

## The genome of *Theobroma cacao*

Xavier Argout<sup>\*1</sup>, Jerome Salse<sup>\*2</sup>, Jean-Marc Aury<sup>\*3,4,5</sup>, Mark J. Guiltinan<sup>\*6,15</sup>, Gaetan Droc<sup>1</sup>, Jerome Gouzy<sup>7</sup>, Mathilde Allegre<sup>1</sup>, Cristian Chaparro<sup>8</sup>, Thierry Legavre<sup>1</sup>, Siela N. Maximova<sup>6</sup>, Michael Abrouk<sup>2</sup>, Florent Murat<sup>2</sup>, Olivier Fouet<sup>1</sup>, Julie Poulain<sup>3,4,5</sup>, Manuel Ruiz<sup>1</sup>, Yolande Roguet<sup>1</sup>, Maguy Rodier-Goud<sup>1</sup>, Jose Fernandes Barbosa-Neto<sup>8</sup>, Francois Sabot<sup>8</sup>, Dave Kudrna<sup>9</sup>, Jetty Silva S. Ammiraju<sup>9</sup>, Stephan C. Schuster<sup>10</sup>, John E. Carlson<sup>11,12</sup>, Erika Sallet<sup>7</sup>, Thomas Schiex<sup>13</sup>, Anne Dievart<sup>1</sup>, Melissa Kramer<sup>14</sup>, Laura Gelley<sup>14</sup>, Zi Shi<sup>15</sup>, Aurélie Bérard<sup>16</sup>, Christopher Viot<sup>1</sup>, Michel Boccara<sup>1</sup>, Ange Marie Risterucci<sup>1</sup>, Valentin Guignon<sup>1</sup>, Xavier Sabau<sup>1</sup>, Michael J. Axtell<sup>17</sup>, Zhaorong Ma<sup>17</sup>, Yufan Zhang<sup>15</sup>, Spencer Brown<sup>18</sup>, Mickael Bourge<sup>18</sup>, Wolfgang Golser<sup>9</sup>, Xiang Song<sup>9</sup>, Didier Clement<sup>1</sup>, Ronan Rivallan<sup>1</sup>, Mathias Tahiri<sup>19</sup>, Joseph Moroh Akaza<sup>19</sup>, Bertrand Pitollat<sup>1</sup>, Karina Gramacho<sup>20</sup>, Angélique D'Hont<sup>1</sup>, Dominique Brunel<sup>16</sup>, Diogenes Infante<sup>21</sup>, Ismael Kebe<sup>18</sup>, Pierre Costet<sup>22</sup>, Rod Wing<sup>9</sup>, W. Richard McCombie<sup>14</sup>, Emmanuel Guiderdoni<sup>1</sup>, Francis Quetier<sup>23</sup>, Olivier Panaud<sup>8</sup>, Patrick Wincker<sup>3,4,5</sup>, Stephanie Bocs<sup>1</sup>, Claire Lanaud<sup>1</sup>.

*\*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 Recherche Agronomique UMR 1095, 63100 Clermont-Ferrand, France

3 CEA, IG, Genoscope, 2 rue Gaston Crémieux CP5702, F-91057 Evry, France

4 CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, F-91057 Evry, France

5 Université d'Evry, F-91057 Evry, France

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

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

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

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

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

11 Penn State University, The School of Forest Resources and the Huck Institutes of the Life Sciences, University Park, PA 16802, USA

12 The Department of Bioenergy Science and Technology (WCU), Chonnam National University, 333 Yongbongro, Buk-Gu, Gwangju, 500-757, Korea

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

14 Cold Spring Harbor Laboratory, NY 11723, USA

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

16 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

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

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

19 Centre National de la Recherche Agronomique (CNRA), B.P. 808, Divo, Côte d'Ivoire

20 Comissão Executiva de Planejamento da Lavoura Cacaueira (CEPLAC), Km 22 Rod. Ilheus Itabuna, Cx. postal 07, Itabuna 45600-00, Bahia, Brazil

