selected for genetic mapping.

The linkage analysis and map construction were performed using JoinMap software version 4.1 (*47*) using LOD threshold of 5 and Kosambi's function (*48*) to calculate genetic distance between two loci. A consensus genetic map was built using the F2 segregating loci as anchor markers in order to merge the two homologous parental linkage groups. The final

high density *C. canephora* map comprises 3,230 loci distributed on 11 linkage groups coded from A to K (**Table S3**). The genetic map covers 1,471 cM, with an average interval between two adjacent mapped markers of 0.46 cM. Only one important genetic gap (12 cM) was observed on linkage group J.

Scaffold anchoring

All available sequence-based markers from the consensus genetic linkage map were BLAST-aligned against the scaffolds. Sequence-based markers were filtered out and only markers presenting a single hit were retained. More precisely, a hit was taken into account if its HSPs showed a minimal identity per cent of 90%, conformed to a maximal distance of 3,000 bp between HSPs, and displayed a cumulated size greater than or superior to 60% of the marker-sequence length. 1,295 markers were unambiguously located on the assembly and used in combination with 1,644 RADseq markers to anchor and orient the scaffolds along the *C. canephora* pseudomolecules. A total of 349 scaffolds covering approximately 364 megabases (Mb) (64% of the assembled genome sequence) were anchored to the 11 *C. canephora* chromosomes, among which 139 representing 290 Mb (51% of the assembled genome) were both anchored and oriented. 98% of the 100 largest scaffolds and 96.4% of scaffolds larger than 1Mb were anchored on chromosomes. As such, the coffee genome assembly can be considered rather complete, with only a small fraction of genes potentially missing from total scaffolds. The overview of the assembly anchoring on the genetic map is reported in **Table S4**.

Pseudomolecules

The 349 anchored scaffolds were joined to generate 11 pseudomolecules **(Figure S3)** that were named according to the linkage group nomenclature. Each scaffold join was denoted with 100 N base pairs. 139 mapped scaffolds have known orientation along the pseudomolecules while the remaining 210 mapped scaffolds were assigned a random orientation. 12,996 scaffolds (totaling 204 Mb) remain unmapped in the current genome release and were grouped arbitrarily into a pseudomolecule named “Un” (for “unknown”), each scaffold being joined by 100 Ns.

Comparing physical to genetic map distance

All available sequence-based markers were mapped onto the 11 pseudomolecules. As can be seen (**Figure S3**), there is considerable variation in the physical:genetic map distance ranging from 67 kb/cM on distal end of pseudomolecule 11 to over 4 Mb/cM on the central part of pseudomolecule 3. In general, crossing-over frequency appears negatively correlated with the density of repeat sequences. The greater the retrotransposon density is, the higher the ratio between physical and genetic distances.

Protein-coding gene annotation

Repeat masking

Most of the genome comparisons were performed with repeat-masked sequences. For this purpose, we searched and sequentially masked several kinds of repeats using the Repeatmasker software (*49*):

- Known plant repeats and transposons available in Repbase (*50*)
- Tandem repeats with the TRF program (*51*)
- 653 known TEs of *C. canephora*

As a result, 38% of the assembled bases were masked.

In addition, RepeatScout (*52*) was run to detect repeats *ab initio*.

Protein mapping

The *Arabidopsis thaliana* (TAIR 10, 2011/01/03 release), *Solanum tuberosum* (http://potatogenomics.plantbiology.msu.edu/, release v3.4), *Solanum lycopersicum* (ftp://ftp.kazusa.or.jp/pub/tomato/) and *Vitis vinifera* proteomes as well as Gentianales proteins contained in UniProt (version available at 23-Jan-2012) were used to detect conserved proteins in the *C. canephora* genome. As Genewise (*53*) is time-greedy, the proteomes were first aligned with the *C. canephora* genome assembly using BLAT (*35*). Subsequently, we extracted genomic regions in which no protein hit had been found by BLAT and realigned proteins with more permissive parameters. Each significant match was then refined using Genewise in order to identify exon/intron boundaries.

Ab initio predictions: SNAP and FGenesh

Two *ab initio* gene prediction softwares: SNAP (*54*) and FGENESH (*55*) were used to predict gene models on the *C. canephora* genome assembly. SNAP was trained with Gmorse gene models.

Coffea public cDNAs

A collection of 255,032 cDNA sequences from the genus *Coffea* (taxon id 13,442, including 174,715 sequences from *C. arabica* and 65,738 sequences from *C. canephora* available in EMBL version 27/01/2012) was first aligned with BLAT onto the assembly, and then only the best matches (with % identity > 90%) for each read were selected. In a second step, each match was extended by 1 kb on each end, and realigned with the cDNA sequence using the Est2genome software (*56*). 90.80% of *C. arabica* ESTs were mapped, with an average % identity of 98.12% (after Est2genome); likewise, 94.15% of *C. canephora* ESTs were mapped, with an average % identity of 98.35%.

Coffea. arabica cDNAs

Two sets of 454 reads were assembled separately using Newbler, generating 6,302 contigs (N50 = 1,043 bp) for *C. arabica* var. Caturra and 13,206 contigs (N50 = 1093 bp) for *C. arabica* var. Bourbon. These contigs were then aligned onto the *C. canephora* genome assembly with BLAT, and only the best matches (with % identity > 90%) for each read were selected. Then, each match was extended by 1 kb on each end, and realigned with the read using the Est2genome software (*56*).

Coffea canephora UniGenes

A *C. canephora* UniGene reference database was built with transcripts from the "SGN coffee UniGene build III" (16,046 clusters, SOL), the "Brazilian Coffee Initiative" (16,801 clusters) (*57*) and the "French *C. canephora* sequencing consortium" (52,683 clusters). The first two resources are the result of Sanger sequencing (about 78,470 ESTs from various tissues), the latter of deep Illumina sequencing (approximately 118 million 76 bp reads of total RNA isolated from leaf, stem and flower tissues of the accession 200-94. The assembly of these 3 resources was made using the CAP3 software (http://seq.cs.iastate.edu/) with default parameters. Redundancy among contigs was investigated using BLAST (*i.e.*, all-against-all sequence comparison). Duplicates (*i.e.*, more than 95% identity on more than 97% of sequence length) were removed, keeping the longest sequences. A total of 56,216 non-redundant clusters were retained with an average length of 663 bp. The UniGene set was first aligned with the *C. canephora* genome assembly using BLAT (*35*). BLAT alignments between translated genomic sequences and translated transcript sequences were performed using default parameters. For each alignment, a minimal percentage identity, set to 90%, was required. For each UniGene sequence, the best match and all matches with a score within

90% of the best match score were retained. Then, to refine BLAT alignments, we used Est2Genome (*56*).

RNA-Seq and Gmorse

RNA-Seq reads from the *C. canephora* accessions 200-94 (stem and flower, young and old leaves) and BD54 (leaves) were obtained by sequencing cDNAs using Illumina technology at either Genoscope (accession 200-94) or MGX-Montpellier (accession BD54). The single-end reads obtained were 76 nucleotides long. Finally, 355,191,128 reads were mapped to the *C. canephora* genome using BWA (*33*) with default parameters. Reads that aligned on exon-exon junctions could not be mapped to the genomic sequence. To improve read mapping, we split 80 bp unmapped reads into two parts of 40 bp and again launched the mapping using SOAP2(*58*). This resulted in the mapping of 32,278,671 fragments of 40 bp. Using the mapped and unmapped reads, we launched the Gmorse software (*59*). We obtained 930,181 transcript models with a plausible coding sequence (CDS greater than 50 amino acids), clustered into 162,177 loci. Furthermore, to generate gene expression data, additional transcriptome sequencing was performed using a set of various samples (**Table S5).**

Integration of resources using GAZE

All resources described here were used to automatically build *C. canephora* gene models using GAZE (*60*). Individual predictions from each of the programs (SNAP, FGENESH, Genewise, Est2genome, Gmorse) were broken down into segments (coding, intron, intergenic) and signals (start codon, stop codon, splice acceptor, splice donor, transcript start, transcript stop).

Exons predicted by *ab initio* software packages – Genewise, Est2genome and Gmorse – were used as coding segments. Introns predicted by Genewise, Est2genome and Gmorse were used as intron segments. Intergenic segments were created from the span of each mRNA, with a negative score (forcing GAZE not to split genes). Predicted repeats were used as introns and intergenic segments, to avoid prediction of genes encoding proteins in such regions. The entire genome was scanned to find signals (splice sites, start and stop codons). In addition, transcript stop signals were extracted from the ends of mRNAs (polyA tail positions).

Each segment extracted from a software output that predicts exon boundaries (such as Genewise, Est2genome, or *ab initio* predictors) was used by GAZE only if GAZE chose the same boundaries. Each segment or signal from a given program was given a value reflecting our confidence in the data, and these values were used as scores for the arcs of the GAZE automaton. All signals were given a fixed score, but segment scores were context sensitive; coding segment scores were linked to the percentage identity (%ID) of the alignment, and intronic segment scores were linked to the %ID of the flanking exons. A weight was assigned to each resource to further reflect its reliability and accuracy in predicting gene models. This weight acts as a multiplier for the score of each information source, before processing by GAZE. Finally, gene predictions created by GAZE were filtered according to their scores and lengths. An additional filter was applied to remove genes corresponding to mitochondrial and chloroplast insertions; chloroplast and mitochondrial sequences were aligned to the genome and to the annotated transcripts, which identified 746 chloroplast and 289 mitochondrial regions in the genome, covering 700,869 and 430,786 bp, respectively (for which GAZE gene predictions were discarded). The final C. *canephora* annotation contains 25,574 gene models (**Table S6**).

Annotation validation

In order to validate the annotation, the 25,574 *C. canephora* proteins were BLASTed against the *Arabidopsis* proteome (TAR10, 27,416 proteins), alongside that of tomato (iTAG2.3, 34,727 proteins) and grape (*V. vinifera*) (UniProt, 29,836 proteins) (*10*). Most of the *Coffea* proteins showed similar lengths to their *Arabidopsis* counterparts. Only 8% of predicted *Coffea* proteins did not find a match in *Arabidopsis*, as compared to 18% of tomato and 14% of grape. These comprise species-specific proteins, but probable misannotations as well (**Figure S4**). Overall, the data suggest that the annotation is of comparable quality to the iTAG tomato and potato annotations (*10*).

As another quality check, 63 families of transcription factors (TF) were annotated using protein-coding sequence and InterPro domain (IPR) scans available in the GreenPhyl database (InterPro v28). Most families showed similar sizes to those of other sequenced plant genomes (**Table S7**), although a notable expansion of the RWP-RK family (IPR003035) was observed.

Functional annotation and categorization of the coffee genome

All coffee gene models were functionally annotated by assigning their associated generic Gene Ontology (GO) terms and Enzyme codes (EC) through the Blast2GO program (*61*), based on homology to proteins from other species as determined by BLAST and the occurrence of INTERPRO functional domains identified by INTERPROSCAN. Annotations were further expanded using ANNEX (*62*). The following settings were used: BLAST searches were conducted for each protein (BLASTX, nr database, HSP cut-off length 33, report 20 hits, maximum e-value $1E^{-10}$), followed by mapping and annotation (e-value hit filter $1E^{-10}$, annotation cut-off 55, GO weight 5, HSP-hit coverage cut-off 20). Note that a gene might have more than one distinct function and therefore might be annotated with more than one GO functional category or EC code. To have a broader overview of functional annotation, the resulting GO terms were also mapped onto the corresponding Plant GO slim terms. We performed significance analyses of differential distributions of GO terms in subsets of coffee genes versus GOs for the entire genome by means of Fisher's exact test. To control for multiple testing, the resulting p-values were corrected by means of Benjamini and Hochberg (*63*).

Identification of clusters of orthologous genes using OrthoMCL

As a prerequisite to comparing gene content of *C. canephora* to other organisms at the whole genome scale, we constructed families of homologous proteins from all sequences of *C. canephora* and a representative sample of eudicot organisms, i.e., *Arabidopsis thaliana* (TAIR, release 10), *Vitis vinifera* (Genoscope, release 1) and *Solanum lycopersicum* (ITAG, release 2.3). The complete dataset represents 118,151 protein coding gene sequences.
We first removed 7,057 highly similar paralogous genes using the CD-HIT algorithm (*64*). Transposable elements were filtered out by using BLASTP (*65*) against RepBase, corresponding to 3,361 sequences. Finally, an all-against-all comparison using BLASTP was performed with an e-value cut-off of 1e-10. Clustering was then performed based on a Markov cluster algorithm (MCL) using OrthoMCL (*66*) (version 1.4) with default parameters.

81,837 of 107,733 protein sequences (76.0%) were clustered into 16,917 orthologous groups. Of the 25,574 protein-coding genes predicted for *C. canephora*, 18,451 were clustered into a total of 13,545 groups. 1,475 genes were clustered into 430 clusters specific to *C. canephora*, of which 1,155 have at least one InterPro domain. A Venn diagram representing these data is shown in **Figure S5**.

Non-coding gene annotations

Sequences of mature plant miRNAs were retrieved from the Plant MicroRNA Database (*67*) and used as queries to search the *C. canephora* genome assembly using the LeARN pipeline (*68*) with default parameters (gap_value=2, mm_value=1, gu_value=0.5, score_threshold=3, min_length_alignment=20 and no_mismatch_positions=10;11).

A list of 92 microRNAs precursors with their predicted hairpin structures were computationally predicted from 33 families based on sequence similarity with known miRNAs in PMRD (**Table S8**). As their identifications were based on the sequence homology with other plants, the miRNA population size in *Coffea canephora* is likely underestimated.

miRNA families identified in *Coffea* are deeply conserved among plant phyla. According to the phylogenetic classification done by Cuperus *et al*. (*69*), eight families were already present in the common ancestor of embryophytes, one in tracheophytes, two in spermatophytes, and 21 in the angiosperms (**Table S9)**. Only one was found to be specific to the eudicots (miR2111, **Table S9**).

Organelle-derived sequences

The exchange and redistribution of genetic material among the mitochondrial, plastid, and nuclear genomes has been a major force in shaping plant genome evolution. Organelle-derived sequences, collectively referred to as nuclear organelle DNA (norgDNA), are abundant in sequenced plant genomes, and organelle to nucleus transfers are still on going. Most norgDNAs are less than 1 kb in length (*70-72*) with several notable exceptions, including a 620 kb fragment of nuclear mitochondria DNA (NUMT) found on chromosome 2 in *Arabidopsis* (*73, 74*), and a 131 kb fragment of nuclear plastid DNA (NUPT) on chromosome 2 in rice (*75*). The turnover of organelle-derived fragments is high; an estimated 80% of the norgDNA insertions are lost within a million years of integration (*76*).

Organelle-derived fragments were identified in the *C. canephora* genome using the published *C. arabica* chloroplast genome (GenBank ID: EF044213.1) and the *Nicotiana tabacum* mitochondrial genome (GenBank ID: BA000042.1), as the *C. canephora* organelle sequences remain unassembled. A total of 2,014 NUPT insertions were identified in the *C. canephora* genome. NUPT fragments are distributed randomly throughout the genome and range in size from 100 bp to 16,000 bp (**Figure S6**). Interestingly, ten fragments are larger than 10kb and 25 fragments are larger than 5 Kb. This contrasts with the relatively small NUPT fragments identified in grape, soybean, *Arabidopsis*, sorghum, and rice, but is comparable to fragments found in maize (*77*). NUPT fragments collectively represent about 0.16% of the *C. canephora* genome, which is comparable to other sequenced plant genomes. We calculated the divergence times of NUPT fragments to estimate their relative age and time of insertion. The synonymous substitution rate per synonymous site (Ks) between the inserted and published chloroplast genome sequences was used with the molecular clock from *Arabidopsis* (7×10^{-9} substitutions per synonymous site per year) to calculate divergence times.

NUPT fragments show a range in insert times from 0-14 million years ago (Mya). Older fragments are either too divergent to identify, or were subjected to shuffling and elimination mechanisms. Thirty per cent of NUPT fragments have an estimated age of less than 2 Mya, suggesting NUPT integrations are frequent and ongoing in the *C. canephora* genome. Longer fragments have the lowest divergence (**Figure S7**), suggesting insertions are quickly fragmented and shuffled following integration.

NUMT fragments were identified using the *N. tabacum* mitochondrial genome. Plant mitochondrial genomes have largely conserved gene content, but differ extensively in size, gene order, repeat composition, and chloroplast genome transfers. Estimates for NUMT

insertions are thus likely to be underestimations because of differences between the tobacco and *C. canephora* mitochondrial genomes. Nonetheless, 559 insertions were identified, summing to over 900 Kb or 0.15% of the assembled genome. NUMT insertions are on average significantly smaller than NUPT insertions, with one notable exception. A 750 Kb cluster of NUMT fragments was identified in unanchored scaffold 158. Organelle insertions preferentially integrate into the highly repetitive pericentric region of chromosomes, likely explaining why this scaffold is unanchored using markers from the genetic map. The cluster of NUMT fragments contains 21 genes and 5 gene fragments, representing around half the estimated gene content of the mitochondrial genome. This NUMT cluster is the largest identified to date in any sequenced plant genome, with the only other large fragment reported being a 620 kb fragment on chromosome 2 in *Arabidopsis* and a 108 kb fragment on chromosome 1 in maize. This may suggest an alternative integration mechanism for organelle-derived sequences in the coffee genome.

Identification, classification and distribution of Transposable Elements in the *C. canephora* genome

De novo TE identification

The REPET TEdenovo package (v2.1) was used to *de novo* identify TEs in *C. canephora* contigs (*78*). 25,217 contigs were used for self-comparison with PILER. After clustering, 6,812 reference sequences were kept and classified according to the REPBASE database (http://www.girinst.org/repbase/) and our manually annotated TE library. Elements were classified according to Wicker et al. (*79*) and each class was investigated to confirm the quality of prediction. On the 6,812 reference sequences, 6,414 fell into Class I elements (94%). We observed that *de novo* detection found a very low number of Class II transposons, few MITEs, and few SINEs, but a huge number of LTR retrotransposons (RLG and RLC) and non-autonomous LTR retrotransposon (RXX, TRIM and LARDS) reference sequences (**Figure S8**). *Gypsy* (RLG) and *Copia* (RLC) sequences were clustered into 189 RLC and 388 RLG clusters according to the rules from Wicker et al. (*79*).

The TEdenovo dataset is composed of 6,812 reference sequences for which 1,393 were classified as "chimeric" (elements with more than one TE classification). These elements could be nested TEs, and they were removed from the dataset for further analysis. Each category found by TEdenovo was investigated and compared to previous TE identifications. Comparisons were first performed with the TE manual library. We found that 77% of the reference sequences predicted by TEdenovo have strong similarities with the TE manual library. Most of them (90%) have identical classification between both databases. Differences in classification were mainly due to unidentified nested elements within manually annotated sequences.

TE annotation along the C. canephora pseudomolecules

We used the 5,363 non-chimeric repeats predicted by TEdenovo to annotate the *C. canephora* pseudomolecules. Results were combined with previous mapping analysis. In total we found that TEs account for half of the coffee pseudomolecules (49.2%, **Table S10**). LTR retrotransposons represent 42% of the sequenced genome. Among them, the *Ty3-Gypsy* family represents the largest part (24.1% of the genome). All known classes of plant TEs are present in the genome, but none of these represent a significant part of it (**Table S10**). The distribution of TEs was plotted and compared to gene density (**Figure S9**). As observed in other plant genomes and already suggested for coffee (*32*), there is an inverse relationship between gene density and repetitive sequences.

Conservation of LTR retrotransposon groups in other plant genomes: new cases of outstanding conservation of Ty1/Copia elements between plant genomes

All LTR retrotransposon groups (with the exception of RLX) were used as queries to perform BLASTN searches against a selection of 33 available sequenced plant genomes retrieved from NCBI and Phytozome as follows: 24 dicotyledonous genomes - *Nicotiana sylvestris*, *Solanum lycopersicum*, (tomato), *Solanum tuberosum* (potato), *Mimulus guttatus*, *Utricularia gibba* (bladderwort), *Vitis vinifera* (grape), *Cucumis sativus, Citrullus lanatus* (watermelon), *Fragaria vesca* (strawberry), *Prunus persica* (peach), *Malus domestica* (apple), *Medicago truncatula, Cicer arietinum* (chickpea), *Lotus japonicus, Glycine max* (soybean), *Phaseolus vulgaris* (common bean), *Populus trichocarpa* (poplar), *Manihot esculenta* (cassava), *Ricinus communis, Theobroma cacao* (cacao), *Carica papaya* (papaya), *Arabidopsis thaliana, Brassica rapa* (rapeseed), and *Citrus clementina* (clementine); 7 monocotyledonous genomes - *Phoenix dactylifera* (date palm), *Elaeis oleifera* (oil palm), *Musa acuminata* (banana), *Zea mays* (maize), *Sorghum bicolor* (sorghum), *Brachypodium distachyon* (false brome), and *Oryza sativa* (rice), and two other genomes: *Amborella trichopoda* (a basal angiosperm) and *Selaginella moellendorffii* (a non-angiosperm). The best BLASTN bit scores for each LTR retrotransposon group used as query are plotted on the y axis of a graphic composed of all ordered LTR retrotransposon groups on the x-axis (**Figure S10**; in blue are all *Copia* groups and in green, all *Gypsy* groups). High BLASTN bit scores (higher than 3000) were observed for some *Ty1-Copia* groups in *Mimulus*, common bean, poplar, castor bean, clementine and banana, while scores for *Ty3/Gypsy* groups remained low for all studied species. The conservation of LTR retrotransposons was studied in detail in the *M. acuminata* genome (*80*). Thirteen *C. canephora* sequences from the *Ty1/Copia* group were found to be conserved with sequences from the banana genome with a score higher than 2000. All of them belong to the Tork clade (**Figure S11**).

Measurement of gene order evolution

The measurement of gene order evolution through the comparison of genomes is complicated by WGDs – whole genome duplication or triplication events. There may be several orthologs in one genome of single-copy genes in another genome, and these paralogs may each have retained or changed their gene order contexts in different ways. Moreover, after WGD there is paralog loss on a massive scale, deleting gene copies by the process of *fractionation* (*81*). This results in the scrambling of gene order (*82*), considering the order of genes along a chromosome as a piecing together of gene *adjacencies*.

We deal with gene order changes in the context of fractionation through a measurement of *excess adjacencies* in the comparison of two genomes (*82, 83*). Examining only the genes with at least one ortholog in the other genome, the numbers of different gene adjacencies (respecting gene orientation) in each genome, a_1 and a_2, respectively, are compared with the total number of different gene adjacencies in the two genomes $a_{1,2}$. If $a_1=a_2=a_{1,2}$, the two genomes are identical, or differ only through polyploidy. If, at the other extreme, the two genomes have gene orders that are completely scrambled, then $a_{1,2}=a_1+a_2$. Then

$$D_1=a_{1,2} - a_1 \text{ and } D_2=a_{1,2} - a_2$$

are measures of genomic divergence, while

$$d_1=D_1/a_{1,2} \text{ and } d_2= D_2/a_{1,2}$$

are normalized measures of genome divergence.

The traditional ways of assessing gene-order evolution, such as breakpoint distance, reversal/translocation distance or double-cut-and-join, lose their motivation – as measures of genome rearrangement – when fractionation is a major cause of gene-order disruption. Excess adjacencies *D* or normalized excess adjacencies *d* are more appropriate measures in this case.

We will refer to two sister genomes, one having undergone a WGD (or triplication) and one having escaped it, since their divergence, as a polyploid and diploid, respectively.

We wish to remove the effects of fractionation on gene-order disruption, so that the effects of rearrangements can be discerned. Rearrangement is the component of gene order change we need to measure, being most regularly associated with genome divergence. Thus, in comparing a diploid genome with a sister polyploid (e.g., hexaploid), the correction is done by identifying each *fractionation interval* I_0, which is a set of contiguous genes in the diploid whose orthologs are also present in single copies only in the hexaploid but partitioned among two or three contiguous subsets (intervals) I_1, I_2 and sometimes I_3 at different locations in the genome. An efficient *consolidation* algorithm makes this task feasible in reasonable computing time (*83*). Then the three subsets I_1, I_2, I_3 in the polyploid are *consolidated*, and they and I_0 are each replaced by a dummy or *virtual* gene, all identically labelled, say as *C*.

Eventually all the single-copy genes in the polyploid are incorporated into virtual genes. For the purposes of gene-order comparison, replacing all the smaller subsets in the polyploid with the identical virtual gene as in the diploid, removes all the excess adjacencies caused by fractionation within the subsets. When this is done, most of the remaining excess adjacencies are indicative of gene-order evolution by rearrangement. There is one important class of exceptions: when I_3 is absent, we can imagine that in the original polyploid there were three copies of all the genes in I_0, arranged in three contiguous intervals, but that through fractionation, one of these intervals has been completely lost. There will be no direct evidence of the position it originally occupied; instead there will be a new adjacency, which will be counted as an excess adjacency, again caused by fractionation and not by rearrangement.

In most cases we can correct this problem using syntenic context to identify the position of the annihilated, or "null", *I3*, and to insert *C* into that position, reducing by 1 the number of excess adjacencies not due to rearrangement.

As an example of applying our method to compare coffee and grape to tomato, we first need to identify orthologs between each of the diploids and tomato (*10*). To do this we use the SYNMAP package (*84*) on the COGE platform (*85*), which combines sequence similarity with syntenic context to produce comparative maps between two genomes, including comparative gene orders for suitably annotated genomes. We then apply the OMG! procedure (*86*) to resolve the small number of output gene groups with more than one member in the diploid genome or more than three in tomato.

We first note in **Table S11** that the severe criteria for both sequence similarity and syntenic context limit the number of genes detected to less than 40% of the total number in the genomes. The remaining genes either show no orthology in the other genome, or have not remained in an evolutionarily stable syntenic context in one or both of the genomes. The fact that more orthologs and tomato paralogs were retained in the coffee comparison is an indication of the more recent shared lineage between these two genomes, and of their earlier divergence from grape. This is bolstered by the fact that gene similarity scores reported by SYNMAP average 74.4% for coffee/tomato, significantly higher ($p < 10^{-20}$) than the 73.5% for grape/tomato, despite the fact that grape is thought to be the most slowly evolving among the sequenced core eudicots (*87*), although not as slow as *Nelumbo nucifera* (sacred lotus), a more basal eudicot (*88*).

Coffee-specific gene family size changes

It is well-known that, in moderately-sized families, genes are gained by tandem gene duplication and lost by deletion (or pseudogenization), by the so-called birth and death (BD) process (*89*). Both birth and death have density-dependent rates (*90*), because the probability of having a gain or loss event depends on the number of members (e.g., the higher the number of gene copies, the higher the probability of having a gene duplication by unequal crossing over). However, such density-dependence of the birth rate imposes some limitations to the applicability of the BD model in genome-wide studies. First, gene families are often defined by automated methods such as OrthoMCL (*66*) that cluster divergent gene families into small orthogroups. In species-specific orthogroups, the BD model cannot be applied (*91*), since no duplications can arise from zero ancestral members (i.e., zero is an absorbing state). Second, the stochastic BD model only accounts for single-gene duplications (*92*), and thus it is not suitable to model WGD events (which are relatively common in plant species). To circumvent these problems, we thus applied the gain and death (GD) stochastic model, as implemented in the BadiRate program (*11*). Unlike the birth rate, the gain rate is density-independent, and accounts for all gene acquisitions, regardless of whether they originated by tandem gene duplication, exon shuffling, WGD, or any other molecular mechanism. Additionally, this stochastic GD model provides the appropriate statistical framework for testing biologically relevant hypotheses.

To detect species-specific rates of gene gain and death (GD), we compared the fit of different branch models to the data: (i) a global rates model (following the BadiRate notation, the GD-GR-ML model), where all lineages share the same GD rates, (ii) a species-specific rates model (GD-SSR-ML), where GD rates might vary in a given focal species, and (iii) a free rates model (GD-FR-ML), where each lineage can exhibit particular GD rates. We evaluated the strength of evidence of the six branch models (a GD-SSR-ML for each one of the four focal species, a GD-FR-ML and a GD-GR-ML) via the weighted Akaike Information Criterion (wAIC). In particular, we considered that a branch model has significant support if its wAIC ratio is 2.7 times higher than the second best fit branch model (*93*). Nevertheless, this test only allows detecting gene families with species-specific GD rates. To determine whether such significantly different rates correspond to expansions or contractions, we compared the number of family members in the extant species with those in its most recent common ancestor (its parent node in a phylogenetic tree). If the net size change is positive, we consider a gene family size change an expansion, and *vice versa*. This analysis was performed using the BadiRate–anc option, which conducts a joint ancestral reconstruction of the gene family sizes at all the internal nodes of the phylogenetic tree.

Since the low amount of phylogenetic information per gene family (relative to the number of parameters) might yield local maxima during the likelihood optimization, we also implemented some modifications to the BadiRate program, including a more comprehensive assessment of the likelihood values on a wider parametric space, and the possibility of using more accurate initial starting values. In particular, since the GD-GR-ML model is a specific case of GD-SSR-ML, we used the best GD-GR-ML estimates as initial values for GD-SSR-ML. Likewise, we used the best GD-SSR-ML estimates as starting values for GD-FR-ML. To accomplish a more thorough exploration of the parametric space, we also performed additional computer replicates, using random and parsimoniously inferred starting values. For each family, the GD-GR-ML model was run 5 times, while the GD-SSR-ML and the GD-FR-ML models three times. The results of the BadiRate analyses on the 16,917 orthogroups identified by OrthoMCL are shown in **Table S12**. The functional categories (plant GO slim and generic GO terms) differentially represented among coffee-specific expanded genes families are displayed in **Tables S13 to S15.**

Identification and classification of putative NBS resistance genes

Hidden Markov models (*94*) and BLAST (*65*) searches were used to identify NBS encoding *R*-genes and their homologues in the coffee genome. Predicted coffee open reading frames (ORFs) were screened using HMMs to search for the Pfam NBS (NB-ARC) family PF00931 domain (http://pfam.sanger.ac.uk). HMMER search 3.0 (http://hmmer.janelia.org) retrieved 399 proteins of high quality (< 1 x E-60). Their sequences were aligned using MAFFT (*95*) and used to construct a coffee-specific NBS HMM using HMMER build 3.0. This permitted the identification of 1496 NBS-candidate proteins (< 1E-2), which were individually analysed. The presence of NBS domains was confirmed using the National Center for Biotechnology Information's (NCBI) Conserved Domains tool (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and the Pfam protein families database version 26.0 (*94*). From the 1496 candidate proteins only 561 proteins were selected as putative NBS resistance genes on the basis of complete NB-ARC domains in NCBI and significant NB-ARC domains in Pfam (**Table S16**). The coffee NBS-encoding gene family accounts for approximately 2.2% of total predicted genes in the *C. canephora* genome. This proportion is higher than that of other angiosperm genomes: 1.8% in *Vitis vinifera* (*9*), 1.4% in *Oryza sativa* (*96*), 0.9% in *Populus trichocarpa* (*97*) and *Theobroma cacao* (*98*), and 0.7% in *Carica papaya* (*99*) and *Arabidopsis thaliana* (*100*).

In order to classify the putative NBS resistance genes, Pfam HMM searches were performed using a TIR model (PF01582) and several LRR models (PF00560, PF07723, PF07725, PF12799, PF13306, PF13516, PF13504, PF13855, PF08263), to detect TIR domains and LRR motifs in the NBS-encoding amino-acid sequences. The presence of TIR domains and LRR motifs was validated using NCBI's Conserved Domains tool and Pfam analysis. CC motifs were detected using the COILS prediction program 2.2 (http://www.ch.embnet.org/software/COILS_form.html) with a *p* score cut-off of 0.9. Predicted NBS genes were then evaluated using the Multiple Expectation Maximization for Motif Elicitation tool (MEME, http://meme.ebi.edu.au/meme/) to show domain and motif organization, and to detect the number of LRRs. MEME was run on predicted ORFs of TIR and non-TIR genes both together and separately. Two hundred and eighty genes (49.9%) had an amino-terminal CC motif in the N-terminal region but only four genes (0.7%) with a TIR domain were detected (**Table S16**). One hundred and forty-five genes (25.8%) presented LRR motifs in the C-terminal domain. TIR motifs are markedly underrepresented in the coffee genome (0.7% of NBS genes) compared to tomato (13.9%), potato (14.9%), grapevine (20.7%), poplar (31.1%) and *Arabidopsis* (70.3%) (**Table S16**). The LRR motifs were also less abundant in the coffee genome than in other eudicots.

Identification of C. canephora caffeine biosynthesis genes

Full length mRNA (complete CDS) sequences of ten previously described *Coffea N*-methyltransferases (*18, 101*) were used to search the *C. canephora* genome using the BLASTN program with an E-value cut-off of 1E-40 and an alignment length higher than 90% of the sequence length. Alignment and manual annotation were performed using the *Coffea* N-methyltransferase proteins validated by McCarthy et al. (*18*) and Yoneyama et al. (*101*) and retrieved from SwissProt.

Multiple amino-acid sequence alignments of NMTs were performed using Muscle (*102*). Maximum likelihood trees were built using MEGA5 (*103*) with the JTT model and 1000 bootstrap replicates.

Phylogenetic analysis of NMTs from coffee and other caffeine producting plants

We used TBLASTN in CoGe using the CcXMT protein as query, and a cutoff of 1E-10, to identify related NMTs in grapevine, peach, poplar, *Arabidopsis*, *Mimulus*, potato, and

Utricularia genomes. Translated CDS were aligned using MUSCLE and manually inspected for completeness; short CDS were excluded from the alignment to optimize our ability to recover meaningful overall phylogenetic patterns. The sequences were then realigned, and trimmed using Gblocks to derive a data matrix for maximum likelihood analysis using the RaxML web server (http://embnet.vital-it.ch/raxml-bb/). The alignment was analyzed both as nucleotide CDS and as translated amino acids. In both cases, 100 bootstrap replicates were specified with a gamma model of rate heterogeneity; for nucleotide data, the GTR model of molecular evolution was specified, whereas for protein data, the JTT model was used.

Microsyntenic analyses of *C. canephora* and *T cacao* NMT genes

Microsyntenic analyses were performed using CoGe (see complete description of methodologies used in Ibarra-Laclette et al. (*104*)) with coffee and cacao gene models as input. CoGe BLAST and GEvo analysis identified 4 large genomic blocks containing at least one NMT coding sequence.

Pairwise estimates of synonymous substitution rates (Ks) among NMT CDS

Since silent substitutions are effectively neutral, they are expected to approximately reflect the background mutation rate. We obtained maximum likelihood Ks estimates using codeml in PAML. Relative time units for all Ks values were based on normalization with the smallest (youngest) Ks value, such that the within-group average for genes labeled blue (Figure 2, main text) was equal to 1.

Positive selection analyses of NMT genes

Two different classes of models were implemented using the codeml program. First were the "branch-specific" models, which permit heterogeneity in ω ratios among selected branches in the phylogeny, previously defined as "foreground" branches (*105*). Second, were the "branch-site" models, which allow ω values to vary in selected branches on the tree while also taking into account heterogeneity in selective pressures throughout the sequences, by defining different codon site-classes with different ω ratios (*106*). In our analyses, each stem-lineage branch leading to caffeine biosynthetic genes in coffee, tea, and cacao (respectively) was considered as the foreground branch under examination. Within the first model class (branch-specific class), we ran the one-ratio model 0, constraining all branches in the tree to the same ω ratio, and then the two-ratios model 2, allowing for the foreground branch to evolve at an independent ratio than the rest of branches. Within the second class (branch-site class), we ran clade model D, allowing a class of sites to evolve under heterogeneous selective pressures between the foreground branch and the rest of the tree (*107*), and then model A (*24*), featuring an extra class of sites under PS with $\omega > 1$ in the foreground branch. These models perform maximum likelihood (ML) estimates of ω ratios and attach a log likelihood (lnL) value to each examined alignment and tree topology. Likelihood Ratio Tests (LRTs) permit comparison of the fit of two nested models by examining the significance of differences between their lnL values (calculated as 2ΔLnL - twice the difference between their lnL values) (*108*). LRTs asymptotically follow a χ^2 distribution with the number of degrees of freedom equal to the differences in number of parameters between the models being compared.

LRT tests were used to compute a p-value for the fitting of the examined dataset to the alternative model being tested. To test for asymmetric sequence evolution, we compared the one-ratio model 0 to the two-ratios model 2, allowing a different ω to be estimated for the foreground branch. To measure divergent selective pressures acting on a significant number of amino acids (codon sites), a test comparing the site-specific null discrete model 3 (which allows the ω ratio to vary among sites while holding ω constant among branches in the tree)

and clade model D was used (*107*). Divergent selective pressures occurring on a substantial fraction of amino acids may be indicative of functional specialization, either under relaxed purifying selection or under PS. To explicitly test for PS as opposed to relaxed purifying selection affecting a few sites in the selected branch, we compared null model A, where ω is fixed at one for the foreground branch as null model, and the model A (*24*). Model A also implements a Bayes Empirical Bayes (BEB) procedure that calculates posterior probabilities of a codon being subjected to PS (*109*).

Supplementary text

Author contributions

L.C.-.P., A.D., G.D., R.G., M.P. and C.Z. contributed equally to this work. A.C.A., A.D.K., G.Gr., D.S., G.Gi., V.A.A., P.W. and P.La. coordinated the study. F.D., D.S., G.Gi., V.A.A., P.W. and P.La. wrote the paper, with important contributions from L.C-P. A.A., P.B., M.B., M.C.C., D.C., L.D., K.L., L.T.L., L.L., B.N., A.P., G.P., D.P., P., M.Ri., Q.Z., A.K.A., B.B., A.D.K., G.Gr., R.J.H., J., R.M., C.N. and S.R. generated data. F.D., L.C.-.P., A.D., G.D., R.G., M.P., C.Z., F.A., G.A., J.M.A., S.B., C.C., A.C., D.C., C.D.S., F.D., P.W., F.D.B., S.D., O.G., T.G., V.G., K.J., V.J., T.J., T.L., J.L., M.L., T.Le., L.T.L., P.Li., A.M., B.N., V.P., D.P., M.Ri., M.Ro., J.R., C.T.D., R.V., Q.Z., X.A., A.D.K., R.M., S.R., D.S., G.Gi., V.A.A., P.W. and P.La. analyzed data.

Gene order evolution after polyploidization

Figure S12 shows the orthologies between genes in coffee, grapevine and tomato. These orthologies are partitioned among twenty-one panels in the figure, representing the twenty-one ancestral core eudicot chromosomes inferred to have resulted from the triplication of an earlier seven-chromosome ancestor (*9*). The one-to-one correspondence between grapevine regions and coffee regions is unequivocal, as is the more recent triplication affecting tomato, which has a three-to-one correspondence with coffee and with grape.

When the data from the pepper genome (*110*) is compared with coffee and grapevine, a very similar pattern emerges, as is evident in **Figure S13**. Pepper clearly has undergone the same triplication as tomato.

We calculated *d* (*4*) between six genomes unaffected by recent WGD (coffee, peach, cacao, grape, papaya and strawberry) and a larger group of core eudicots including these six as well as seven recent WGD descendants (*Mimulus*, *Utricularia*, tomato, *Arabidopsis*, *Medicago*, soybean and poplar). In the comparisons involving the latter group, a consolidation analysis was applied to remove excess adjacencies due to fractionation. The results, in terms of non-shared adjacencies in each pair of genomes are displayed in **Table S17**, and well as in **Figure 1D** (main text).

