The genome of *Theobroma cacao*

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

Nature Genetics: doi: 10.1038/ng.736

Table of contents

Nature Genetics: doi: 10.1038/ng.736

Nature Genetics: doi: 10.1038/ng.736

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[1]. This material is now conserved in the International Cocoa Genbank (ICG, Trinidad) and was recently characterized by Motilal et al[2]. 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[3]. 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[4]. 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[5]. Besides theobromine and caffeine, the flavan-3-ols epicatechin, catechin, procyanidin B-2, procyanidin B-5, procyanidin C-1, [epicatechin-(4β-8)]3-epicatechin, and [epicatechin- (4β-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[6]. 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.

Nature Genetics: doi: 10.1038/ng.736

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[7] 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[8].

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 DMRAX2 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[9]. 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[7] from high molecular weight nuclear DNA using modifications recently described for *Oryza sativa*[8]. 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,

Nature Genetics: doi: 10.1038/ng.736

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[8]. 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^{-1}), X-gal (20 mg mL^{-1}) 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[10].

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^{-1}. 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 95C, 5 sec at 55°C, and 2.5 min at 60°C)[11]. After reaction cleanup (Cleanseq, Agencourt), reactions were separated on a 3730xl ABI DNA analyzer. Sequences were base called using the program Phred[12]. 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).

Nature Genetics: doi: 10.1038/ng.736

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[13] 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[14]. Short-read sequences were aligned on the cocoa genome assembly using the SOAP[15] 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[16]. *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^2) and one internal standard (~0.5 cm^2) were chopped with a razor blade in a Petri dish with 800

Nature Genetics: doi: 10.1038/ng.736

µL of cold Galbraith nuclear isolation buffer[17] supplemented with 10 mM sodium metabisulfite, 1% polyvinylpyrrolidone 10,000 and 5 µg/mL RNAse. The suspension was passed through a 48 µm mesh nylon filter. The nuclei were stained with 50 µg/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.*[18]), was calculated and expressed in Mbp using the conversion factor 1 pg DNA = 978 Mbp[19]. 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[9,20-22]. 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[23].

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

Nature Genetics: doi: 10.1038/ng.736

used the mapped markers to anchor and orientate the scaffolds along the *T. cacao* pseudomolecules. Among the 1259 markers, 1192 (94.3%) have a unique Blat hit in the scaffolds using a decreasing 90% to 80% identity threshold. The overview of the assembly anchoring on the genetic map is reported in Supplementary Table 5 and Supplementary Figure 3 for each cocoa linkage group.

Transposable elements

TE annotation

De novo identification of Long terminal repeat (LTR) retrotransposons followed a three-step approach. The first step was the identification of candidates integrating the results of LTR_finder[24], LTRharvest[25], and in-house software that searches for reverse transcriptase signatures and LTRs. The second step consisted of manual curation of the candidates, keeping only one reference element per family. The last step was the annotation of all retrotransposons using an in-house software based on BLASTN.

Class II elements were detected through BLASTX against Repbase[26] proteins. These anchors were used to search for transposase occurrences in the scaffolds and they were later extended to include the inverted repeats. Furthermore, miniature inverted repeat TEs (MITEs) were identified using MUST (http://csbl1.bmb.uga.edu/ffzhou/MUST/).

A BLASTN[27] "walking" approach was used to construct repeated elements from the 3,220,522 unassembled reads (454 and BAC ends). This approach was based on the identification of highly repeated sequences (seeds) and a BLASTN analysis against the repeated sequences database. After the identification of an element, the related sequences were eliminated from the database.

Southern blots analysis

DNA extraction was performed from fresh leaves according to Lanaud et al.[20] and digested with restriction enzyme *Hind*III. Sequences for probe synthesis were analyzed for PCR primer design using Primer3 software. For the *Copia*-like LTR retrotransposon *Gaucho*, the primer pair used to amplify a DNA fragment of 968 bp was: (forward) 5'-TTTCGCTGTGACGAAAGATG-3' and (reverse) 5'-ACGCTGTCTTGGGTACATCC-3'. For the tandem repeat *ThCen*, the primers pair used to amplify a DNA fragment of 160 bp was: (forward) 5'-CATGCCTTCGAAAGTCC-3' and (reverse) 5'-TGGACTTTTCTTCTCAATCG-3'.

RFLP analysis

Genomic DNA (2 µg) was completely restricted with 10 U of *Hind*III (New England Biolabs), fractionated on 0.8% agarose gels and transferred to Hybond-N+ (Amersham) nylon membranes using the alkali blotting protocol as described in its users' manual. The blots were hybridized for 24 h at 65°C in a solution containing 5x SSC, 5x Denhardt's, 0.5% SDS, 100 µg/mL fragmented and denatured herring sperm DNA and ^{32}P-labeled probe. The membranes was then washed with 2x SSC 0.5x SDS (2x 30 min) and 0.5x SSC 0.1x SDS (30 min) and exposed with autoradiography film (48 hours).

