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

France Denoued,1,2,3 Lorenzo Carretero-Paulet,4 Alexis Dereeper,4 Gaëtan Droé,5 Romain Aprea,7 Marco Pietrella,8 Chunjing Zheng,9 Adriana Alberti,3 François Anthony,5 Giuseppe Aprea,2 Jean-Marc Aury,1 Pascal Bento,5 Maria Bernard,1 Stéphanie Bocs,9 Claudein Campa,7 Alberto Cenci,6 Marie-Christine Combes,5 Dominique Courzioullat,11 Corinne Da Silva,1 Loretta Daddiego,12 Fabien De Bellis,6 Stéphane Dussert,5 Olivier Garsmeur,6 Thomas Gayraud,14 Katharina Jahn,9,13,14 Véronique Jamilloux,13 Thierry Joët,9 Karine Labadie,1 Tianying Lan,4,16 Julie Leclercq,6 Maud Lepellely,11 Thierry Leroy,7 Lei-Ting Li,17 Pablo Librado,13 Loredana Lopez,12 Adriana Muñoz,19,20 Benjamin Noel,1 Alberto Pallavicini,5 Gaetano Perrotta,12 Valérie PonceT,7 David Pot,6 Priyono,22 Michel Rigoureau,11 Mathieu Rouard,10 Julio Rozas,18 Christine Tranchant-Dubreuil,7 Robert VanBuren,17 Qiong Zhang,17 Alan C. Andrade,23 Xavier Argout,6 Benoît Bertrand,24 Alexandre de Kochko,8 Giorgio Graziosi,21,23 Robert J Henry,26 Jayarama,27 Ray Ming,17 Chifumi Nagai,28 Steve Rounsley,29 David Sankoff,6,7,9,17,28 Patrick Wincker,1,2,3,10 Philippe Laschermes5

Coffee is a valuable beverage crop due to its characteristic flavor, aroma, and the stimulating effects of caffeine. We generated a high-quality draft genome of the species Coffea canephora, which displays a conserved chromosomal gene order among asterid angiosperms. Although it shows no sign of the whole-genome triplication identified in Solanaceae species such as tomato, the genome includes several species-specific gene family expansions, among them N-methyltransferases (NMTs) involved in caffeine production, defense-related genes, and alkaloid and flavonoid enzymes involved in secondary compound synthesis. Comparative analyses of caffeine NMTs demonstrate that these genes expanded through sequential tandem duplications independently of genes from cacao and tea, suggesting that caffeine in eudicots is of polyphyletic origin.

With more than 2.25 billion cups consumed every day, coffee is one of the most important crops on Earth, cultivated across more than 11 million hectares. Coffee belongs to the Rubiaceae family, which is part of the Euasterid I clade and the fourth largest family of angiosperms, consisting of more than 11,000 species in 600 genera (1). We sequenced Coffea canephora (2n = 2x = 22 chromosomes), an outcrossing, highly heterozygous diploid, and one of the parents of C. arabica (2n = 4x = 44 chromosomes), which was derived from hybridization between C. canephora and C. eugenioides (2). A total of 54.4 million Roche 454 single and mate-pair reads and 163,905 Sanger bacterial artificial chromosome–end reads were generated from a doubled haploid accession, representing ~30× coverage of the 710-Mb genome (3). Additional Illumina sequencing data (60×) were used to improve the assembly (table S1) (4). The resulting assembly consists of 25,216 contigs and 13,345 scaffolds with a total length of 568.6 Mb (80% of 710 Mb), including 97 Mb (17%) of intercontig gaps. Eighty percent of the assembly is in 635 scaffolds, and the scaffold N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) is 1.26 Mb (table S2). A high-density genetic map covering 349 scaffolds and comprising ~64% of the assembly (364 Mb) and 86% of the annotated genes was anchored to the 11 C. canephora chromosomes (4). More than 96% of the scaffolds larger than 1 Mb were anchored (Fig. IA).

We annotated 25,574 protein-coding genes (4) (table S6), 92 microRNA precursors, and 2573 organellar-to-nucleolar genome transfers (4). Transposable elements account for ~50% of the genome (4), of which ~85% are long terminal repeat (LTR) retrotransposons. Large-scale comparison between C. canephora LTR retrotransposons and those of reference plant genomes shows outstanding conservation of several Copia groups across distantly related genera, suggesting that horizontal mobile element transfers may be more frequent than generally recognized (5–8).