Disease resistance-related genes

Plant resistance to a range of pathogenic organisms (bacteria, fungi, insects, nematodes, oomycetes and viruses) is conferred by a diverse group of disease resistance proteins (*13, 111*). Classification of these proteins is based primarily on predicted domains and motifs. One of the largest families of resistance proteins encodes NBS domains (*12*). The NBS proteins are subdivided into different classes based on the structure of their N- and C-terminal domains (*112*). The N-terminal domain contains either a Coiled-Coil (CC) motif, a Toll/Interleukin-1 Receptor (TIR) motif or a sequence without obvious CC or TIR motifs. The C-terminal domain occurs either with or without Leucine-Rich Repeats (LRRs) (*13*). The NBS domain functions by binding and/or hydrolyzing ATP (*113, 114*). The LRR domain interacts with the products of pathogen AVR genes directly or indirectly and hence is thought to function primarily in the recognition of the presence of pathogens (*115*). TIR domains are involved in resistance specificity determination and signalling (*116*). Together, the domains of NBS-LRR proteins function to directly or indirectly detect pathogen effectors and activate defence signal transduction in plants (*117, 118*).

Genomic organization of putative NBS resistance genes

Three hundred and forty-eight putative NBS resistance genes (*4*) (62.0%) were mapped onto the 11 *C. canephora* chromosomes. The remaining genes were situated on unanchored scaffolds (ChrUn). Of the 348 anchored NBS genes, 244 (70.1%) were found on five chromosomes (1, 3, 5, 8, 11), with a maximum of 54 genes (15.5%) on chromosome 11 (**Table S18**). By contrast, chromosomes 9 and 10 contain only two (0.6%) and ten (2.9%) anchored NBS genes, respectively. Unlike poplar (*119*) and watermelon (*120*), no chromosome lacks NBS-encoding genes in coffee. Gene clusters were searched for in the coffee genome according to Holub's (*121*) definition, i.e., using a region containing four or more genes within 200 kb or less. From the 348 anchored NBS genes, 169 reside in 30 gene clusters that are unevenly distributed over the chromosomes (**Table S18**). The richest chromosome (chr3) contains seven gene clusters, representing 39 NBS genes. No gene clusters were detected on chromosomes 9 and 10 (although singletons were found). The other ten gene clusters were detected among the unanchored NBS genes (48). Cluster size varied between 45.7 Kb (CL18 with 5 genes) and 657.1 Kb (CL23 with 13 genes). The average number of genes in a cluster is 5.4 (14 clusters with four genes, 12 clusters with five, 7 clusters with six, 3 clusters with seven, 2 clusters with eight, and 1 each with nine or 13 genes). Similarly to other plant genomes (*100, 119, 122, 123*), the clusters are composed of combinations of CNL-, NL- and N-encoding genes. No TIR gene was found in a cluster. Aside from gene clusters, there are 47 tightly linked doublets of NBS genes and 27 triplets. Therefore, a total of 392 NBS genes reside either in a gene cluster or in a tandem array. In all, 169 singletons are dispersed over the coffee genome. The ratio of singletons to the total number of NBS genes (30.1%) is similar to that in rice (30.3%) and poplar (32.5%), but higher than that in grapevine (16.8%) and *Arabidopsis* (26.4%)(*119*).

Orthology relationships of putative NBS resistance genes

The 561 NBS proteins from *C. canephora* were assigned to orthogroups using OrthoMCL (*4*). A total of 530 proteins (94.5%) were assigned to 40 orthogroups. These orthogroups are unevenly distributed on the coffee chromosomes, ranging from two genes on chromosome 9 to 65 genes on chromosome 3 (**Table S19**). ORTHOMCL7 is the most important orthogroup, with 125 coffee NBS genes. Eleven orthogroups appear to be specific to the coffee genome, *i.e.*, are not found in the other genomes included in the analysis. Moreover five of these orthogroups (ORTHOMCL4795, ORTHOMCL4821, ORTHOMCL15067, ORTHOMCL15152, ORTHOMCL15218) only contain coffee NBS genes. These findings suggest an expansion of NBS genes in the coffee genome. For the 40 clusters of NBS genes, 22 fit entirely in a single orthogroup, representing a total of 123 NBS genes (56.7% of clustered NBS genes). Sixteen of these clusters (92 NBS genes) are assigned to four orthogroups (ORTHOMCL5, ORTHOMCL7, ORTHOMCL15, ORTHOMCL19). The NBS genes that are clustered within orthogroups may have evolved by duplication and divergence of linked gene families. Such a mechanism has probably played a major role in the evolution of chromosome 8 since 36 NBS genes out of a total of 42 NBS genes located on this chromosome are found in gene clusters.

The SH3 gene conferring resistance to coffee leaf rust

Coffee leaf rust caused by the biotrophic fungus *Hemileia vastatrix* is a major disease that greatly limits *C. arabica* production in almost all growing countries around the world. The resistance of coffee plants is controlled by at least nine *R* genes called SH1-SH9, either singly or in combination (*124*). Organization and evolution of the *SH3* gene has recently been studied (*39*). The corresponding sequence was used to identify the putative *SH3 C. canephora* gene by BLAST search (*65*). Four NBS genes from chromosome 3 and one from chromosome

6 were selected with a *p* score cut-off of 1 x E-10. The NBS gene Cc03g04710 shows the highest identity with the *SH3* sequence (97.5%) and is therefore considered to be the putative *SH3* gene. Cc03g04710 was assigned to ORTHOMCL4821 (**Table S19**).

Phenylpropanoid pathway

The number of studies that have focused on the relationship between plant-derived foods containing phenolic compounds (flavonoids/hydroxycinnamic acids and esters) with high antioxidant activity and human health (*125-128*) has steadily increased over the last few years. One important group of antioxidants is the chlorogenic acid isomers (CGAs), which are soluble hydroxycinnamic acid esters (HCEs) found at remarkably high levels in the coffee bean (**Figure S14**).

The main CGAs present in green coffee beans consist of esters formed between one or two transcinnamic acids (caffeic or ferulic acid) and quinic acid (*129*), and belong to three classes, each containing three isomers differing in the position of their acyl residues. They include the monocaffeoylquinic acids (CQAs), the dicaffeoylquinic acids (diCQAs) and the feruloylquinic acids (FQAs). Their total content was found to vary from 7 to 14.4% dry matter in *C. canephora* (*130-134*) and from 3.4 to 4.8% in *C. arabica*(*131*). Due to their high concentration, CGAs found in coffee are important for at least two key points. The first one is the growing list of their health benefits; indeed, these compounds have been attributed several pharmaceutical properties, such as antioxidant activity (*135, 136*) and inhibition of the HIV-1 integrase (*137-139*). The second point is related to their corresponding degradation products generated by coffee bean roasting, including chlorogenic acid lactones (*140, 141*) and likely phenolic compounds such as guaiacol and 4-vinylguaiacol (*142*), which are key flavor compounds of coffee. Therefore, due to the high concentrations and roles attributed to CGAs and their related degradation products for health benefits but also for coffee quality, it will be interesting to learn more about how their synthesis/accumulation is controlled on a genetic level. Several authors (*27, 133, 134, 143-146*) have already characterized at least one copy of a number of the key genes encoding proteins involved in CGA metabolism (PAL, C4H, 4CL, HCT/HQT, C3'H, CCoAOMT), and/or their transcriptional abundance in several tissues.

With the sequencing of the coffee genome, we had the opportunity to identify all the genes from each of the six distinct families cited above. Search for genes involved in the initial phenylpropanoid pathway (PPP) and in the metabolism of CGAs was done by running the BLAST program using query sequence data from known coffee plant proteins but also from biochemically and genetically characterized proteins from other model plants whose genomes were already sequenced (poplar (*97*), *Arabidopsis* (*147*), tomato (*10*)). Complementary analyses were then conducted to validate the functional annotation of the candidates, including their alignment with full length plant proteins and the analysis of the corresponding phylogenetic trees, but also by analyzing their protein sequence analysis through the "Batch Web CD-Search Tool", an NCBI's interface to searching the Conserved Domain Database (*148-150*).

The genome-wide search for coffee genes involved in the general PPP and in the metabolism of CGAs led to the identification of 25 genes encoding proteins belonging to 6 distinct families; *i.e.*, PAL, C4H, 4CL, HCT/HQT, C3'H and CCoAOMT (**Tables S20 and S21**). We compared coffee gene copy numbers with those from the genomes of poplar, *Arabidopsis* and tomato (**Table S21)**. The most obvious peculiarity was the presence of one copy of the HQT gene in both tomato and coffee, whereas no copy of this gene had been found in poplar and *Arabidopsis*, which is consistent with previous observations (*151, 152*). The presence of at least one copy of HQT in addition to at least one HCT seems to characterize plant species accumulating hydroxycinnamoyl-quinate esters (CGA), particularly 5-caffeoylquinic acid (CQA), such as tomato (*151, 153*) (*SlHQT*, accession number

CAE46933), tobacco (*153-155*) (*NtHQT*, accession number CAE46932), sweet potato (*156*) (*IbHQT*, accession number BAJ14795), and cardoon (*157*) (*CynCardHQT*, accession number ABK79689). The absence of the gene encoding HQT in *Arabidopsis*, which does not accumulate CGAs, suggests that this gene plays a central role in CQA synthesis. Expression of genes encoding HQT and C3'H1 in both perisperm and endosperm of coffee seeds seemed to be correlated, and were exhausted at the beginning of fruit formation when CQA accumulates (**Figure S15**). In parallel, genes encoding HCT and C3'H2 are weakly expressed, suggesting that these genes function in tandem and are not devoted to CGA biosynthesis (**Figure S14**). An increased *CCoAOMT1* expression (Cc02p18970) appears later (at 180 DAP), which could indicate the beginning of lignification based on degradation of the accumulated CGAs. In the case of *Populus*, a species accumulating hydroxycinnamoyl-quinate esters, the lack of a HQT gene seems to be offset by the expansion of the HCT gene family (*152, 158*) represented by 7 genes in this species. Looking at **Figure S16**, it can also be noticed that the seven poplar HCT protein sequences can be grouped into 2 subgroups based on the highly conserved "HXXXD" sequence. Indeed, as can been shown by the alignment, whereas *Populus trichocarpa* PtrHCT1 and PtrHCT6 possess the "HHAAD" highly conserved motif found in biochemically characterized HCTs from other species having the ability to use both shikimate and quinate as substrates, the five other *Populus trichocarpa* HCTs (PtrHCT2-PtrHCT5 and PtrHCT7) possess an unusual motif, "HI/TLA/GD", that is intermediate between the highly conserved motif "HHAAD" found in the HCT proteins and the highly conserved "HT/NLSD" motif found more specifically in the HQT proteins. A phylogenetic tree built based on an alignment including all of these mentioned proteins (**Figure S17**) is consistent with the classification of the 7 HCT genes of *Populus trichocarpa* into 2 subgroups; whereas PtrHCT1 and PtrHCT6 cluster within the phylogenetic clade containing the HCT enzymes from other plants, PtrHCT2-PtrHCT5 and PtrHCT7 cluster separately, between the two distinct clusters formed between the HQT and HCT proteins, seeming to indicate that these five poplar HCTs might have common substrate specificities with both known HCT and HQTs, perhaps explaining why this species is able to accumulate CGA, contrary to *Arabidopsis*. The consistency between this gene family's expansion/specificities with the difference of metabolism between *Arabidopsis* and coffee/tomato/poplar specifically for CGA accumulation seems to be further strengthened by the observation that the same expansion of genes is also observed for C3'H, a second key enzyme for CGA synthesis. Indeed, whereas coffee, tomato and poplar possess two to four genes encoding C3'H proteins, only one was found in *Arabidopsis* genome.

Given the functional importance that can be attributed to HCT and HQT, we found it essential to go further in order to identify all the other members belonging to the BAHD acyltransferases (benzylalcohol acetyl-, anthocyanin-O-hydroxy-cinnamoyl-, anthranilate-N-hydroxy-cinnamoyl/benzoyl-, deacetylvindoline acetyltransferases), members of which are involved in the biosynthesis of other important plant compounds. These proteins have the ability to acylate plant secondary metabolites whose products include small volatile esters, hydroxycinnamic acids esters, modified anthocyanins, and constitutive defense compounds and phytoalexins (*159*). Using the bioinformatics tools, parameters and approach described in **Table S22**, our genome-wide search identified a total of 25 BAHDs, all containing both the "HXXXD" and "DFGWG" conserved motifs generally found in known BAHDs and including the above discussed HCT and HQT members (The locus IDs and genome location of all the 25 members can be found in **Table S23**). Then, because it is particularly difficult to predict the specific function of each individual BAHD in the modification of secondary metabolites, we aligned the protein sequences from each of the 25 coffee BAHDs with full-length plant BAHDs that have been biochemically or genetically characterized (**Figure S16**). Based on the work published by D'Auria et al. (*159*), the protein alignment presented in

Figure S16, and the related phylogenetic tree (**Figure S17)**, we found that the 25 coffee candidate BAHDs clustered as follows: three in Clade I, involved in the modification of phenolic glucosides such as anthocyanins; eleven in Clade III, involved in the modification of alkaloid compounds and in volatile ester biosynthesis; and eleven, including coffee HCT and HQT, in Clade V, involved in the formation of p-coumaryl shikimate/quinate esters, which are required for HCE synthesis and are intermediates in the lignin biosynthesis pathway, but also in volatile biosynthesis.

Caffeine N-methyltransferases (NMTs) in *Coffea* and their evolution

N-methyltransferases (NMTs) are the principal enzymes involved in caffeine biosynthesis (*160, 161*). Caffeine is a purine alkaloid accumulated in coffee beans and is the most characteristic secondary metabolite of the plant. Its evolutionary role is not clear. Some authors cite an insect repellent or insecticidal effect, but this effect is not strongly supported by the literature (*162*). Another role often invoked is allelopathy, which prevents the germination and/or growth of competing species (*16*).

Identification of C. canephora caffeine biosynthesis genes and comparison to available CDS

Only three different methylation steps are required for caffeine synthesis from xanthosine, which is considered the first precursor. The best supported pathway proceeds through 7-methylxanthosine and theobromine (3,7-methylxanthine) to yield caffeine (1,3,7-methylxanthine). Each of these compounds results from a methylation of its immediate precursor (**Figure 2A**, main text). Alternative pathways have been proposed, through theophylline or through paraxanthine (*19*).

Full length mRNA (complete CDS) sequences of ten previously described *Coffea N*-methyltransferases (*18, 101*) were used to search the *C. canephora* genome (*4*). Several genes presented a conserved sequence homology and structure with previously confirmed NMT genes (**Table S24**, **Figures S18-20**).

Multiple amino-acid sequence alignments of NMTs were performed using Muscle (*102*). Maximum likelihood trees were built using MEGA5 (*103*) with the JTT model and 1000 bootstrap replicates. Three of the genes belong to each of the three known subfamilies: one on chromosome 9 (named *CcXMT*) to the xanthosine methyltransferase (*XMT*), one located on an unanchored scaffold ("ChrUn" pseudomolecule) (*CcMXMT*) related to 7-methylxanthine methyltransferase (*MXMT*), and one on chromosome 1 (CcDXMT) related to 3,7-dimethylxanthine methyltransferase (*DXMT*) (**Figure S20**). One additional gene was found to be similar to the methyltransferase-like (MTL) subgroup defined by Ogawa et al. (*19*). These authors determined that MTLs showed high similarity to *MXMT*, and that the motifs found in MTLs make it highly probable that they possess methyltransferase activity, although perhaps not participating in caffeine biosynthesis.

Phylogenetic position of *C. canephora* NMTs

NMT genes involved in caffeine biosynthesis are rare in the plant kingdom, since only few plants are known to accumulate caffeine. Among these, only the genome of *Theobroma cacao* has been extensively analysed, but not, for example, tea (*Camellia sinensis*), guarana (*Paulina cupana*), cola (*Cola acuminata* and *C. nitida*) nor yerba mate (*Ilex paraguariensis*). Only for tea, and some related *Camellia* species, are there a few sequences corresponding to this pathway in the NCBI database. All of these plants belong to different and distant families: Rubiaceae for coffee, Malvaceae for cacao and cola, Theaceae for tea, Sapindaceae for guarana, and Aquifoliaceae for yerba mate. A common origin and conservation of the pathway is not evident, since many other plants would have been expected to have retained

caffeine biosynthesis during their evolution instead of losing it (*20*). Instead, it seems more probable that convergent evolution has resulted in the selection of a common pathway.

To understand their evolutionary relationships, NMT coding sequences identified in *C. canephora* and in other caffeine-producing plants (*Camellia sp.* and *Theobroma cacao*) (*101*) were placed into a larger phylogenetic context through simultaneous analysis with NMTs mined from a number of core eudicot genomes (*4*). Results are shown in **Figure S21** (nucleotides) and **S22** (amino acids).

With both forms of data, two *Camellia sinensis* caffeine-biosynthetic NMTs grouped with CDS from the cacao genome, including the characterized caffeine synthase gene *TcBS1* (Tc10_g001820). However, given the lack of a tea genome to mine multiple NMTs from, we cannot conclude with reasonable certainty that the sister group relationship between cacao and tea caffeine synthases is correct. In contrast, in both trees, the coffee NMTs with caffeine biosynthetic function appear to be only distantly related to the genes encoding the cacao and tea enzymes. Indeed, there are NMTs from peach, poplar, grapevine, potato, and even ones from cacao and coffee, that are more closely related to the coffee caffeine synthase enzymes. Provisionally then, we can specify that caffeine biosynthetic activity evolved at least twice from within the NMT genes and eudicot genomes sampled.

For a focused view of phylogenetic relationships of NMTs in coffee, cacao, and tea, including for further molecular evolutionary work, we carefully annotated all members of ORTHOMCL170, an orthogroup identified(*4*) that exclusively consists of coffee NMTs and includes all known genes with caffeine synthase function (*18, 19*). We similarly re-predicted gene models for cacao genes determined via BLAST searches of its genome against TcBCS1, which is known to be involved in caffeine biosynthesis. Previously predicted gene models from coffee and cacao were refined using GeneWise (*53*), using the *C. canephora* XMT and TcBCS1 proteins, respectively, and the corresponding genomic sequences. In 6 cases, coffee NMTs were repredicted as full or near-full length, and many other gene models were improved to some degree. Cacao gene models could not be further improved.

Despite our re-prediction efforts, several *C. canephora* NMT genes in ORTHOMCL170 still comprised incomplete NMT sequences, although they contained complete ORFs and SAM dependent carboxyl methyltransferase domains (InterPro IPR005299; **Table S25**). In the absence of evidence to the contrary, and since there are no possibly gene-model interrupting assembly gaps immediately surrounding them (**Figure S23-24**), we infer these gene fragments to be true pseudogenes.

Focused phylogenetic analysis (**Figure 2**, main text) of (i) repredicted *C. canephora* nucleotide sequences belonging to ORTHOMCL170, (ii) several sequences previously reported to be involved in caffeine biosynthesis in different tea species (*163*), and (iii) the re-predicted cacao sequences, was performed using PhyML with the GTR model with gamma shape parameter = 0.7148, 8 substitution categories, empirical nucleotide frequencies, best of NNI and SPR, and 1000 bootstrap replicates. Four clades were retrieved, 3 corresponding to the species-specific clades containing genes with (or with putative) caffeine synthase activity, and a fourth corresponding to coffee NMTs of unknown function, the latter sister to coffee caffeine synthases (**Figure 2**, main text).

NMT expression profiles

The expression profiles of all coffee NMTs (defined as such by the coffee gene models and determined to have at least the SAM dependent carboxyl methyltransferase domain with INTERPRO; see main text) are given in **Table S25**. The three main NMT enzymes of the caffeine biosynthesis pathway (*CcXMT*, *CcMXMT*, and *CcDXMT*) are very strongly expressed in leaf and pericarp. Interestingly, the caffeine synthase-like gene *CcMTL* also shows high expression in several surveyed tissues, possibly suggesting an important role

in caffeine synthesis. Other than these genes, only *CcNMT3* has appreciable expression. There could be some root expression in *CcNMT8*, and some root or leaf expression in other genes, but their expression may be not significantly different from background. There is very little expression in most of the NMTs from the expanded orthogroup, suggesting that they could be pseudogenes.

Genome structural analysis of C. canephora and Theobroma cacao NMT genes

In order to verify conclusions on orthology versus paralogy generated from our phylogenetic analysis, we also carefully analysed the genomic contexts of the *C. canephora* and *T. cacao* NMTs (*4*). Two scaffolds, from chromosomes 1 and 9, contained 4 NMT genes each within close proximity of one another, indicating a series of tandem duplications (**Figures S23-S24**). Another region on chromosome 1 displayed 2 tandem NMT genes. A fourth scaffold from chromosome 2 contained a single NMT gene. None of the scaffolds were syntenic to each other, indicating that the NMTs in these regions did not descend from pre-gamma hexaploidy ancestral blocks, but rather that the NMT tandem arrays evolved *in situ*.

Genome context analysis of the region on chromosome 9 revealed two regions of the cacao genome that are syntenic to it, but with no NMTs contained within the cacao region (**Figure S25**). The tandem array on coffee chromosome 9 includes the known caffeine biosynthetic gene *CcXMT*, as well as two other genes that show significant expression in transcriptome analysis, *CcNMT3* and *CcMTL*, the latter previously suspected to be involved in caffeine biosynthesis as well. Considerable synteny is observed between coffee and two neighboring blocks on cacao chromosome 6, but these cacao regions do not contain any NMT genes. As such, confirming the paralogy predictions from phylogenetic analysis, the tandem NMT array on coffee chromosome 9 is evolutionarily independent from NMTs existing in the cacao genome.

Similar synteny analyses of the *C. canephora* region of chromosome 1 containing a tandem NMT array against the cacao genome confirmed that regions containing NMT arrays have syntenic counterparts in cacao, but that no NMT genes are found in these cacao regions. As such the paralogy predictions from phylogenetic analysis are confirmed for these coffee NMTs as well. The *CcDXMT* gene, which directly converts theobromine or paraxanthine to caffeine, is located within a tandem array on chromosome 1 that includes several NMTs found in ORTHOMCL170 (**Figure S27**). *CcDXMT* appears nested within the same phylogenetic group as the other caffeine-biosynthetic NMTs that lie on chromosome 9 (**Figure S27**; **Figure 2**, main text), suggesting that it translocated away from its founding tandem array, which includes *CcXMT*, *CcMTL*, and *CcNMT3* (**Figure S25**). Although the region around *CcDXMT* shows synteny with a region in cacao, the cacao genome lacks an NMT in this homologous block, and lacks synteny altogether to the region of the coffee block that contains other ORTHOMCL170 NMTs (**Figure S26**). The block containing the other tandem NMT array lying in a different region of *C. canephora* chromosome 1 (**Figure S27**) shows no synteny to the cacao genome. As a corollary to the analysis above, the tandem array on chromsome 9 and one of those chromosome 1 are syntenically linked to three distinct cacao blocks devoid of NMTs (**Figure S28**), supporting their common ancestry in a single genomic region.

We similarly examined the *T. cacao* genome for caffeine synthase-related NMTs. We discovered one tandem pair of NMTs that included the functionally characterized gene *TcBCS1*, and several chromosomal blocks containing single NMTs (**Figure S29**). When searching the coffee genome for synteny to the cacao region surrounding *TcBCS1*, we found that this cacao region shows synteny to 4 regions of coffee (**Figure S30**). It is clear from the HSP coverage that most of the block containing *TcBCS1* has clear homologies in the coffee genome. The immediate region around *TcBCS1* appears (with considerable gene deletion)

syntenic to with *C. canephora* scaffold 7. *TcBCS1* has an NMT ortholog on that scaffold, gene model number GSCOCT00037185001 (Cc01g00110). This gene annotates to SAM dependent NMTs in *Arabidopsis*, and importantly, it is not one of the various coffee NMTs, caffeine-related or otherwise, that we have so far analysed. We infer that this gene and *TcBCS1* descend from a common gene (i.e., that they are orthologs), with the cacao gene eventually evolving into a caffeine synthase, and the coffee ortholog perhaps maintaining the ancestral function. As such, it appears that any potential cacao ortholog of the founding caffeine synthase gene in coffee may have been deleted over evolutionary time (**Figure S26**), while the ancestral gene from which cacao caffeine synthase (*TcBCS1*) evolved is retained in coffee, but with a non-caffeine-related function. The independent evolution of caffeine biosynthesis in coffee and cacao is therefore verified by genome structure as well as gene phylogeny.

Model for the evolution of tandem NMT arrays in C. canephora

From the analysis above, we can conclude the independent evolutionary origin and diversification of caffeine biosynthetic enzymes in coffee. In order to provide relative timing of the duplication events that gave rise to the coffee caffeine synthases and related NMTs, we performed all pairwise estimates of synonymous substitution rates (Ks) among CDS (**Figure S31**). We calculated relative ages for the crown-group (tandem array) radiations as well as the duplication events that seeded them; the former were based on within-lineage Ks averages, and the latter on between-group averages (see **Figure 2**, main text, and **Figure S31**) (*4*). It was important for these calculations that Ks averages were calculated by phylogenetic group rather than by block membership, since the latter may not well reflect the phylogenetic history of genes contained within such blocks. As such, *CcDXMT* was included in the clade otherwise found on the red block, and *CcNMT19* was included with the clade otherwise found on the blue block.

As described above, microsynteny analyses of ORTHOMCL170 showed that some known and putative coffee caffeine synthase genes, *CcXMT*, *CcMTL*, and *CcNMT3*, form a tight assemblage of co-expressed tandem duplicates reminiscent of a metabolic gene cluster. Given that some plant metabolic gene clusters have been shown to be of relatively recent origin (*23*), we sought to further unravel the role of gene duplication in the expansion of the coffee NMT gene family. As described above, ORTHOMCL170 includes three tandem arrays that we infer to have likely evolved in close proximity to each other. As shown in Figure 2E (main text), the three main clades in ORTHOMCL170 are distributed among a minimum of three genomic blocks; however, clade membership does not always reflect block membership. In one parsimonious scenario, some phylogenetically recent tandem duplicates were redistributed away from their close relatives via even more recent block rearrangements. One such block movement appears to have shifted *CcDXMT* away from its ancestral array, the putative metabolic cluster described above, where its participation in caffeine synthesis could have been even more integrated by common regulatory control. The functionally characterized *TcBCS1* gene is also accompanied by a tandem duplicate, but the pair evolved independently from and is not syntenic to the tandem arrays of *C. canephora* (**Figure S29**).

Positive selection analyses of NMT genes

We also sought to examine the possible role that molecular evolutionary pressures may have played in the multiple origins of caffeine biosynthesis described in the previous sections. For this purpose, we generated a codon alignment of (i) caffeine biosynthetic NMTs from coffee (including *CcMTL* and *CcNMT3*, based on the above arguments), tea, and cacao, plus (ii) sequences of 7 *Arabidopsis* genes that form a well defined clade near these caffeine synthase NMTs in our large-scale phylogenetic analysis. Among these *Arabidopsis* genes are

several with well-characterized benzoic, salicylic and nicotinic NMT functions (AT5G38020, AT2G14060, and AT3G11480). We reconstructed a phylogeny of these NMTs using PhyML, as in (*4*) (**Figure S32**). Specifically, we used the codon alignment and tree (with its branch lengths), to test for signatures of asymmetric evolutionary rates, divergent selection pressures, and positive Darwinian selection by implementing different codon substitution models in a maximum likelihood framework.

It is widely accepted that selection acts differentially over non-synonymous (amino acid-changing; d_N) substitutions compared to synonymous (silent; d_S) ones. Accordingly, the ratio between non-synonymous and synonymous rates ($\omega = d_N / d_S$) provides a measure of the rate of evolution (or strength of selection) acting on genes, and yields insights on the molecular evolutionary mechanisms of protein diversification and functional specialization. Most proteins are subjected to selective constraints to conserve their structure and maintain their function. Therefore, most changes in protein-coding sequences are deleterious. Consequently, synonymous changes are generally unaffected by selection, but non-synonymous changes are often purged by purifying selection, yielding ω values $<< 1$. When synonymous and non-synonymous substitutions accumulate at the same rate ($\omega \sim 1$), a gene is expected to be under no selective pressure, such as in pseudogenes, and the gene is said to evolve under strict neutrality (*89, 164, 165*). Finally, certain regions of genes will evolve under positive selection (PS), being subjected to higher rates of non-synonymous substitutions, thus resulting in $\omega > 1$. Estimation of ω ratios across a set of genes can thus provide signatures of functional specialization and adaptive evolution. Estimation of ω values was performed by means of the codeml program from the PAML v4.4 package (*166*) on the basis of multiple alignments of codon sequences and a tree topology.

Two different classes of models were implemented using the codeml program(*4*). Likelihood Ratio Tests (LRTs) tests were used to compute a p-value for the fitting of the examined dataset to the alternative model being tested (*4*). Tests of asymmetric evolution (significant for all foreground branches; **Table S26**) revealed that caffeine biosynthetic NMTs evolve under stronger selective constraints (showing lower ω values) than their *Arabidopsis* counterparts. This finding is consistent with the shorter branch lengths for at least coffee and tea in the tree in **Figure S32**. On the other hand, none of the tests for divergent selective pressures were significant. Despite caffeine biosynthetic NMTs having been subjected to strong purifying selection during most of their evolution, positive selection may have fixed specific amino acid changes key for acquiring caffeine biosynthetic capacity. Therefore, we explicitly tested for positive selection in the 3 foreground branches. Significant results were only found for the coffee branch, and a few amino acid changes were identified as fixed by positive selection. These changes are specific to coffee except for one site that is shared with cacao and *Arabidopsis* NMTs. All PS amino acid changes were conserved among coffee sequences except position 277 of XMT, occupied by a C in all sequences except DXMT, where the residue is R. PS amino acids were then mapped onto the corresponding positions for both XMT and DXMT, for which 3D structures were available in the literature (**Table S27**). One position, T-94 of XMT (T-93 of DXMT), was found at the NMT dimer interface.

Fatty Acid Desaturases

Fatty acid desaturases (FAD) introduce double-bonds in the carbon chain of fatty acids (FA) (*167*). Desaturase activity therefore primarily controls the level of unsaturation of membrane (phospholipids) and storage (triacylglycerols) lipids. FAD mostly differ for substrate preference (e.g. 18:0, 18:1 or 18:2), the position at which the double-bond is created (e.g. Δ4, Δ7, Δ9, Δ12 or Δ15) and sub-cellular localization (e.g. plastid or endomembranes).

The level of unsaturation of FA also determines the thermal and oxidative stability of vegetable oils and food products (*168*). Coffee beans contain large amounts of oil (10-18 % of

the dry matter, depending on the species and genotype) (*169*). Almost half of FA stored in the coffee endosperm (which represents ca. 99% of the mature bean mass) are polyunsaturated (*26, 27*), making roasted coffee beans highly susceptible to rancidity, which constitutes a significant constraint for the coffee industry (*170, 171*).

Linoleic acid (18:2) is the major polyunsaturated FA in the coffee bean, accounting for *ca*. 45% of total FA in almost all coffee species and varieties studied so far (*25, 26*). The microsomal oleate desaturase FAD2 (1-acyl-2-oleoyl-sn-glycero-3-phosphocholine Δ12-desaturase) is the key enzyme responsible for the production of 18:2 in seeds. In comparison with *Arabidopsis* (*167*), coffee displays a much higher number of copies of microsomal Δ12 desaturases (FAD2) (**Figure S33A**). Five FAD2 copies were localized on chromosome 1 (chr1) and one copy on chromosome 6 (chr6). Chr1 copies separate in two groups of genes very closely distributed (Cc01g05140, 05170 and 05180, and, Cc01g03650 and 03680, respectively. This suggests that the five chr1 FAD2 copies arose from one or two events of small-scale duplications. For almost all coffee FAD genes, the intron/exon structure was highly similar to that of *Arabidopsis* orthologs, except for three chr1 FAD2 paralogs (Cc01g05140, 03650 and 03680) that show reduced CDS length, two of which also display alternative intron/exon structures.

Our RNA-Seq data suggest transcriptional specialization for two of the six *FAD2* copies, with *CcFAD2.3* being actively transcribed in the endosperm during coffee seed development (Figure S30A). *CcFAD2.3* peak transcript abundance coincides with the dramatic increase in 18:2 content that occurs during seed development at the perisperm-endosperm transition (*27*) (**Figure S33A** and **S33C**).

The very low intra- and inter-specific variability for seed 18:2 content within the genus *Coffea* (*25, 26*) (**Figure S33B**) challenges the opportunity to improve this trait through conventional breeding approaches. In this respect, seed-specific transcriptional specialization of *CcFAD2.3* may offer an advantageous way for lowering the coffee bean 18:2 content and subsequent susceptibility to rancidity through genetic engineering. Finally, in contrast with *CcFAD2.3*, but analogously to some of the coffee NMTs, transcription of the three chromosome 1 *FAD2* paralogues with reduced coding sequence length (*CcFAD2.1*, *CcFAD2.4* and *CcFAD2.5*) was very low or nil in all tissues analyzed (**Figure S33A**), indicating possible loss of function during evolution.

Supplementary Figures

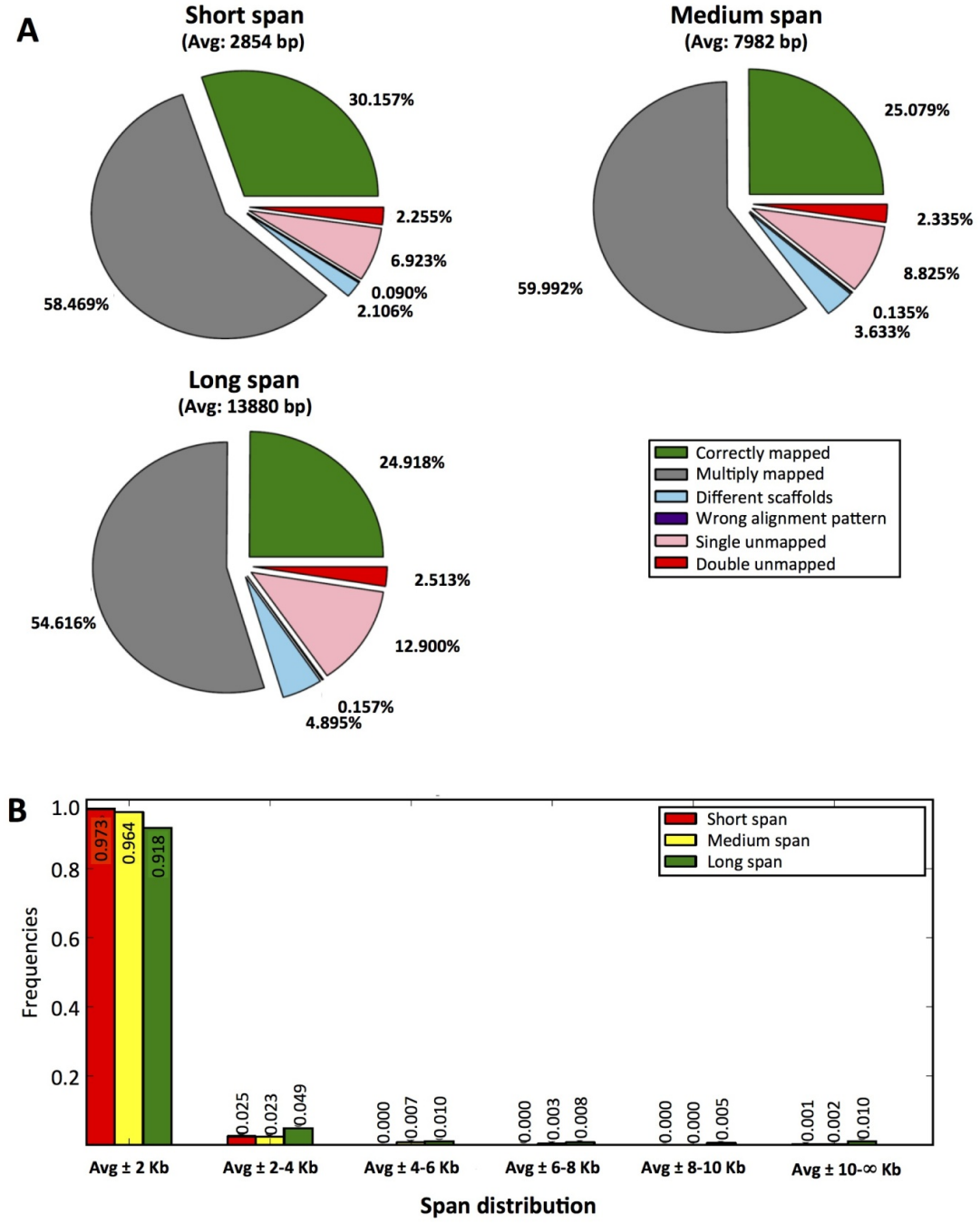
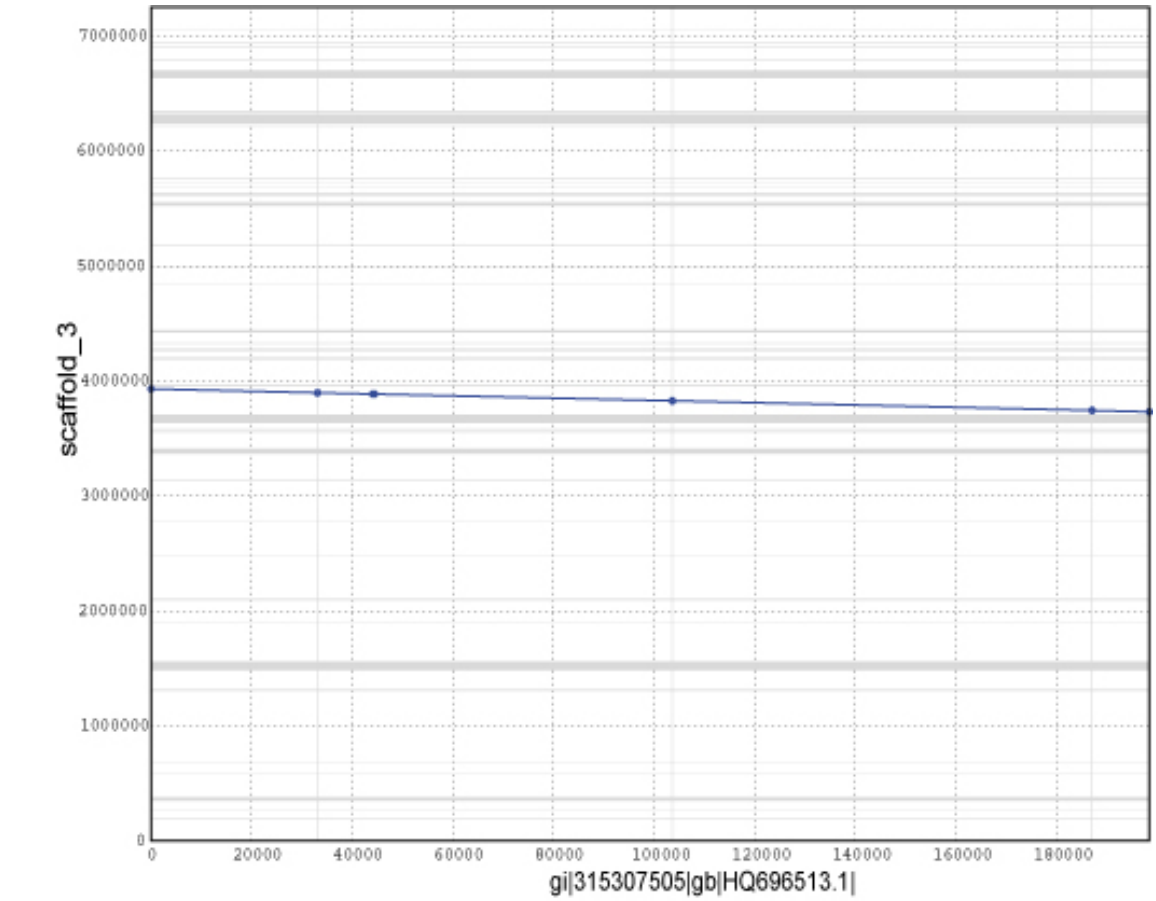


Figure S1. **Genome assembly validation. A.** mapping of mate-pair sequences on the genome assembly. Only 0.12% of the mate pairs mapped are in the wrong orientation on the same scaffold, indicating structural correctness of the assembly. **B.** Actual span distribution of the pairs on the assembly. Approximately 95% of the mate pairs showed a span within ± 2 Kb of the calculated average.