Nature Genetics: doi: 10.1038/ng.736

Fluorescence In situ Hybridization (FISH) of TE probes

FISH was performed on mitotic metaphase spreads prepared from meristem root tip cells of an Amelonado cocoa genotype as described by D'Hont et al[28]. The probes were labeled with Alexa-488 dUTP and Alexa-594 dUTP by random priming (Fisher Bioblock Scientific). The hybridization mixture (50 µL per slide) consisted of 50% formamide, 10% dextran sulphate, 2x SSC, 1% SDS and 2 µg/mL labeled probe. The slides were denatured in a solution of 70% formamide in 2× SSC at 80°C for 3 min. The denatured probe was placed on the slide and hybridization was performed overnight in a moist chamber at 37°C. After hybridization, slides were washed for 10 min in 2x SSC, 0.5x SSC, 0.1x SSC at 42°C. The slides were mounted in Vectashield antifade solution (Vector Laboratories) containing 2.5 µg/mL DAPI as counterstaining. The slides were examined with a Leica DMRAX2 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 Volocity software (Perkin Elmer).

Protein coding gene annotations

Transcriptome

Two EST collections were used to support the annotations.

• *454 EST from Criollo B97-61/B2 transcriptomes*

RNA was extracted from stems, mature and young leaves, vegetative buds, flowers, and floral cushions of the genotype B97-61/B2 according to Argout et *al.*[29]. cDNA was synthesized with the SMARTer™ PCR cDNA Synthesis Kit. The 454 sequencing generated 992 821 raw reads from which 715 457 cleaned reads were identified for assembly. To assemble the 454 data, we used a modified version of the ESTtik[30] pipeline based on TGICL[31]. We started by cleaning the raw sequences: (1) they were trimmed by removing vector sequence using the Vecscreen software against the Univec database and (2) low complexity sequences were masked using the Mdust program (http://compbio.dfci.harvard.edu/tgi/software/). Any reads composed of more than 85% low complexity regions were discarded. We compared cleaned reads to the comprehensive non-coding RNA sequence database fRNAdb v3.4[32] using BLASTN reads with an E-value below 1E-20 were discarded. Finally, we assembled reads longer than 120 bp using the TGICL software[31]. During the clustering step, two sequences were clustered together if their overlap length was above 60 bp with an identity percentage above 94%. Then, two sequences of a cluster were assembled if their overlap length was above 60 bp with an identity percentage above 95%. With this method, 38,737 contigs were assembled.

• *Sanger EST from different cocoa genotype transcriptomes*

We also used a previous collection of 149,650 ESTs, enriched in full length cDNA, and corresponding to 48,594 unique transcripts. These ESTs were sequenced using the Sanger method in the context of an international project and assembled with ESTtik[30] Among the 149,650 ESTs, 2,850 ESTs were produced from a pure homozygous Criollo originated from Belize, similar to the Criollo genotype B97-61/B2 and 47,800 ESTs were produced from hybrids between Criollo and Forastero genotypes.

Nature Genetics: doi: 10.1038/ng.736

Protein coding gene model predictions

Gene model predictions have been produced using the integrative gene prediction package EuGene[33]. It allows integrating various sources of information including statistical and similarity information. The full set of similarities against the scaffolds used in all *T. cacao* predictions includes:
Similarities to available *T. cacao* ESTs, obtained using GenomeThreader[34].
Similarities to proteins from SwissProt[35], TAIR[36], Malvacae Genbank extraction[37], with high confidence genes models of *Glycine max*[38] and peptides reconstructed from *T. cacao* ESTs using Prot4EST[39], searched using NCBI-BLASTN[27].
Similarities to EST from *A. thaliana, Gossypium, Vitis vinifera and Citrus* Genbank extraction[37] and *T. cacao* transcriptome, searched using NCBI-TBLASTX.

To train the statistical models for *T. cacao*, a set of high-confidence sequences was built as follows. A version of EuGene previously trained on *Medicago. truncatula* was applied using all previous similarities as evidence. Only predicted gene models that were fully covered by EST alignments and without any 'N' in their genomic sequence were retained. This resulted in a collection of 11,853 coding sequences (CDSs; total length of 11,732,053 bp), 41,844 intron sequences (13,624,230 bp), 6,991 3' UTR sequences (3,158,283 bp) and 5,639 5' UTR sequences (1,481,325 bp). Statistical models of DNA composition (Interpolated Markov Models) for *T. cacao* were trained on these regions.

To train SpliceMachine for translation start prediction, we extracted from this same set of fully EST-supported predictions, all regions around the predicted ATG that also corresponded to the alignment of the N-terminal region (20 AA) of a *Malvacae* or Swissprot protein sequence. Sequences containing 'N' were removed. Following redundancy filtering, we obtained a set of 317 positive examples. 10,000 negative examples were built from the reverse strand of the same regions. The ratio between the number of positive and negative examples is usual: each validated transcript provides just one positive example (the ATG) but its reverse complement usually contains many more occurrences of the ATG 3-mer defining negative examples.

