Deciphering the genome structure and paleohistory of *Theobroma cacao*

Xavier Argout*[1], Jerome Salse*[2], Jean Marc Aury*[3], Gaetan Droc[1], Jerome Gouzy[4], Mathilde Allegre[1], Cristian Chaparro[5], Thierry Legavre[1], Mark Guiltinan[6], Siela Maximova[6], Michael Abrouk[2], Florent Murat[2], Olivier Fouet[1], Julie Poulain[3], Manuel Ruiz[1], Yolande Roguet[1], Maguy Rodier-Goud[1], Jose Fernandes Barbosa-Neto[5], Francois Sabot[5], Dave Kudrna[7], Jetty Siva S. Ammiraju[7], Stephan C. Schuster[8], John E. Carlson[9], Erika Sallet[4], Schiex T.[10], Anne Dievart[1], Melissa Kramer[11], Laura Gelley[11], Shi Z.[12], Aurélie Bérard[13], Christopher Viot[1], Michel Boccara[1], Ange Marie Risterucci[1], Valentin Guignon[1], Xavier Sabau[1], Axtell MJ.[14], Ma Z.[14], Zhang Y.[12], Spencer Brown[15], Mickael Bourge[15], Wolfgang Golser[7], Xiang Song[7], Didier Clement[1], Ronan Rivalan[1], Mathias Tahi[16], Joseph Moroh Akaza[16], Bertrand Pitollat[1], Karina Gramacho[17], Angélique D'Hont[1], Dominique Brunel[13], Diogenes Infante[18], Ismael Kebe[16], Pierre Costet[19], Rod Wing[7], W. Richard McCombie[11], Emmanuel Guiderdoni[1], Francis Quetier[20], Olivier Panaud[5], Patrick Wincker[3], Stephanie Sidibe-Bocs[1], Claire Lanaud[1].

**These authors contributed equally to this work*

1 CIRAD - Biological Systems Department – UMR DAP TA A 96/03- 34398, Montpellier, cedex 5- France

2 Institut National de la Recherché Agronomique UMR 1095, 63100 Clermont-Ferrand, France

3 Genoscope (CEA) and UMR 8030 CNRS-Genoscope-Université d'Evry, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France

4 INRA-CNRS LIPM Laboratoire des Interactions Plantes Micro-organismes, BP 52627, 31326 Castanet Tolosan Cedex, France

5 UMR 5096 CNRS-IRD-UPVD, Laboratoire Génome et Développement des Plantes, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

6 Penn State University, Department of Horticulture and the Huck Institutes of the Life Sciences, University Park, PA 16802, USA

7 Arizona Genomics Institute and School of Plant Sciences, University of Arizona, Tucson AZ 85721, USA

8 Penn State University, Department of Biochemistry and Molecular Biology, University Park, PA 16802, USA

9 Penn State University, The School of Forest Resources and the Huck Institutes of the Life Sciences, University Park, PA 16802, USA and The Department of Bioenergy Science and Technology (WCU), Chonnam National University, 333 Yongbongro, Buk-Gu, Gwangju, 500-757, Korea

10 Unité de Biométrie et d'Intelligence Artificielle (UBIA), UR875 INRA, F-31320 Castanet Tolosan France

11 Cold Spring Harbor Laboratory, NY 11723, USA

12 Penn State University, Plant Biology Graduate Program and the Huck Institutes of the Life Sciences, University Park, PA 16802, USA

13 INRA, UR 1279 Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, Centre National de Génotypage, 2, rue Gaston Crémieux, CP5724, 91057 Evry, France

14 Penn State University, Bioinformatics and Genomics Ph.D. Program &Department of Biology, University Park, PA 16802, USA

15 Institut des Sciences du Végétal, UPR 2355, CNRS, 91198 Gif-Sur-Ivette, France

16 Centre national de la recherche agronomique (CNRA), B.P. 808, Divo, Côte d'Ivoire

17 CEPLAC, Km 22 Rod. Ilheus Itabuna, Cx. postal 07, Itabuna 45600-00, Bahia, Brazil

18 Centro Nacional de Biotecnología Agrícola, Instituto de Estudios Avanzados, Caracas 1015-A, Venezuela

19 Chocolaterie VALRHONA, 8, quai du général de Gaulle, 26600 Tain l'Hermitage, France

20 Département de Biologie, Université d'Evry Val d'Essonne, 25 boulevard François Mitterrand, 91025 Evry, France

ABSTRACT

We sequenced and assembled the genome of *Theobroma cacao*, an economically important tropical fruit tree crop that is the source of chocolate. The assembly corresponds to 76% of the estimated genome size and contains almost all previously described genes, with 82% of them anchored on the 10 *T. cacao* chromosomes. Analysis of this sequence information highlighted specific expansion of some gene families during evolution, for example flavonoid-related genes. It also provides a major source of candidate genes for *T. cacao* improvement. Based on the inferred paleohistory of the *T. cacao* genome, we propose an evolutionary scenario whereby the ten *T. cacao* chromosomes were shaped from an ancestor through eleven chromosome fusions. The *T. cacao* genome can be considered as the simplest living relic of higher plant evolution identified to date.