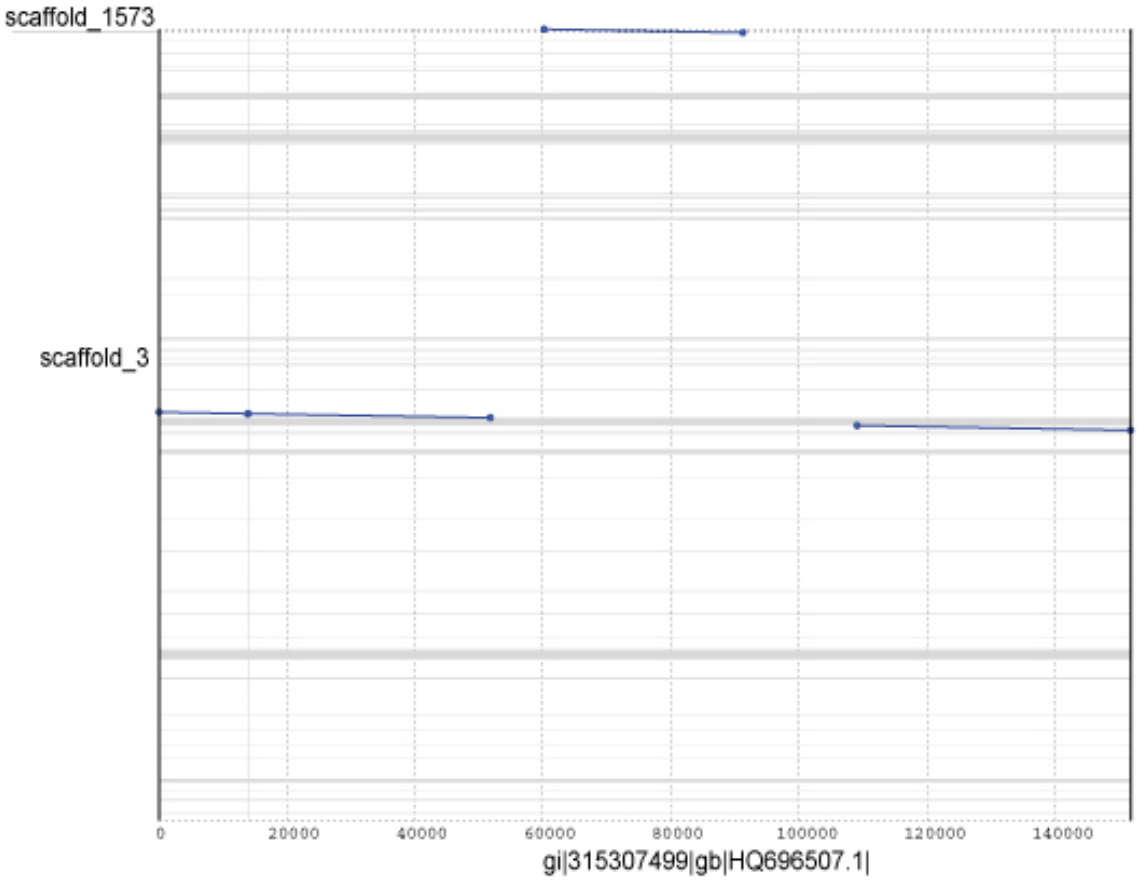


Figure S2 (continued)

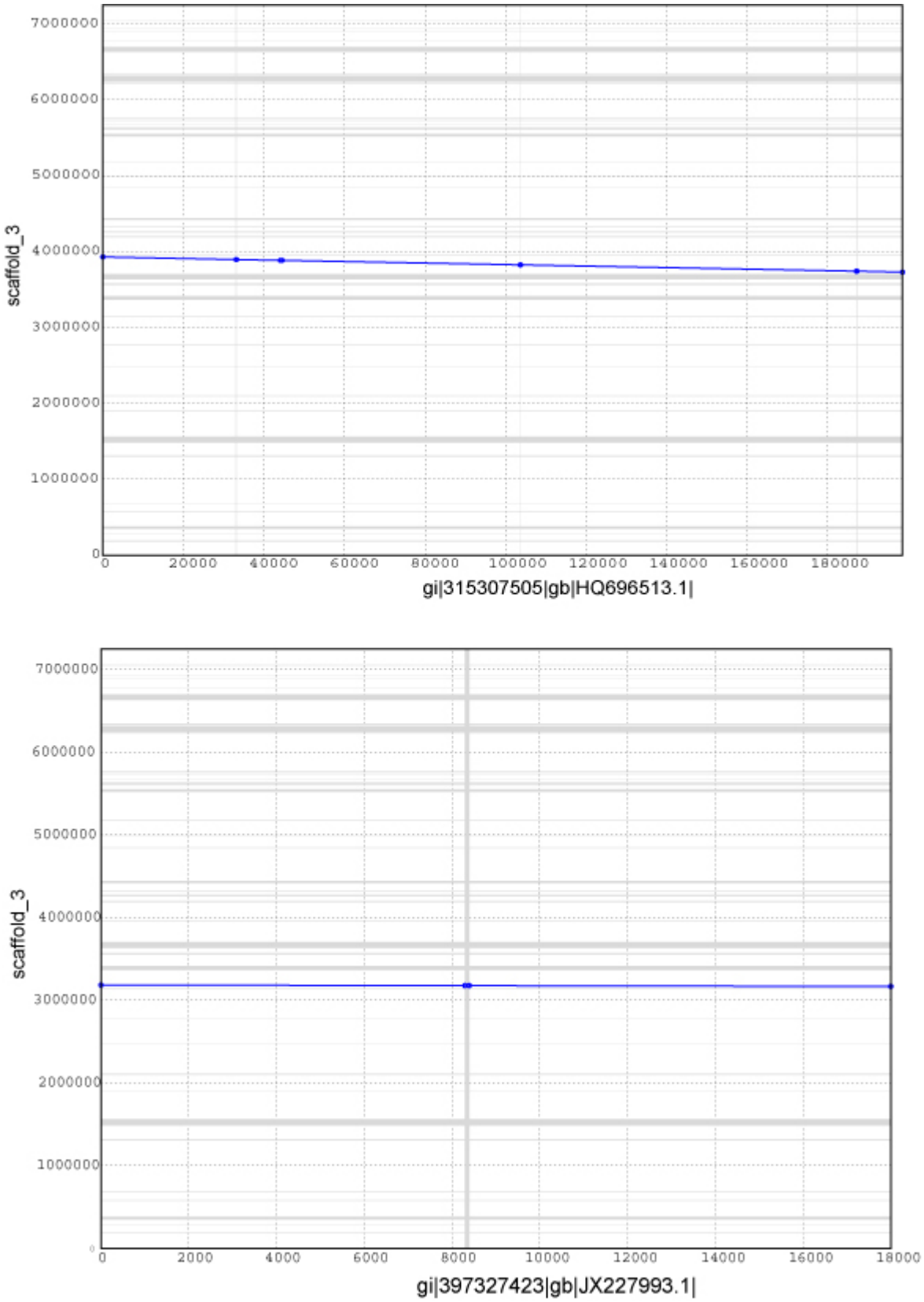


Figure S2 (continued)

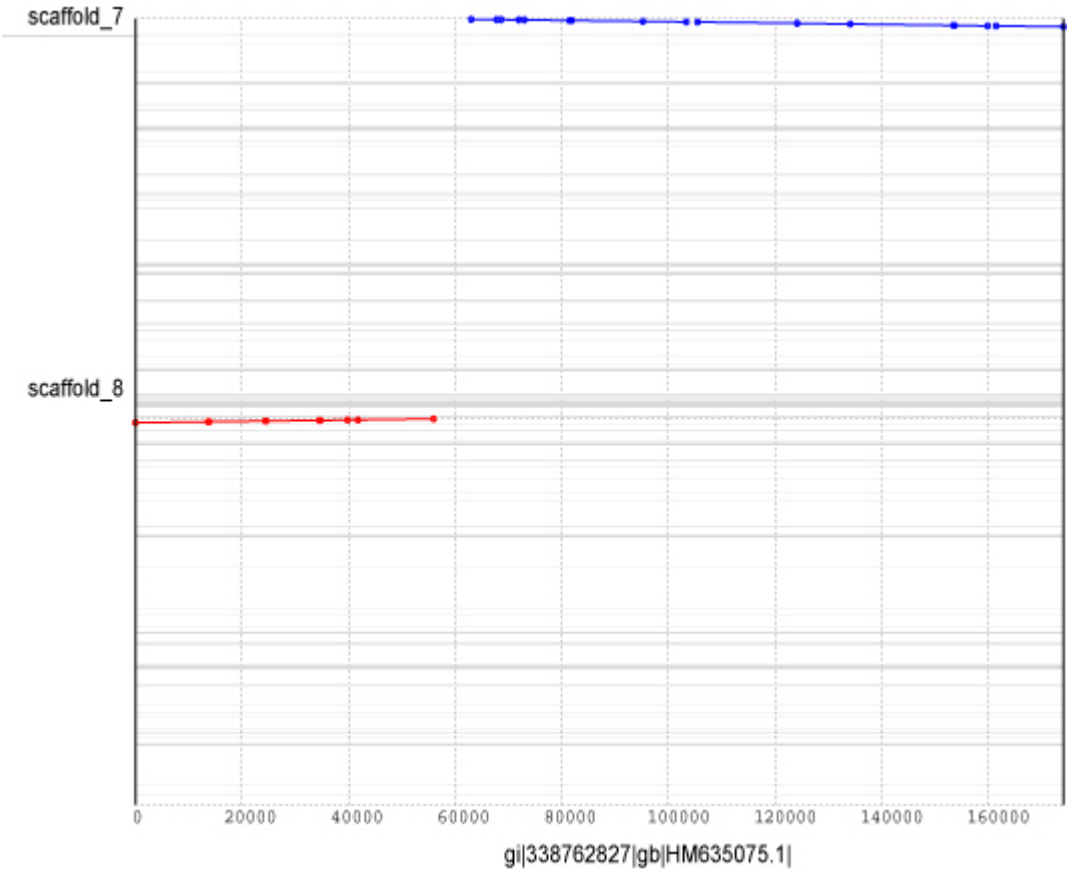


Figure S2 (continued)

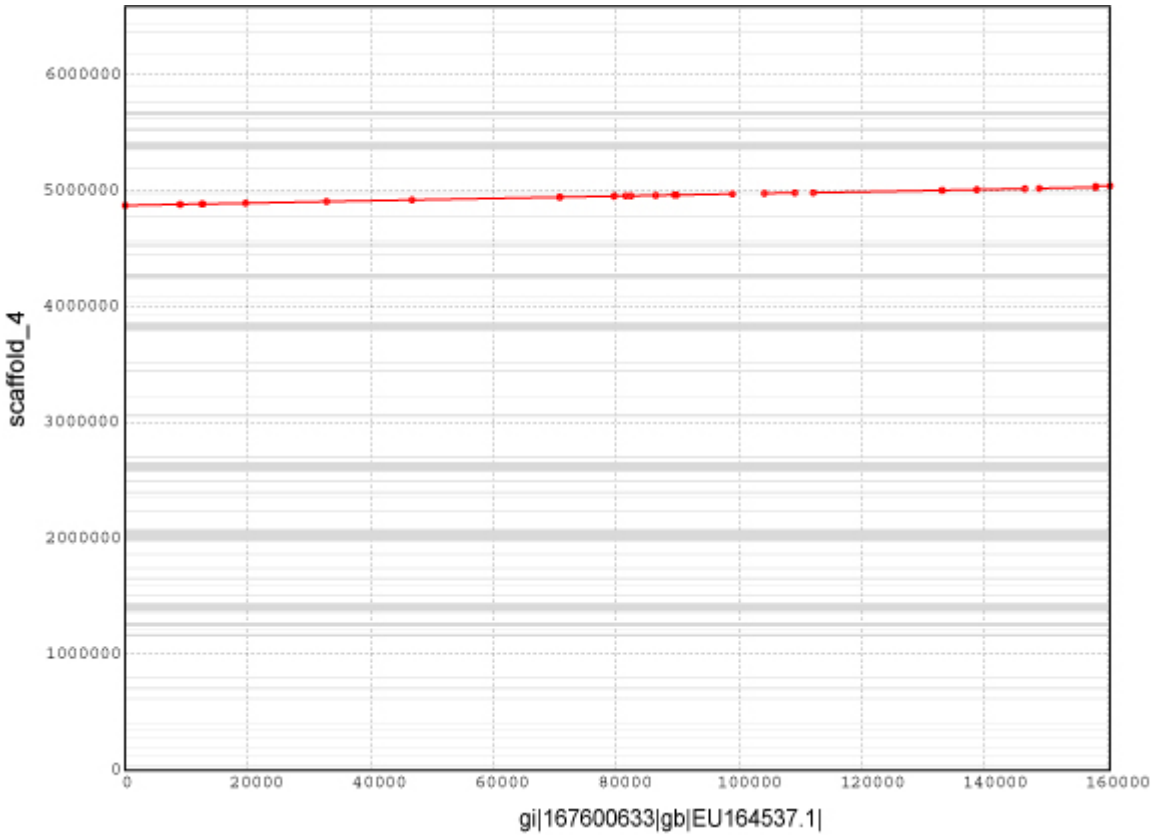


Figure S2 : Dotplot representing the positions of the matches obtained using MUMmer (nucmer) between *Coffea canephora* BACs (on the x axis) and the assembled scaffolds (on the y axis). The inter-contig gaps are displayed in grey on the y axis.

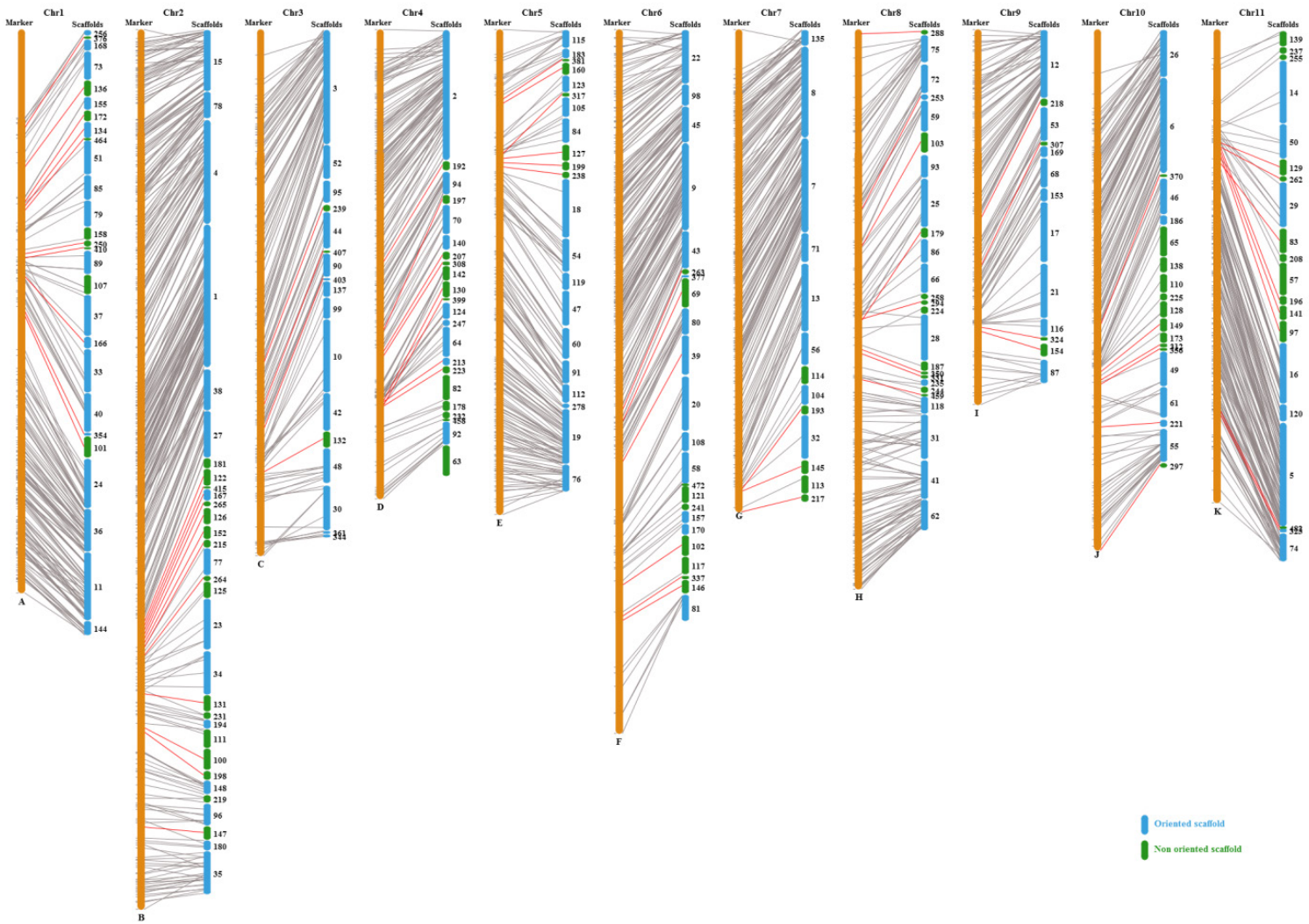


Figure S3: ***C. canephora*** **pseudomolecules.** Assembled scaffolds were anchored in the eleven linkage groups (A to K, orange) with the corresponding genetic markers (black bars). Lines connect linkage map markers and the DNA pseudomolecule (grey lines denote consistent data whereas red lines indicate markers with an approximate genetic location). Oriented scaffolds are represented in blue and non-oriented scaffolds are in green.

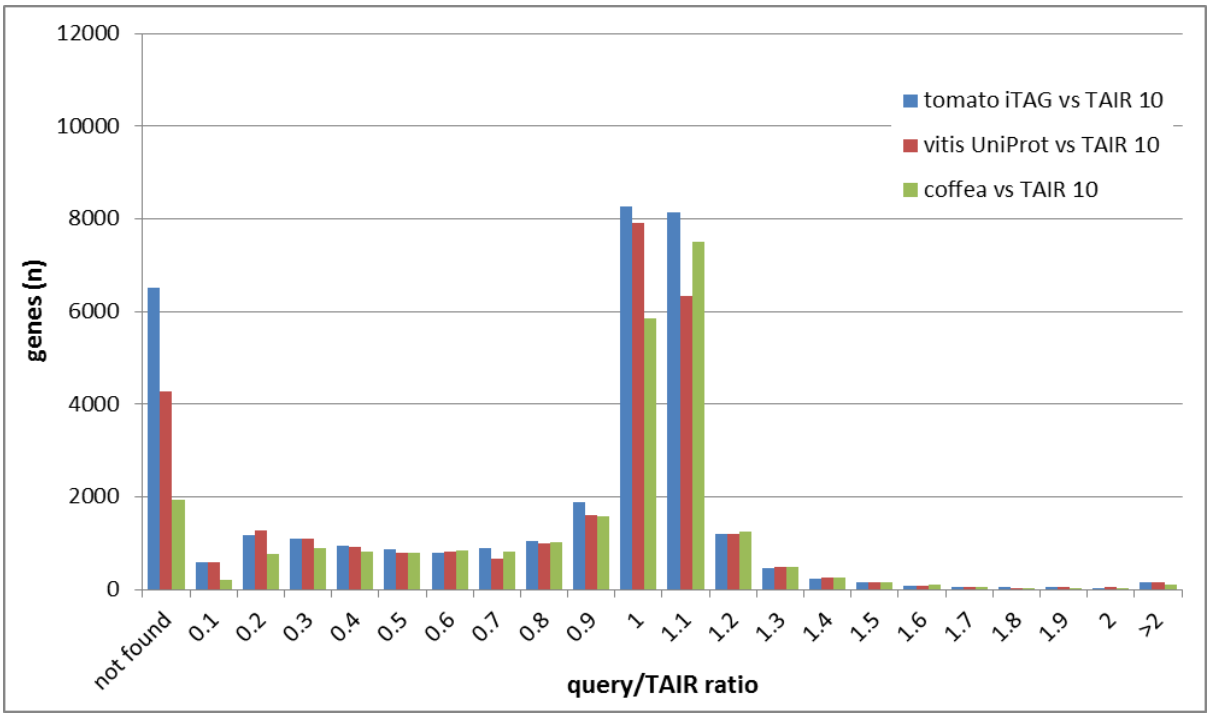


Figure S4. Pairwise comparison of protein length ratios (X axis) of coffee, tomato and grape proteins used as queries, with their best BLAST hit in the *Arabidopsis* TAIR10 annotation. On the extreme left, proteins without an *Arabidopsis* match are shown. *A. thaliana* (TAIR10): 27,416 genes, *S. lycopersicum (*iTAG2.3): 34,727 genes, *V. vinifera* (UniProt): 29,836 genes, *C. canephora*: 25,574 genes. "Not found" refers to lack of orthologs above the threshold (1E-3).

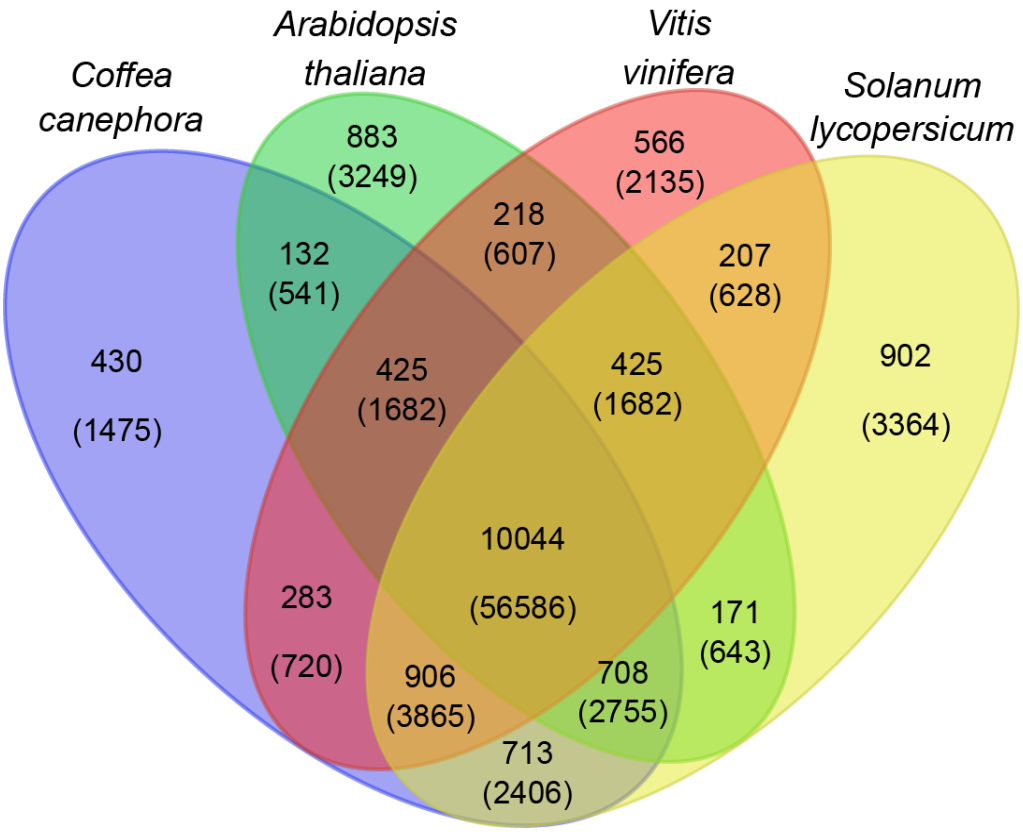


Figure S5. **Venn diagram showing the distribution of shared gene families among *Coffea canephora*, *Arabidopsis thaliana*, *Solanum lycopersicum* and *Vitis vinifera*.** Numbers in parentheses indicate the number of genes in each cluster.

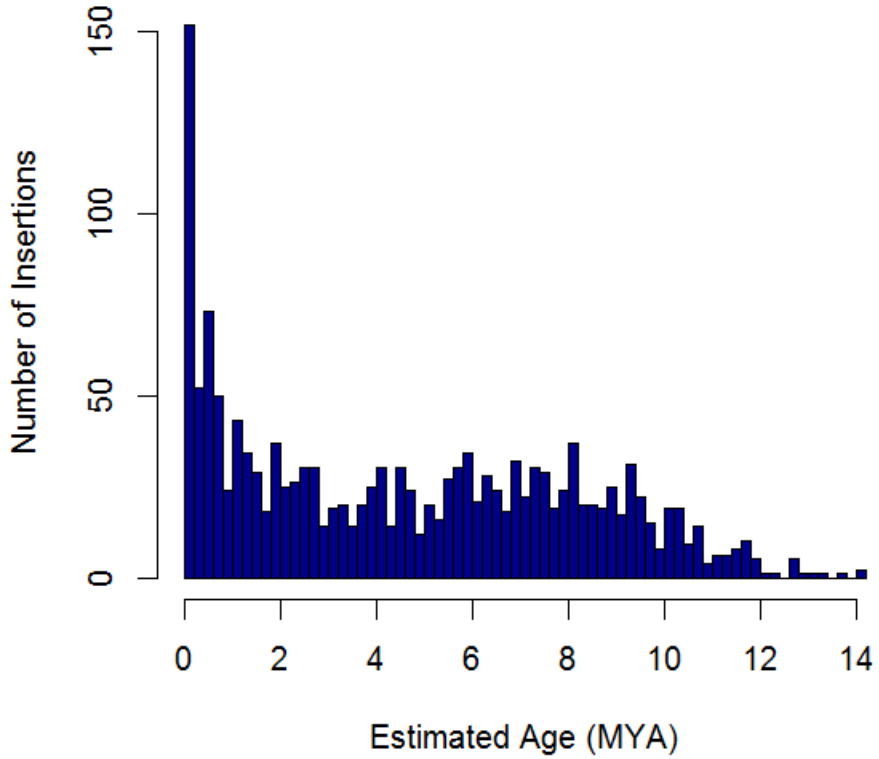


Figure S6. Distribution of the estimated insertion time of NUPT fragments in the *C. canephora* genome. A histogram of the estimated age (Mya) calculated using the Ks of NUPT vs. the chloroplast genome against the number of insertions.

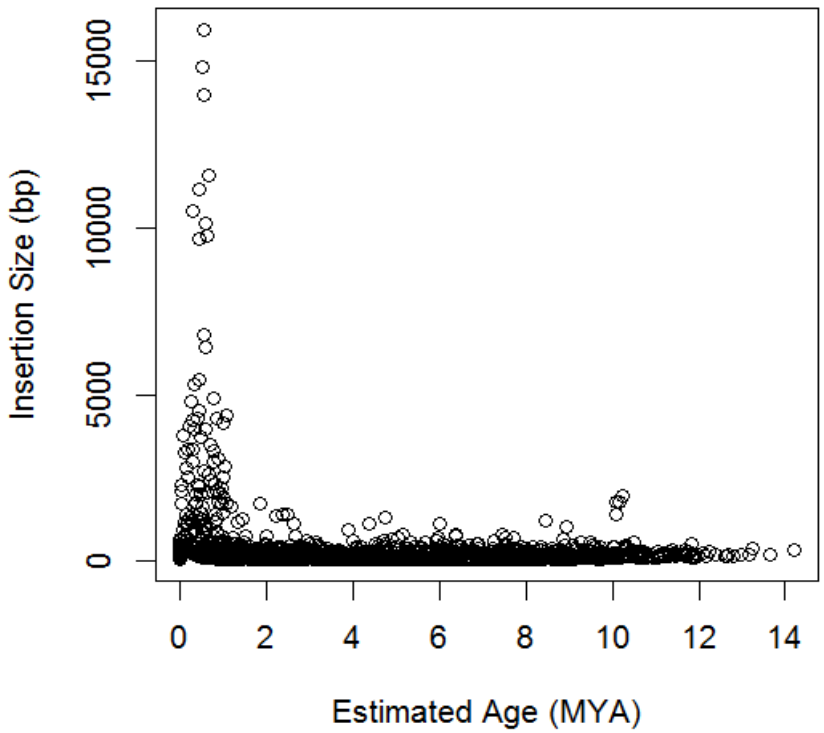


Figure S7. Longer NUPT fragments integrated more recently in the *C. canephora* genome. The insert size of the NUPT fragments is plotted against the estimated age.

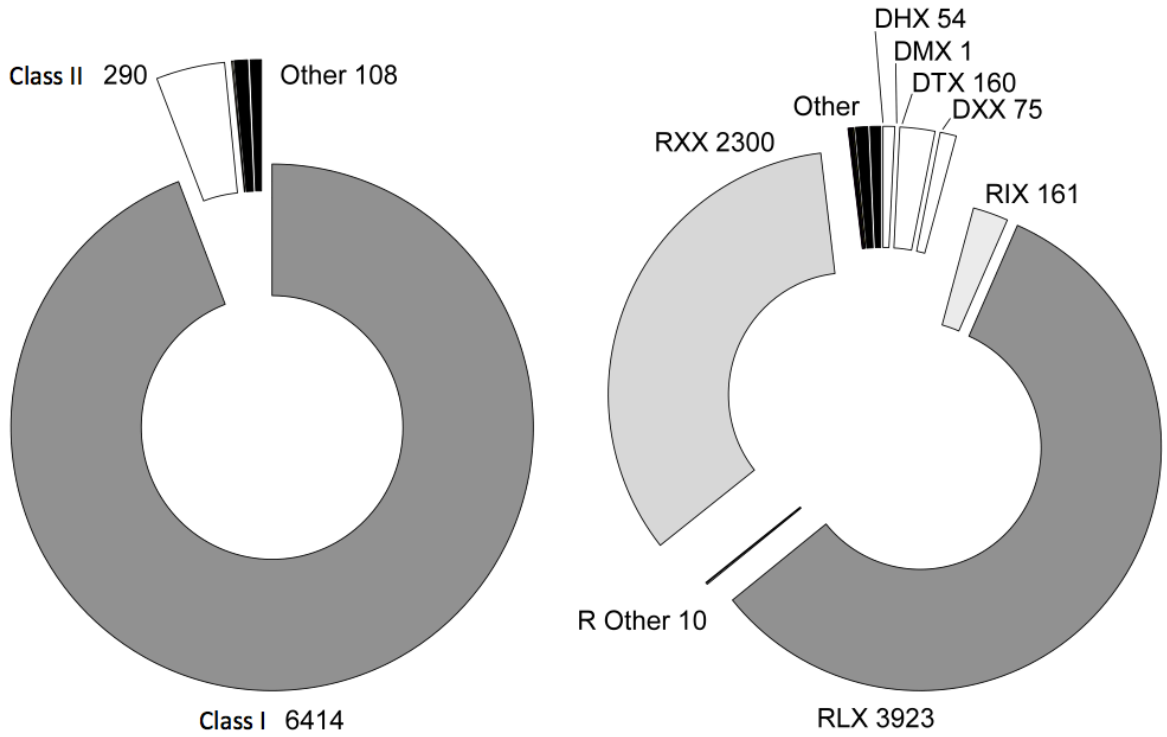


Figure S8. **Classification of the 6,812 predicted TEs from *TEdenovo* REPET.** DHX: Helitron; DTX: Transposon TIRs; DXX: MITE; RIX: LINE; RLX: LTR retrotransposons (RLG and RLC); RXX: TRIM and LARDS.

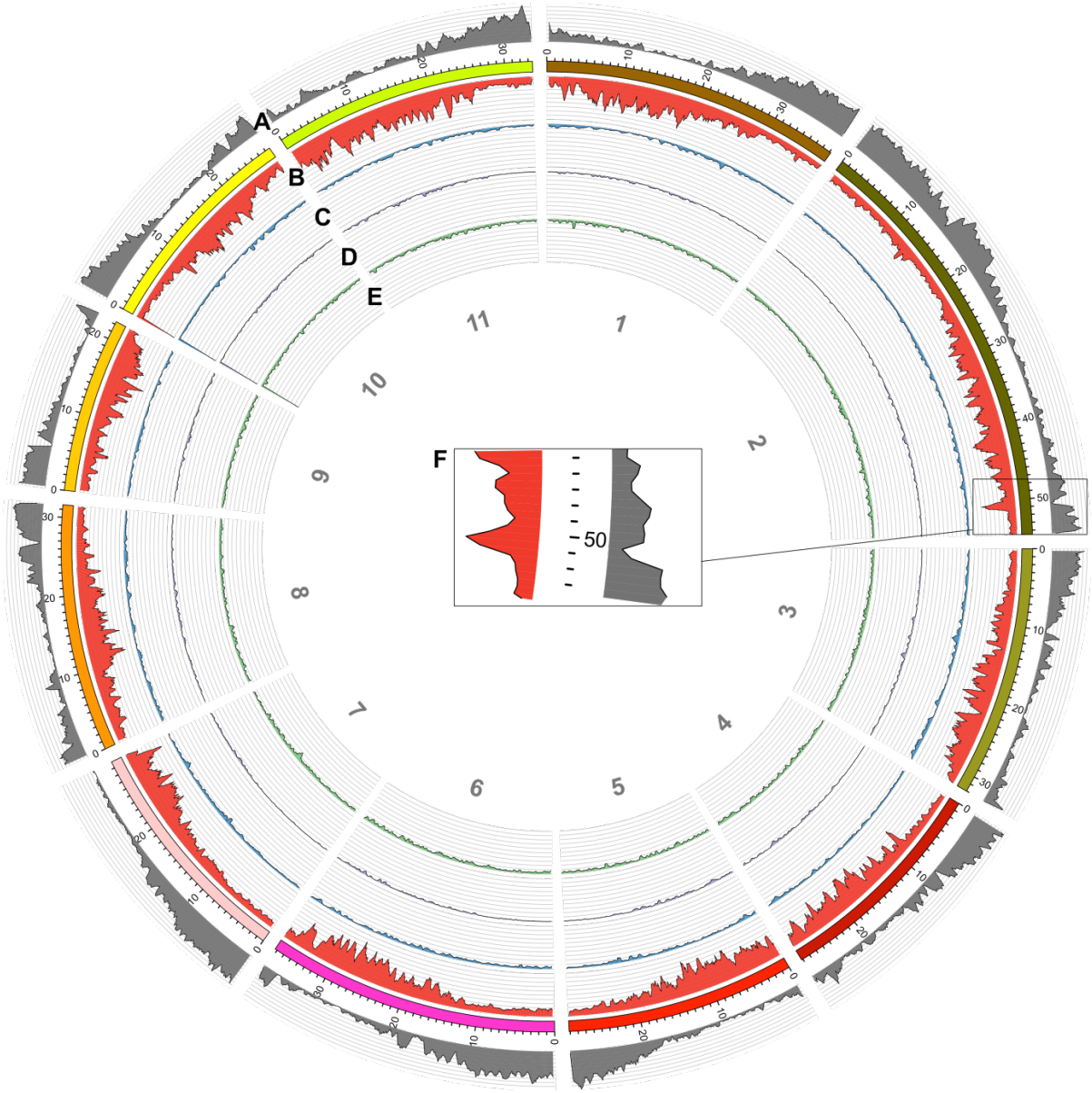


Figure S9. Distribution of genomic features along the 11 pseudomolecules of *C. canephora*. ChrUn, not shown, represents unanchored contigs. Distribution of gene densities (**A**, in grey), LTR retrotransposons (**B**, in red), non-LTR retrotransposons (**C**, in blue), DNA transposons (**D**, in purple) and unclassified repeats (**E**, in green) are figured. The distribution of LTR retrotransposons is generally negatively correlated with gene density (**F**). Gene density and TE densities are not at the same scale.