Structurally, the coffee genome shows no sign of a whole-genome polyploidization in its linkage since the γ triplication at the origin of the core eudicots (9) (Fig. 1B). Coffee contains exactly three paralogous regions for each of the seven pre-γ ancestral chromosomes (Fig. 1B). Coffee chromosomal regions show unique one-to-one correspondences with grapevine chromosomes (Fig. 1C and fig. S12) and a one-to-three correspondence with the tomato genome, which underwent a second lineage-specific triplication during its evolutionary history (10). Although grapevine, a rosid, is the most conservative core eudicot in terms of integrity of gross chromosomal structure, coffee displays less gene-order divergence to all other rosids, despite being an asterid itself (9). Coffee also shows little syntenic divergence relative to other sequenced asterids (Fig. 1D, table S17, and supplementary text). To classify gene families in the C. canephora genome, we ran OrthoMCL on inferred protein sequences from coffee, grapevine, tomato, and Arabidopsis (4), generating 18,917 groups of orthologous genes (Fig. S5). To examine coffee-specific gene family expansions with potential adaptive significance, we fit branch models that were implemented in BadiRate (II) to these orthogroups (4). In the coffee lineage, 202 orthogroups
clustering 1270 genes were supported as expanded (Akaike information criterion > 2.7). Among gene ontology (GO) terms annotating these, 98 out of 4300 generic terms were significantly over- or underrepresented (table S14). Most GOs enriched (p < 0.001) in coffee relative to tomato, grapevine, and peach belong to two main functional categories: defense response and metabolic process, the later including different catalytic activities (table S15).

Among defense response functions, there is a clear expansion of nucleotide binding site disease-resistance genes (D2, I3) in the C. canephora genome (4). Most genes that grouped together within single orthogroups were tandemly arrayed, suggesting that R genes evolved by tandem duplication and divergence of linked gene families (supplementary text). Several gene functions involved in secondary metabolite biosynthesis are significantly expanded in the C. canephora genome, including enzymes associated with the production of phenylpropanoids such as flavonoids and isoflavones (naringenin 3-dioxygenase, isoflavone 2'-hydroxylase), alkaloids (strictosidine synthase, tropine dehydrogenase), monoterpenes (e.g., menthol dehydrogenase), and caffeine [N-methyltransferases (NMTs)] (Fig. 2). For example, indole alkaloids such as the monomer alkaloid yohimbine and antimalarial drug quinine are prominent secondary compounds of the coffee family and its parent order, Gentianales (14), and the GO term indole biosynthetic process was highly enriched (P < 0.001) in coffee relative to tomato, grapevine, and Arabidopsis.

Caffeine is a purine alkaloid synthesized by several eudicot plants, including coffee, cacao (Theobroma cacao), and tea (Camellia sinensis) (Fig. 2). Caffeine is synthesized in both coffee leaves, where it has insecticidal properties (15), and fruits and seeds, where it inhibits seed germination of competing species (16). The late steps in caffeine biosynthesis are mediated by a series of NMTs (Fig. 2A) (17).

Among coffee-expanded genes, NMT activity is one of the more highly enriched GO terms (table S15). A single gene family (ORTHOMCL170) clusters 23 genes in coffee, but none in grapevine, tomato, or Arabidopsis (table S12), and this cluster contains genes encoding known enzymes of the caffeine biosynthetic pathway (18, 19). Maximum likelihood (ML) phylogenetic analysis of ORTHOMCL170 with tea and cacao NMTs that have similar activities reveals species-specific gene clades (Fig. 2C). We analyzed these relationships in a broader evolutionary context by including genome-wide samples of NMTs from coffee, cacao, and other eudicot species. ML trees show that the genes encoding the closest Arabidopsis NMT relatives of coffee caffeine biosynthetic enzymes are involved in benzoic, salicylic, and nictitonic functions (4) (supplementary text). Caffeine biosynthetic NMTs from coffee nested within a gene clade distinct from those of cacao or tea, which group together as sister lineages. Thus, a
The principal caffeine biosynthetic pathway involves three methyltransferase steps: xanthosine methylation, theobromine synthase, and caffeine synthase. Three NMTs (xanthosine methyltransferase, theobromine synthase, caffeine synthase) are involved in the biosynthesis of caffeine in coffee beans. The NMTs are encoded by tandem arrays that have undergone duplication and diversification events. These events have led to the evolution of caffeine biosynthesis in the coffee plant, providing an example of convergent evolution of secondary metabolic pathways encoded by tandemly duplicated genes.