21 Centro Nacional de Biotecnología Agrícola, Instituto de Estudios Avanzados (IDEA), Caracas 1015-A, Venezuela

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

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

# **Table of contents**

Supplementary Notes.....	4
Origin of the Criollo genotype B97-61/B2 subjected to sequencing and requirement for cocoa bean fermentation to generate chocolate quality precursors .....	4
Origin of the Criollo genotype B97-61/B2 subjected to sequencing .....	4
Requirement for cocoa bean fermentation to generate chocolate qualities .....	4
High molecular weight DNA preparation.....	5
Isolation of cell nuclei on cocoa leaves .....	5
Isolation of nuclear DNA .....	5
Purification of nuclear DNA .....	5
Construction of BAC libraries.....	5
Genomic sequencing and assembly.....	6
Genome sequencing.....	6
Genome assembly.....	7
Automatic error corrections with Solexa/Illumina reads .....	7
Genome size evaluations.....	7
Estimation of nuclear DNA content by flow cytometry .....	7
Genome size variations among <i>T. cacao</i> genotypes, <i>Theobroma</i> species and the closely related genus <i>Herrania</i> .....	8
Anchoring the assembly on the high-density genetic map .....	8
Transposable elements.....	9
TE annotation .....	9
Southern blots analysis .....	9
RFLP analysis.....	9
Fluorescence In situ Hybridization (FISH) of TE probes .....	10
Protein coding gene annotations.....	10
Transcriptome.....	10
Protein coding gene model predictions.....	11
Homology search and functional annotation.....	11
Filtering of protein coding genes tagged as transposable element genes or as false positives .....	12
Construction of families of homologous polypeptides and identification of cocoa subfamily-specific polypeptides .....	12
Non-coding gene annotations and target prediction.....	13
<i>Theobroma cacao</i> rRNA annotation .....	13
<i>Theobroma cacao</i> microRNA annotation .....	13
<i>Theobroma cacao</i> microRNA target prediction .....	14
Identification of LRR-LRK genes in the <i>T. cacao</i> genome .....	14
Characterization of <i>T. cacao</i> genes orthologous to NBS-encoding genes.....	15
Classification of the predicted genes encoding NBS domains.....	15
NBS domain motif description.....	15
Total number and organization of <i>T. cacao</i> genes orthologous to NBS-encoding genes .....	15
Phylogenetic analysis of NBS domains .....	16
Identification of NPR1 genes in the cocoa genome .....	16
Genome distribution of <i>T. cacao</i> genes orthologous to NBS, LRR-LRK and NPR1-like genes and comparative mapping with QTLs related to disease resistance in <i>T. cacao</i> .....	17
Genome distribution of lipid and flavonoid orthologous genes and comparative mapping with QTLs for traits related to fat and flavonoids .....	18

Cacao genome synteny, duplication, evolution and paleohistory. ....	18
Arabidopsis, grape, poplar, soybean, papaya sequence databases. ....	18
Synteny and duplication analysis. ....	19
Distribution of $K_S$ distances (MYA scale) for paralogous and orthologous gene pairs ...	19
Supplementary Tables .....	20
Supplementary Figures .....	38
Supplementary References.....	69

## **Supplementary Note**

### **Origin of the Criollo genotype B97-61/B2 subjected to sequencing and requirement for cocoa bean fermentation to generate chocolate quality precursors**

#### ***Origin of the Criollo genotype B97-61/B2 subjected to sequencing***

An expedition was undertaken in 1994 to collect ancient Criollo material in the Maya mountains from Belize<sup>1</sup>. This material is now conserved in the International Cocoa Genbank (ICG, Trinidad) and was recently characterized by Motilal et al<sup>2</sup>. These authors assessed the relationships of these Criollo germplasms with other cocoa accessions and determined their putative ancestral contribution to the Trinitario hybrid group. One of these Belizean Criollo genotypes (B97-61/B2) was chosen for the sequencing of its genome. Cocoa clones are generally self-incompatible and highly heterozygous. Criollo genotypes are self-compatible and the B97-61/B2 clone is highly homozygous, facilitating the genome assembly. Its homozygosity level was first estimated at 93% by genotyping with 130 microsatellite markers and at 99.9% by genotyping with 795 single-nucleotide polymorphisms (SNPs) using the Illumina Golden Gate system.

#### ***Requirement for cocoa bean fermentation to generate chocolate qualities***

Fermentation of the fresh cocoa beans that are surrounded by a pectinaceous pulp is an important step in producing quality chocolate. This is a natural and complex process mediated by a large number of fungi and bacteria, which are mechanically inoculated onto the pods when they are cut and handled during harvest. The microorganism population composition varies during the progression of the fermentation<sup>3</sup>. The time and duration of fermentation depend on the type of cocoa and the region where it is grown, but involves the stacking of cocoa beans in a pile or a box, with successive turning of the pile or the box during three to seven days. Early in the process, the sugars are converted to ethanol and lactic acid due to the action of yeast and lactic acid bacteria; later, ethanol is oxidized to acetic acid by acetic acid bacteria.