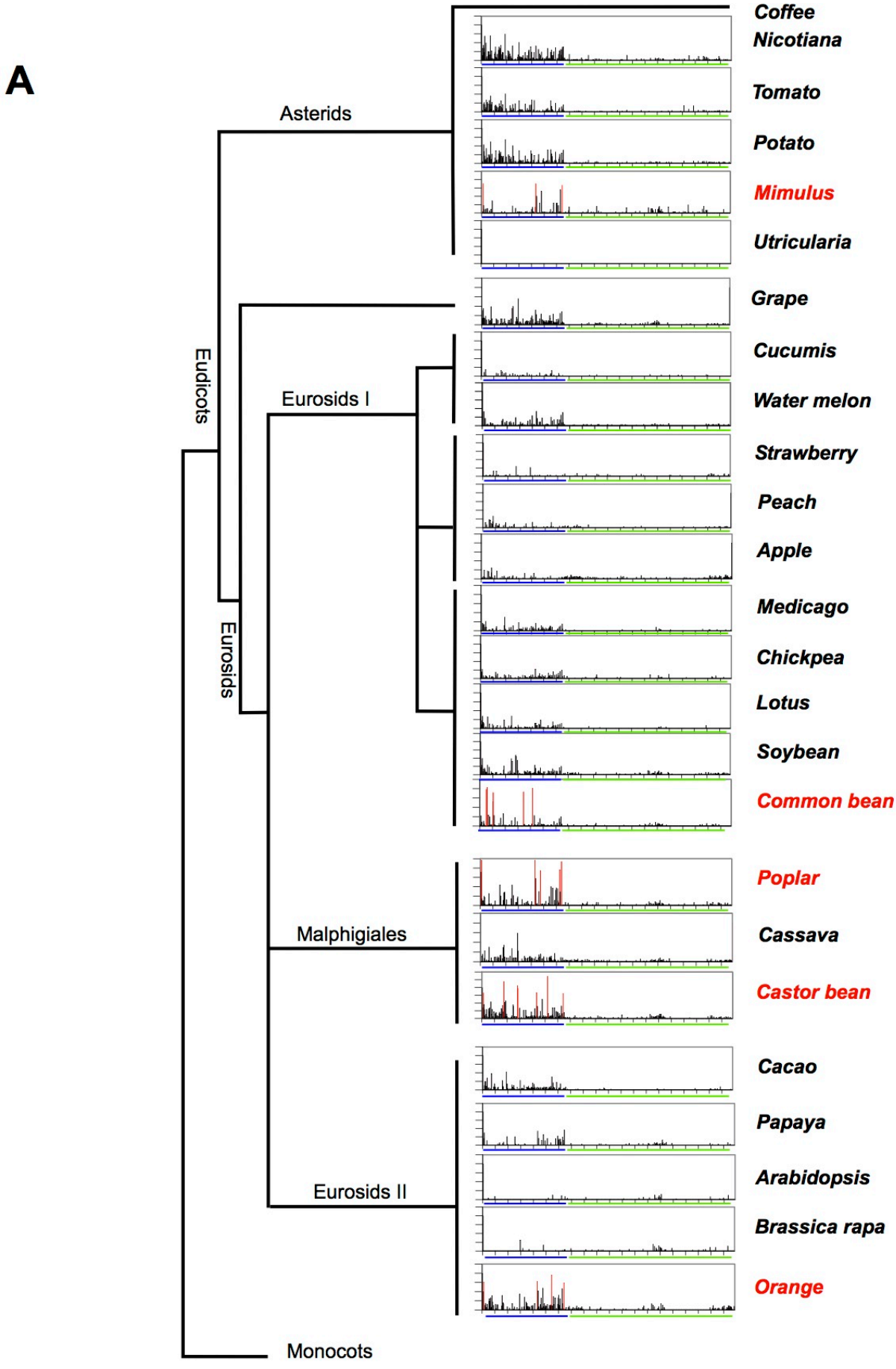


Figure S10 (continued)

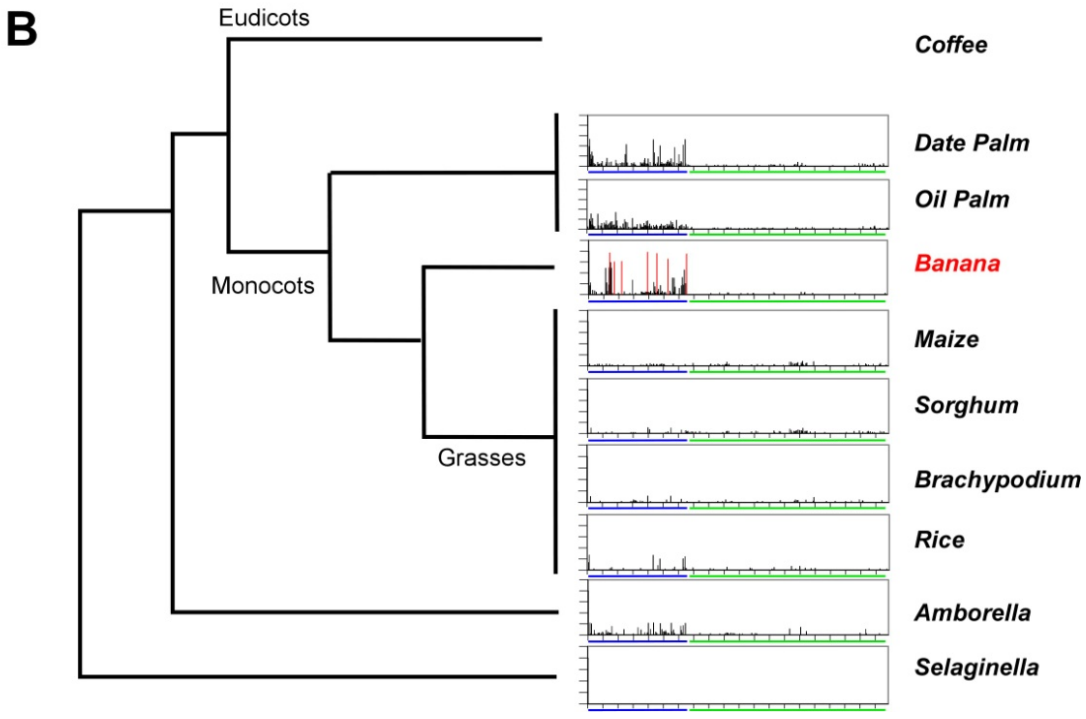


Figure S10. **Conservation of Coffee *Ty1/Copia* (188, blue section) and *Ty3/Gypsy* (387, green section) LTR retrotransposon groups across a selection of 33 sequenced plant genomes** (**A**, dicotyledonous genomes; **B**, monocotyledonous and non-dicotyledonous genomes). BLASTN bitscores are plotted for each group and with the same order for each genome. Bitscores higher that 3000 are indicated in red. Studied species are ordered according to plant phylogeny.

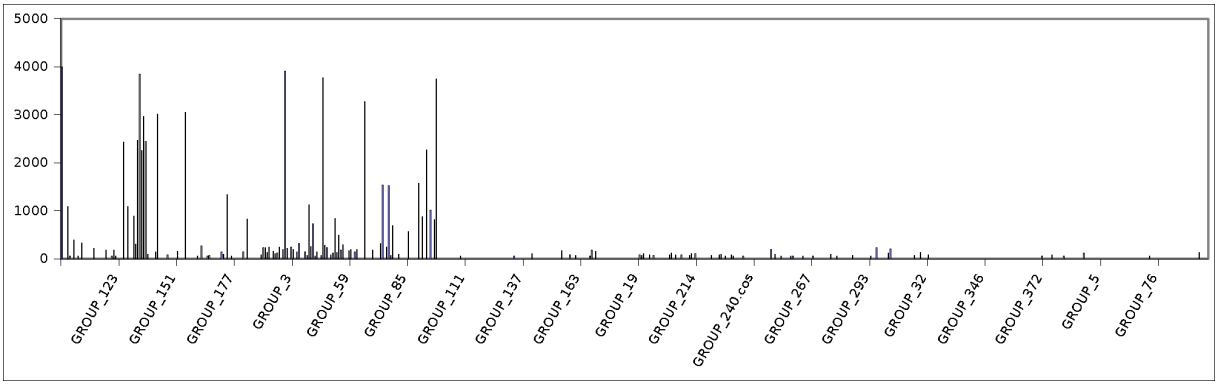


Figure S11. **Conservation of the 188 *Ty1/Copia* and 387 *Ty3/Gypsy* LTR retrotransposon groups from *C. canephora* to those from the *Musa* genome.** Vertical bars represent the best BLASTN bit score result of each coffee LTR retrotransposon group against the *Musa* genome.

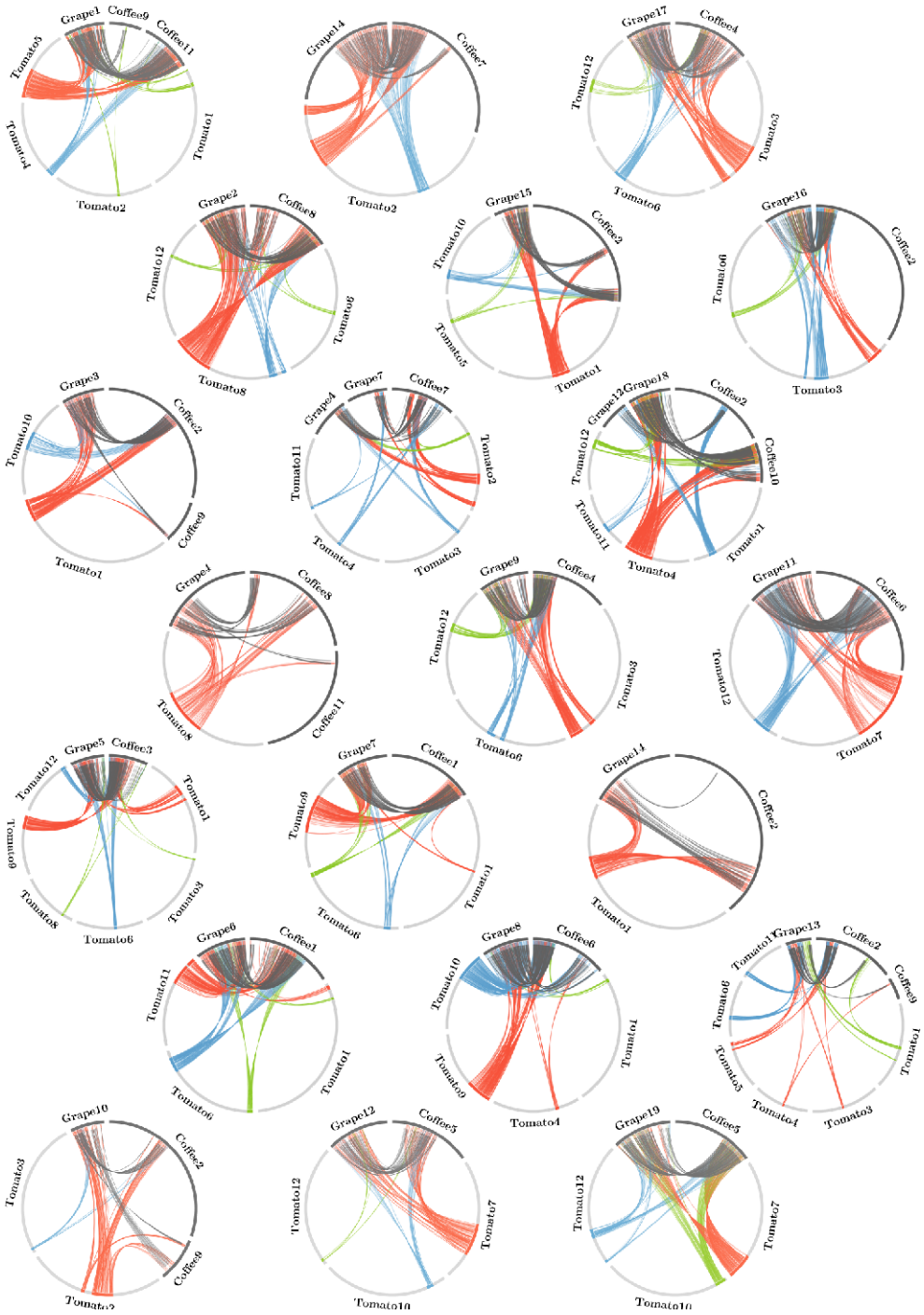


Figure S12. **Global synteny among the coffee, tomato and grape genomes.** Each row of three circles represents one of the seven pre-gamma core eudicot chromosomes. The three circles in each row represent the three copies of that chromosome after the gamma triplication, corresponding in most cases to a single grape chromosome. The red, blue and green homology lines connect the largest, second largest and smallest homeologous regions in tomato resulting from the *Solanum* hexaploidization. The large disproportion in the sizes of these regions is consistent with an initial tetraploidy reflected in the "blue" and "green" subgenomes, followed by a period of fractionation, and a subsequent incorporation of the red subgenome, which has had less time to fractionate, and/or manifests subgenome dominance.

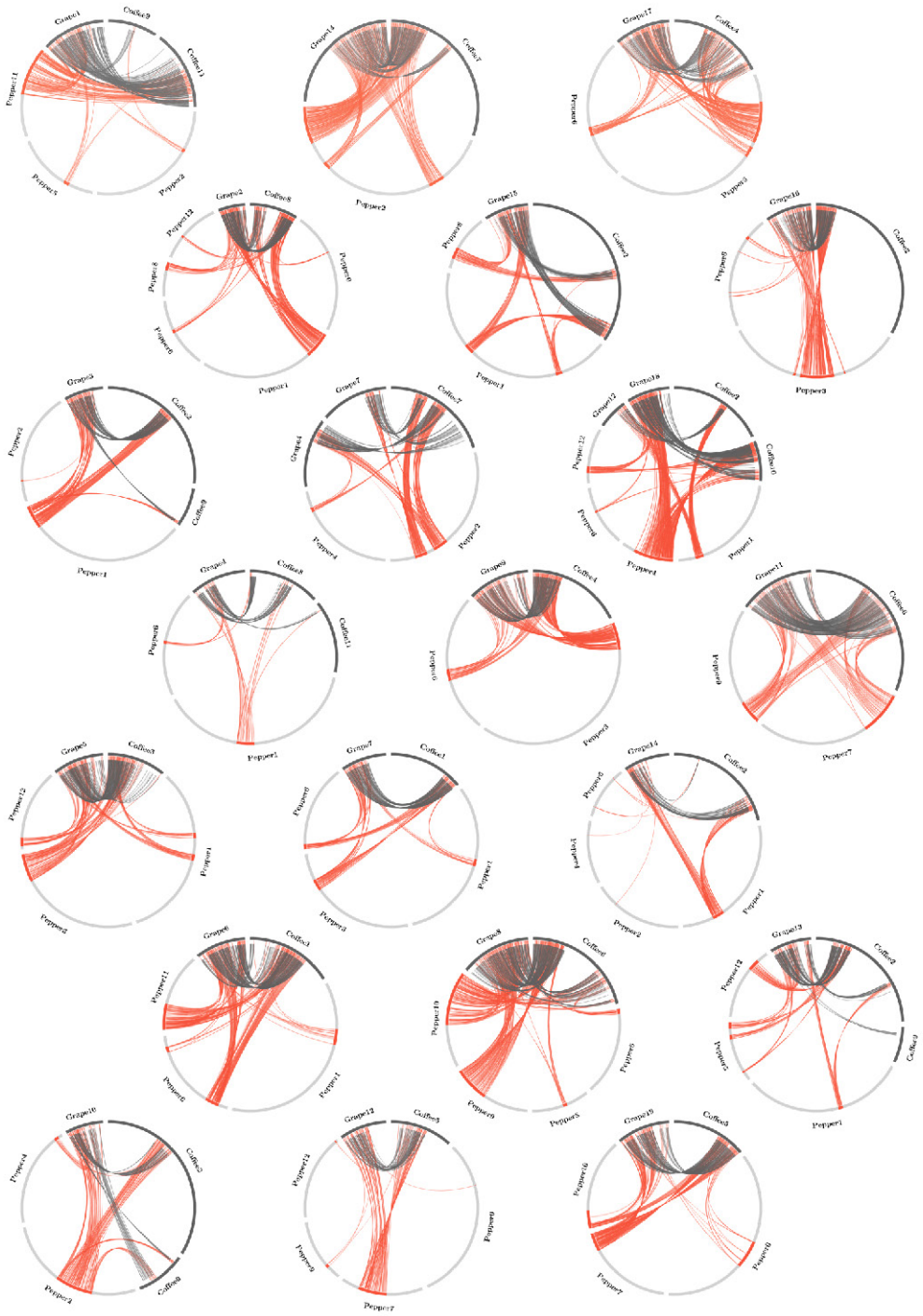


Figure S13: Global synteny among the coffee, pepper and grape genomes, constructed in the same way as **Figure S11**. Unlike the tomato genome, there is insufficient paralogy in the pepper genome to identify the three homeologous regions with confidence in many cases.

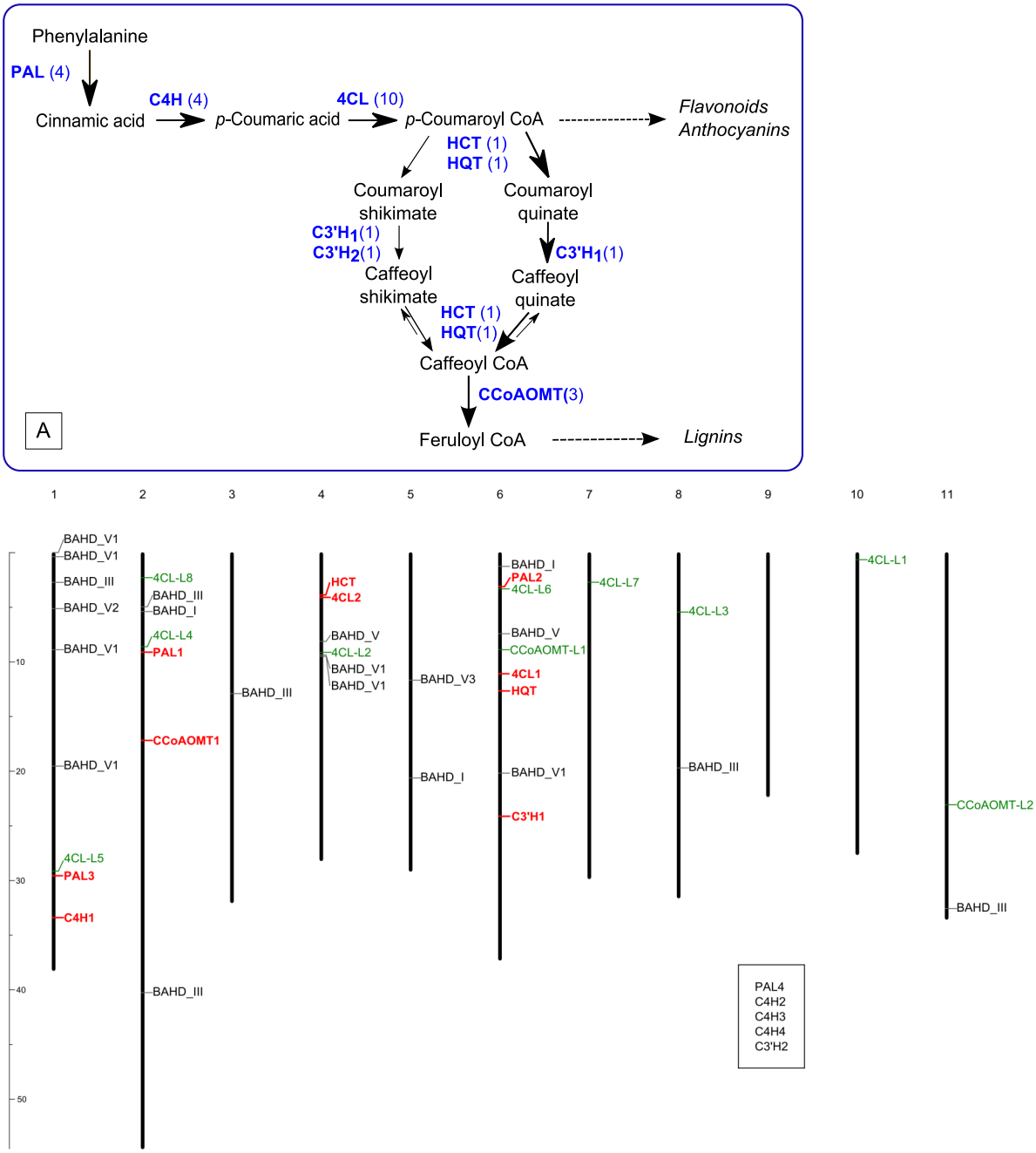


Figure S14. Hydroxycinnamic acid ester metabolism. A. Proposed pathway in *C. canephora*. Each enzyme is annotated with the number of corresponding UniGenes shown in parentheses. PAL, phenylalanine ammonia lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-hydroxycinnamoyl CoA ligase/4-coumarate-CoA ligase; HCT, hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase; HQT, hydroxycinnamoyl CoA quinate hydroxycinnamoyltransferase; C3'H (1, 2), p-coumarate 3'-hydroxylase; CCoAOMT, CaffeoylCoA-*O*-methyltransferase. **B.** Chromosomal distribution of genes involved in hydroxycinnamic acid ester metabolism on the scaffolds forming the 11 pseudomolecules of *C. canephora.* BAHD, benzylalcohol acetyl-, anthocyanin-*O*-hydroxy-cinnamoyl-, anthranilate-*N*-hydroxy-cinnamoyl/benzoyl-,deacetylvindoline acetyltransferase. In red, genes previously located using COS markers; in green the other members of the gene families involved in the pathway; in black, members of clades from BAHD superfamily. Unanchored genes are indicated in the right.

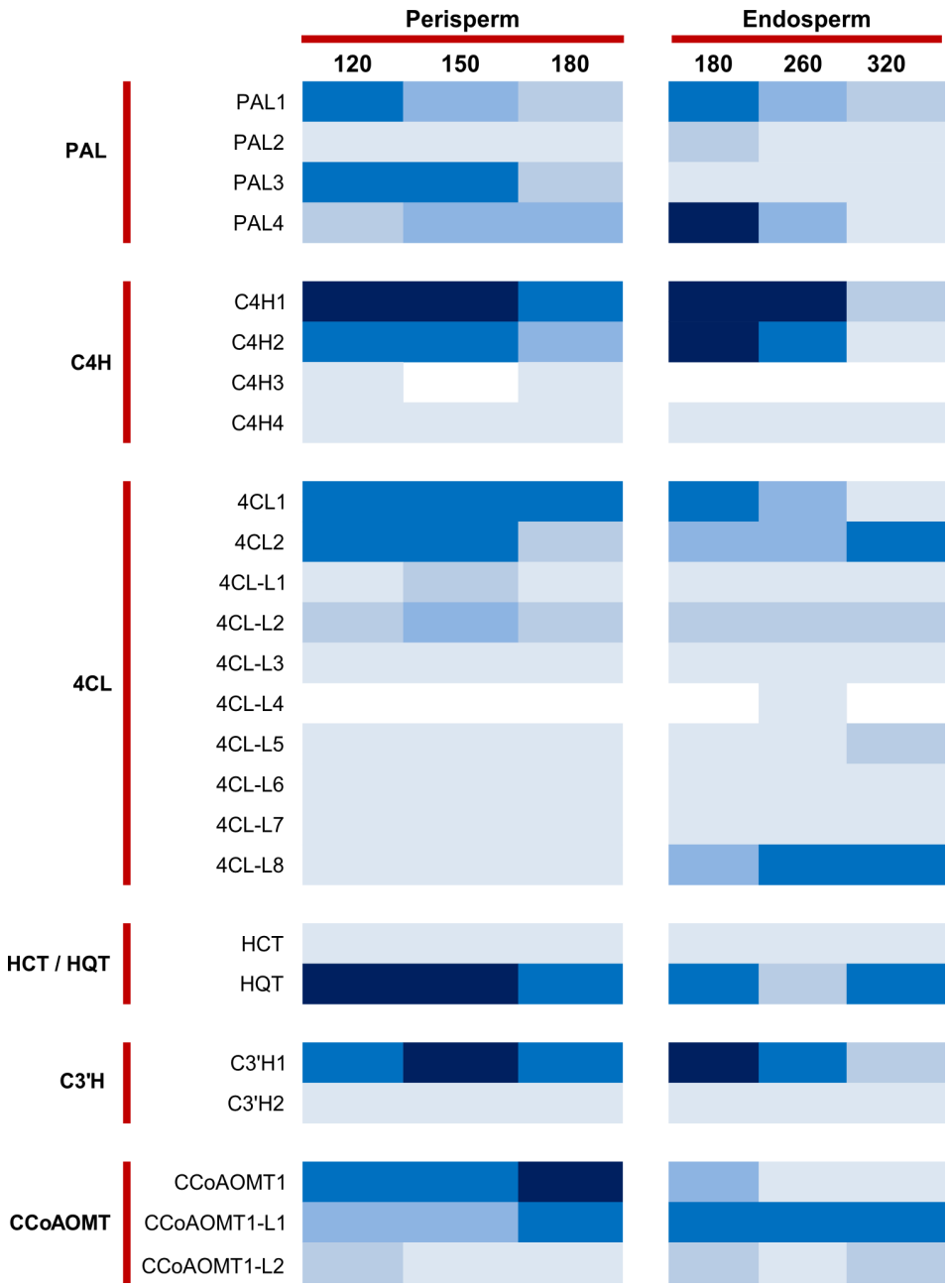


Figure S15. Expression pattern of PPP genes during fruit development from 120 to 320 days after pollination. The intensity of the blue color indicates the level of expression (white : no expression; lightly blue: 1< RPKM<20; pale blue: 12<RPKM<50; blue: 51<RPKM<100; intense blue: 101<RPKM<500, marine blue: RPKM > 500).

Figure S16 (Page 1/5)

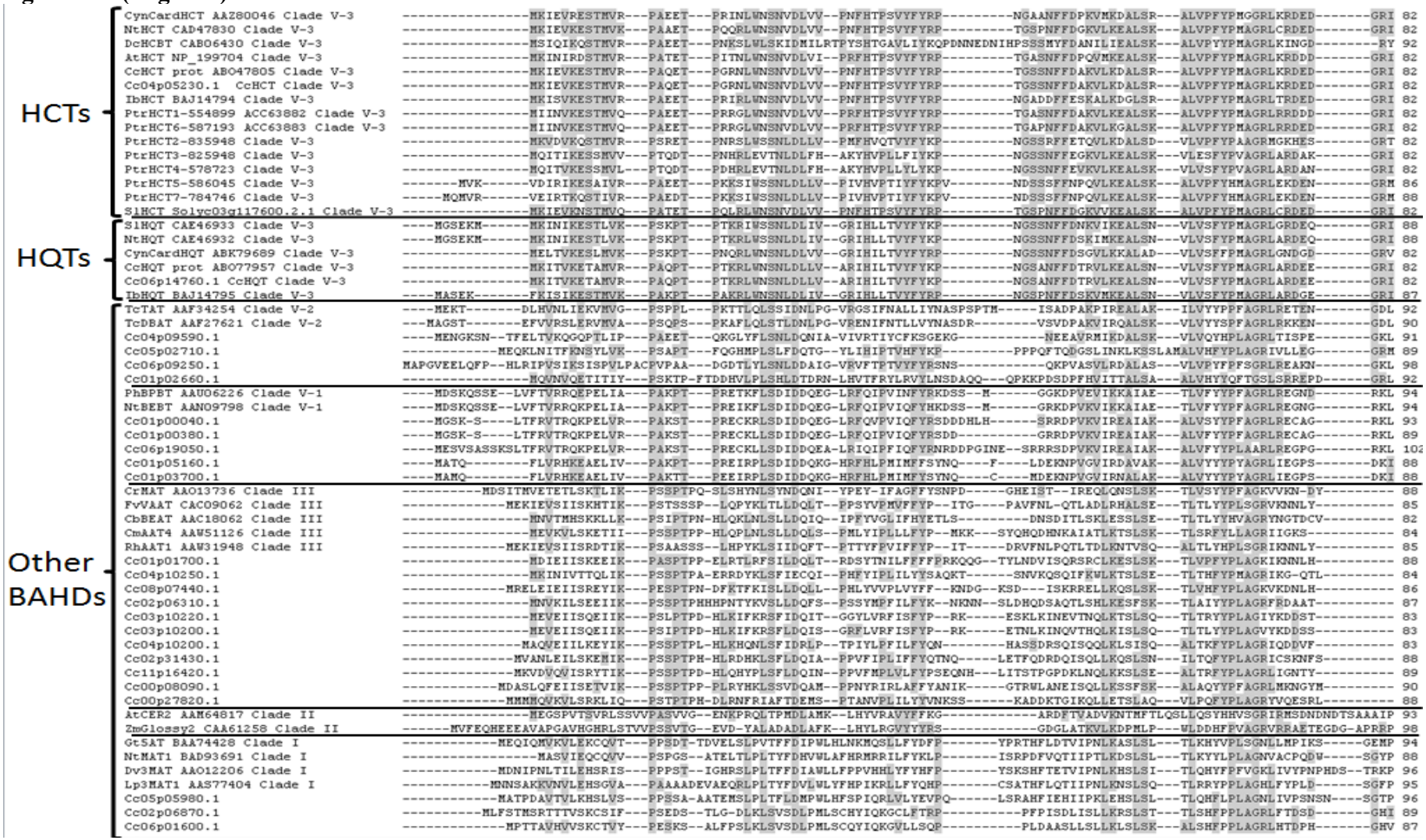


Figure S16 (Page 2/5)

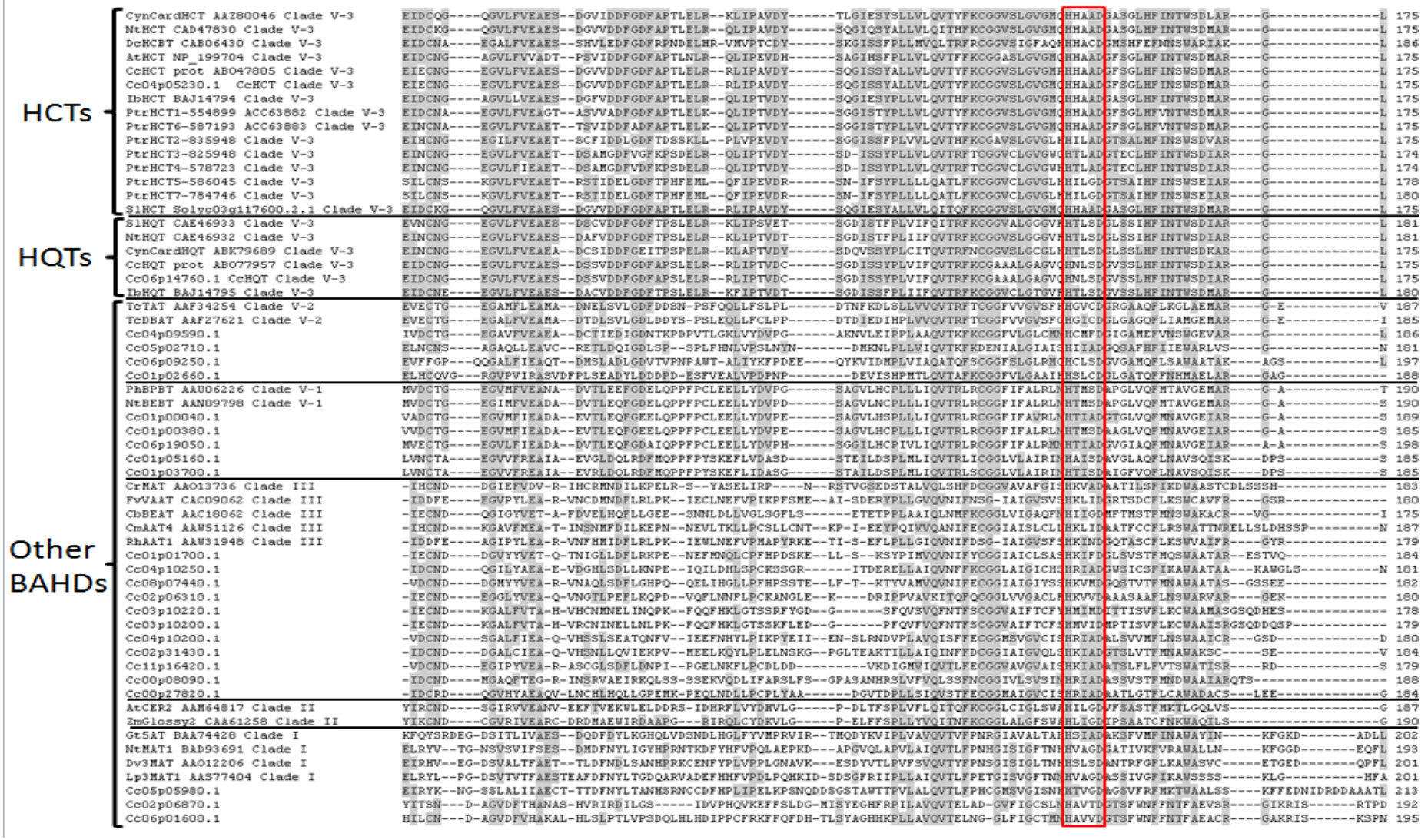


Figure S16 (Page 3/5)

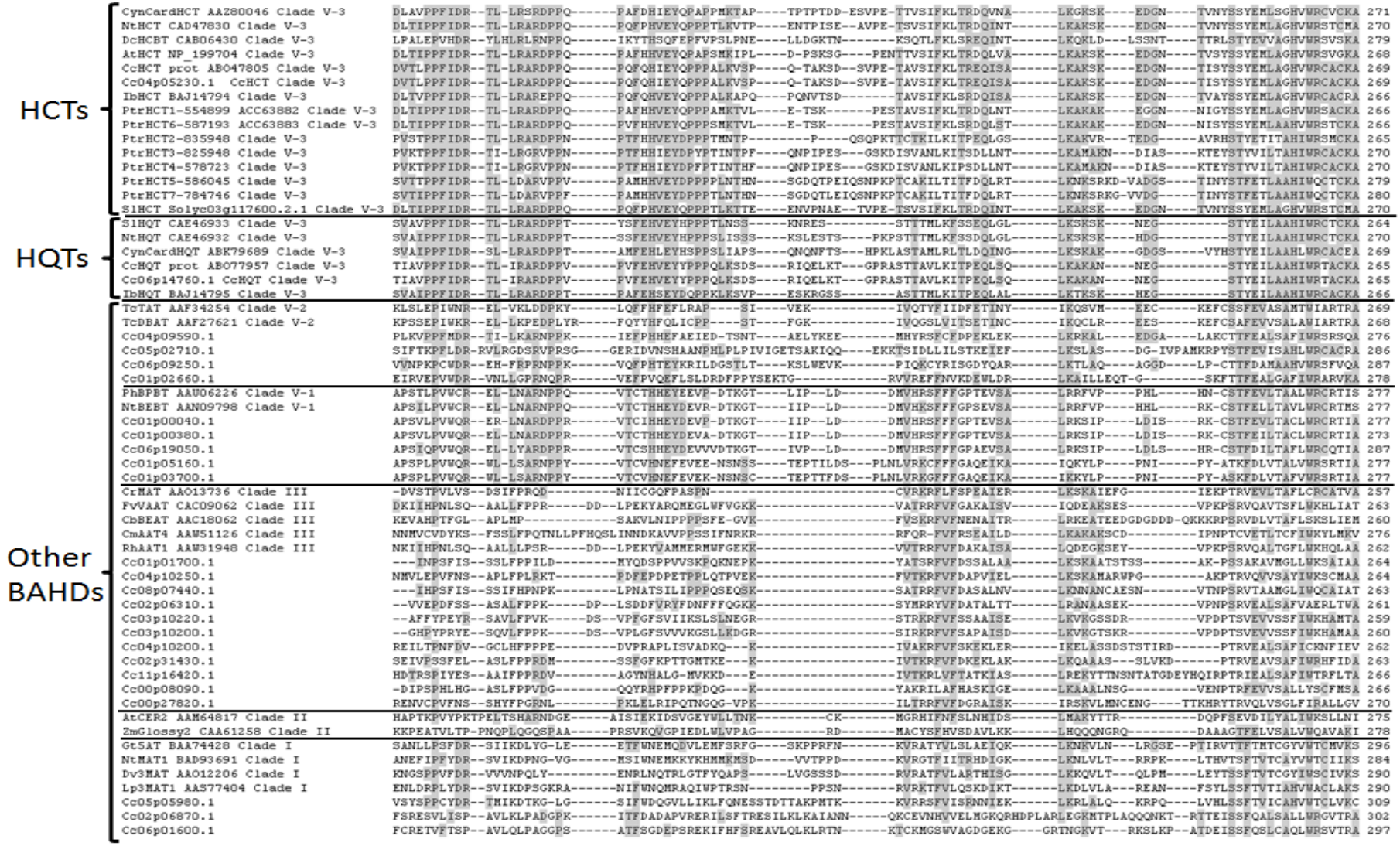


Figure S16 (Page 4/5)

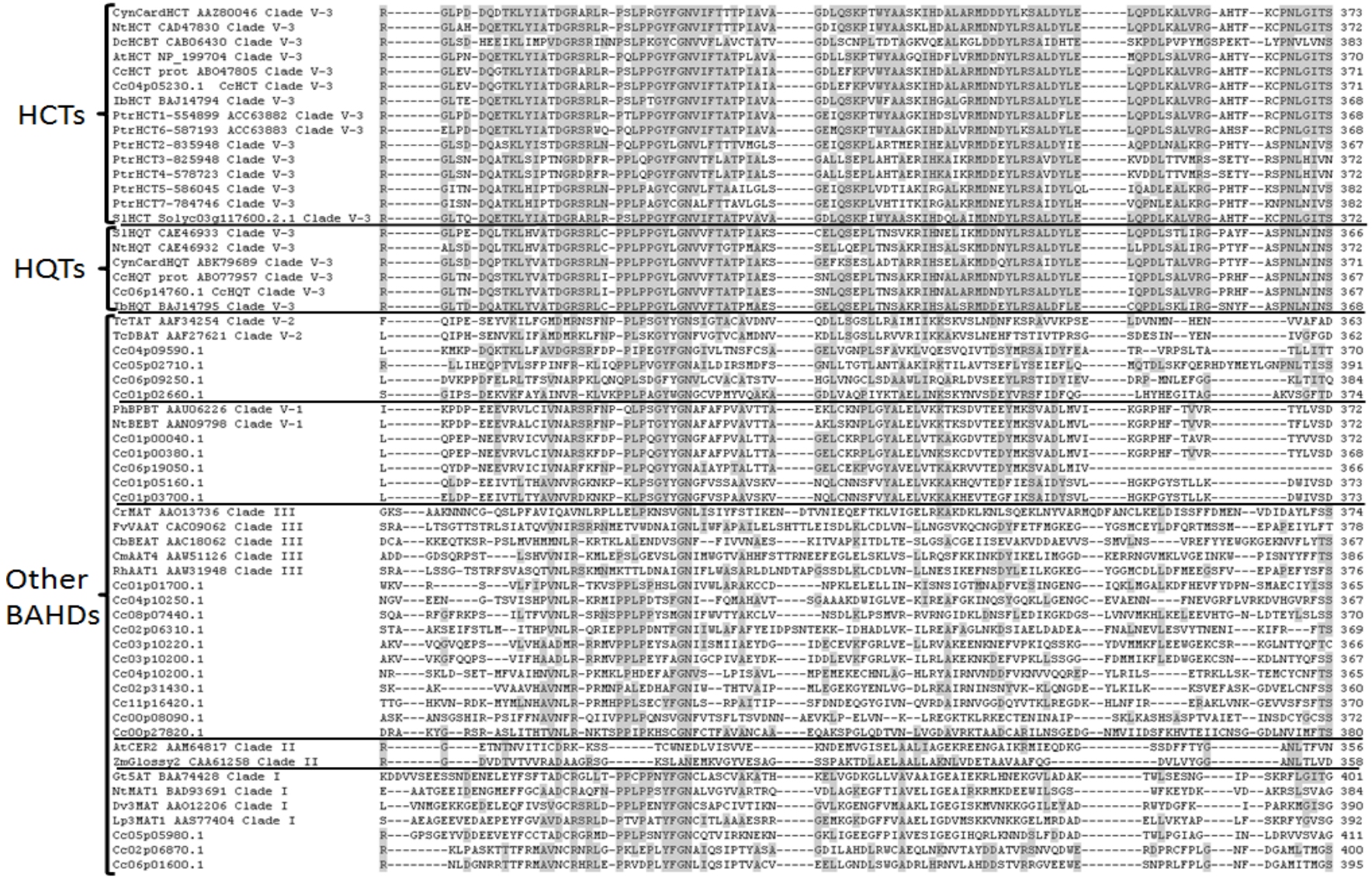


Figure S16 (Page 5/5)

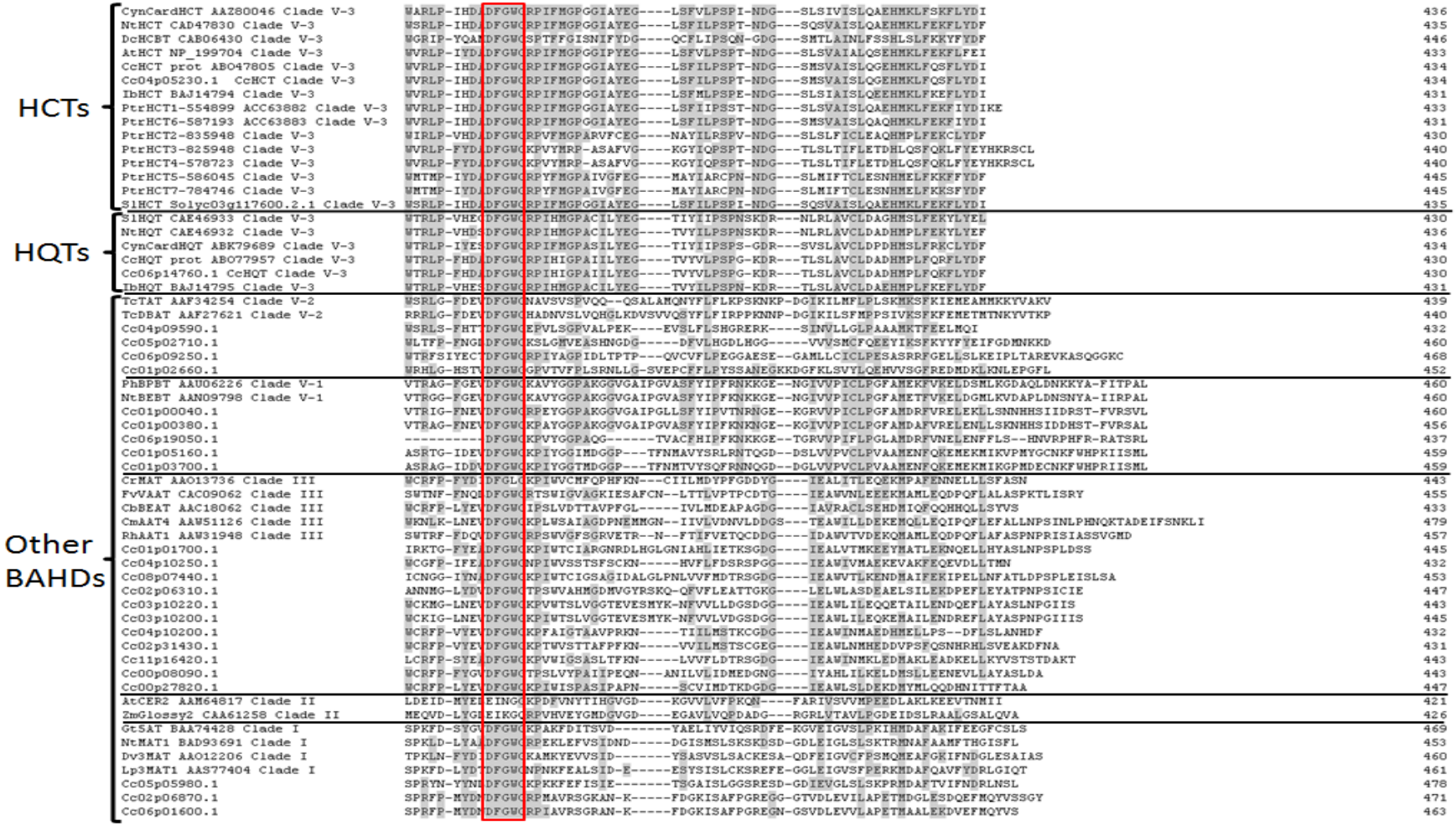


Figure S16. Multiple alignment of all the 25 BAHD candidate proteins identified in the coffee genome (from Table S21) with genetically or biochemically characterized BAHD acyltransferasess (including some of those used in (*159*). The amino acid sequences were aligned using the MEGALIGN program from Lasergene 9 Core suite (DNASTAR ®) with CLUSTAL W method and default parameters (Gap penalty:10; Gap Length penalty: 0.20; Protein weight matrix: Gonnet 250). The residues shaded with solid light grey are those that match the residues found in the majority of the sequences.