Genomic functional diversification via tandem duplication may have helped shape other aspects of coffee bean chemical composition. Linoleic acid, which is produced by the oleate desaturase FAD2, is the major polyunsaturated fatty acid in the coffee bean (25, 26), where it contributes to aroma composition and flavor retention after roasting (4). Coffee has six FAD2 genes compared with one in Arabidopsis, and most of these have arisen from tandem duplications on chromosome 1 (fig. S33). RNA sequencing data suggest transcriptional specialization for two of the six FAD2 copies, with CcFAD2.3 being actively involved in the production of linoleic acid.
transcribed in developing endosperm (supplementary text). Peak transcript abundance coincides with the dramatic increase in linoleic acid content that occurs during seed development at the perisperm-endosperm transition (27).

Our analysis of the adaptive genomic landscape of <i>C. canephora</i> identifies the convergent evolution of caffeine biosynthesis among plant lineages and establishes coffee as a reference species for understanding the evolution of genome structure in asterid angiosperms.

**REFERENCES AND NOTES**

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

**ACKNOWLEDGMENTS**

We acknowledge the following sources for funding: AnR-08-GENM-022-001 (to P.L.); AnR-09-GENM-014-002 (to P.W.); AnR-09-GENM-015-002 (to C.J.B.); AnR-09-GENM-014-001 (to P.L.).

We thank P. Facella (ENEA) for Roche 454 sequencing and the College of Arts and Sciences, University at Buffalo (to V.A.A.).

We thank P. Facella (ENEA) for Roche 454 sequencing and the College of Arts and Sciences, University at Buffalo (to V.A.A.). We thank P. Facella (ENEA) for Roche 454 sequencing and the College of Arts and Sciences, University at Buffalo (to V.A.A.). We thank P. Facella (ENEA) for Roche 454 sequencing and the College of Arts and Sciences, University at Buffalo (to V.A.A.). We thank P. Facella (ENEA) for Roche 454 sequencing and the College of Arts and Sciences, University at Buffalo (to V.A.A.).

**SUPPLEMENTARY MATERIALS**

<http://www.sciencemag.org/content/345/6201/suppl/DC1>

**REFERENCES**

(28–75)

28 April 2014. accepted 29 July 2014

10.1126/science.1255274

---

**GENOME EDITING**

**Prevention of muscular dystrophy in mice by CRISPR/Cas9–mediated editing of germline DNA**

Chengzu Long,1,*, John R. McNally,1,*, John M. Shelton,2 Alex A. Mireault,1 Rhonda Bassel-Duby,3 Eric N. Olson†

Duchenne muscular dystrophy (DMD) is an inherited X-linked disease caused by mutations in the gene encoding dystrophin, a protein required for muscle fiber integrity. DMD is characterized by progressive muscle weakness and a shortened life span, and there is no effective treatment. We used clustered regularly interspaced short palindromic repeat/Cas9 (CRISPR/Cas9)–mediated genome editing to correct the dystrophin gene (<i>Dmd</i>) mutation in the germ line of <i>mice</i>, a model for DMD, and then monitored muscle structure and function. Genome editing produced genetically mosaic animals containing 2 to 100% correction of the <i>Dmd</i> gene. The degree of muscle phenotypic rescue in mosaic mice exceeded the efficiency of gene correction, likely reflecting an advantage of the corrected cells and their contribution to regenerating muscle. With the anticipated technological advances that will facilitate genome editing of postnatal somatic cells, this strategy may one day allow correction of disease-causing mutations in the muscle tissue of patients with DMD.

<sup>1</sup>Department of Molecular Biology and Hamon Center for Regenerative Science and Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
<sup>2</sup>Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
<sup>3</sup>These authors contributed equally to this work. To whom correspondence should be addressed. E-mail: eric.olson@utsouthwestern.edu

---

<sup>†</sup>To whom correspondence should be addressed. E-mail: eric.olson@utsouthwestern.edu