This fermentation process is accompanied by changes in pH and the rise of the temperature of the stack<sup>4</sup>. The fermentation products permeate the cotyledons, killing the embryo and producing biochemical reactions that induce changes both in the structure of the seed at the subcellular level, and in the metabolites present in the beans. The changes influence the aroma and develop the aroma precursors in the fermented seeds<sup>5</sup>. Besides theobromine and caffeine, the flavan-3-ols epicatechin, catechin, procyanidin B-2, procyanidin B-5, procyanidin C-1, [epicatechin-(4 $\beta$ -8)]3-epicatechin, and [epicatechin-(4 $\beta$ -8)]4-epicatechin are among the key compounds contributing to the bitter taste as well as the astringent mouth feel imparted upon consumption of roasted cocoa<sup>6</sup>. A complexity of aromatic terpene and lipid metabolites also contribute greatly to the flavor of cocoa. In addition, there is a strong influence of both the environment and the genetic origins of cocoa beans on flavor development.

## High molecular weight DNA preparation

High molecular weight DNA was prepared following isolation of nuclei prepared from cocoa leaves of B97-61/B2 according to the following protocols:

### *Isolation of cell nuclei on cocoa leaves*

Isolation of nuclei was carried out as previously described<sup>7</sup> with the following exceptions: (1) the amount of starting tissue was lowered to 0.5 g / 10 mL NIBM buffer to avoid clogging during the filtration steps; (2) the steps of filtration with Miracloth (CALBIOCHEM®) were replaced by five successive filtrations with nylon filters (SEFAR NITEX®) with decreasing mesh diameters: 250 µm, 100 µm, 50 µm and two times 11 µm; and (3) to reduce organelle contamination in the nuclei preparations, nuclei isolation buffer containing 0.5 % TritonX-100 was used during the nuclei washing steps<sup>8</sup>.

The quality of extraction was monitored by epifluorescence microscopy by assessing the number of nuclei (blue) compared to the chloroplasts (red) and cellular debris (green). A mixture of 10 µL of nuclei solution and 10 µL 4',6-diamidino-2-phenylindole (DAPI) 1.5 µg/mL was prepared and placed on a glass slide layered with coverslip. The slides were then examined with a Leica DM RAX2 fluorescence microscope and the images of blue, red and green fluorescence were acquired separately with a cooled high resolution black and white CCD camera. The camera was interfaced to a PC running the Velocity® software (Perkin Elmer).

### *Isolation of nuclear DNA*

The extraction of nuclear DNA followed a protocol using a MATAB buffer already described for isolation of genomic DNA<sup>9</sup>. The only changes were on the first and last steps: (1) there was no crushing of tissue, the starting material was 500 µL of nuclei solution for 2 mL of extraction buffer per tube; (2) DNA was resuspended with 300 µL of TE (10 mM Tris-HCl and 1 mM EDTA, pH 8.0).

### *Purification of nuclear DNA*

DNA purification followed the protocol from the Nucleobond® PC 20 kit (Macherey-Nagel) with the following modifications: the culture and lysis of cells was replaced by a crude DNA solution. To adjust the salt concentrations and pH, a 1 mL mixture of 200 µL of crude DNA (20 µg DNA maximum), 450 µL of water and 350 µL S3 buffer + RNase (buffer kit) was prepared. This solution was homogenized on an oscillating table for a minimum of 1 hour. This DNA preparation was then shared among the several collaborating laboratories involved in these sequencing activities: Genoscope (France), The Pennsylvania State University (USA) and Cold Spring Harbor Laboratory (USA).

## Construction of BAC libraries

Two *T. cacao* BAC libraries were constructed at the Arizona Genomic Institute following established methods<sup>7</sup> from high molecular weight nuclear DNA using modifications recently described for *Oryza sativa*<sup>8</sup>. Young leaves from an adult plant of *T. cacao*, variety Criollo B97 61/B-2, were provided by the Cocoa Research Unit at The University of the West Indies,

Trinidad. Nuclei were isolated and collected in agarose plugs. DNA digestions were performed with varying amounts of *Hind*III or *Eco*RI to identify the appropriate partial digestion conditions for selection of large size restriction fragments followed by ligation to pAGIBAC1 vector (a modified pIndigoBAC536Blue with an additional *Swa*I site<sup>8</sup>. Ligation products were transformed into DH10B T1 phage-resistant *Escherichia coli* cells (Invitrogen, Carlsbad, CA) and plated on LB agar that contained chloramphenicol (12.5 µg mL<sup>-1</sup>), X-gal (20 mg mL<sup>-1</sup>) and IPTG (0.1 M). Clones were robotically transferred to barcoded 384-well plates containing LB freezing medium. After incubation for 18 h, plates were backfilled to replace blank wells, replicated and frozen at -80°C. The *Hind*III library was named TC\_CBa and the *Eco*RI library was named TC\_CBb. Both libraries are available to the public from the Arizona Genomics Institute Resource Center<sup>10</sup>.