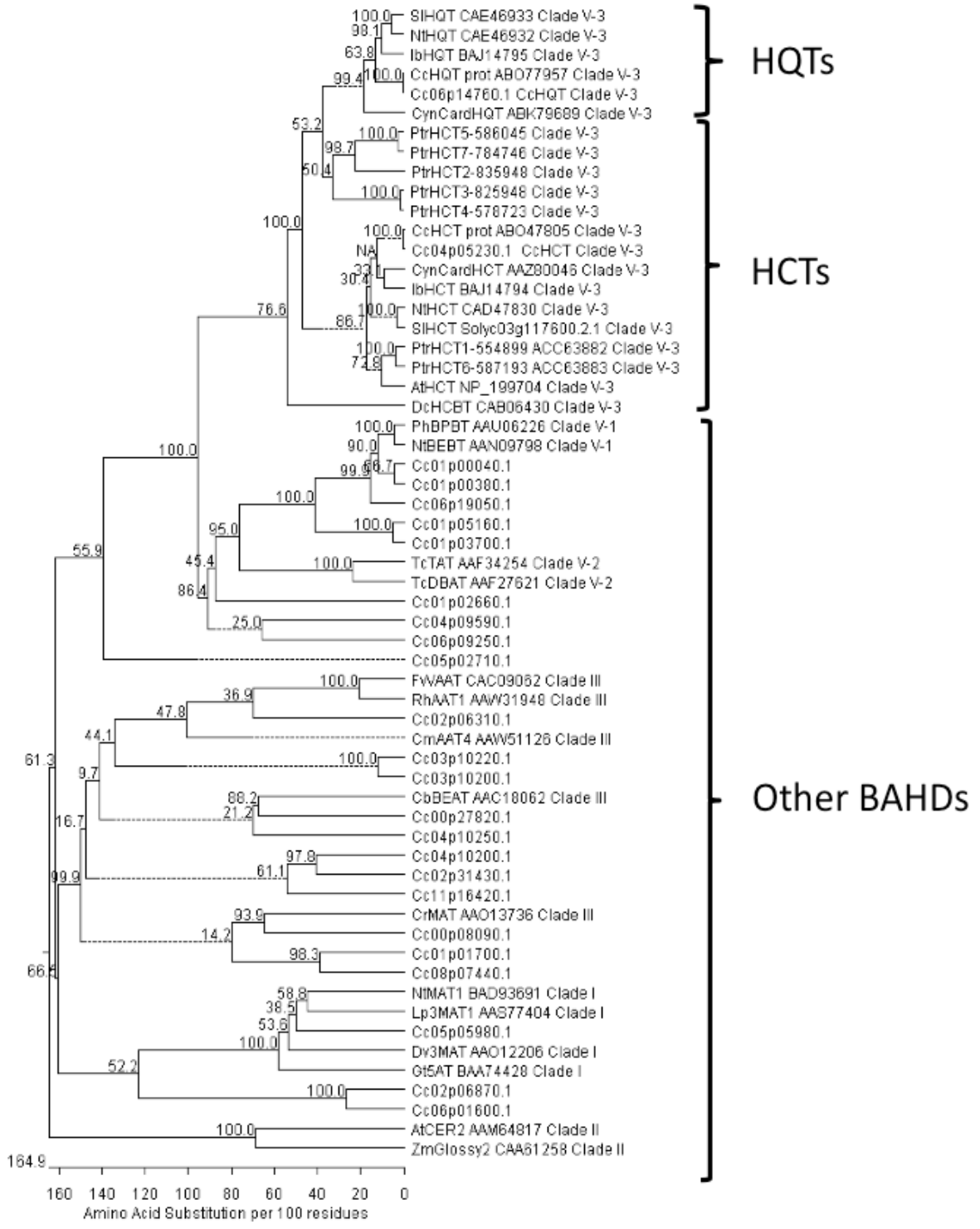


Figure S17. Phylogenetic tree built based on the protein alignment shown in Figure S16. The amino acid sequences were aligned using CLUSTAL W method and MEGALIGN software from Lasergene 9 Core suite (DNASTAR ®). A neighbour-joining tree was generated and bootstrapped at default settings of 1000 trials and a random seed of 111.

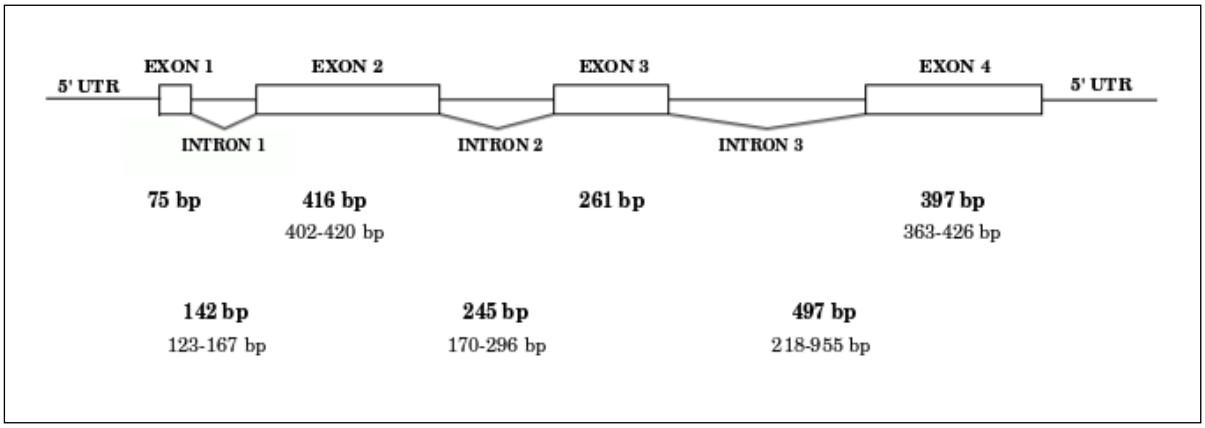


Figure S18. N-methyltransferase gene structure in *C. canephora.* Average, minimum and maximum sizes are given in bp for each intron and exon of full-length genes of the caffeine pathway.

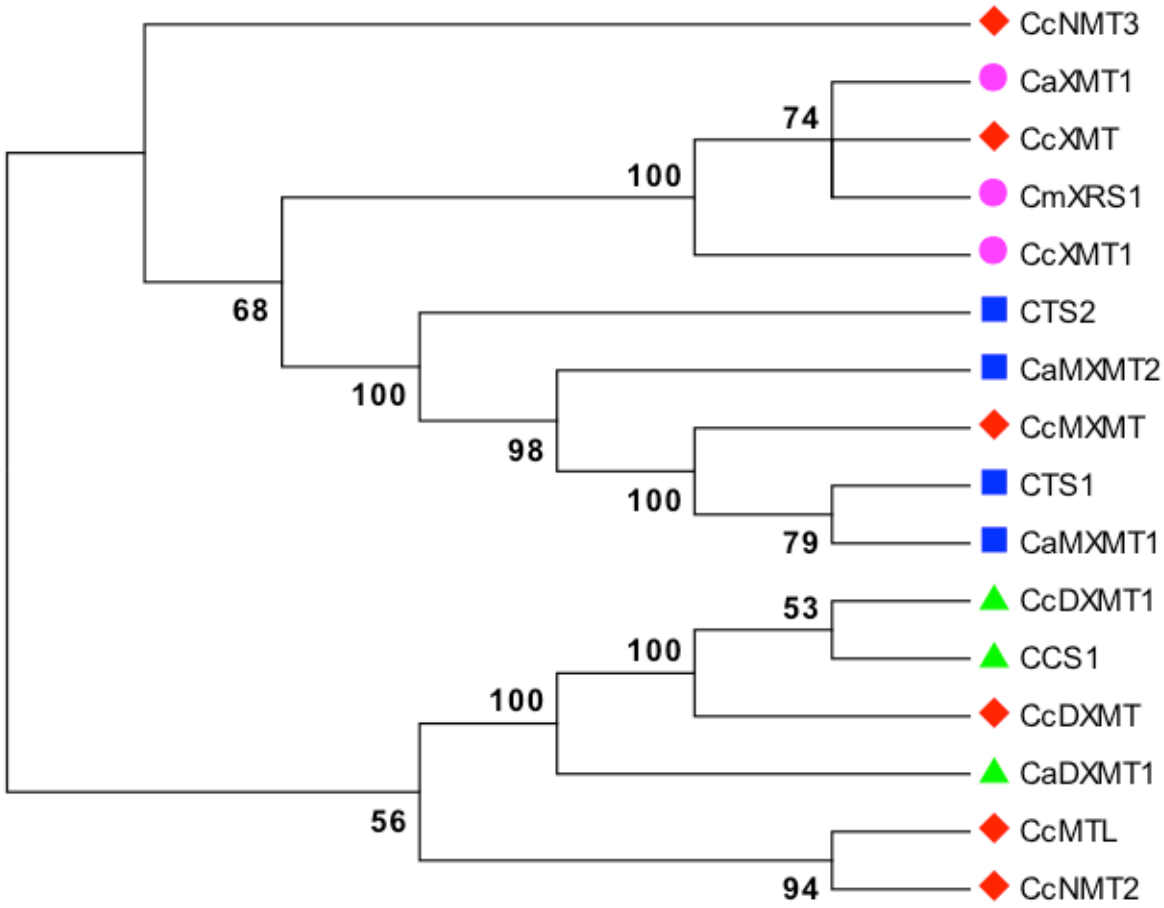


Figure S19. Molecular phylogenetic analysis of *C. canephora* NMT coding sequences (CDS). Phylogenetic relationships of the six putative full-length NMT CDS identified in the *C. canephora* genome (red) and the NMTs from *C. canephora* or *C. arabica* (different colors per subfamily) analysed by Yoneyama et al. (*101*) and McCarthy et al. (*18*). Sources of the sequences are as follows: *CaXMT1*, AB048793; *CmXRS1*, AB034699; *CcXMT1*, DQ422954; *CTS2*, AB054841; *CaMXMT2*, AB084126; *CTS1*, AB034700; *CaMXMT1*, AB084794; *CcDXMT1*, DQ422955; *CCS1*, AB086414; *CaDXMT1*, AB084125. The tree was inferred using maximum likelihood and the bootstrap values were inferred from 1000 replicates.

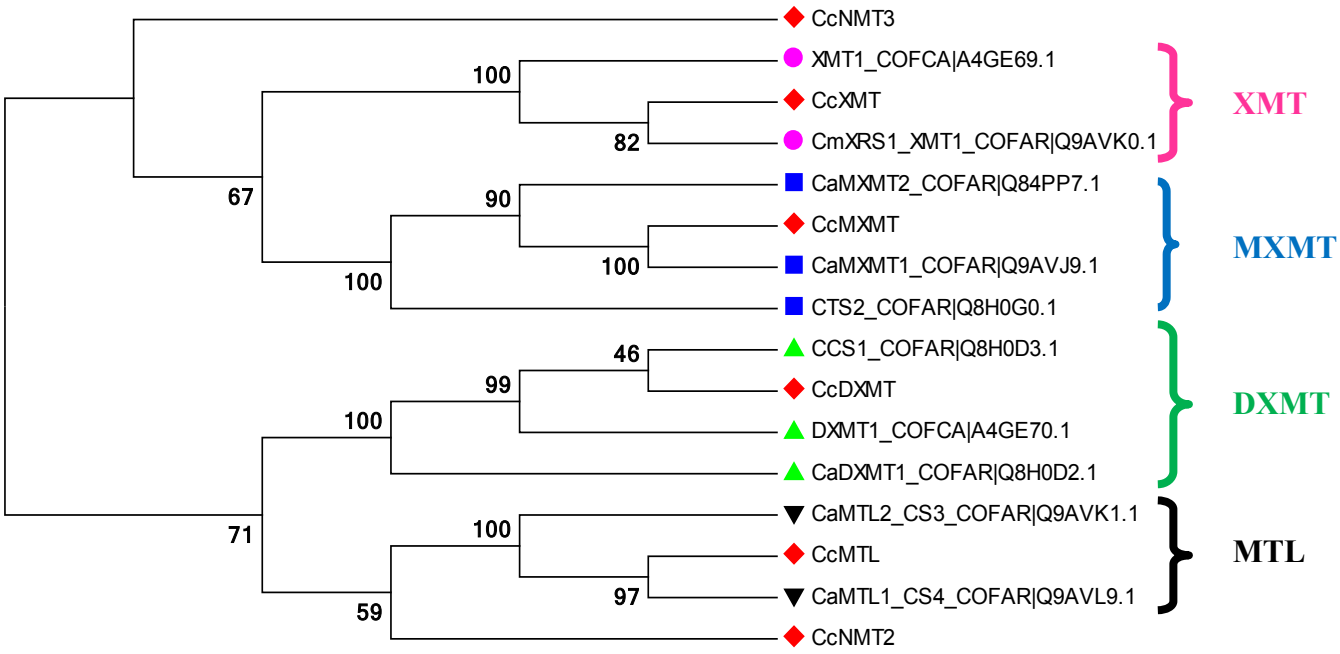


Figure S20. Molecular phylogenetic analysis of *C. canephora* NMT proteins.
Phylogenetic relationships of the six putative full-length NMT proteins identified in the *C. canephora* genome (red diamonds). The consensus tree was inferred using maximum likelihood and the bootstrap values estimated from 1,000 replicates. The color code is as in **Figure S19**.

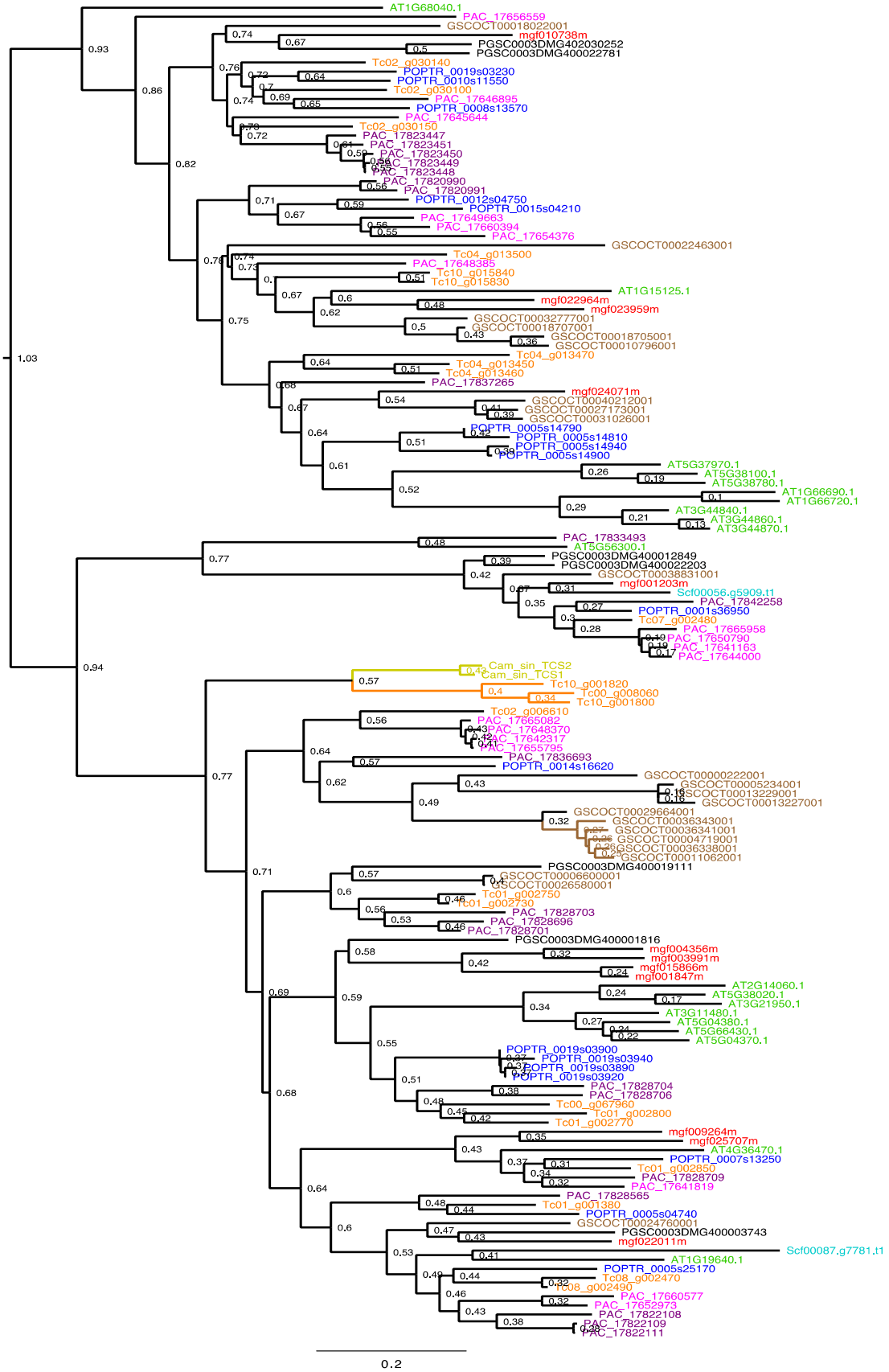


Figure S21. Caffeine biosynthesis has a phylogenetically independent origin in coffee; inferences from a large analysis of NMT nucleotide sequences from multiple eudicots.

The maximum likelihood tree shown is based on a codon alignment of CDS extracted from CoGe, with the exception of 2 *Camellia sinensis* CDS from GenBank (*TCS1* and *TCS2*; AB031280.1 and AB031281.1, respectively). Names follow gene model IDs. Colored branches mark caffeine biosynthetic clades; brown = coffee, yellow = tea, orange = cacao. Sequence names are also colored according to species: brown = coffee, yellow = tea, orange = cacao, purple = grapevine, pink = peach, blue = poplar, green = *Arabidopsis*, red = *Mimulus*, black = potato, and cyan = *Utricularia*.

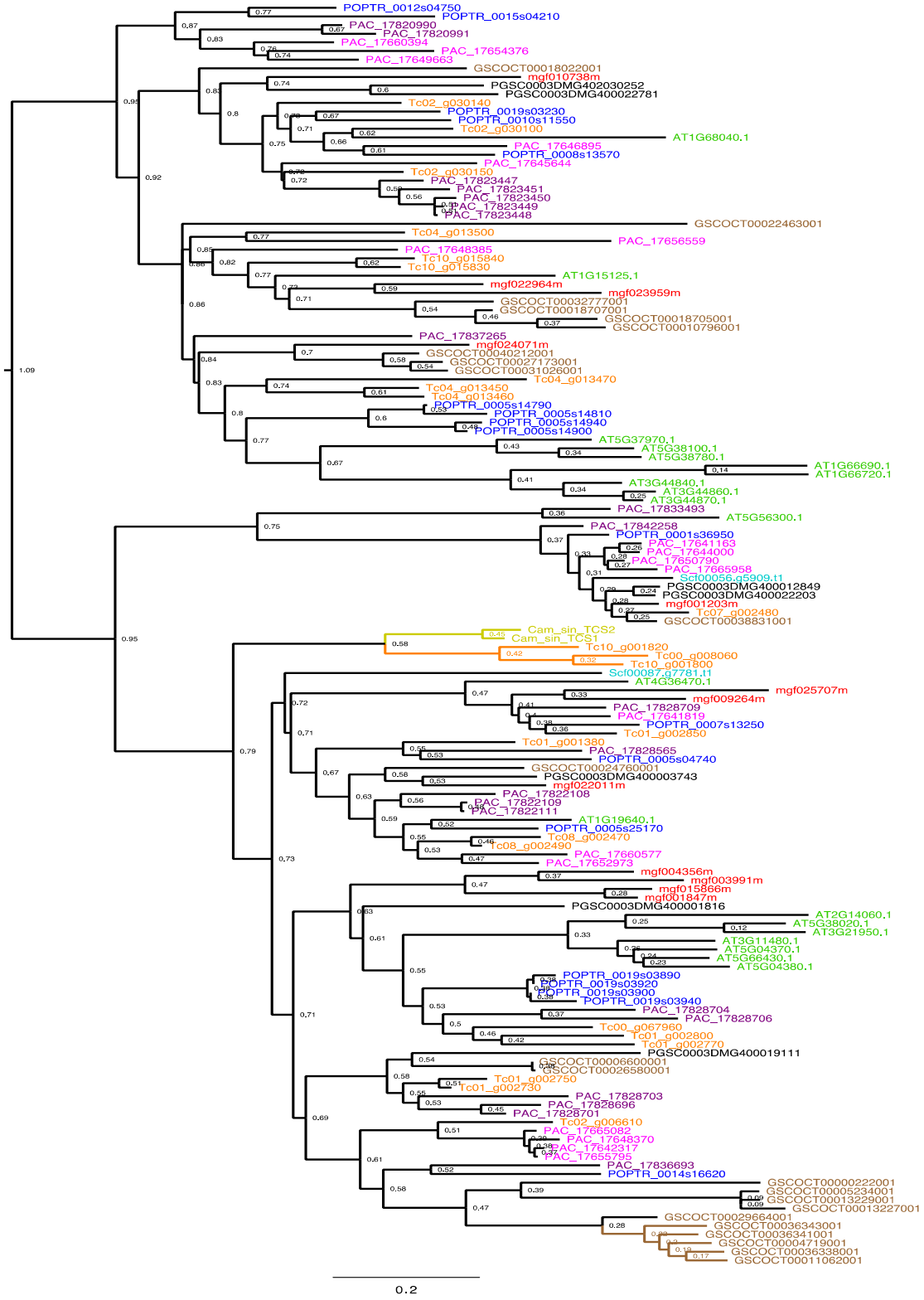


Figure S22. Caffeine biosynthesis has a phylogenetically independent origin in coffee; inferences from a large analysis of NMT amino acid sequences from multiple eudicots. The maximum likelihood tree shown is based on translated and then aligned CDS sequences extracted from CoGe, with the exception of 2 *Camellia sinensis* CDS from GenBank (*TCS1* and *TCS2*). Colors are as in **Figure S21**.

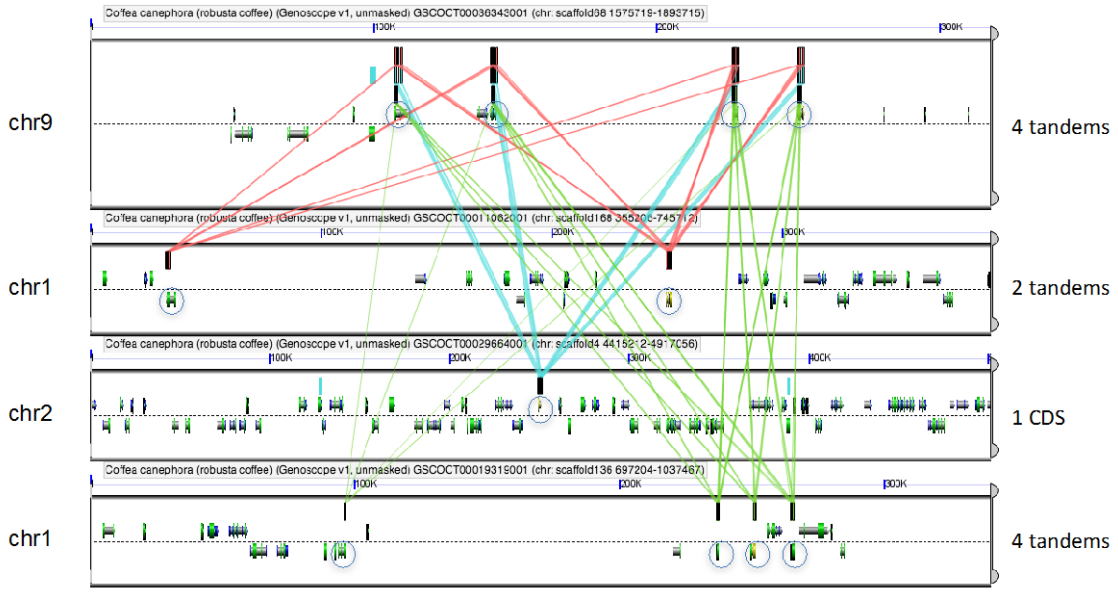


Figure S23. **Four scaffolds from *C. canephora* containing at least one NMT coding sequence (circled), with three of these blocks showing tandem NMT arrays**. Gene models (in green) lie at the centres of each block; high-scoring HSPs lie above these models, with different colored lines connecting them for clarity among the four scaffolds. Despite these shared HSPs, the regions are not syntenic to one another.

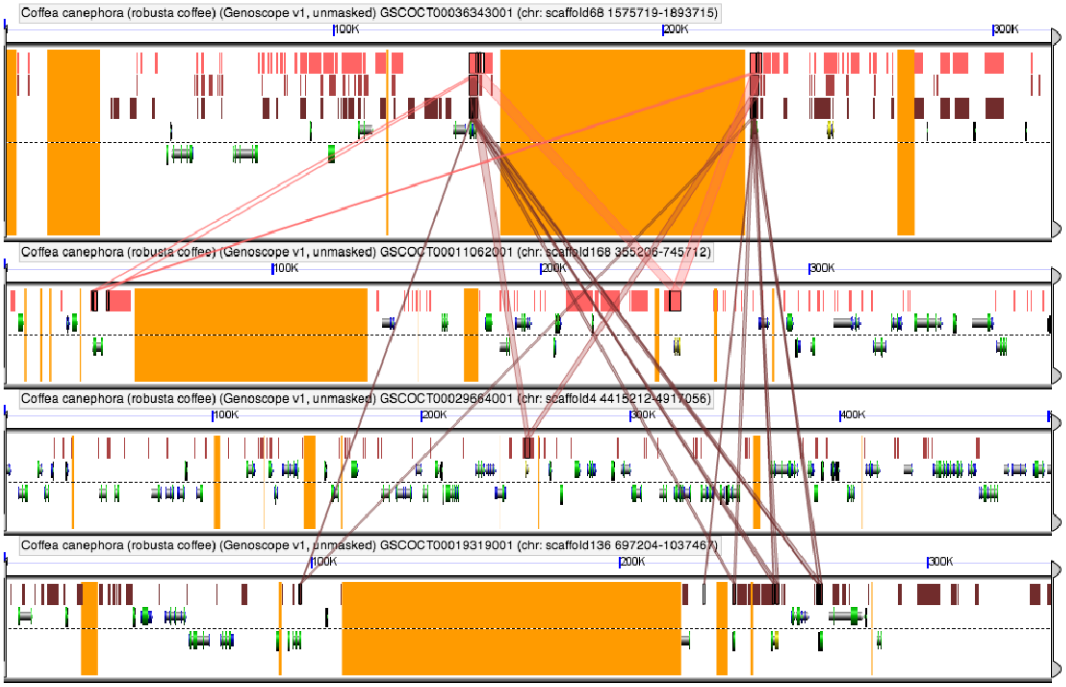


Figure S24. **Unmasked view of the 4 *C. canephora* blocks shown in Figure S21.** Although there are several large and small gaps in the assembly (orange), none of the NMT gene models (joined by lines between high-scoring HSPs) appear directly affected.

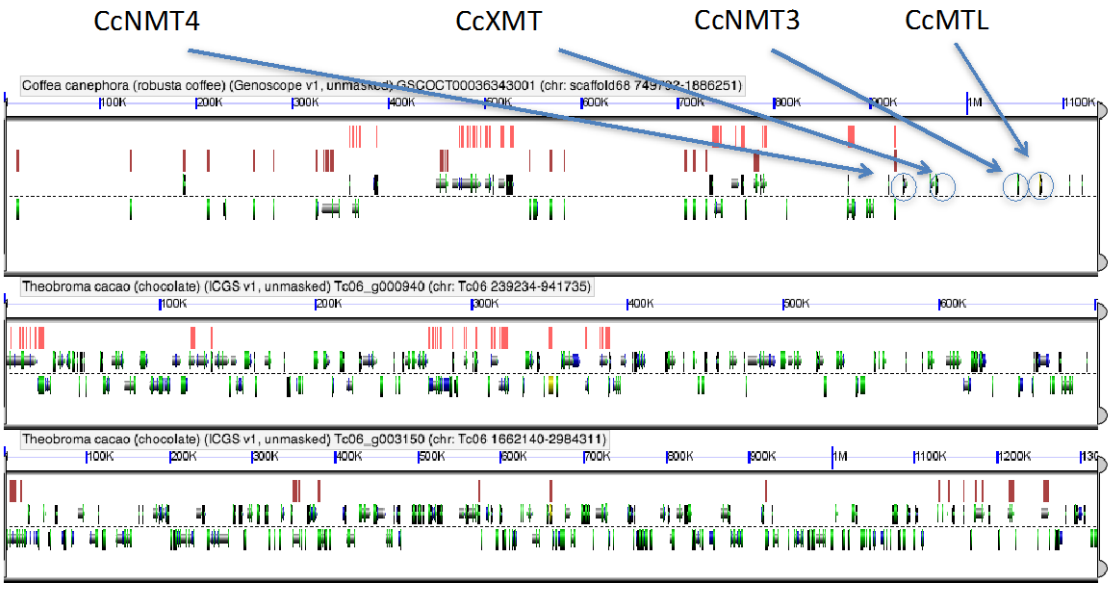


Figure S25. **Microsynteny analysis of the cacao genome compared to the an NMT-containing region of *C. canephora* chromosome 9.** The rows of red and brown HSPs on the top block, from coffee, have numerous syntenic hits to the cacao blocks below. However, the NMTs (circled) on the coffee block (top) are just outside the region of synteny. Gene models and HSPs are figured as in **Figure S23**.

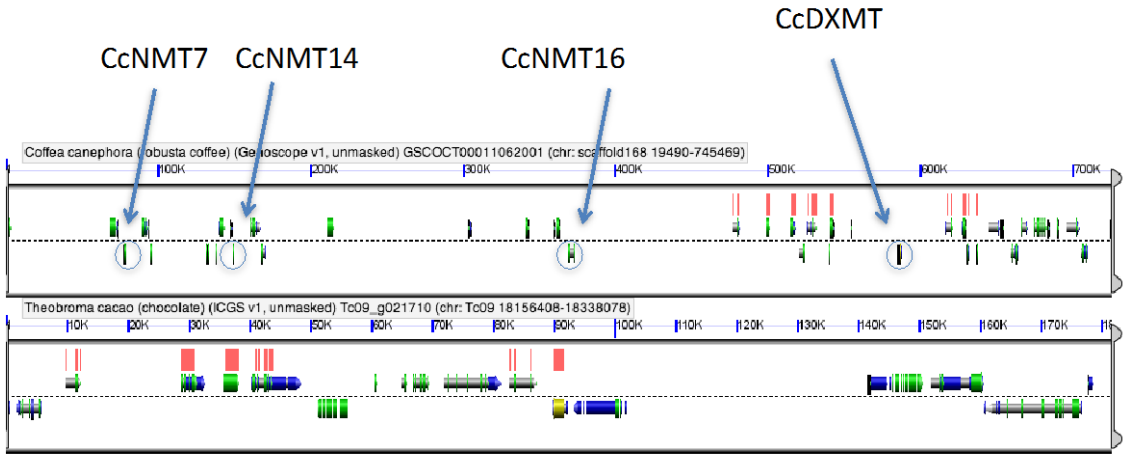


Figure S26. **The *C. canephora* region directly surrounding the *CcDXMT* gene shares synteny with a cacao block, but a cacao ortholog is missing.** *CcDXMT* is part of a tandem array including other ORTHOMCL170 NMTs that lie upstream in a region without synteny to cacao.

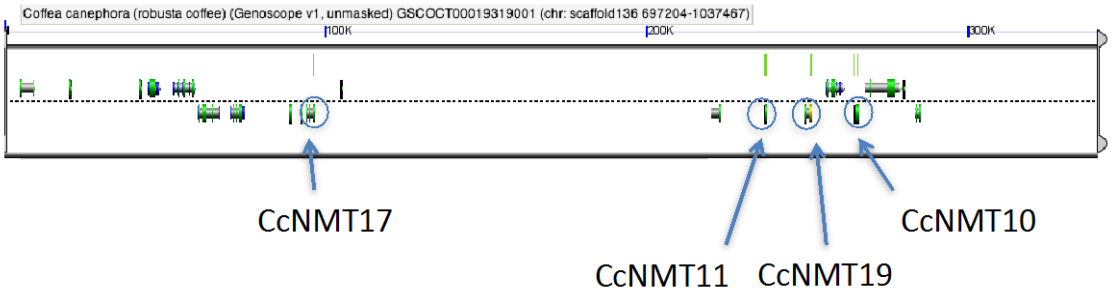


Figure S27. **Several non-caffeine *C. canephora* NMTs lie in a distinct tandem array on chromosome 1.**

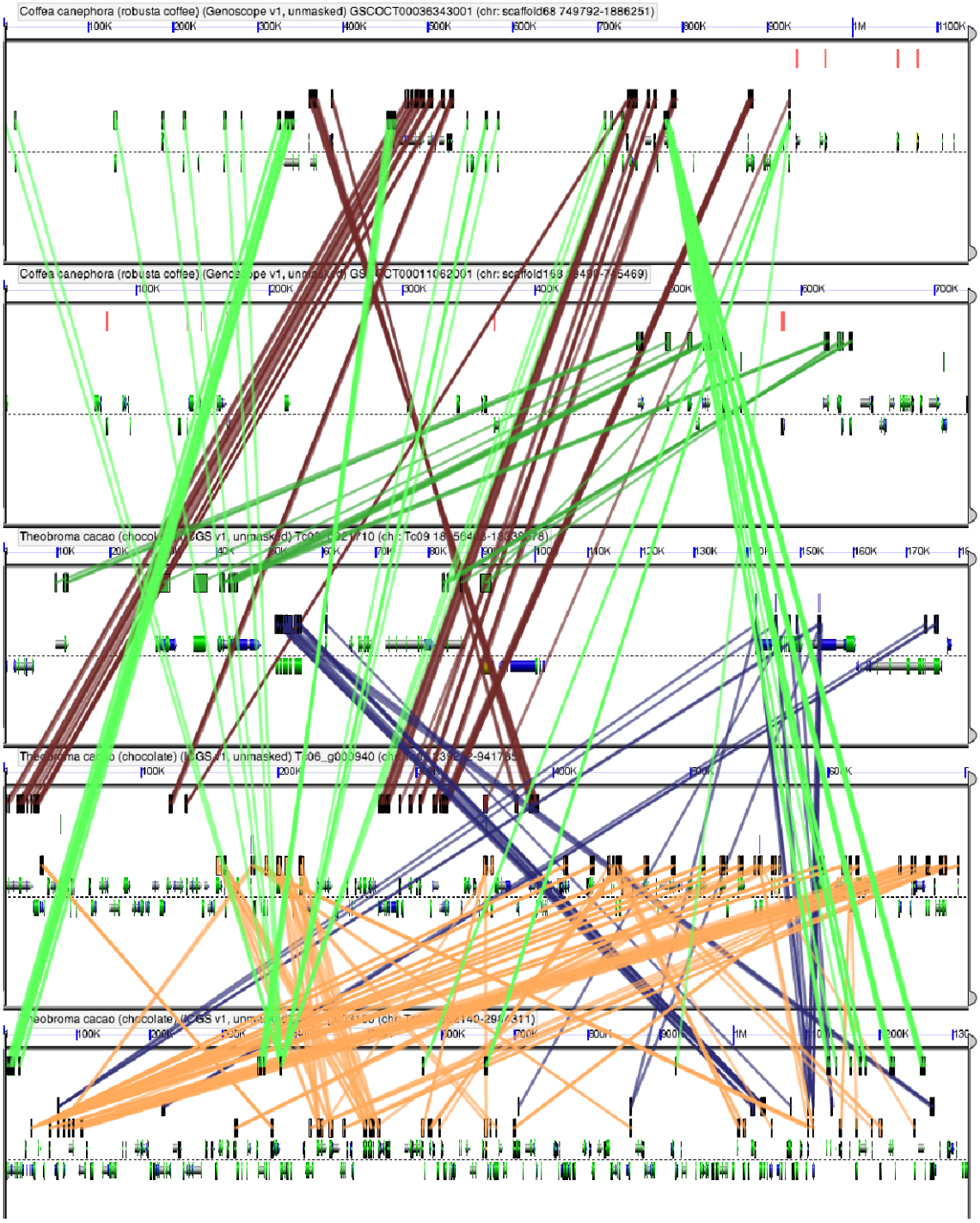


Figure S28. Tandem NMT arrays in coffee share a common genomic context with NMT-free regions of the cacao genome. The tandem array on coffee chromosome 9 and one of those chromosome 1 are syntenically linked to three distinct cacao blocks devoid of NMTs, supporting their common ancestry in a single genomic region.