INTRODUCTION *Theobroma cacao* L., is a diploid tree fruit species (2n = 2x = 20; Davie, 1933) that is endemic to the South American rainforests. Its seeds are used in a wide range of products, the most popular being chocolate. There is also an increasing appreciation of its value for environmental preservation because the cocoa tree can be cultivated under forest shade, allowing for land rehabilitation and enrichment of biodiversity, while providing income for many subsistence farmers[1] . *T. cacao* was first domesticated by Mesoamerican natives approximately 3000 years ago[2]. Criollo, the first domesticated variety, provides white beans highly appreciated for making fineflavored aromatic chocolate (Supplementary Data 1 online). Relics of the ancestral Criollo, first cultivated by Olmec or Mayan people, can still be encountered in old Mesoamerican plantations or in forests where Mayan people lived[3]. Our genome sequence is derived from a Belizean Criollo plant collected in the Mayan mountains[4] (Supplementary Data 1). Cocoa culture has now expanded to all humid tropical countries, providing 3.7 million of tons of cocoa annually (http://www.icco.org/economics/market.aspx) and income for millions of small-scale farmers. However these farmers face increasing threats from fungal diseases and insect pests, which are globally responsible for 30% of harvest losses (http://www.dropdata.org/cocoa/cocoa_prob.htm). Like many other tropical crops, knowledge of *T. cacao* genetics and genomics is limited. Therefore, to accelerate progress in cocoa breeding and understanding of its biochemistry, we sequenced and analyzed the genome of a Belizean Criollo genotype (B97-61/B2), which has large white beans suitable for producing a high quality and fine-flavored chocolate. This genotype is suitable for high quality genome sequence assembly because it is highly homozygous as a result of the many generations of self-fertilization that occurred naturally during the domestication process.

CONCLUSION *Theobroma cacao* is the first long-generation-time, tropical tree fruit crop that has been sequenced. We were able to assemble 76% of its genome and identify 28,798

protein-coding genes among which 23,529 (82%) could be anchored in the ten cocoa chromosomes. We found that 682 gene families are specific to *T. cacao*. Only 20% of the genome consisted of transposable elements, a significantly lower percentage than in other genomes of similar size. The analysis of specific gene families that are potentially linked to cocoa qualities and disease resistance, two important traits for cocoa consumption and cultivation, revealed that particular expansion or reduction of some gene families had occurred during evolution. The mapping of these gene families along the cocoa chromosomes and comparison with the genome regions involved in trait variation (QTLs) constitutes an invaluable source of candidate genes for further functional studies that aim to discover the specific genes directly involved in trait variation. This genome sequence will facilitate a better understanding of trait elaboration and will accelerate *T. cacao* selection through efficient marker-assisted selection and exploitation of genetic resources. This study has highlighted the close evolutionary distance of the *T. cacao* genome from the eudicot putative ancestor, showing a limited number of recombinations between ancestral chromosomes, as was also observed in grape[7]. *T. cacao*, which has only ten pairs of chromosomes, is easily propagated by both sexual and vegetative methods, and can be transformed; it represents a new and perhaps the simplest model to study evolutionary processes, gene functions and tree fruit crop genetics and biochemistry. The large amount of information generated by this project dramatically changes the status of this tropical plant and its potential interest for the scientific community. We hope this situation will encourage greater investment in research with *Theobroma cacao*, the "food of the Gods" whose magic flavor has spread worldwide since the time of the Maya and Aztec civilizations, and whose continued study will benefit developing countries for which cocoa is of high economic importance.

Acknowledgements We would like to thank CIRAD, the Agropolis foundation, the Région Languedoc Roussillon, Agence Nationale de la Recherche (ANR), Valrhona and the Venezuelan Ministry of Science, Technology and Industry for their financial contribution to this project. We thank the Toulouse Midi-Pyrénées bioinformatic platform for providing us with computational resources. Activities at The Pennsylvania State University were supported by a gift from the Hershey Corp., by funds from the American Cocoa Research Institute Endowment, and through support from the Schatz Center for Tree Molecular Genetics in The School of Forest Resources. Acquisition of an Illumina sequencer by the CSHL was supported by National Science Foundation grant DBI0923128 to WRM. We would like to thank F. C. Baurens, Y. Jiao, O. Garsmeur and C. dePamphilis for helpful advice and assistance with bioinformatics.