Characteristics, quality assessment and estimated genome coverage of the two BAC libraries were determined and are summarized in Supplementary Table 1. A representative subset of BAC clones from each library was assembled to allow confident determinations of % chloroplast clones (which are a major contamination concern), % non-insert clones and the average insert size. To estimate average insert sizes, 5 µL aliquots of subset BAC plasmid DNA were digested with 5 U of *Not*I enzyme for 3 hrs at 37°C. The digestion products were separated by pulsed-field gel electrophoresis (CHEF-DRIII system, Bio-Rad) in a 1% agarose gel in 0.5x TBE buffer. Electrophoresis was carried out for 16 hours at 14°C with an initial switch time of 5 sec, a final switch time of 15 sec, in a voltage gradient of 6V cm<sup>-1</sup>. The observed cloned inserts were compared to those of the MidRange I PFG Marker (New England Biolabs) (Supplementary Fig 1). The average insert size of BAC clones from each library was determined to be: TC\_CBa 135 kb; TC\_CBb 137 kb (Supplementary Fig 2). The % non-insert containing clones was determined by the number of clones observed that showed a vector band without an insert band in the PFGE display. No empty clones were observed in either library (Supplementary Fig 1).

The % chloroplast content was determined from the number of clone end sequences that displayed high confidence BLAST similarities to the *Arabidopsis thaliana* or *Oryza sativa* chloroplast genomic sequences. Plasmid DNA (5 µL) was reacted with vector sequencing primers, T7 and BES\_HR primer (CAC TCA TTA GGC ACC CCA), BigDye terminator V.3, dNTPs, and sequencing buffer in a total volume of 12 µL followed by 150 cycles of PCR (10 sec at 95°C, 5 sec at 55°C, and 2.5 min at 60°C)<sup>11</sup>. After reaction cleanup (Cleanseq, Agencourt), reactions were separated on a 3730xl ABI DNA analyzer. Sequences were base called using the program Phred<sup>12</sup>. Following BLAST analysis, no chloroplast sequences were found in either library.

The estimated genome coverage of each BAC library, based upon the current genome size of 430 Mbp for *T. cacao* B97-61/B2 genotype, and the average BAC insert sizes that we determined, were 5.14x for TC\_CBa and 8.04x for TC\_CBb (Supplementary Table 1).

## Genomic sequencing and assembly

### 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 (Supplementary Table 2).

### **Genome assembly**

Sanger and 454 reads were assembled with Newbler version 2.3. From the initial 26,519,827 reads 80,65% (21,387,691) were assembled. We obtained 25,912 contigs that were linked into 4,792 scaffolds. The contig N50 was 19.8 kb, and the scaffold N50 was 473.8 kb (Supplementary Table 4). The cumulative scaffold size was 326.9 Mb, about 24% smaller than the estimated genome size of 430 Mb. The *T. cacao* cDNA unigene resources (see below) were aligned with the assembly using the Blat<sup>13</sup> algorithm with default parameters and only the best match was kept for each unigene. The high coverage of the genome was confirmed by the alignment and the assembly contains 97.8% of the 38,737 cocoa unigenes.

### **Automatic error corrections with Solexa/Illumina reads**

One way to improve the 454 assembly is to complement it with another type of data with a different bias in error type, as described previously<sup>14</sup>. Short-read sequences were aligned on the cocoa genome assembly using the SOAP<sup>15</sup> software (with a seed size of 12 bps and a maximum gap size allowed of 3 bp per read). 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, (2) the quality of the considered bases, the previous and the next one were above 20, and (3) the remaining sequences (before and after) around the error 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 inside 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 and 70% of the reads located at that position agreed.

Since we only allow reads uniquely mapped and reads mapped with a maximum of two mismatches and three indels, several regions were devoid of Illumina tags. In a first step, one or several errors were corrected, and during subsequent iterations of the strategy, regions that were devoid of Illumina reads were also covered. We therefore decided to iterate the previous strategy during several cycles until no new errors were found. Four cycles were required (the first cycle corrected 45,061 errors, the second 4,310, the third 1,044 errors and the fourth, 299 errors).