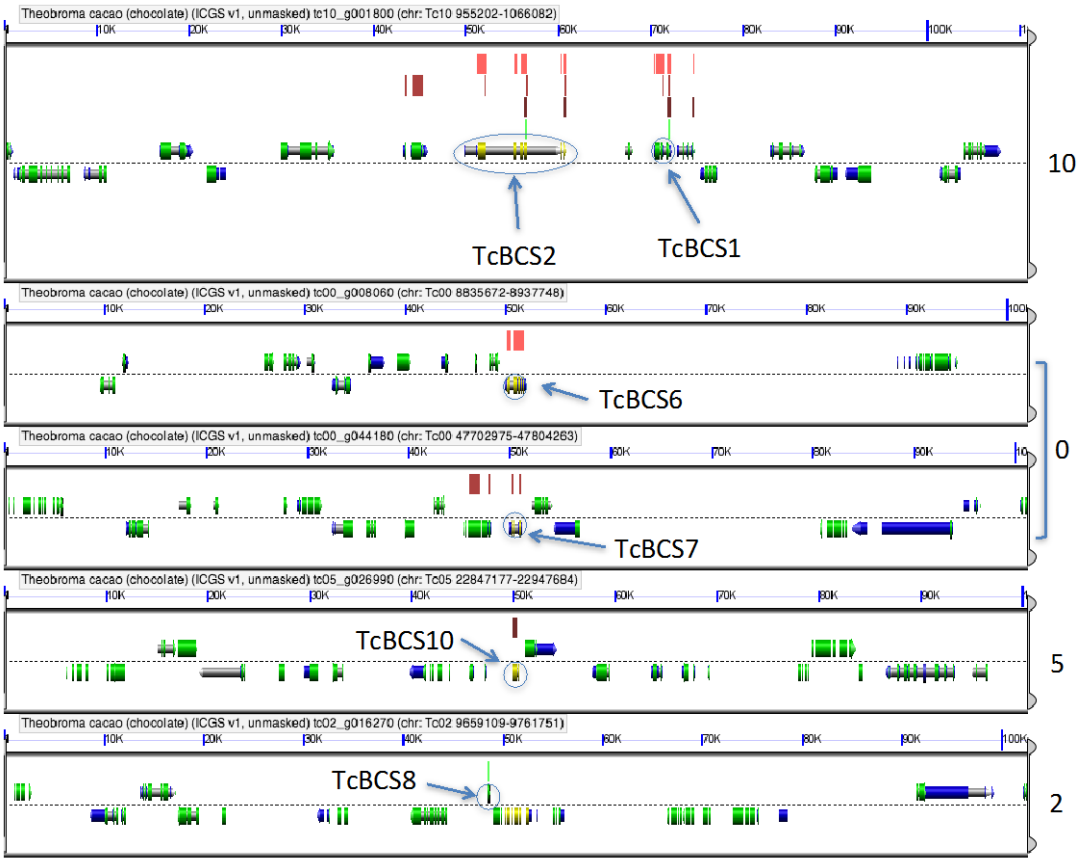


Figure S29. ***Theobroma cacao*** **genome blocks containing caffeine biosynthesis and related NMTs.** *TcBCS1* on chromosome 10 (as shown by numbering to the right), which is the experimentally confirmed caffeine synthase gene in cacao, has a tandem duplicate, *TcBCS2*. Other cacao NMTs exist as singletons (without cross synteny) on other blocks.

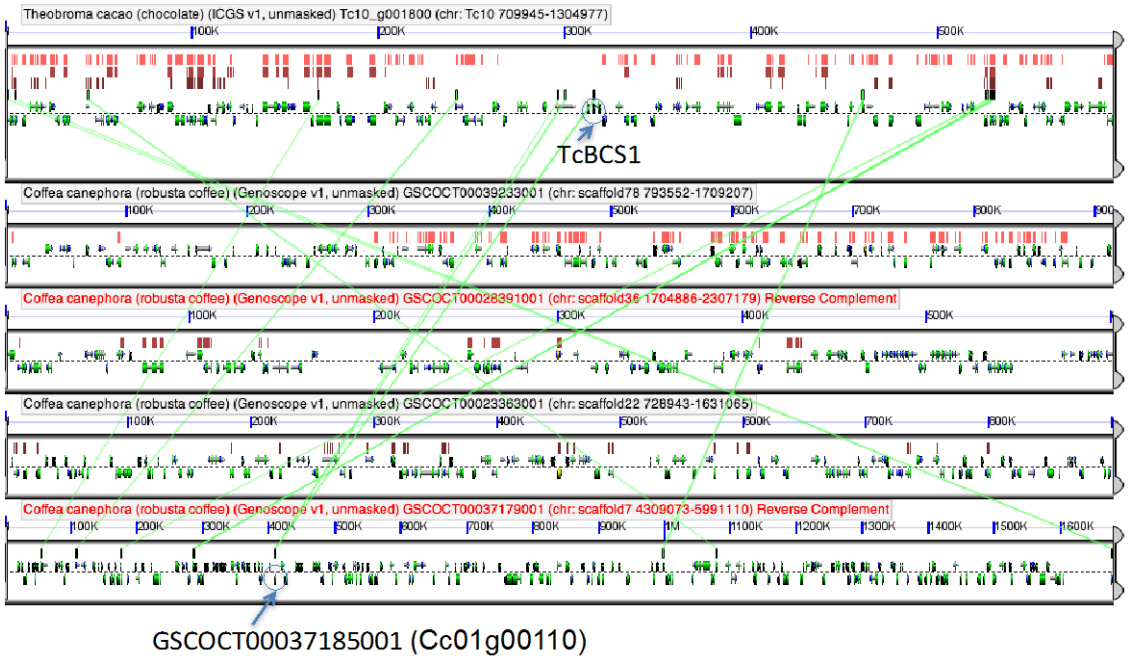


Figure S30. **The experimentally determined gene for caffeine biosynthesis in *T. cacao*, *TcBCS1*, lies in a genomic region with synteny to 4 regions of the coffee genome.** HSPs above the gene models in the cacao block (top) show considerable intercalated synteny. One of the coffee blocks (bottom) contains the *TcBCS1* ortholog, GSCOCT00037185001, a non-caffeine-related NMT that has not been considered otherwise in our analyses.

A

CcNMT3											
CcMTL	0.1089										
CcXMT	0.0856	0.0936									
CcNMT4	0.1511	0.1466	0.1422								
CcNMT7	1.8557	1.9862	1.8758	1.9852							
CcNMT14	1.8598	2.0914	1.9017	2.0491	0.0452						
CcNMT16	1.7302	1.9087	1.743	1.8643	0.0732	0.0402					
CcDXMT	0.0692	0.0847	0.0479	0.1209	1.8428	1.861	1.7756				
CcNMT10	1.9194	2.1463	1.9444	2.0994	0.3501	0.3384	0.3528	2.1074			
CcNMT17	2.3899	2.3879	2.176	2.5437	0.565	0.5378	0.5137	2.2593	0.3142		
CcNMT19	1.8598	2.0914	1.9017	2.0491	0.0452	0	0.0402	1.861	0.3384	0.5378	
	CcNMT3	CcMTL	CcXMT	CcNMT4	CcNMT7	CcNMT14	CcNMT16	CcDXMT	CcNMT10	CcNMT17	CcNMT19

Pairwise Ks between coffee NMTs

B

red	0.10507		
blue		0.040666667	
green			0.3142
	red	blue	green

Average Ks within phylogenetic group (colored according to block location of seed gene)

C

0.44175	blue/green
2.00224	red/blue or green

Average pairwise Ks between phylogenetic groups

Figure S31: Synonymous substitution (Ks) rates permit relative dating of duplications within and among the *C. canephora* NMT tandem arrays. A, Codeml in PAML was used to calculate pairwise Ks values for the complete coding sequences figured. Colors of gene names indicate original tandem array memberships, i.e., phylogenetic relationships; where the physical relationships of some of these array members have moved, the gene names are shown with backgrounds colored according to modern block membership, as in Figure. 2 (main text). For example, the *CcDXMT* gene name is shown in red since it evolved within the tandem array also containing *CcNMT4*, *CcNMT3*, *CcMTL*, and *CcXMT*, but because the gene remained behind after translocation of the red block, its cell is shaded blue. Pairwise Ks values. when colored, are shown as such according to phylogenetic relationships of NMT duplicates, regardless of their modern physical placements. **B,** Average Ks values *within* each phylogenetic group, colored according the original ("seed") duplicates that founded each tandem array, as shown in Figure 2 (main text). These values, when converted to relative times, are interpreted to represent the approximate timing of duplication events *within* arrays, i.e., when their tandem diversifications occurred. **C,** Ks values averaged *between* the different tandem arrays, with the relative divergence times inferred reflecting approximately when each array was founded by a given "seed" duplicate. Colorations are as in B.

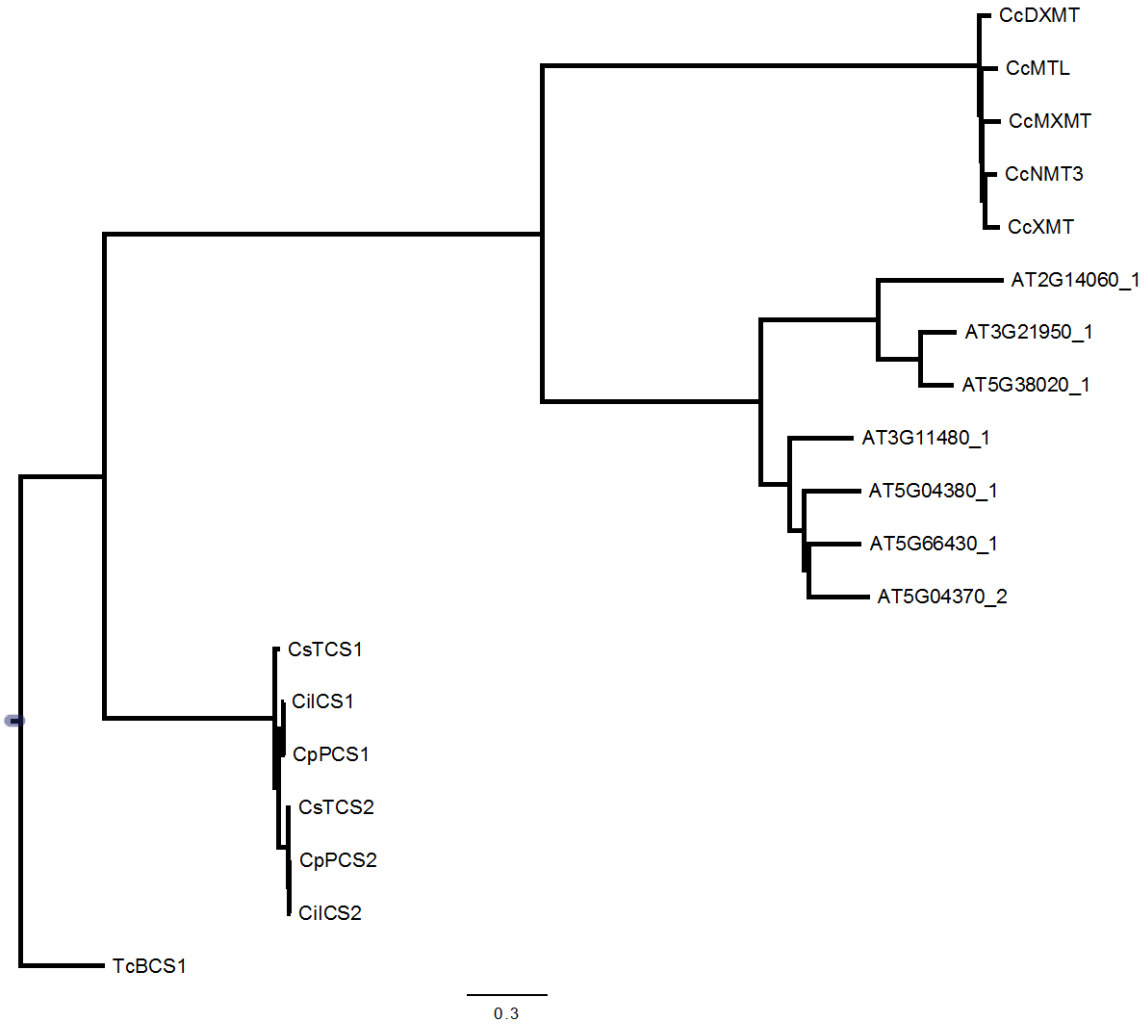


Figure S32. Maximum likelihood phylogeny of known caffeine biosynthetic genes in coffee, tea, and cacao, plus 2 linked and coordinately-expressed NMTs from coffee and 7 genes with non-caffeine function from *Arabidopsis*.

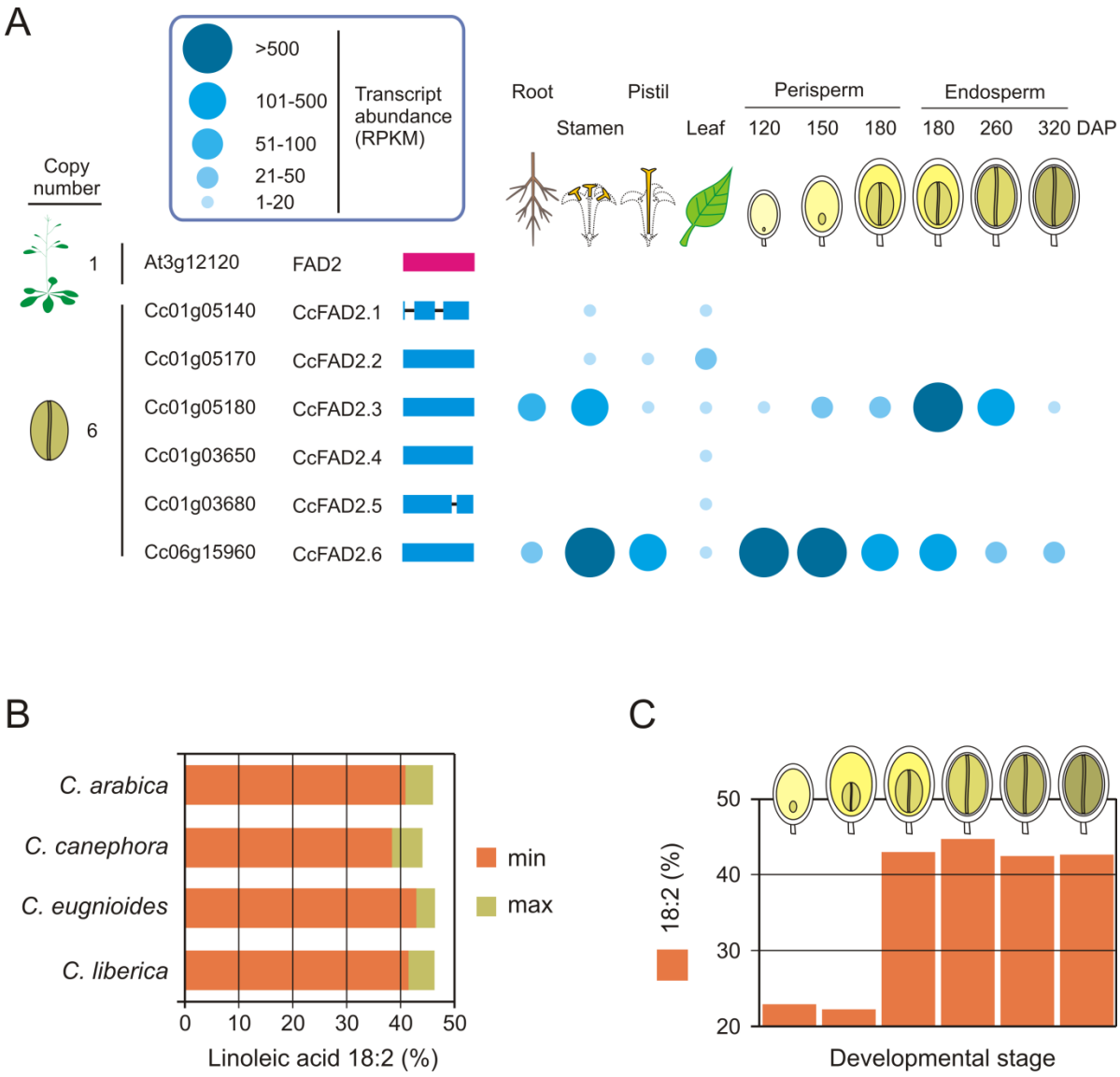


Figure S33. Genome-wide analysis of fatty acid (FA) desaturases in coffee. The 15 FA desaturase genes identified in the genome of *C. canephora* are classified according to (from left to right): their subfamily, which specifies the position of the double bond introduced in the carbon chain, their sub-cellular localization, their chromosome localization, as specified by gene code. Products of the reaction catalyzed by the different enzymes are indicated between brackets. For each enzyme, the number of gene copies identified in the coffee genome is compared with that of *Arabidopsis*. The structure of each coffee FA desaturase gene (blue-colored exons) is also compared with that of an *Arabidopsis* ortholog (magenta-colored exons) representative of the gene family. Exon and intron length is strictly proportional to nucleotide sequence size. Gene expression levels in various tissues (root, stamen, pistil, leaf, perisperm and endosperm), and at various developmental stages of the perisperm and the endosperm (days after pollination, DAP), are indicated by blue-colored spots, the size and color-intensity of which specifies the number of reads per kilobase and million reads (RPKM). Gene expression levels in leaves sampled on *C. canephora* plants cultivated at four temperature regimes (18/14, 23/19, 28/24 and 33/29°C, day/night temperatures) are also given on the right with colored bars. For each gene displaying > 10 RPKM, color intensity indicates normalized transcript abundance from 0% (white) to 100% (dark red).

Supplementary Tables

Table S1. Raw sequencing data overview.

	Number of reads	Number of bases	Coverage	Fragment size (bp)
Roche/454 single end reads	28,725,267	10,759,034,874	15.15	NA
Roche/454 single end long reads	12,403,671	5,855,380,726	8.25	NA
Roche/454 mate pairs reads	4,113,325	1,325,854,339	1.87	13,600
Roche/454 mate pairs reads	4,803,363	1,563,848,456	2.20	7,800
Roche/454 mate pairs reads	4,370,296	1,227,506,974	1.73	2,800
Sanger BAC-ends	143,605	193,347,146	0.27	130,000-170,000
Illumina single end reads	56,483,059	4,114,934,000	5.80	NA
Illumina paired-end reads	226,531,636	38,641,750,000	54.43	300-600

Table S2. Assembly statistics.

	Raw assembly		Final Assembly	
	Contigs	Scaffolds	Contigs	Scaffolds
Number	91,439	13,345	25,216	13,345
Cumulative size (Mb)	475.6	569.4	471.3	568.6
Average size (kb)	5.2	42.6	18.7	42.6
N50 size (kb)	14.8	1,261	51.1	1,261
N50 number	8,509	108	2,290	108
N80 size (kb)	4.3	65.2	15.5	65.3
N80 number	26,145	637	7,259	635
Largest size (kb)	193.8	9,035	817.6	9,028

Table S3. *C. canephora* genetic map characteristics. The eleven linkage groups (A to K) were described according to their size (cM), number of markers, marker density and origin of the markers (RADs, RFLPs, SSRs and SNPs).

Linkage group	Size (cM)	Nb. Markers	Density	RADs	RFLPs, SSRs, SNPs
A	136	320	0.43	157	163
B	221	520	0.43	304	216
C	127	235	0.54	136	97
D	111	274	0.41	149	125
E	117	266	0.44	140	126
F	169	349	0.49	168	181
G	124	305	0.41	156	149
H	137	271	0.51	168	103
I	93	180	0.52	98	82
J	124	286	0.44	156	130
K	112	224	0.50	115	109
Total	1471	3230	0.46	1747	1483

Table S4. Overview of the anchoring of the assembly on the *C. canephora* linkage groups.

Pseudomolecule (Linkage group)	Number of scaffolds	Size (Mb)	No. of genes per pseudomolecule	Gene density (genes/Mb)
1 (A)	34	38.2	2198	57.5
2 (B)	42	54.5	4000	73.4
3 (C)	29	32.0	1632	51.0
4 (D)	35	28.2	1727	61.2
5 (E)	33	29.1	1661	57.1
6 (F)	31	37.3	2839	76.1
7 (G)	21	29.8	2146	72.0
8 (H)	39	31.6	1718	54.4
9 (I)	26	22.3	1094	49.1
10 (J)	34	27.6	1653	59.9
11 (K)	25	33.5	1753	52.3
Un	12996	205.6	3603	17.5

Table S5. Description of the RNA-Seq data used for gene expression analysis. Vegetative and reproductive tissues (root, leaves, pistil, stamen) were taken from greenhouse-grown *C. canephora* plants. To investigate the effect of the growing temperature on the transcriptome, plants of *C. canephora* were cultivated for two months in four sets of contrasted growing conditions with different diurnal/nocturnal temperatures: 18/14, 23/19, 28/24 and 33/29°C. Perisperm and endosperm samples were prepared from fruits collected on field-grown *C. canephora* plants at different developmental stages: 120, 150 and 180 days after pollination (DAP) for perisperm, and 180, 260 and 320 DAP for endosperm. The Illumina reads were aligned using BWA (version 0.7.2, BWA-MEM algorithm) against the *C. canephora* 25,574 protein-coding gene models used as reference transcriptome.

Samples	Conditions/ stage	Replicate	Type of sequences	No of raw reads	No. of cleaned reads	No. of unambiguously mapped reads
Leaf		1	2 x 100 nt	32,000,000	#	22,352,608
		2		19,581,904	#	13,641,913
Root		1	2 x 100 nt	30,541,254	#	22,126,471
Stamen		1	2 x 100 nt	21,296,792	#	13,754,195
Pistil		1	2 x 100 nt	27,345,728	#	16,910,294
Perisperm	120 DAP	1	2 x 100 nt	194,776,548	151,136,192	114,892,270
		2	2 x 100 nt	204,104,288	161,970,384	99,499,872
Perisperm	150 DAP	1	2 x 100 nt	158,526,444	61,024,346	46,747,685
		2	2 x 100 nt	173,119,798	90,379,264	69,046,975
		3	2 x 100 nt	192,120,378	75,067,146	56,656,825
Perisperm	180 DAP	1	2 x 100 nt	138,025,566	74,103,706	41,889,236
		2	2 x 100 nt	193,989,000	19,435,420	14,180,533
		3	2 x 100 nt	200,185,866	108,682,774	64,340,964
Endosperm	180 DAP	1	2 x 100 nt	159,700,240	125,787,798	92,083,762
		2	2 x 100 nt	162,151,272	64,488,618	47,951,545
		3	2 x 100 nt	170,412,800	88,995,062	65,414,285
Endosperm	260 DAP	1	2 x 100 nt	157,423,006	146,130,104	99,858,416
		2	2 x 100 nt	201,027,848	125,616,384	89,230,431
		3	2 x 100 nt	214,070,658	169,367,600	115,829,425
Endosperm	320 DAP	1	2 x 100 nt	138,424,748	92,766,848	53,977,028
		2	2 x 100 nt	179,230,196	96,372,172	71,310,083
		3	2 x 100 nt	214,387,026	148,779,860	110,650,806
Leaf	18/14°C	1	1 x 100 nt	36,484,088	#	26,730,866
		2	1 x 100 nt	37,583,402	#	27,149,887
Leaf	23/19°C	1	1 x 100 nt	37,498,939	#	27,324,383
		2	1 x 100 nt	36,596,537	#	26,734,851
Leaf	28/24 °C	1	1 x 100 nt	30,477,145	#	22,372,506
		2	1 x 100 nt	21,182,110	#	15,543,751
Leaf	33/29°C	1	1 x 100 nt	38,619,894	#	28,514,134
		2	1 x 100 nt	35,669,174	#	26,405,098

Table S6. Annotation metrics.

Number of genes	25,574
Number of intronless genes	5004
Gene size (mean : median)	3684 : 2788
No. of exons/ gene (mean : median)	5.1 : 4
CDS size (mean: median)	1206 : 1002
Coding nucleotides	30,830,841 (5.4%)
Number of introns	104,944
Intron size (mean : median)	483 : 208
% contigs with >=1 gene (% of nt in those contigs)	16.6% (82.3%)

Table S7. Transcription factor distribution in *Coffea* and other sequenced plant genomes. The gene copy number of each TF family was compared among 14 genomes for 63 Transcription factors. We used a unique five-letter species code for each organism. Transcription factors have been classified using protein-coding sequences and InterPro domains (IPR) scans available in the GreenPhyl database (InterPro v28). The classification was carried out according to the TAP classification rules proposed by Lang et al.(*172*) However, instead of PFAM, we used the corresponding InterPro domains (all sources). Some TFs with no defined IPR/PFAM (i.e., Trihelix, HRT, DBP, GARP_G2-like, VOZ, Alfin-like) were classified using HMMsearch (v2.3.2 and cut-off 1e-05) based on HMM files created from MSAs downloaded on PlnTFDB (v3.0).

TF	cofca	sollc	arath	carpa	glyma	medtr	poptr	ricco	vitvi	musac	bradi	orysa	sorbi	zeama
ABI3/VP1	47	76	74	33	63	94	119	42	21	41	42	60	61	46
Alfin-like	6	11	7	4	19	5	10	5	8	20	12	10	16	21
AP2/EREBP	103	170	146	93	350	136	209	115	128	275	129	167	158	210
ARF	17	22	22	10	57	28	37	18	20	52	41	27	27	38
ARID	10	14	10	8	20	14	13	10	10	12	9	9	8	11
AS2/LOB	33	47	43	35	86	32	57	34	45	72	25	36	36	43
BBR/BPC	6	6	7	2	10	2	15	5	3	12	3	4	5	4
BES1	7	9	8	5	14	4	16	7	6	12	7	6	8	10
bHLH	121	163	158	100	326	119	212	132	118	284	135	177	179	203
bZIP	40	58	61	37	125	50	80	42	41	123	74	84	86	100
C2C2_CO-like	8	13	17	8	20	10	15	10	12	24	20	17	16	20
C2C2_Dof	28	33	36	19	79	26	41	23	26	74	27	30	29	42
C2C2_GATA	17	26	27	20	48	29	32	16	16	47	21	23	27	34
C2C2_YABBY	6	9	6	7	17	7	12	6	7	25	8	8	8	13
C2H2	92	161	151	128	294	143	219	119	114	242	139	195	158	208
C3H	68	81	74	47	148	60	99	55	65	113	62	73	66	75
CAMTA	4	6	5	2	13	4	7	4	4	6	7	6	6	6
CCAAT_HAP2	8	10	10	5	21	5	12	6	5	15	7	11	10	12
CPP	5	4	8	5	11	6	13	6	6	8	9	11	8	10
CSD	2	5	4	3	7	1	4	5	3	6	4	3	2	4
DBP	1	1	1	0	3	1	4	1	3	3	3	3	2	6
E2F/DP	8	8	8	5	14	6	10	6	6	10	7	9	10	17
EIL	4	9	6	4	12	12	7	4	4	17	6	9	7	9
FHA	15	18	17	13	31	16	31	19	14	20	20	20	17	15
GARP_ARR-B	3	6	11	5	14	4	9	5	7	10	5	5	7	7
GARP_G2-like	41	70	40	34	79	30	60	32	36	97	44	44	43	52
GeBP	2	11	22	4	9	4	8	4	3	7	14	13	15	21
GRAS	50	54	33	39	111	61	102	48	48	73	45	60	75	74
GRF	7	14	10	9	22	5	21	12	9	21	6	13	11	17
HB_KNOX	8	10	9	7	29	7	12	7	10	18	7	12	9	14
HB	52	74	68	38	142	52	98	47	53	102	55	59	55	79
HD-Zip	19	26	21	16	52	14	32	16	20	59	22	26	23	33
HRT	1	1	2	2	1	2	1	1	1	1	1	1	1	2

TF	cofca	sollc	arath	carpa	glyma	medtr	poptr	ricco	vitvi	musac	bradi	orysa	sorbi	zeama
LFY	1	1	1	1	3	1	1	1	1	1	1	2	1	1
LIM	9	15	13	7	30	13	17	10	14	21	11	10	9	14
MADS	55	105	110	151	158	92	89	41	61	80	53	76	76	66
mTERF	25	30	35	23	56	22	55	35	25	30	38	34	38	28
MYB	109	155	148	89	322	121	223	123	157	314	90	140	124	175
MYB-related	51	68	59	42	120	52	76	46	47	84	43	54	70	86
NAC	63	102	114	79	173	74	169	96	82	172	89	148	125	149
NZZ	0	1	1	1	0	0	2	0	2	0	0	0	0	0
PLATZ	7	21	12	9	28	11	20	11	10	26	15	15	15	14
Pseudo ARR-B	4	5	5	4	10	2	6	6	5	11	5	5	3	4
RB	1	1	1	1	3	1	1	2	2	2	2	2	4	4
RWP-RK	**28**	**10**	**14**	**5**	**23**	**8**	**19**	**10**	**8**	**17**	**15**	**13**	**13**	**17**
S1Fa-like	1	1	3	1	4	4	2	1	2	3	1	2	1	1
SAP	1	1	1	1	1	1	1	1	1	1	0	0	0	0
SBP	14	17	17	11	45	13	31	15	18	58	18	19	18	30
Sigma70-like	6	7	6	3	12	4	7	5	6	9	7	6	6	9
Sir2	2	2	2	2	4	2	2	2	2	2	2	2	2	3
SRS	4	9	11	4	22	10	11	5	4	9	5	5	5	9
TAZ	5	9	10	6	7	9	9	5	4	14	4	7	7	6
TCP	18	36	24	22	55	15	33	22	19	46	21	22	28	42
TEA	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Tify	13	21	18	11	33	21	25	14	15	51	21	21	19	34
Trihelix	18	21	22	23	56	14	41	23	27	35	15	16	19	27
TUB	8	11	11	7	23	14	14	7	17	22	12	15	13	14
ULT	3	3	2	10	31	3	3	2	1	3	1	2	1	2
VOZ	2	2	2	2	6	2	4	2	2	5	2	3	4	4
Whirly	3	2	3	2	6	1	3	2	2	3	2	2	3	2
WRKY	49	81	73	50	176	74	108	58	59	153	74	104	95	123
zf_HD	11	22	17	10	50	15	21	11	12	27	15	15	15	21
Zinc finger, MIZ type	3	4	3	1	6	4	4	2	4	6	5	4	4	3
Total	1427	2009	1957	1393	3939	1687	2752	1497	1559	3302	1681	2005	2026	2496

Table S7 (continued)

Note: Gene copy number may differ for some species and/or transcription factors from the Lang et al. study. These differences can be explained by the use of different genome releases. In addition, the use of INTERPRO domains instead of PFAM domains can also explain some deviations. It points out that the exact gene number for a TF for recently sequenced genomes is not stable and has to be cautiously considered with regards to the source of data. Nevertheless, the data remained globally consistent and permitted to compare their distribution across multiple genomes.

Table S8. Conserved miRNA families found in *Coffea canephora.*

microRNA family	number of paralogous loci	mature sequence
156	6	ahy-miR156b, aly-miR156g, aly-miR156f, vvi-miR156h
159	5	acb-miR159, ahy-miR159
160	3	aqc-miR160a, csi-miR160, ptc-miR160h
162	1	aly-miR162a
164	1	aly-miR164a
166	5	aly-miR166a
167	3	aly-miR167a, ath-miR167m
169	5	aly-miR169b, mtr-miR169c, mtr-miR169f
170	2	ace-miR170a, aly-miR170
171	7	ghr-miR171, aqc-miR171a, aly-miR171b, sbi-miR171f, vvi-miR171b
172	5	vvi-miR172b, vun-miR172, ppd-miR172b, gma-miR172c, aqc-miR172b
390	3	aly-miR390a
393	1	bdi-miR393a
394	1	ahy-miR394
395	9	aly-miR395d, bdi-miR395a
396	5	gcl-miR396a, aly-miR396b, aly-miR396a, ghr-miR396a
397	1	aly-miR397a
398	2	aly-miR398a, aqc-miR398b
399	5	aqc-miR399, aly-miR399a, aly-miR399d, bdi-miR399b, bdi-miR399b
408	1	ppt-miR408b
529	1	ppt-miR529d
miR1120	1	hpr-miR1120
miR2111a	1	aly-miR2111a
miRf10005-akr	1	ath-miRf10005-akr
miRf10239-akr	1	ath-miRf10239-akr
miRf10258-akr	1	ptc-miRf10258-akr
miRf10421-akr	1	ptc-miRf10421-akr
miRf10467-akr	1	ptc-miRf10467-akr
miRf10574-akr	3	ath-miRf10574-akr
miRf10982-akr	1	ptc-miRf10982-akr
miRf11151-akr	5	ptc-miRf11151-akr
miRf11215-akr	1	osa-miRf11215-akr
miRf12256-akr	3	ptc-miRf12256-akr

acb : Arabidopsis cebennensis; ace : Allium cepa ; Ahy: arachis hypogaea ; Aly : Arabidopsis lyrata ; aqc: Aquilegia coerulea ; ath: Arabidopsis thaliana ; bdi: Brachypodium distachyon ; csi : citrus sinensis ; ghr: Gossypium hirsutum ; gma: Glycine max ; hpr : Henrardia persica ; mtr: Medicago truncatula ; osa : oryza sativa ; ppd : . Populus trichocarpa x Populus deltoides ; ptc: Populus trichocarpa ; sbi: Sorghum bicolor ; vun : Vigna unguiculata ; vvi: Vitis vinifera

Table S9. The numbers of miRNA loci in *Coffea canephora* compared with those of some other plants with sequenced genomes.

			Number of paralogous loci		
microRNA family	**Common ancestor found in**	***Coffea canephora***	***Populus trichocarpa***	***Vitis vinifera***	***Arabidopsis thaliana***
156	Embryophytes	6	11	9	14
159	Embryophytes	5	6	3	3
160	Embryophytes	3	8	6	3
166	Embryophytes	5	17	8	7
171	Embryophytes	7	14	9	6
390	Embryophytes	3	4	1	2
395	Embryophytes	9	10	14	6
408	Embryophytes	1	1	1	1
396	Tracheophytes	5	7	4	2
397	Spermatophytes	1	3	2	2
398	Spermatophytes	2	3	3	3
162	Angiosperms	1	3	1	2
164	Angiosperms	1	6	4	3
167	Angiosperms	3	8	5	5
169	Angiosperms	5	32	25	15
170	Angiosperms	2	0	0	3
172	Angiosperms	5	9	4	7
393	Angiosperms	1	4	2	2
394	Angiosperms	1	2	3	2
399	Angiosperms	5	12	9	6
529	Angiosperms	1	0	0	0
miR1120	Angiosperms	1	0	0	0
miRf10005-akr	Angiosperms	1	1	0	1
miRf10239-akr	Angiosperms	1	1	0	1
miRf10258-akr	Angiosperms	1	1	0	1
miRf10421-akr	Angiosperms	1	1	0	1
miRf10467-akr	Angiosperms	1	1	0	1
miRf10574-akr	Angiosperms	3	1	0	1
miRf10982-akr	Angiosperms	1	1	0	1
miRf11151-akr	Angiosperms	5	1	0	1
miRf11215-akr	Angiosperms	1	1	0	0
miRf12256-akr	Angiosperms	3	1	0	0
miR2111a	Eudicots	1	0	0	1

Table S10. Percentage of identified Transposable Elements along *C. canephora* pseudomolecules. ChrUn is composed of unanchored contigs.

Pseudomolecule	RLG	RLC	RLX	RIX	RSX	DTX	DXX	DHX	XXX	Total %
chr1	17.37	6.13	9.58	1.82	0.08	2.70	1.19	0.78	2.06	41.71
chr2	14.16	5.14	8.07	1.35	0.08	3.07	0.95	0.41	2.01	35.24
chr3	18.42	6.44	8.70	1.73	0.06	2.95	1.32	1.12	2.07	42.81
chr4	16.04	6.51	9.07	1.65	0.08	2.83	1.06	0.54	1.96	39.74
chr5	19.03	7.02	9.48	2.10	0.07	2.74	1.23	0.71	1.98	44.36
chr6	15.58	5.72	8.90	1.54	0.08	3.06	0.99	0.58	1.91	38.36
chr7	15.41	5.36	8.67	1.40	0.08	2.91	1.06	0.85	2.21	37.95
chr8	16.26	6.17	10.10	1.93	0.08	3.04	1.17	0.71	2.12	41.58
chr9	20.56	6.53	10.24	2.11	0.08	2.68	1.69	0.44	1.87	46.2
chr10	16.27	6.20	8.74	1.79	0.07	3.14	1.32	0.49	1.93	39.95
chr11	19.08	7.10	10.88	2.02	0.07	2.69	1.02	0.63	1.87	45.36
ChrUn	37.38	8.15	14	1.73	0.04	1.57	0.2	0.49	1.09	64.65
Total %	24.14	6.84	10.92	1.73	0.07	2.43	0.80	0.59	1.68	**49.2**

Table S11. Examples of statistics on orthologs and paralogs retained for the comparisons after detection by SynMap. Percentages are out of total number of genes in genome.

genes in coffee	9964 (39.0%)	genes in grape	9169 (34.8%)
genes in tomato	12,540 (36.1%)	genes in tomato	11,478 (33.1%)
single-copy	7680	single-copy	7122
pairs	1992	pairs	1785
triples	292	triples	262

Table S12. Results of the BadiRate analysis of gene family expansion and contractions. The fit of the different branch models of gene turnover to the 16,917 orthogroups defined by OrthoMCL is shown. For each gene family, the table shows the fit of six branch models of gene turnover (GR, ForeGroundTomato, ForeGroundArabidopsis, ForeGroundGrape, ForeGroundArabidopsis, FR) to the number of genes at the extant species ("Family sizes" column), including their number of parameters, their likelihoods, and their weighted AICs (wAIC). For each family, the table also reports the best-fit branch model ("best-fit model" column), as well as the support of this best-fit model ("support of best fit model"). Best-fit model supports higher than 2.7 (according to the wAIC criterion) were considered significant (shaded cells). In cases where the best-fit model includes lineage-specific GD rates (i.e., the foreground branch models), the table also reports the inferred net family size change in the focal lineage ("Net Gene Number Change in ForeGround Lineage" column). **See separate excel file.**

Table S13. Plant GO slim terms differentially distributed among genes belonging to coffee-specific expanded and contracted families, according to the BadiRate analyses. For each GO term, the table shows the number of GO counts among the coffee-specific expanded genes families (‘sample counts’ column), the total number of GO counts in the genome (‘background’ column), the Fisher’s exact test of differential GO term distribution (raw and adjusted *p*-values, by Benjamini-Hochberg), as well as the direction of the differential representation in the sample (over- or under-represented). **See separate excel file.**

Table S14. Plant generic GO terms differentially distributed among genes belonging to coffee-specific expanded and contracted families, according to the BadiRate analyses. For each GO term, the table shows the number of GO counts among the coffee-specific expanded genes families (‘sample counts’ column), the total number of GO counts in the genome (‘background’ column), the Fisher’s exact test of differential GO term distribution (raw and adjusted p-values, by Benjamini-Hochberg), as well as the direction of the differential representation in the sample (over- or under-represented). **See separate excel file.**

Table S15: Summary of selected GO terms enriched among gene families expanded in coffee. In gray, those significant after Benjamini-Hochberg correction. B and M indicate biological process and molecular function GO class, respectively. Sample counts and Genome counts indicate the total number of genes annotated with particular GO terms among expanded families (of which there are 1270 genes total) versus the entire coffee genome (which contains a total of 25574 genes)

GO Class	GO Description	Sample counts	Genome counts	p-values
B	defense response	160	355	4.21E-127
M	catalytic activity	49	350	5.22E-21
M	transferase activity	38	303	2.69E-15
B	metabolic process	48	495	6.59E-15
M	methyltransferase activity	20	153	4.54E-09
M	hydrolase activity	22	303	9.00E-06
M	oxidoreductase activity	32	558	1.20E-05
B	secondary metabolic process	4	5	3.19E-05
M	sinapoyltransferase activity	4	5	3.19E-05
M	strictosidine synthase activity	6	24	5.42E-05
B	monoterpenoid biosynthetic process	4	9	0.000168234
M	(-)-menthol dehydrogenase activity	4	9	0.000168234
B	indole biosynthetic process	7	48	0.000260469
M	tropine dehydrogenase activity	4	11	0.000309558
M	naringenin 3-dioxygenase activity	3	9	0.002277388
B	terpenoid biosynthetic process	6	54	0.002542178
M	isoflavone 2'-hydroxylase activity	2	12	0.039907858
B	alkaloid biosynthetic process	4	54	0.044268512

Table S16. NBS-encoding *R*-genes in coffee, potato(*173*), tomato(*10*), poplar(*174*) and *Arabidopsis*(*100*).

Predicted domains	Coffee	Potato	Tomato	Grapevine	Poplar	*Arabidopsis*
Non-TIR genes						
CC-NBS-LRR	79	65	118	200	119	51
CC-NBS	239	24	17	26	19	4
NBS-LRR	63	177	39	12	71	4
NBS	169	104	51	0	27	1
Other NBS	7	0	4	186	41	2
TIR genes						
TIR-NBS-LRR	3	37	17	90	64	83
TIR-NBS	1	12	5	14	13	21
Other NBS	0	16	15	7	48	43
Total	**561**	**435**	**266**	**535**	**402**	**209**

Table S17. Normalized proportion (x100) of non-shared adjacencies, reflecting gene order divergences among core eudicot species, corrected for fractionation effects.
Yellow highlighting indicates that coffee is closest to asterids, while peach and cacao are closest to rosids.

	coffee	peach	cacao	grape	papaya	strawberry
Mimulus	18	22	23	26	35	33
Utricularia	35	38	38	40	47	43
tomato	23	27	29	30	39	36
coffee		23	23	27	32	45
grape	27	26	25		33	49
Arabidopsis	26	26	25	30	33	34
papaya	32	31	32	33		50
cacao	23	22		25	32	45
Medicago	28	28	31	33	39	33
soybean	27	25	27	30	35	35
strawberry	45	40	45	49	50	
peach	23		22	26	31	40
poplar	18	17	18	21	25	25

Table S18. Distribution of NBS-encoding *R*-genes and gene clusters in the coffee genome. A gene cluster is a region containing four or more genes within 200 Kb or less(*121*).

	Chromosomes													
Predicted domains	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Subtotal anchored genes**	**Non-anchored genes**	**Total**
Non-TIR genes														
CNL	2	4	14	6	5	0	2	7	0	0	0	40	29	69
CN	20	11	22	12	15	4	5	11	0	3	19	122	77	199
NL	4	2	5	1	5	4	2	9	0	0	3	35	29	64
N	10	6	27	4	10	9	7	14	2	5	23	117	63	180
Other	4	1	2	1	3	0	11	0	0	2	7	31	14	45
TIR genes														
TNL	0	0	0	0	0	0	0	1	0	0	0	1	0	1
TN	0	0	0	0	0	0	0	0	0	0	2	2	1	3
Total	**40**	**24**	**70**	**24**	**38**	**17**	**27**	**42**	**2**	**10**	**54**	**348**	**213**	**561**
Clusters	3	2	7	2	3	2	2	5	0	0	4	30	10	40
No. genes	20	9	39	10	18	10	9	36	0	0	18	169	48	217

Table S19. Distribution of orthogroups containing NBS genes in the coffee genome.

	Chromosomes													
Orthogroup	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Subtotal anchored genes**	**Non-anchored genes**	**Total**
ORTHOMCL5		12		2				6				20	15	35
ORTHOMCL6								1			1	2	4	6
ORTHOMCL7	27	1	21				1					50	75	125
ORTHOMCL15	7	2	10	2	10		18			2	2	53	25	78
ORTHOMCL19		1		1		2		29			3	36	23	59
ORTHOMCL23			1			5						6	7	13
ORTHOMCL42				16		1					1	18	8	26
ORTHOMCL43	1			1	6							8	15	23
ORTHOMCL54	2		11			1		1		1	16	32	9	41
ORTHOMCL67			6			4					6	16	1	17
ORTHOMCL138			1		1					4	7	13	1	14
ORTHOMCL142			5		11		1	1				18	2	20
ORTHOMCL151		1										1	0	1
ORTHOMCL256	2				2							4	11	15
ORTHOMCL398						1						1	0	1
ORTHOMCL456							1					1	0	1
ORTHOMCL473					1			1				2	2	4
ORTHOMCL648			1								4	5	0	5
ORTHOMCL1009									1			1	2	3
ORTHOMCL1016					1		3					4	0	4
ORTHOMCL1063											1	1	0	1
ORTHOMCL1357		2										2	0	2
ORTHOMCL1497										1		1	0	1
ORTHOMCL2767									1		3	4	1	5
ORTHOMCL4266					3							3	0	3
ORTHOMCL4371						1						1	0	1
ORTHOMCL4795							1				4	5	0	5

ORTHOMCL4821			4		1							5	0	5
ORTHOMCL5239			2									2	0	2
ORTHOMCL5797												0	1	1
ORTHOMCL5830		1										1	0	1
ORTHOMCL5841											1	1	0	1
ORTHOMCL9205					1							1	0	1
ORTHOMCL9700							1					1	0	1
ORTHOMCL12442						1						1	0	1
ORTHOMCL15067											1	1	1	2
ORTHOMCL15152											2	2	0	2
ORTHOMCL15177			1									1	0	1
ORTHOMCL15218			2									2	0	2
ORTHOMCL16510				1								1	0	1
Total	**39**	**20**	**65**	**23**	**37**	**16**	**26**	**39**	**2**	**8**	**52**	**327**	**203**	**530**

Table S20. List of *C. canephora* proteins orthologous to known plant proteins involved in the initial phenylpropanoid pathway (PPP) and hydroxycinnamic acid ester (HCE) metabolism.

Gene	Query Enzyme *(Accession number)*	Gene Copy Numbers				*Coffea canephora* Locus Numbers				
		Poplar*	Arabidopsis**	Tomato**	Coffee**					
PAL	PtrPAL1 *(ACC63888)*	5	4	16	4	Cc00p20750.1	Cc01p10880.1	Cc02p11230.1	Cc06p03980.1	
C4H	PtrC4H1 *(ACC63873)*	3	1	1	4	Cc00p10050.1	Cc00p00170.1	Cc01p15360.1	Cc00p05190.1	
4CL	Ptr4CL1 *(EEF03042)*	17	13	14	10	Cc01p10510.1	Cc02p02860.1	Cc04p05510.1	Cc04p09970.1	Cc06p04280.1
						Cc06p13370.1	Cc07p03940.1	Cc08p03990.1	Cc10p00840.1	Cc02p10610.1
HCT	CcHCT *(ABO47805)*	7	1	1	1	Cc04p05230.1				
HQT	SlHQT *(CAE46933)*	0	0	1	1	Cc06p14760.1				
C3'H	PtrC3H3 *(ACC63870)*	4	1	2	2	Cc00p08150.1	Cc06p20390.1			
CCoAOMT	CcCCoAOMT1 *(ABO77959)*	6	2	28	3	Cc02p18970.1	Cc06p11010.1	Cc11p06680.1		

* For poplar (*Populus trichocarpa*) (Pt), the gene copy numbers reported in the table are those published by Shi et al. in 2006(*152*).

** Gene copy numbers were determined using the BLASTP program(*65, 175*) and full-length protein sequences as queries. One example of query used for each family is listed in the second column, and corresponds to a published sequence from an enzyme genetically or biochemically characterized. In addition, for each family, other query sequences (not listed) were used to confirm the results returned using the listed query enzymes. Then, in order to validate the annotation of the candidates, the hits with the best E values (the BLAST searches were launched with e-value set at 2e-10) and percentages of identity were first aligned with known enzymes, and then subjected to the "Batch Web CD-Search Tool", an NCBI's interface to searching the Conserved Domain Database(*148-150*) with an expect value set at 0.01 (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi). The alignments were performed using ClustalW through MegAlign software (Lasergene, DNASTAR version 9.1.0). For *Arabidopsis thaliana* (At) and tomato, the *Arabidopsis* proteins (TAIR 10 release) database and the tomato genome "ITAG Release 2.3 predicted proteins (SL2.40)", were queried from Sol Genomics Network (http://solgenomics.net/tools/BLAST), respectively. For coffee, the Coffee Genome Hub Assembly V1.0 database was queried. For each enzyme family in coffee, the gene copy numbers found and their corresponding locus numbers (locus ID) are indicated.

Table S21. Details for all 25 candidate genes involved in the PPP initial reactions and HCE synthesis identified in the coffee genome (Table S20). The details include: the Gene Family they belong to, their General Function, Their Locus ID in the Coffee Genome Hub Assembly V1.0 database, the names attributed, the locations of the genes in the *C. canephora* Double Haploid (CcDH) genome (Chromosome number, Start and End locations, and Scaffold number), their Unique Name/Function, and the name of the corresponding enzyme in either the *C. canephora* diploid genome or *C. arabica* tetraploid genome for the genes that were characterized prior to the Coffee Genome Sequencing Project.

Gene family	General function	Locus ID	Gene Name	Chr.	Start	End	Scaffold Number	Unique Name/ Function	Correspond to an isoenzyme of
Phenylalanine ammonia-lyase (PAL)	Phenylpropanoid Metabolism/ Initial Reactions	Cc02p11230.1	CcDH_PAL1	chr2	9150460	9154515	Scaffold 4	GSCOCT00029438001~	CcPAL1 GenBank Accession AAN32866
		Cc06p03980.1	CcDH_PAL2	chr6	3129379	3131052	Scaffold 22	GSCOCT00023699001~	CcPAL2 GenBank Accession AEO94540
		Cc01p10880.1	CcDH_PAL3	chr1	29560890	29564593	Scaffold 24	GSCOCT00024202001~	CcPAL3 GenBank Accession AEO94541
		Cc00p20750.1	CcDH_PAL4	chr0	132648726	132651946	Scaffold 1586	GSCOCT00001383001~	
Trans-cinnamate 4-monooxygenase (C4H)	Phenylpropanoid Metabolism/ Initial Reactions	Cc01p15360.1	CcDH_C4H1	chr1	33302670	33305477	Scaffold 11	GSCOCT00016571001~	CcC4H SEQ ID NO:23 Patent WO/2007/044992
		Cc00p10050.1	CcDH_C4H2	chr0	84613032	84621393	Scaffold 511	GSCOCT00003814001~	
		Cc00p05190.1	CcDH_C4H3	chr0	40646841	40647707	Scaffold 257	GSCOCT00000319001~	
		Cc00p00170.1	CcDH_C4H4	chr0	314366	316236	Scaffold 67	GSCOCT00036093001~	
4-coumarate-CoA ligase (4CL)	Phenylpropanoid Metabolism/ Initial Reactions	Cc06p13370.1	CcDH_4CL1	chr6	11098526	11101894	Scaffold 9	GSCOCT00041408001~	Cc4CL1 SEQ ID NO:25 Patent WO/2007/044992
		Cc04p05510.1	CcDH_4CL2	chr4	4111158	4120611	Scaffold 2	GSCOCT00022309001~	Cc4CL2 SEQ ID NO:26 Patent WO/2007/044992
		Cc10p00840.1	CcDH_4CL-L1	chr10	688109	694486	Scaffold 26	GSCOCT00024543001~	
		Cc04p09970.1	CcDH_4CL-L2	chr4	9135366	9152764	Scaffold 14	GSCOCT00042531001~	
		Cc08p03990.1	CcDH_4CL-L3	chr8	5495865	5499040	Scaffold 59	GSCOCT00034199001~	
		Cc02p10610.1	CcDH_4CL-L4	chr2	8656646	8658672	Scaffold 4	GSCOCT00029510001~	
		Cc01p10510.1	CcDH_4CL-L5	chr1	29144630	29147854	Scaffold 24	GSCOCT00024147001~	
		Cc06p04280.1	CcDH_4CL-L6	chr6	3373703	3376971	Scaffold 22	GSCOCT00023742001~	
		Cc07p03940.1	CcDH_4CL-L7	chr7	2736366	2738614	Scaffold 8	GSCOCT00039644001~	
		Cc02p02860.1	CcDH_4CL-L8	chr2	2323248	2326781	Scaffold 15	GSCOCT00020155001~	
Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT)	Phenylpropanoid Metabolism/ Chlorogenic Acids* Metabolism	Cc04p05230.1	CcDH_HCT	chr4	3908595	3911935	Scaffold 2	GSCOCT00022270001~ Hydroxycinnamoyl-	CcHCT GenBank Accession ABO47805
Hydroxycinnamoyl-CoA quinate hydroxycinnamoyl transferase (HQT)	Phenylpropanoid Metabolism/ Chlorogenic Acids* Metabolism	Cc06p14760.1	CcDH_HQT	chr6	12651706	12655746	Scaffold 9	GSCOCT00041214001~ Hydroxycinnamoyl-	CcHQT GenBank Accession ABO77957
Monooxygenase/p-coumarate 3-hydroxylase (C3'H)	Phenylpropanoid Metabolism/ Chlorogenic Acids* Metabolism	Cc06p20390.1	CcDH_C3'H1	chr6	24122683	24126973	Scaffold 20	GSCOCT00022969001~	CcC3'H1 GenBank Accession ABB83676
		Cc00p08150.1	CcDH_C3'H2	chr0	68519309	68524322	Scaffold 385	GSCOCT00000289001~	CcC3'H2 GenBank Accession ABB83677
Caffeoyl-CoA O-Methyltransferase (CCoAOMT)	Phenylpropanoid Metabolism/ Chlorogenic Acids* Metabolism/ Cell Wall Biosynthesis (Lignin)	Cc02p18970.1	CcDH_CCoAOMT1	chr2	17197993	17199471	Scaffold 1	GSCOCT00014459001~	CcCCoAOMT1 GenBank Accession ABO77959
		Cc06p11010.1	CcDH_CCoAOMT-L1	chr6	8925295	8926763	Scaffold 9	GSCOCT00041707001~	CaCCoAOMT-L1 SEQ ID NO:15 Patent
		Cc11p06680.1	CcDH_CCoAOMT-L2	chr11	23108496	23112473	Scaffold 16	GSCOCT00020692001~	CcCCoAOMT-L2 SEQ ID NO:16 Patent

Table S22. BAHD proteins involved in the acylation of plant secondary metabolites* to produce compounds such as volatile esters and chlorogenic acid isomers.

Gene	Query Enzymes *(Accession numbers)*	Gene Copy Numbers				*Coffea canephora* Locus Numbers					
		Poplar	Arabidopsis	Tomato	Coffee						
HQT (BAHD)	SlHQT (CAE46933) CcHQT (ABO77957)	0	0	1	1	Cc06p14760.1					
HCT (BAHD)		7	1	1	1	Cc04p05230.1					
Other BAHDs		65	29	40	23	Cc01p00040.1	Cc01p00380.1	Cc06p19050.1	Cc01p05160.1	Cc01p03700.1	Cc01p02660.1
						Cc04p09590.1	Cc06p09250.1	Cc05p02710.1	Cc02p06310.1	Cc03p10220.1	Cc03p10200.1
						Cc01p01700.1	Cc08p07440.1	Cc04p10200.1	Cc02p31430.1	Cc11p16420.1	Cc00p08090.1
						Cc00p27820.1	Cc04p10250.1	Cc05p05980.1	Cc02p06870.1	Cc06p01600.1	

The genome-wide searches for BAHD enzymes in four plants (poplar, *Arabidopsis*, tomato and coffee) were conducted in two steps. First, a BLASTP 2.2 was performed using SlHQT and CcHQT as queries with an E-value set at 2e-10, and Blosum62 as the matrix. For poplar, the poplar genome was queried through the Plant GDB Website at www.plantgdb.org, by choosing PtPep_Populus predicted proteins (v2.0) as the database. For *Arabidopsis* (TAIR10 release) and tomato (ITAG Release 2.3; SL2.40), the BLAST searches were launched using the BLASTP program hosted at Sol Genomics Network (http://solgenomics.net/tools/BLAST). For coffee, the Coffee Genome Hub Assembly V1.0 database was queried. Then, for each search, all hits returned using these parameters were analysed individually, and only those containing both the "**HxxxD**" and "**DFGWG**" highly conserved domains, found in nearly all the functionally characterized enzymes(*159*), were considered and counted. In the case of the coffee genome, 57 hits were initially returned by the BLAST search. After analysis for the presence of both the "**HxxxD**" and "**DFGWG**" domains, 25 BAHDs, including the *HQT* and *HCT* genes, were kept for further analysis. These 25 BAHDs are not only candidate enzymes involved in the metabolism of hydroxycinnamic acids and esters (including chlorogenic acid isomers, candidate molecules for human health and coffee quality), but also in the biosynthesis of compounds such as volatile esters, which play an important role in coffee quality and flavor, or in the modification of compounds such as anthocyanins. The 25 proteins were then aligned with genetically or biochemically characterized BAHD acyltransferases (including some of those used by D'Auria) using ClustalW through the MegAlign software (Lasergene, DNASTAR version 9.1.0) (**Figure S16**). Then, a phylogenetic tree was built based on this alignment using the same software (bootstrapping parameters were as follows: Number of trials: 1000; Random seed: 111) (**Figure S17)**. As can be shown by the phylogenetic tree, and referring to the publication of D'Auria et al (*159*), the phylogenetic distances established and the grouping of the members by clade can help to predict the function of the BAHD enzymes found in coffee genome. For example, sequence Cc06p19050.1 is grouped with characterized members from clade V-I, a subclade containing enzymes capable of forming the volatile benzenoid ester benzylbenzoate, or responsible for making specific odour in the related plants. A second example is sequence Cc02p06310.1, which clusters within a major subgroup of clade III consisting of acetyltransferases involved in volatile ester biosynthesis in flowers and ripening fruits.

*: Acylation of plant secondary metabolites including those generated through the phenylpropanoid pathway to produce products such as small volatile esters (but also modified anthocyanins, as well as constitutive defense compounds and phytoalexins).

Table S23. Detailed description of the 25 candidate BAHDs identified in the coffee genome. Details include the genome location of all the genes (Locus ID, Chromosome Number, Start and End locations, Scaffold Number) and their function attributed by automated genome annotation.

Locus ID	Chr.	Start	End	Scaffold number	Function
Cc00p08090.1	chr0	68112243	68113574	Scaffold 383	GSCOCT00006943001~ Putative Deacetylvindoline O-acetyltransferase~ DAT~ complete
Cc00p27820.1	chr0	174426642	174427985	Scaffold 4202	GSCOCT00006800001~ Putative Vinorine synthase~ ACT~ complete
Cc01p00040.1	chr1	37175	39395	Scaffold 256	GSCOCT00012436001~ Benzyl alcohol O-benzoyltransferase~ HSR201~ complete
Cc01p00380.1	chr1	397146	399824	Scaffold 3887	GSCOCT00013680001~ Benzyl alcohol O-benzoyltransferase~ HSR201~ complete
Cc01p01700.1	chr1	2731546	2732883	Scaffold 73	GSCOCT00037859001~ Putative Salutaridinol 7-O-acetyltransferase~ SALAT~ complete
Cc01p02660.1	chr1	5127048	5129138	Scaffold 172	GSCOCT00013192001~ Putative Taxadien-5-alpha-ol O-acetyltransferase~ TAT~ complete
Cc01p03700.1	chr1	8933561	8936458	Scaffold 51	GSCOCT00033062001~ Putative Benzyl alcohol O-benzoyltransferase~ HSR201~ complete
Cc01p05160.1	chr1	19533095	19535391	Scaffold 166	GSCOCT00011474001~ Putative Benzyl alcohol O-benzoyltransferase~ HSR201~ complete
Cc02p06310.1	chr2	4978892	4980235	Scaffold 78	GSCOCT00039158001~ Putative Vinorine synthase~ ACT~ complete
Cc02p06870.1	chr2	5431740	5433155	Scaffold 78	GSCOCT00039240001~ Putative BAHD acyltransferase DCR~ DCR~ complete
Cc02p31430.1	chr2	40272906	40274201	Scaffold 34	GSCOCT00027565001~ Putative Vinorine synthase~ ACT~ complete
Cc03p10200.1	chr3	12885358	12886695	Scaffold 44	GSCOCT00031025001~ Putative Vinorine synthase~ ACT~ complete
Cc03p10220.1	chr3	12937641	12938972	Scaffold 44	GSCOCT00031029001~ Putative Vinorine synthase~ ACT~ complete
Cc04p05230.1	chr4	3908595	3911935	Scaffold 2	GSCOCT00022270001~ Hydroxycinnamoyl-Coenzyme A shikimate/quinate hydroxycinnamoyltransferase~ HCT~ complete
Cc04p09590.1	chr4	8153822	8155482	Scaffold 2	GSCOCT00022851001~ Omega-hydroxypalmitate O-feruloyl transferase~ HHT1~ complete
Cc04p10200.1	chr4	9412879	9414177	Scaffold 94	GSCOCT00042562001~ Putative Vinorine synthase~ ACT~ complete
Cc04p10250.1	chr4	9504032	9505330	Scaffold 94	GSCOCT00042568001~ Putative Vinorine synthase~ ACT~ complete
Cc05p02710.1	chr5	11712999	11714381	Scaffold 48	GSCOCT00020896001~ Putative Hydroxycinnamoyl-Coenzyme A shikimate/quinate hydroxycinnamoyltransferase~ HCT~ complete
Cc05p05980.1	chr5	20657717	20659153	Scaffold 60	GSCOCT00035195001~ Putative Malonyl-coenzyme A:anthocyanin 3-O-glucoside-6"-O-malonyltransferase~ 3MAT~ complete
Cc06p01600.1	chr6	1306113	1307504	Scaffold 22	GSCOCT00023385001~ Putative BAHD acyltransferase DCR~ DCR~ complete
Cc06p09250.1	chr6	7454894	7456300	Scaffold 9	GSCOCT00041938001~ Putative Omega-hydroxypalmitate O-feruloyl transferase~ HHT1~ complete
Cc06p14760.1	chr6	12651706	12655746	Scaffold 9	GSCOCT00041214001~ Hydroxycinnamoyl-Coenzyme A shikimate/quinate hydroxycinnamoyltransferase~ HCT~ complete
Cc06p19050.1	chr6	20211753	20215397	Scaffold 39	GSCOCT00028954001~ Benzyl alcohol O-benzoyltransferase~ HSR201~ complete
Cc08p07440.1	chr8	19736745	19738106	Scaffold 28	GSCOCT00025372001~ Putative Vinorine synthase~ ACT~ complete
Cc11p16420.1	chr11	32580671	32582002	Scaffold 74	GSCOCT00038181001~ Putative Salutaridinol 7-O-acetyltransferase~ SALAT~ complete

Table S24. Known caffeine-biosynthetic and related N-methyltransferase (NMT) genes in *C. canephora*

						Sizes								
Gene ID	**Gene model (alias)**	**Gene Name**	**Chr**	**Start**	**Stop**	**Gene (bp)**	**Protein (aa)**	**Exon 1**	**Exon 2**	**Exon 3**	**Exon 4**	**Intron 1 length/ phase**	**Intron 2 length/ phase**	**Intron 3 length/ phase**
Cc01g00720	**GSCOCT00011062001**	***CcDXMT***	Chr1	1 210 494	1 212 499	2 006	386	75	417	261	402	146/0	257/0	448/0
Cc09g06950	**GSCOCT00036343001**	***CcMTL***	Chr9	8 151 260	8 153 271	2 012	386	75	420	261	402	144/0	170/0	540/0
Cc09g06960	**GSCOCT00036341001**	***CcNMT3***	Chr9	8 174 351	8 176 414	2 064	387	75	420	261	402	136/0	255/0	515/0
Cc09g06970	**GSCOCT00036338001**	***CcXMT***	Chr9	8 259 623	8 261 616	1 994	374	75	420	261	363	136/0	255/0	484/0
Cc00g24720	**GSCOCT00004719001**	***CcMXMT***	ChrUn	155 539 857	155 541 685	1 829	380	75	417	261	384	123/0	259/0	322/0

Table S25. Summary of examined NMT sequences from coffee, cacao and tea species, with expression profiles. Background colors indicate the three main synteny groups among coffee NMTs are indicated with purple, green and blue symbols. See separate excel file.

Table S26. Results of the asymmetric evolution, divergent selection and positive selection tests. See separate excel file.

Table S27. Description of the positively selected aminoacids. See separate excel file.

References and Notes

1. E. Robbrecht, J. F. Manen, 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 (2006).

2. P. Lashermes, M. C. Combes, J. Robert, P. Trouslot, A. D'Hont, F. Anthony, A. Charrier, Molecular characterisation and origin of the *Coffea arabica* L. genome. *Mol. Gen. Genet.* **261**, 259–266 (1999). Medline doi:10.1007/s004380050965

3. M. Noirot, V. Poncet, P. Barre, P. Hamon, S. Hamon, A. de Kochko, Genome size variations in diploid African *Coffea* species. *Ann. Bot. (London)* **92**, 709–714 (2003). Medline doi:10.1093/aob/mcg183

4. Materials and methods are available as supplementary materials on *Science* Online.

5. S. Schaack, C. Gilbert, C. Feschotte, Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* **25**, 537–546 (2010). Medline doi:10.1016/j.tree.2010.06.001

6. A. Roulin, B. Piegu, P. M. Fortune, F. Sabot, A. D'Hont, D. Manicacci, O. Panaud, Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae. *BMC Evol. Biol.* **9**, 58 (2009). Medline doi:10.1186/1471-2148-9-58

7. M. El Baidouri, M. C. Carpentier, R. Cooke, D. Gao, E. Lasserre, C. Llauro, M. Mirouze, N. Picault, S. A. Jackson, O. Panaud, Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Res.* **24**, 831–838 (2014). Medline doi:10.1101/gr.164400.113

8. C. Moisy, A. H. Schulman, R. Kalendar, J. P. Buchmann, F. Pelsy, The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. *Theor. Appl. Genet.* **127**, 1223–1235 (2014). Medline doi:10.1007/s00122-014-2293-z

9. O. Jaillon, J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pè, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quétier, P. Wincker; French-Italian Public Consortium for Grapevine Genome Characterization, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007). Medline doi:10.1038/nature06148

10. S. Sato, S. Tabata, H. Hirakawa, E. Asamizu, K. Shirasawa, S. Isobe, T. Kaneko, Y. Nakamura, D. Shibata, K. Aoki, M. Egholm, J. Knight, R. Bogden, C. Li, Y. Shuang, X. Xu, S. Pan, S. Cheng, X. Liu, Y. Ren, J. Wang, A. Albiero, F. Dal Pero, S. Todesco, J. Van Eck, R. M. Buels, A. Bombarely, J. R. Gosselin, M. Huang, J. A. Leto, N. Menda, S.

Strickler, L. Mao, S. Gao, I. Y. Tecle, T. York, Y. Zheng, J. T. Vrebalov, J. M. Lee, S. Zhong, L. A. Mueller, W. J. Stiekema, P. Ribeca, T. Alioto, W. Yang, S. Huang, Y. Du, Z. Zhang, J. Gao, Y. Guo, X. Wang, Y. Li, J. He, C. Li, Z. Cheng, J. Zuo, J. Ren, J. Zhao, L. Yan, H. Jiang, B. Wang, H. Li, Z. Li, F. Fu, B. Chen, B. Han, Q. Feng, D. Fan, Y. Wang, H. Ling, Y. Xue, D. Ware, W. Richard McCombie, Z. B. Lippman, J.-M. Chia, K. Jiang, S. Pasternak, L. Gelley, M. Kramer, L. K. Anderson, S.-B. Chang, S. M. Royer, L. A. Shearer, S. M. Stack, J. K. C. Rose, Y. Xu, N. Eannetta, A. J. Matas, R. McQuinn, S. D. Tanksley, F. Camara, R. Guigó, S. Rombauts, J. Fawcett, Y. Van de Peer, D. Zamir, C. Liang, M. Spannagl, H. Gundlach, R. Bruggmann, K. Mayer, Z. Jia, J. Zhang, Z. Ye, G. J. Bishop, S. Butcher, R. Lopez-Cobollo, D. Buchan, I. Filippis, J. Abbott, R. Dixit, M. Singh, A. Singh, J. Kumar Pal, A. Pandit, P. Kumar Singh, A. Kumar Mahato, V. Dogra, K. Gaikwad, T. Raj Sharma, T. Mohapatra, N. Kumar Singh, M. Causse, C. Rothan, T. Schiex, C. Noirot, A. Bellec, C. Klopp, C. Delalande, H. Berges, J. Mariette, P. Frasse, S. Vautrin, M. Zouine, A. Latché, C. Rousseau, F. Regad, J.-C. Pech, M. Philippot, M. Bouzayen, P. Pericard, S. Osorio, A. Fernandez del Carmen, A. Monforte, A. Granell, R. Fernandez-Muñoz, M. Conte, G. Lichtenstein, F. Carrari, G. De Bellis, F. Fuligni, C. Peano, S. Grandillo, P. Termolino, M. Pietrella, E. Fantini, G. Falcone, A. Fiore, G. Giuliano, L. Lopez, P. Facella, G. Perrotta, L. Daddiego, G. Bryan, M. Orozco, X. Pastor, D. Torrents, M. G. M. van Schriek, R. M. C. Feron, J. van Oeveren, P. de Heer, L. daPonte, S. Jacobs-Oomen, M. Cariaso, M. Prins, M. J. T. van Eijk, A. Janssen, M. J. J. van Haaren, S.-H. Jo, J. Kim, S.-Y. Kwon, S. Kim, D.-H. Koo, S. Lee, C.-G. Hur, C. Clouser, A. Rico, A. Hallab, C. Gebhardt, K. Klee, A. Jöcker, J. Warfsmann, U. Göbel, S. Kawamura, K. Yano, J. D. Sherman, H. Fukuoka, S. Negoro, S. Bhutty, P. Chowdhury, D. Chattopadhyay, E. Datema, S. Smit, E. G. W. M. Schijlen, J. van de Belt, J. C. van Haarst, S. A. Peters, M. J. van Staveren, M. H. C. Henkens, P. J. W. Mooyman, T. Hesselink, R. C. H. J. van Ham, G. Jiang, M. Droege, D. Choi, B.-C. Kang, B. Dong Kim, M. Park, S. Kim, S.-I. Yeom, Y.-H. Lee, Y.-D. Choi, G. Li, J. Gao, Y. Liu, S. Huang, V. Fernandez-Pedrosa, C. Collado, S. Zuñiga, G. Wang, R. Cade, R. A. Dietrich, J. Rogers, S. Knapp, Z. Fei, R. A. White, T. W. Thannhauser, J. J. Giovannoni, M. Angel Botella, L. Gilbert, R. Gonzalez, J. Luis Goicoechea, Y. Yu, D. Kudrna, K. Collura, M. Wissotski, R. Wing, H. Schoof, B. C. Meyers, A. Bala Gurazada, P. J. Green, S. Mathur, S. Vyas, A. U. Solanke, R. Kumar, V. Gupta, A. K. Sharma, P. Khurana, J. P. Khurana, A. K. Tyagi, T. Dalmay, I. Mohorianu, B. Walts, S. Chamala, W. Brad Barbazuk, J. Li, H. Guo, T.-H. Lee, Y. Wang, D. Zhang, A. H. Paterson, X. Wang, H. Tang, A. Barone, M. Luisa Chiusano, M. Raffaella Ercolano, N. D'Agostino, M. Di Filippo, A. Traini, W. Sanseverino, L. Frusciante, G. B. Seymour, M. Elharam, Y. Fu, A. Hua, S. Kenton, J. Lewis, S. Lin, F. Najar, H. Lai, B. Qin, C. Qu, R. Shi, D. White, J. White, Y. Xing, K. Yang, J. Yi, Z. Yao, L. Zhou, B. A. Roe, A. Vezzi, M. D'Angelo, R. Zimbello, R. Schiavon, E. Caniato, C. Rigobello, D. Campagna, N. Vitulo, G. Valle, D. R. Nelson, E. De Paoli, D. Szinay, H. H. de Jong, Y. Bai, R. G. F. Visser, R. M. Klein Lankhorst, H. Beasley, K. McLaren, C. Nicholson, C. Riddle, G. Gianese, S. Sato, S. Tabata, L. A. Mueller, S. Huang, Y. Du, C. Li, Z. Cheng, J. Zuo, B. Han, Y. Wang, H. Ling, Y. Xue, D. Ware, W. Richard McCombie, Z. B. Lippman, S. M. Stack, S. D. Tanksley, Y. Van de Peer, K. Mayer, G. J. Bishop, S. Butcher, N. Kumar Singh, T. Schiex, M. Bouzayen, A. Granell, F. Carrari, G. De Bellis, G. Giuliano, G. Bryan, M. J. T. van Eijk, H. Fukuoka, D. Chattopadhyay, R. C. H. J. van Ham, D. Choi, J. Rogers, Z. Fei, J. J. Giovannoni, R.

Wing, H. Schoof, B. C. Meyers, J. P. Khurana, A. K. Tyagi, T. Dalmay, A. H. Paterson, X. Wang, L. Frusciante, G. B. Seymour, B. A. Roe, G. Valle, H. H. de Jong, R. M. Klein Lankhorst; Tomato Genome Consortium, The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012). Medline doi:10.1038/nature11119

11. P. Librado, F. G. Vieira, J. Rozas, BadiRate: Estimating family turnover rates by likelihood-based methods. *Bioinformatics* **28**, 279–281 (2012). Medline doi:10.1093/bioinformatics/btr623

12. S. H. Hulbert, C. A. Webb, S. M. Smith, Q. Sun, Resistance gene complexes: Evolution and utilization. *Annu. Rev. Phytopathol.* **39**, 285–312 (2001). Medline doi:10.1146/annurev.phyto.39.1.285

13. L. McHale, X. Tan, P. Koehl, R. W. Michelmore, Plant NBS-LRR proteins: Adaptable guards. *Genome Biol.* **7**, 212 (2006). Medline doi:10.1186/gb-2006-7-4-212

14. F. Gleason, R. Chollet, *Plant Biochemistry* (Jones and Bartlett, Sudbury, MA, 2011).

15. J. A. Nathanson, Caffeine and related methylxanthines: Possible naturally occurring pesticides. *Science* **226**, 184–187 (1984). Medline doi:10.1126/science.6207592

16. A. Pacheco, J. Pohlan, M. Schulz, Allelopathic effects of aromatic species intercropped with coffee: Investigation of their growth stimulation capacity and potential of caffeine uptake in Puebla, Mexico. *Allelopathy J.* **21**, 39–56 (2008).

17. H. Ashihara, H. Sano, A. Crozier, Caffeine and related purine alkaloids: Biosynthesis, catabolism, function and genetic engineering. *Phytochemistry* **69**, 841–856 (2008). Medline doi:10.1016/j.phytochem.2007.10.029

18. A. A. McCarthy, J. G. McCarthy, The structure of two *N*-methyltransferases from the caffeine biosynthetic pathway. *Plant Physiol.* **144**, 879–889 (2007). Medline doi:10.1104/pp.106.094854

19. M. Ogawa, Y. Herai, N. Koizumi, T. Kusano, H. Sano, 7-Methylxanthine methyltransferase of coffee plants. Gene isolation and enzymatic properties. *J. Biol. Chem.* **276**, 8213–8218 (2001). Medline doi:10.1074/jbc.M009480200

20. E. Pichersky, E. Lewinsohn, Convergent evolution in plant specialized metabolism. *Annu. Rev. Plant Biol.* **62**, 549–566 (2011). Medline doi:10.1146/annurev-arplant-042110-103814

21. B. Field, A. E. Osbourn, Metabolic diversification—Independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543–547 (2008). Medline doi:10.1126/science.1154990

22. M. Matsuno, V. Compagnon, G. A. Schoch, M. Schmitt, D. Debayle, J. E. Bassard, B. Pollet, A. Hehn, D. Heintz, P. Ullmann, C. Lapierre, F. Bernier, J. Ehlting, D. Werck-Reichhart, Evolution of a novel phenolic pathway for pollen development. *Science* **325**, 1688–1692 (2009). Medline doi:10.1126/science.1174095

23. B. Field, A. S. Fiston-Lavier, A. Kemen, K. Geisler, H. Quesneville, A. E. Osbourn, Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 16116–16121 (2011). Medline doi:10.1073/pnas.1109273108

24. J. Zhang, R. Nielsen, Z. Yang, Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005). Medline doi:10.1093/molbev/msi237

25. D. Villarreal, A. Laffargue, H. Posada, B. Bertrand, P. Lashermes, S. Dussert, Genotypic and environmental effects on coffee (*Coffea arabica* L.) bean fatty acid profile: Impact on variety and origin chemometric determination. *J. Agric. Food Chem.* **57**, 11321–11327 (2009). Medline doi:10.1021/jf902441n

26. S. Dussert, A. Laffargue, A. de Kochko, T. Joët, Effectiveness of the fatty acid and sterol composition of seeds for the chemotaxonomy of *Coffea* subgenus *Coffea*. *Phytochemistry* **69**, 2950–2960 (2008). Medline doi:10.1016/j.phytochem.2008.09.021

27. T. Joët, A. Laffargue, J. Salmona, S. Doulbeau, F. Descroix, B. Bertrand, A. de Kochko, S. Dussert, Metabolic pathways in tropical dicotyledonous albuminous seeds: *Coffea arabica* as a case study. *New Phytol.* **182**, 146–162 (2009). Medline doi:10.1111/j.1469-8137.2008.02742.x

28. O. Maurin, A. P. Davis, M. Chester, E. F. Mvungi, Y. Jaufeerally-Fakim, M. F. Fay, Towards a phylogeny for *Coffea* (Rubiaceae): Identifying well-supported lineages based on nuclear and plastid DNA sequences. *Ann. Bot. (London)* **100**, 1565–1583 (2007). Medline doi:10.1093/aob/mcm257

29. P. Lashermes, E. Couturon, A. Charrier, Doubled haploids of *Coffea canephora*: Development, fertility and agronomic characteristics. *Euphytica* **74**, 149–157 (1994). doi:10.1007/BF00033781

30. C. O. Agwanda, P. Lashermes, P. Trouslot, M.-C. Combes, A. Charrier, Identification of RAPD markers for resistance to coffee berry disease, *Colletotrichum kahawae*, in arabica coffee. *Euphytica* **97**, 241–248 (1997). doi:10.1023/A:1003097913349

31. G. Carrier, S. Santoni, M. Rodier-Goud, A. Canaguier, A. de Kochko, C. Dubreuil-Tranchant, P. This, J.-M. Boursiquot, L. Le Cunff, An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *Am. J. Bot.* **98**, e13–e15 (2011). Medline doi:10.3732/ajb.1000371

32. A. Dereeper, R. Guyot, C. Tranchant-Dubreuil, F. Anthony, X. Argout, F. de Bellis, M. C. Combes, F. Gavory, A. de Kochko, D. Kudrna, T. Leroy, J. Poulain, M. Rondeau, X. Song, R. Wing, P. Lashermes, BAC-end sequences analysis provides first insights into coffee (*Coffea canephora* P.) genome composition and evolution. *Plant Mol. Biol.* **83**, 177–189 (2013). Medline doi:10.1007/s11103-013-0077-5

33. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). Medline doi:10.1093/bioinformatics/btp324

34. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010). Medline doi:10.1186/gb-2010-11-10-r106

35. W. J. Kent, BLAT—The BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). Medline doi:10.1101/gr.229202. Article published online before March 2002

36. J. M. Aury, C. Cruaud, V. Barbe, O. Rogier, S. Mangenot, G. Samson, J. Poulain, V. Anthouard, C. Scarpelli, F. Artiguenave, P. Wincker, High quality draft sequences for

prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008). Medline doi:10.1186/1471-2164-9-603

37. R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, J. Wang, De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010). Medline doi:10.1101/gr.097261.109

38. A. Cenci, M. C. Combes, P. Lashermes, Differences in evolution rates among eudicotyledon species observed by analysis of protein divergence. *J. Hered.* **104**, 459–464 (2013). Medline doi:10.1093/jhered/est025

39. A. F. Ribas, A. Cenci, M. C. Combes, H. Etienne, P. Lashermes, Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. *BMC Genomics* **12**, 240 (2011). Medline doi:10.1186/1471-2164-12-240

40. R. Guyot, F. Lefebvre-Pautigny, C. Tranchant-Dubreuil, M. Rigoreau, P. Hamon, T. Leroy, S. Hamon, V. Poncet, D. Crouzillat, A. de Kochko, Ancestral synteny shared between distantly-related plant species from the asterid (*Coffea canephora* and *Solanum* Sp.) and rosid (*Vitis vinifera*) clades. *BMC Genomics* **13**, 103 (2012). Medline doi:10.1186/1471-2164-13-103

41. R. Guyot, M. de la Mare, V. Viader, P. Hamon, O. Coriton, J. Bustamante-Porras, V. Poncet, C. Campa, S. Hamon, A. de Kochko, 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 (2009). Medline doi:10.1186/1471-2229-9-22

42. A. L. Delcher, S. L. Salzberg, A. M. Phillippy, Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* **Chapter 10**, Unit 10.3 (2003). Medline

43. D. Crouzillat, M. Rigoreau, M. R. Priyono, paper presented at the ASIC 24th International Conference on Coffee Science, San José, Costa Rica, 11 to 16 November 2012.