### **Genome size evaluations**

The genome size of the sequenced cocoa clone, B97-61/B2, was estimated by flow cytometry. In order to check a potential relationship between genome size and transposable elements, the genome size was also estimated for a panel of cocoa genotypes from various genetic origins, and for representatives of related wild species from the same genus, *Theobroma*, or from a closely related genus, *Herrania*. (Supplementary Table 3)

### **Estimation of nuclear DNA content by flow cytometry**

The total DNA amount was assessed by flow cytometry according to Marie and Brown<sup>16</sup>. *Lycopersicon esculentum* cv. Roma (2C = 1.99 pg, 40.0% GC) and *Petunia hybrida* cv. PxPc6 (2C = 2.85 pg, 41.0% GC) were used as internal standards. Leaves of studied species (~2 cm<sup>2</sup>) and one internal standard (~0.5 cm<sup>2</sup>) were chopped with a razor blade in a Petri dish with 800

$\mu\text{L}$  of cold Galbraith nuclear isolation buffer<sup>17</sup> supplemented with 10 mM sodium metabisulfite, 1% polyvinylpyrrolidone 10,000 and 5  $\mu\text{g}/\text{mL}$  RNase. The suspension was passed through a 48  $\mu\text{m}$  mesh nylon filter. The nuclei were stained with 50  $\mu\text{g}/\text{mL}$  propidium iodide, a DNA-intercalating fluorochrome.

DNA content of 5000-10,000 stained nuclei was determined for each sample using a CyFlow® SL3 flow cytometer (Partec, Sainte Geneviève des Bois, France) with a 532 nm green solid state laser (100 mW). Using forward- and side-scatter to gate nuclei, fluorescence emission of propidium iodide was collected through a 590 nm long pass filter. The nuclear DNA value was calculated using the linear relationship between the fluorescent signals from the G0-G1 peaks of the unknown specimen and the known internal standard. The supplementary compounds in the buffer avoid interference from browning or tanning: only in the case of *T. grandiflora* was it necessary to make repeat preparations to obtain stable preparations. A further indicator of reliability was the observed linearity (2.00) between 2C and 4C nuclei of the internal standards. *L. esculentum* was a satisfactory internal standard in all cases. The monoploid C-value, 1C, (according to Greilhuber *et al.*<sup>18</sup>), was calculated and expressed in Mbp using the conversion factor 1 pg DNA = 978 Mbp<sup>19</sup>. Means were analyzed with a two-way T-test and grouped according to Bonferroni.

### ***Genome size variations among T. cacao genotypes, Theobroma species and the closely related genus Herrania***

Significant differences appear among these accessions of *T. cacao* (Supplementary Table 3). The B97-61/B2 genotype being sequenced has  $2C = 2x = 0.88$  pg, a haploid genome of 430 Mbp. The 2C values of the *T. cacao* accessions ranged from 0.84 pg to 1.01 pg. One species, *T. microcarpa*, within the genus has clearly a smaller genome ( $2C = 0.73$  pg). Two have relatively large genomes at the top end of the range, *T. speciosa* and *T. grandiflora* (both  $2C = 1.02$  pg). The related *Herrania* spp. cover a similar range of genome sizes ( $2C = 0.69$ – $1.05$  pg).

### **Anchoring the assembly on the high-density genetic map**

Maps of two progenies were used to establish a consensus map suitable for anchoring the assembly:

- A F1 progeny of 256 individuals, located at the Centre National de Recherche Agronomique (CNRA, Divo, Ivory Coast) which resulted from the cross of 2 heterozygous genotypes: UPA402, an Upper Amazon Forastero from Peru, and UF676, a Trinitario (hybrid between Forastero and Criollo) selected in Costa Rica. This progeny was used previously to establish the reference cocoa map, on which all available markers are progressively mapped<sup>9,20-22</sup>. The last map established included 600 codominant SSR and RFLP markers.
- A F2 progeny of 136 individuals, located at Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC, Itabuna, Brazil), recently produced from a cross between 2 heterozygous parents: ICS1, a Trinitario selected in Trinidad, and Scavina6, an Upper Amazon Forastero.

New simple sequence repeat (SSR) and SNP markers were mapped in these 2 progenies, and a consensus map including 1,259 markers was established<sup>23</sup>.

We used the stand alone Blat software<sup>13</sup> to align markers of the genetic map against the scaffolds. Only uniquely aligned markers with a cutoff of 80% identity were retained. We



































































































