44. F. Lefebvre-Pautigny, F. Wu, M. Philippot, M. Rigoreau, Priyono, M. Zouine, P. Frasse, M. Bouzayen, P. Broun, V. Pétiard, S. D. Tanksley, D. Crouzillat, High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. *Tree Genet. Genomes* **6**, 565–577 (2010). doi:10.1007/s11295-010-0272-3

45. C. Lin, L. A. Mueller, J. McCarthy, D. Crouzillat, V. Pétiard, S. D. Tanksley, Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theor. Appl. Genet.* **112**, 114–130 (2005). Medline doi:10.1007/s00122-005-0112-2

46. F. Wu, L. A. Mueller, D. Crouzillat, V. Pétiard, S. D. Tanksley, Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: A test case in the euasterid plant clade. *Genetics* **174**, 1407–1420 (2006). Medline doi:10.1534/genetics.106.062455

47. J. W. Van Ooijen, *JoinMap® 4.0 Software for the Calculation of Genetic Linkage Maps in Experimental Populations* (Kyazma B.V., Wageningen, Netherlands, 2006).

48. D. Kosambi, The estimation of map distances from recombination values. *Ann. Eugen.* **12**, 172–175 (1944). doi:10.1111/j.1469-1809.1943.tb02321.x

49. N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4.10 (2004). Medline

50. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005). Medline doi:10.1159/000084979

51. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). Medline doi:10.1093/nar/27.2.573

52. A. L. Price, N. C. Jones, P. A. Pevzner, De novo identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005). Medline doi:10.1093/bioinformatics/bti1018

53. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004). Medline doi:10.1101/gr.1865504

54. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004). Medline doi:10.1186/1471-2105-5-59

55. V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7** (suppl. 1), 1–12 (2006). Medline doi:10.1186/gb-2006-7-s1-s10

56. R. Mott, EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13**, 477–478 (1997). Medline

57. R. O. Vidal, J. M. Mondego, D. Pot, A. B. Ambrósio, A. C. Andrade, L. F. Pereira, C. A. Colombo, L. G. Vieira, M. F. Carazzolle, G. A. Pereira, A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. *Plant Physiol.* **154**, 1053–1066 (2010). Medline doi:10.1104/pp.110.162438

58. R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, J. Wang, SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009). Medline doi:10.1093/bioinformatics/btp336

59. F. Denoeud, J. M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, F. Artiguenave, Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008). Medline doi:10.1186/gb-2008-9-12-r175

60. K. L. Howe, T. Chothia, R. Durbin, GAZE: A generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res.* **12**, 1418–1427 (2002). Medline doi:10.1101/gr.149502

61. A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005). Medline doi:10.1093/bioinformatics/bti610

62. S. Myhre, H. Tveit, T. Mollestad, A. Laegreid, Additional gene ontology structure for improved biological reasoning. *Bioinformatics* **22**, 2020–2027 (2006). Medline doi:10.1093/bioinformatics/btl334

63. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

64. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012). Medline doi:10.1093/bioinformatics/bts565

65. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997). Medline doi:10.1093/nar/25.17.3389

66. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003). Medline doi:10.1101/gr.1224503

67. Z. Zhang, J. Yu, D. Li, Z. Zhang, F. Liu, X. Zhou, T. Wang, Y. Ling, Z. Su, PMRD: Plant microRNA database. *Nucleic Acids Res.* **38** (suppl. 1), D806–D813 (2010). Medline doi:10.1093/nar/gkp818

68. C. Noirot, C. Gaspin, T. Schiex, J. Gouzy, LeARN: A platform for detecting, clustering and annotating non-coding RNAs. *BMC Bioinformatics* **9**, 21 (2008). Medline doi:10.1186/1471-2105-9-21

69. J. T. Cuperus, N. Fahlgren, J. C. Carrington, Evolution and functional diversification of *MIRNA* genes. *Plant Cell* **23**, 431–442 (2011). Medline doi:10.1105/tpc.110.082784

70. M. Ricchetti, C. Fairhead, B. Dujon, Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**, 96–100 (1999). Medline doi:10.1038/47076

71. T. Mourier, A. J. Hansen, E. Willerslev, P. Arctander, The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Mol. Biol. Evol.* **18**, 1833–1837 (2001). Medline doi:10.1093/oxfordjournals.molbev.a003971

72. E. Richly, D. Leister, NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* **21**, 1081–1084 (2004). Medline doi:10.1093/molbev/msh110

73. X. Lin, S. Kaul, S. Rounsley, T. P. Shea, M. I. Benito, C. D. Town, C. Y. Fujii, T. Mason, C. L. Bowman, M. Barnstead, T. V. Feldblyum, C. R. Buell, K. A. Ketchum, J. Lee, C. M. Ronning, H. L. Koo, K. S. Moffat, L. A. Cronin, M. Shen, G. Pai, S. Van Aken, L. Umayam, L. J. Tallon, J. E. Gill, M. D. Adams, A. J. Carrera, T. H. Creasy, H. M. Goodman, C. R. Somerville, G. P. Copenhaver, D. Preuss, W. C. Nierman, O. White, J. A. Eisen, S. L. Salzberg, C. M. Fraser, J. C. Venter, Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–768 (1999). Medline doi:10.1038/45471

74. R. M. Stupar, J. W. Lilly, C. D. Town, Z. Cheng, S. Kaul, C. R. Buell, J. Jiang, Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: Implication of potential sequencing errors caused by large-unit repeats.

Proc. Natl. Acad. Sci. U.S.A. **98**, 5099–5103 (2001). Medline doi:10.1073/pnas.091110398

75. The Rice Chromosome 10 Sequencing Consortium, In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* **300**, 1566–1569 (2003). Medline doi:10.1126/science.1083523

76. M. Matsuo, Y. Ito, R. Yamauchi, J. Obokata, The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell* **17**, 665–675 (2005). Medline doi:10.1105/tpc.104.027706

77. M. Michalovova, B. Vyskot, E. Kejnovsky, Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: Size, relative age and chromosomal localization. *Heredity* **111**, 314–320 (2013). Medline doi:10.1038/hdy.2013.51

78. T. Flutre, E. Duprat, C. Feuillet, H. Quesneville, Considering transposable element diversification in de novo annotation approaches. *PLOS ONE* **6**, e16526 (2011). Medline doi:10.1371/journal.pone.0016526

79. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H. Schulman, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007). Medline doi:10.1038/nrg2165

80. A. D'Hont, F. Denoeud, J. M. Aury, F. C. Baurens, F. Carreel, O. Garsmeur, B. Noel, S. Bocs, G. Droc, M. Rouard, C. Da Silva, K. Jabbari, C. Cardi, J. Poulain, M. Souquet, K. Labadie, C. Jourda, J. Lengellé, M. Rodier-Goud, A. Alberti, M. Bernard, M. Correa, S. Ayyampalayam, M. R. Mckain, J. Leebens-Mack, D. Burgess, M. Freeling, D. Mbéguié-A-Mbéguié, M. Chabannes, T. Wicker, O. Panaud, J. Barbosa, E. Hribova, P. Heslop-Harrison, R. Habas, R. Rivallan, P. Francois, C. Poiron, A. Kilian, D. Burthia, C. Jenny, F. Bakry, S. Brown, V. Guignon, G. Kema, M. Dita, C. Waalwijk, S. Joseph, A. Dievart, O. Jaillon, J. Leclercq, X. Argout, E. Lyons, A. Almeida, M. Jeridi, J. Dolezel, N. Roux, A. M. Risterucci, J. Weissenbach, M. Ruiz, J. C. Glaszmann, F. Quétier, N. Yahiaoui, P. Wincker, The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**, 213–217 (2012). Medline doi:10.1038/nature11241

81. R. J. Langham, J. Walsh, M. Dunn, C. Ko, S. A. Goff, M. Freeling, Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* **166**, 935–945 (2004). Medline doi:10.1534/genetics.166.2.935

82. D. Sankoff, C. Zheng, Fractionation, rearrangement and subgenome dominance. *Bioinformatics* **28**, i402–i408 (2012). Medline doi:10.1093/bioinformatics/bts392

83. K. Jahn, C. Zheng, J. Kováč, D. Sankoff, A consolidation algorithm for genomes fractionated after higher order polyploidization. *BMC Bioinformatics* **13** (suppl. 19), S8 (2012). Medline

84. E. Lyons, B. Pedersen, J. Kane, M. Freeling, The value of nonmodel genomes and an example using SYNMAP within COGE to dissect the hexaploidy that predates rosids. *Trop. Plant Biol.* **1**, 181–190 (2008). doi:10.1007/s12042-008-9017-y

85. E. Lyons, M. Freeling, How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* **53**, 661–673 (2008). Medline doi:10.1111/j.1365-313X.2007.03326.x

86. C. Zheng, K. M. Swenson, E. Lyons, D. Sankoff, in *Algorithms in Bioinformatics*, T. M. Przytycka, M.-F. Sagot, Eds. [Eleventh International Workshop on Algorithms in Bioinformatics (WABI), Lecture Notes in Computer Science Series, Springer, Berlin, Heidelberg, 2011)], vol. 6833, pp. 364–375.

87. A. H. Paterson, M. Freeling, H. Tang, X. Wang, Insights from the comparison of plant genome sequences. *Annu. Rev. Plant Biol.* **61**, 349–372 (2010). Medline doi:10.1146/annurev-arplant-042809-112235

88. R. Ming, R. VanBuren, Y. Liu, M. Yang, Y. Han, L. T. Li, Q. Zhang, M. J. Kim, M. C. Schatz, M. Campbell, J. Li, J. E. Bowers, H. Tang, E. Lyons, A. A. Ferguson, G. Narzisi, D. R. Nelson, C. E. Blaby-Haas, A. R. Gschwend, Y. Jiao, J. P. Der, F. Zeng, J. Han, X. J. Min, K. A. Hudson, R. Singh, A. K. Grennan, S. J. Karpowicz, J. R. Watling, K. Ito, S. A. Robinson, M. E. Hudson, Q. Yu, T. C. Mockler, A. Carroll, Y. Zheng, R. Sunkar, R. Jia, N. Chen, J. Arro, C. M. Wai, E. Wafula, A. Spence, Y. Han, L. Xu, J. Zhang, R. Peery, M. J. Haus, W. Xiong, J. A. Walsh, J. Wu, M. L. Wang, Y. J. Zhu, R. E. Paull, A. B. Britt, C. Du, S. R. Downie, M. A. Schuler, T. P. Michael, S. P. Long, D. R. Ort, J. William Schopf, D. R. Gang, N. Jiang, M. Yandell, C. W. dePamphilis, S. S. Merchant, A. H. Paterson, B. B. Buchanan, S. Li, J. Shen-Miller, Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* **14**, R41 (2013). Medline doi:10.1186/gb-2013-14-5-r41

89. M. Nei, Selectionism and neutralism in molecular evolution. *Mol. Biol. Evol.* **22**, 2318–2342 (2005). Medline doi:10.1093/molbev/msi242

90. J. H. Matis, T. R. Kiffe, *Lecture Notes in Statistics* (Springer, New York, 2000).

91. T. De Bie, N. Cristianini, J. P. Demuth, M. W. Hahn, CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006). Medline doi:10.1093/bioinformatics/btl097

92. M. Spencer, E. Susko, A. J. Roger, Modelling prokaryote gene content. *Evol. Bioinform. Online* **2**, 157–178 (2006). Medline

93. K. P. Burnham, D. R. Anderson, *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (Springer, New York, 2002).

94. R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, A. Bateman, The Pfam protein families database. *Nucleic Acids Res.* **36** (suppl. 1), D281–D288 (2008). Medline doi:10.1093/nar/gkm960

95. K. Katoh, K. Kuma, H. Toh, T. Miyata, MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005). Medline doi:10.1093/nar/gki198

96. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005). Medline doi:10.1038/nature03895

97. G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Déjardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjärvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leplé, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouzé, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, D. Rokhsar, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006). Medline doi:10.1126/science.1128691

98. X. Argout, J. Salse, J. M. Aury, M. J. Guiltinan, G. Droc, J. Gouzy, M. Allegre, C. Chaparro, T. Legavre, S. N. Maximova, M. Abrouk, F. Murat, O. Fouet, J. Poulain, M. Ruiz, Y. Roguet, M. Rodier-Goud, J. F. Barbosa-Neto, F. Sabot, D. Kudrna, J. S. Ammiraju, S. C. Schuster, J. E. Carlson, E. Sallet, T. Schiex, A. Dievart, M. Kramer, L. Gelley, Z. Shi, A. Bérard, C. Viot, M. Boccara, A. M. Risterucci, V. Guignon, X. Sabau, M. J. Axtell, Z. Ma, Y. Zhang, S. Brown, M. Bourge, W. Golser, X. Song, D. Clement, R. Rivallan, M. Tahi, J. M. Akaza, B. Pitollat, K. Gramacho, A. D'Hont, D. Brunel, D. Infante, I. Kebe, P. Costet, R. Wing, W. R. McCombie, E. Guiderdoni, F. Quetier, O. Panaud, P. Wincker, S. Bocs, C. Lanaud, The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011). Medline doi:10.1038/ng.736

99. B. W. Porter, M. Paidi, R. Ming, M. Alam, W. T. Nishijima, Y. J. Zhu, Genome-wide analysis of *Carica papaya* reveals a small *NBS* resistance gene family. *Mol. Genet. Genomics* **281**, 609–626 (2009). Medline doi:10.1007/s00438-009-0434-x

100. B. C. Meyers, A. Kozik, A. Griego, H. Kuang, R. W. Michelmore, Genome-wide analysis of NBS-LRR–encoding genes in *Arabidopsis*. *Plant Cell* **15**, 809–834 (2003). Medline doi:10.1105/tpc.009308

101. N. Yoneyama, H. Morimoto, C. X. Ye, H. Ashihara, K. Mizuno, M. Kato, Substrate specificity of *N*-methyltransferase involved in purine alkaloids synthesis is dependent upon one amino acid residue of the enzyme. *Mol. Genet. Genomics* **275**, 125–135 (2006). Medline doi:10.1007/s00438-005-0070-z

102. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). Medline doi:10.1093/nar/gkh340

103. K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007). Medline doi:10.1093/molbev/msm092

104. E. Ibarra-Laclette, E. Lyons, G. Hernández-Guzmán, C. A. Pérez-Torres, L. Carretero-Paulet, T. H. Chang, T. Lan, A. J. Welch, M. J. Juárez, J. Simpson, A. Fernández-Cortés, M. Arteaga-Vázquez, E. Góngora-Castillo, G. Acevedo-Hernández, S. C. Schuster, H. Himmelbauer, A. E. Minoche, S. Xu, M. Lynch, A. Oropeza-Aburto, S. A. Cervantes-Pérez, M. de Jesús Ortega-Estrada, J. I. Cervantes-Luevano, T. P. Michael, T. Mockler, D. Bryant, A. Herrera-Estrella, V. A. Albert, L. Herrera-Estrella, Architecture and evolution of a minute plant genome. *Nature* **498**, 94–98 (2013). Medline doi:10.1038/nature12132

105. Z. Yang, Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998). Medline doi:10.1093/oxfordjournals.molbev.a025957

106. Z. Yang, R. Nielsen, Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002). Medline doi:10.1093/oxfordjournals.molbev.a004148

107. J. P. Bielawski, Z. Yang, A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.* **59**, 121–132 (2004). Medline doi:10.1007/s00239-004-2597-8

108. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994). Medline

109. Z. Yang, W. S. Wong, R. Nielsen, Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118 (2005). Medline doi:10.1093/molbev/msi097

110. C. Qin, C. Yu, Y. Shen, X. Fang, L. Chen, J. Min, J. Cheng, S. Zhao, M. Xu, Y. Luo, Y. Yang, Z. Wu, L. Mao, H. Wu, C. Ling-Hu, H. Zhou, H. Lin, S. González-Morales, D. L. Trejo-Saavedra, H. Tian, X. Tang, M. Zhao, Z. Huang, A. Zhou, X. Yao, J. Cui, W. Li, Z. Chen, Y. Feng, Y. Niu, S. Bi, X. Yang, W. Li, H. Cai, X. Luo, S. Montes-Hernández, M. A. Leyva-González, Z. Xiong, X. He, L. Bai, S. Tan, X. Tang, D. Liu, J. Liu, S. Zhang, M. Chen, L. Zhang, L. Zhang, Y. Zhang, W. Liao, Y. Zhang, M. Wang, X. Lv, B. Wen, H. Liu, H. Luan, Y. Zhang, S. Yang, X. Wang, J. Xu, X. Li, S. Li, J. Wang, A. Palloix, P. W. Bosland, Y. Li, A. Krogh, R. F. Rivera-Bustamante, L. Herrera-Estrella, Y. Yin, J. Yu, K. Hu, Z. Zhang, Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5135–5140 (2014). Medline doi:10.1073/pnas.1400975111

111. G. B. Martin, A. J. Bogdanove, G. Sessa, Understanding the functions of plant disease resistance proteins. *Annu. Rev. Plant Biol.* **54**, 23–61 (2003). Medline doi:10.1146/annurev.arplant.54.031902.135035

112. B. C. Meyers, A. W. Dickerman, R. W. Michelmore, S. Sivaramakrishnan, B. W. Sobral, N. D. Young, Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *Plant J.* **20**, 317–332 (1999). Medline doi:10.1046/j.1365-313X.1999.t01-1-00606.x

113. W. I. Tameling, S. D. Elzinga, P. S. Darmin, J. H. Vossen, F. L. Takken, M. A. Haring, B. J. Cornelissen, The tomato R gene products I-2 and MI-1 are functional ATP binding

proteins with ATPase activity. *Plant Cell* **14**, 2929–2939 (2002). Medline doi:10.1105/tpc.005793

114. W. I. Tameling, J. H. Vossen, M. Albrecht, T. Lengauer, J. A. Berden, M. A. Haring, B. J. Cornelissen, F. L. Takken, Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation. *Plant Physiol.* **140**, 1233–1245 (2006). Medline doi:10.1104/pp.105.073510

115. J. G. Ellis, G. J. Lawrence, J. E. Luck, P. N. Dodds, Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* **11**, 495–506 (1999). Medline doi:10.1105/tpc.11.3.495

116. J. E. Luck, G. J. Lawrence, P. N. Dodds, K. W. Shepherd, J. G. Ellis, Regions outside of the leucine-rich repeats of flax rust resistance proteins play a role in specificity determination. *Plant Cell* **12**, 1367–1377 (2000). Medline doi:10.1105/tpc.12.8.1367

117. S. M. Collier, P. Moffett, NB-LRRs work a "bait and switch" on pathogens. *Trends Plant Sci.* **14**, 521–529 (2009). Medline doi:10.1016/j.tplants.2009.08.001

118. A. R. Friedman, B. J. Baker, The evolution of resistance genes in multi-protein plant resistance systems. *Curr. Opin. Genet. Dev.* **17**, 493–499 (2007). Medline doi:10.1016/j.gde.2007.08.014

119. S. Yang, X. Zhang, J. X. Yue, D. Tian, J. Q. Chen, Recent duplications dominate NBS-encoding gene expansion in two woody species. *Mol. Genet. Genomics* **280**, 187–198 (2008). Medline doi:10.1007/s00438-008-0355-0

120. S. Guo, J. Zhang, H. Sun, J. Salse, W. J. Lucas, H. Zhang, Y. Zheng, L. Mao, Y. Ren, Z. Wang, J. Min, X. Guo, F. Murat, B. K. Ham, Z. Zhang, S. Gao, M. Huang, Y. Xu, S. Zhong, A. Bombarely, L. A. Mueller, H. Zhao, H. He, Y. Zhang, Z. Zhang, S. Huang, T. Tan, E. Pang, K. Lin, Q. Hu, H. Kuang, P. Ni, B. Wang, J. Liu, Q. Kou, W. Hou, X. Zou, J. Jiang, G. Gong, K. Klee, H. Schoof, Y. Huang, X. Hu, S. Dong, D. Liang, J. Wang, K. Wu, Y. Xia, X. Zhao, Z. Zheng, M. Xing, X. Liang, B. Huang, T. Lv, J. Wang, Y. Yin, H. Yi, R. Li, M. Wu, A. Levi, X. Zhang, J. J. Giovannoni, J. Wang, Y. Li, Z. Fei, Y. Xu, The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* **45**, 51–58 (2013). Medline doi:10.1038/ng.2470

121. E. B. Holub, The arms race is ancient history in *Arabidopsis*, the wildflower. *Nat. Rev. Genet.* **2**, 516–527 (2001). Medline doi:10.1038/35080508

122. Q. Xu, L. L. Chen, X. Ruan, D. Chen, A. Zhu, C. Chen, D. Bertrand, W. B. Jiao, B. H. Hao, M. P. Lyon, J. Chen, S. Gao, F. Xing, H. Lan, J. W. Chang, X. Ge, Y. Lei, Q. Hu, Y. Miao, L. Wang, S. Xiao, M. K. Biswas, W. Zeng, F. Guo, H. Cao, X. Yang, X. W. Xu, Y. J. Cheng, J. Xu, J. H. Liu, O. J. Luo, Z. Tang, W. W. Guo, H. Kuang, H. Y. Zhang, M. L. Roose, N. Nagarajan, X. X. Deng, Y. Ruan, The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66 (2013). Medline doi:10.1038/ng.2472

123. H. Wan, W. Yuan, K. Bo, J. Shen, X. Pang, J. Chen, Genome-wide analysis of NBS-encoding disease resistance genes in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. *BMC Genomics* **14**, 109 (2013). Medline doi:10.1186/1471-2164-14-109

124. A. J. Bettencourt, C. J. Rodrigues Jr., “Principles and practice of coffee breeding for resistance to rust and other diseases,” in *Coffee: Agronomy*, vol. 4, R. J. Clarke, R. Macrae, Eds. (Elsevier, London, 1988), pp. 199–234.

125. M. A. Lila, From beans to berries and beyond: Teamwork between plant chemicals for protection of optimal human health. *Ann. N. Y. Acad. Sci.* **1114**, 372–380 (2007). Medline doi:10.1196/annals.1396.047

126. F. Ververidis, E. Trantas, C. Douglas, G. Vollmer, G. Kretzschmar, N. Panopoulos, Biotechnology of flavonoids and other phenylpropanoid-derived natural products. Part II: Reconstruction of multienzyme pathways in plants and microbes. *Biotechnol. J.* **2**, 1235–1249 (2007). Medline doi:10.1002/biot.200700184

127. H. Han, B. K. Baik, Antioxidant activity and phenolic content of lentils (*Lens culinaris*), chickpeas (*Cicer arietinum* L.), peas (*Pisum sativum* L.) and soybeans (*Glycine max*), and their quantitative changes during processing. *Int. J. Food Sci. Technol.* **43**, 1971–1978 (2008). doi:10.1111/j.1365-2621.2008.01800.x

128. E. O. Cuevas-Rodríguez, V. P. Dia, G. G. Yousef, P. A. García-Saucedo, J. López-Medina, O. Paredes-López, E. Gonzalez de Mejia, M. A. Lila, Inhibition of pro-inflammatory responses and antioxidant capacity of Mexican blackberry (*Rubus* spp.) extracts. *J. Agric. Food Chem.* **58**, 9542–9548 (2010). Medline doi:10.1021/jf102590p

129. M. N. Clifford, Chlorogenic acids and other cinnamates–nature, occurrence, dietary burden, absorption and metabolism. *J. Sci. Food Agric.* **80**, 1033–1043 (2000). doi:10.1002/(SICI)1097-0010(20000515)80:7<1033::AID-JSFA595>3.0.CO;2-T

130. C.-L. Ky, J. Louarn, B. Guyot, A. Charrier, S. Hamon, M. Noirot, Relations between and inheritance of chlorogenic acid contents in an interspecific cross between *Coffea pseudozanguebariae* and *Coffea liberica* var “dewevrei”. *Theor. Appl. Genet.* **98**, 628–637 (1999). doi:10.1007/s001220051114

131. C.-L. Ky, J. Louarn, S. Dussert, B. Guyot, S. Hamon, M. Noirot, Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica* L. and *C. canephora* P. accessions. *Food Chem.* **75**, 223–230 (2001). doi:10.1016/S0308-8146(01)00204-7

132. C. Bertrand, M. Noirot, S. Doulbeau, A. de Kochko, S. Hamon, C. Campa, Chlorogenic acid content swap during fruit maturation in *Coffea pseudozanguebariae*: Qualitative comparison with leaves. *Plant Sci.* **165**, 1355–1361 (2003). doi:10.1016/j.plantsci.2003.07.002

133. M. Lepelley, G. Cheminade, N. Tremillon, A. Simkin, V. Caillet, J. McCarthy, Chlorogenic acid synthesis in coffee: An analysis of CGA content and real-time RT-PCR expression of HCT, HQT, C3H1, and CCoAOMT1 genes during grain development in *C. canephora*. *Plant Sci.* **172**, 978–996 (2007). doi:10.1016/j.plantsci.2007.02.004

134. Y. Koshiro, M. C. Jackson, R. Katahira, M. L. Wang, C. Nagai, H. Ashihara, Biosynthesis of chlorogenic acids in growing and ripening fruits of *Coffea arabica* and *Coffea canephora* plants. *Z. Naturforsch. C* **62**, 731–742 (2007). Medline

135. Y. Kono, K. Kobayashi, S. Tagawa, K. Adachi, A. Ueda, Y. Sawa, H. Shibata, Antioxidant activity of polyphenolics in diets: Rate constants of reactions of chlorogenic acid and

caffeic acid with reactive species of oxygen and nitrogen. *Biochim. Biophys. Acta* **1335**, 335–342 (1997). Medline doi:10.1016/S0304-4165(96)00151-1

136. M. D. del Castillo, J. M. Ames, M. H. Gordon, Effect of roasting on the antioxidant activity of coffee brews. *J. Agric. Food Chem.* **50**, 3698–3703 (2002). Medline doi:10.1021/jf011702q

137. B. McDougall, P. J. King, B. W. Wu, Z. Hostomsky, M. G. Reinecke, W. E. Robinson Jr., Dicaffeoylquinic and dicaffeoyltartaric acids are selective inhibitors of human immunodeficiency virus type 1 integrase. *Antimicrob. Agents Chemother.* **42**, 140–146 (1998). Medline

138. W. E. Robinson Jr., M. Cordeiro, S. Abdel-Malek, Q. Jia, S. A. Chow, M. G. Reinecke, W. M. Mitchell, Dicaffeoylquinic acid inhibitors of human immunodeficiency virus integrase: Inhibition of the core catalytic domain of human immunodeficiency virus integrase. *Mol. Pharmacol.* **50**, 846–855 (1996). Medline

139. W. E. Robinson Jr., M. G. Reinecke, S. Abdel-Malek, Q. Jia, S. A. Chow, Inhibitors of HIV-1 replication that inhibit HIV integrase. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6326–6331 (1996). Medline doi:10.1073/pnas.93.13.6326

140. A. Farah, T. de Paulis, L. C. Trugo, P. R. Martin, Effect of roasting on the formation of chlorogenic acid lactones in coffee. *J. Agric. Food Chem.* **53**, 1505–1513 (2005). Medline doi:10.1021/jf048701t

141. A. Farah, T. de Paulis, D. P. Moreira, L. C. Trugo, P. R. Martin, Chlorogenic acids and lactones in regular and water-decaffeinated arabica coffees. *J. Agric. Food Chem.* **54**, 374–381 (2006). Medline doi:10.1021/jf0518305

142. R. Dorfner, T. Ferge, A. Kettrup, R. Zimmermann, C. Yeretzian, Real-time monitoring of 4-vinylguaiacol, guaiacol, and phenol during coffee roasting by resonant laser ionization time-of-flight mass spectrometry. *J. Agric. Food Chem.* **51**, 5768–5773 (2003). Medline doi:10.1021/jf0341767

143. C. Campa, M. Noirot, M. Bourgeois, M. Pervent, C. L. Ky, H. Chrestin, S. Hamon, A. de Kochko, Genetic mapping of a caffeoyl-coenzyme A 3-O-methyltransferase gene in coffee trees. Impact on chlorogenic acid content. *Theor. Appl. Genet.* **107**, 751–756 (2003). Medline doi:10.1007/s00122-003-1310-4

144. V. Mahesh, J. J. Rakotomalala, L. Le Gal, H. Vigne, A. de Kochko, S. Hamon, M. Noirot, C. Campa, Isolation and genetic mapping of a *Coffea canephora* phenylalanine ammonia-lyase gene (*CcPAL1*) and its involvement in the accumulation of caffeoyl quinic acids. *Plant Cell Rep.* **25**, 986–992 (2006). Medline doi:10.1007/s00299-006-0152-3

145. V. Mahesh, R. Million-Rousseau, P. Ullmann, N. Chabrillange, J. Bustamante, L. Mondolot, M. Morant, M. Noirot, S. Hamon, A. de Kochko, D. Werck-Reichhart, C. Campa, Functional characterization of two *p*-coumaroyl ester 3′-hydroxylase genes from coffee tree: Evidence of a candidate for chlorogenic acid biosynthesis. *Plant Mol. Biol.* **64**, 145–159 (2007). Medline doi:10.1007/s11103-007-9141-3

146. M. Lepelley, V. Mahesh, J. McCarthy, M. Rigoreau, D. Crouzillat, N. Chabrillange, A. de Kochko, C. Campa, Characterization, high-resolution mapping and differential

expression of three homologous *PAL* genes in *Coffea canephora* Pierre (Rubiaceae). *Planta* **236**, 313–326 (2012). Medline doi:10.1007/s00425-012-1613-2

147. The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000). Medline doi:10.1038/35048692

148. A. Marchler-Bauer, S. H. Bryant, CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* **32** (suppl. 2), W327–W331 (2004). Medline doi:10.1093/nar/gkh454

149. A. Marchler-Bauer, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, S. H. Bryant, CDD: Specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.* **37** (suppl. 1), D205–D210 (2009). Medline doi:10.1093/nar/gkn845

150. A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S. H. Bryant, CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39** (suppl. 1), D225–D229 (2011). Medline doi:10.1093/nar/gkq1189

151. C. Clé, L. M. Hill, R. Niggeweg, C. R. Martin, Y. Guisez, E. Prinsen, M. A. Jansen, Modulation of chlorogenic acid biosynthesis in *Solanum lycopersicum*; consequences for phenolic accumulation and UV-tolerance. *Phytochemistry* **69**, 2149–2156 (2008). Medline doi:10.1016/j.phytochem.2008.04.024

152. R. Shi, Y. H. Sun, Q. Li, S. Heber, R. Sederoff, V. L. Chiang, Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: Transcript abundance and specificity of the monolignol biosynthetic genes. *Plant Cell Physiol.* **51**, 144–163 (2010). Medline doi:10.1093/pcp/pcp175

153. R. Niggeweg, A. J. Michael, C. Martin, Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nat. Biotechnol.* **22**, 746–754 (2004). Medline doi:10.1038/nbt966

154. L. Hoffmann, S. Maury, F. Martz, P. Geoffroy, M. Legrand, Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. *J. Biol. Chem.* **278**, 95–103 (2003). Medline doi:10.1074/jbc.M209362200

155. L. Hoffmann, S. Besseau, P. Geoffroy, C. Ritzenthaler, D. Meyer, C. Lapierre, B. Pollet, M. Legrand, Silencing of hydroxycinnamoyl-coenzyme A shikimate/quinate hydroxycinnamoyltransferase affects phenylpropanoid biosynthesis. *Plant Cell* **16**, 1446–1465 (2004). Medline doi:10.1105/tpc.020297

156. W. Zheng, M. N. Clifford, Profiling the chlorogenic acids of sweet potato (*Ipomoea batatas*) from China. *Food Chem.* **106**, 147–152 (2008). doi:10.1016/j.foodchem.2007.05.053

157. G. Sonnante, R. D'Amore, E. Blanco, C. L. Pierri, M. De Palma, J. Luo, M. Tucci, C. Martin, Novel hydroxycinnamoyl-coenzyme A quinate transferase genes from artichoke are involved in the synthesis of chlorogenic acid. *Plant Physiol.* **153**, 1224–1238 (2010). Medline doi:10.1104/pp.109.150144

158. C. J. Tsai, S. A. Harding, T. J. Tschaplinski, R. L. Lindroth, Y. Yuan, Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in *Populus*. *New Phytol.* **172**, 47–62 (2006). Medline doi:10.1111/j.1469-8137.2006.01798.x

159. J. C. D'Auria, Acyltransferases in plants: A good time to be BAHD. *Curr. Opin. Plant Biol.* **9**, 331–340 (2006). Medline doi:10.1016/j.pbi.2006.03.016

160. H. Ashihara, X. Q. Zheng, R. Katahira, M. Morimoto, S. Ogita, H. Sano, Caffeine biosynthesis and adenine metabolism in transgenic *Coffea canephora* plants with reduced expression of *N*-methyltransferase genes. *Phytochemistry* **67**, 882–886 (2006). Medline doi:10.1016/j.phytochem.2006.02.016

161. K. Mizuno, S. Kurosawa, Y. Yoshizawa, M. Kato, Essential region for 3-N methylation in N-methyltransferases involved in caffeine biosynthesis. *Z. Naturforsch. C* **65**, 257 (2010).

162. P. Araque, H. Casanova, C. Ortiz, B. Henao, C. Pelaez, Insecticidal activity of caffeine aqueous solutions and caffeine oleate emulsions against *Drosophila melanogaster* and *Hypothenemus hampei*. *J. Agric. Food Chem.* **55**, 6918–6922 (2007). Medline doi:10.1021/jf071052b

163. M. Kato, K. Mizuno, A. Crozier, T. Fujimura, H. Ashihara, Caffeine synthase gene from tea leaves. *Nature* **406**, 956–957 (2000). Medline doi:10.1038/35023072

164. K. A. Franklin, P. H. Quail, Phytochrome functions in *Arabidopsis* development. *J. Exp. Bot.* **61**, 11–24 (2010). Medline doi:10.1093/jxb/erp304

165. S. Ohno, *Evolution by Gene Duplication* (Springer, New York, 1970).

166. Z. Yang, PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput.Aappl. Biosci.* **13**, 555–556 (1997). Medline

167. Y. Li-Beisson, B. Shorrosh, F. Beisson, M. X. Andersson, V. Arondel, P. D. Bates, S. Baud, D. Bird, A. Debono, T. P. Durrett, R. B. Franke, I. A. Graham, K. Katayama, A. A. Kelly, T. Larson, J. E. Markham, M. Miquel, I. Molina, I. Nishida, O. Rowland, L. Samuels, K. M. Schmid, H. Wada, R. Welti, C. Xu, R. Zallot, J. Ohlrogge, Acyl-lipid metabolism. *Arabidopsis Book* **11**, e0161 (2013). Medline doi:10.1199/tab.0161

168. E. N. Frankel, *Lipid Oxidation* (The Oily Press, High Wycombe, UK, ed. 2, 2005).

169. S. Dussert, N. Chabrillange, G. Rocquelin, F. Engelmann, M. Lopez, S. Hamon, Tolerance of coffee (*Coffea* spp.) seeds to ultra-low temperature exposure in relation to calorimetric properties of tissue water, lipid composition, and cooling procedure. *Physiol. Plant.* **112**, 495–504 (2001). Medline doi:10.1034/j.1399-3054.2001.1120406.x

170. M. T. L. Kreuml, D. Majchrzak, B. Ploederl, J. Koenig, Changes in sensory quality characteristics of coffee during storage. *Food Sci. Nutr.* **1**, 267–272 (2013). Medline doi:10.1002/fsn3.35

171. A. T. Toci, V. J. Neto, A. G. Torres, A. Farah, Changes in triacylglycerols and free fatty acids composition during storage of roasted coffee. *LWT Food Sci. Technol.* **50**, 581–590 (2013). doi:10.1016/j.lwt.2012.08.007

172. D. Lang, B. Weiche, G. Timmerhaus, S. Richardt, D. M. Riaño-Pachón, L. G. Corrêa, R. Reski, B. Mueller-Roeber, S. A. Rensing, Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: A timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503 (2010). Medline doi:10.1093/gbe/evq032

173. R. Lozano, O. Ponce, M. Ramirez, N. Mostajo, G. Orjeda, Genome-wide identification and mapping of NBS-encoding resistance genes in *Solanum tuberosum* group phureja. *PLOS ONE* **7**, e34775 (2012). Medline doi:10.1371/journal.pone.0034775

174. A. Kohler, C. Rinaldi, S. Duplessis, M. Baucher, D. Geelen, F. Duchaussoy, B. C. Meyers, W. Boerjan, F. Martin, Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* **66**, 619–636 (2008). Medline doi:10.1007/s11103-008-9293-9

175. S. F. Altschul, J. C. Wootton, E. M. Gertz, R. Agarwala, A. Morgulis, A. A. Schäffer, Y. K. Yu, Protein database searches using compositionally adjusted substitution matrices. *FEBS J.* **272**, 5101–5109 (2005). Medline doi:10.1111/j.1742-4658.2005.04945.x