Statistical models for splice sites were built from the spliced EST alignments obtained from GenomeThreader[34]. After redundancy filtering, 20,000 positive examples were extracted from these alignments and 200,000 negative examples were extracted from the opposite strand of the same regions. SpliceMachine was systematically trained using the same context size as for *M. truncatula* training. The ratio between positive and negative examples is expected for the same reason as above.

The EuGene combiner was then used to build a consensus *T. cacao* annotation integrating the previous statistical models and the previous similarities using the same evidence weighting as in the *M. truncatula*-trained version of EuGene. 50,582 genes were predicted.

Homology search and functional annotation

To identify putative homologies to known protein sequences, we performed BLASTP for each predicted coding sequence against the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases[35]. Based on three parameters: (1) Qcov (Query coverage = length high-scoring segment pair (HSP)/length query), (2) Scov (Subject coverage = length HSP/length subject) and (3) identity, we kept only the best result to assign a putative function to a polypeptide (Fig. 1). We have used a reciprocal best-hit-based search approach to identify putative

Nature Genetics: doi: 10.1038/ng.736

orthologs from different species (Supplementary Table 9). Additional information was obtained by protein-signature scanning. InterProScan was used for sequence comparison to the InterPro database[40].

Filtering of protein coding genes tagged as transposable element genes or as false positives

There is no perfect strategy or program either to predict genes either to separate plant genes from transposable element genes (TEG). So we made empiric choices, the first one was to not mask the sequence before predicting genes but rather to filter TEG afterward. We also knew that as in all prediction, there were false positives and false negatives; so we decided to apply empiric analyses with empiric thresholds to remove false positives. We did nothing for the false negatives. In the next versions of the genome, the filtering steps should be refined. The predicted genes filtering was done after the functional annotation and other analyses described here. Filters based on nucleic and protein sequence comparisons were all applied to the 50,582 predicted genes. Then, the genes that were never eliminated by any filter were kept to constitute the final set of 28,798 protein coding genes. 17,342 transposable element genes (TEG) were tagged as follow:

8,744 TEGs were tagged using BLASTP of the predicted polypeptides against Repbase [30 Jurka 2005] polypeptides (repbase1405_aaSeq_cleaned_TE.fa http://www.girinst.org/server/RepBase/index.php) with Qcov higher than 70%.

Of the 41,838 remaining genes, 330 were removed using Megablast of the predicted CDS against the *Copia*-like LTR retrotransposon Gaucho.

Of the 41,508 remaining genes, 4875 were discarded using Megablast of the predicted CDS against the 67,575 TEs annotated on the assembly (Table 1) or the 2,036 TE families identified in Supplementary Table 6 with Qcov higher than 70%.

Of the 36,633 remaining genes, 3,393 were removed based on keywords in the product and Interpo domains.

Of the 33240 remaining genes, 4439 genes were tagged as false positive among predicted polypeptides of length lower than 100 aa that have no similarity found either using BLASTP of the polypeptide against Swiss-Prot with a Qcov higher than 70% and identity higher than 30% or using TBLASTN against the *T. cocoa* EST contigs with a Qcov higher than 70% and identity higher than 70%.

Finally of the 28,801 remaining genes, three were removed because they were overlapping rRNA genes (manual checking).

Thus overall, 28,798 protein coding gene models were retained of which 23,529 were mapped on the 10 pseudomolecules.

Construction of families of homologous polypeptides and identification of cocoa subfamily-specific polypeptides

As a prerequisite to comparing gene content of *T. cacao* to other organisms at the whole-genome scale, we constructed families of homologous proteins from all sequences from cocoa and a wide phylogenetic range of eudicot organisms such as *A. thaliana, V. vinifera, Populus trichocarpa* and *Glycine max*.

We first removed highly similar paralogous genes using the CD-HIT algorithm[41]. Then, we performed an all-against-all comparison using BLASTP, and alignments with a Qcov and Scov lower than 80% and an identity lower than 30% were retained. Finally, BLAST results were fed into the stand alone OrthoMCL program using a default MCL inflation parameter of

Nature Genetics: doi: 10.1038/ng.736

1.5[42]. 103,580 of 147,507 protein sequences (70.2%) were clustered into 18,154 ortholog groups (Supplementary Table 9). 2,053 genes were clustered into 682 clusters specific to cocoa. A Venn diagram representing these data is shown in Figure 3.

In comparative genomics, BLAST is commonly applied to infer homology relationships between sets of genes from different organisms[43]. One widespread method is based on Best BLAST Mutual Hits (BBMH). If the two sequences have reciprocal relationships, then they are presumed to be the most similar to each other and are deemed putative orthologs. We used BBMH to define cocoa orthologs of genes in the other organisms (Supplementary Table 9). Gene ontology (GO) terms were assigned to the five proteomes using the Blast2GO software[44] (Figure S6-S7).

Non-coding gene annotations and target prediction

Theobroma cacao rRNA annotation

T. cacao rRNA genes were predicted by aligning the *A. thaliana* 25S, 18S and 5.8S rRNA against the scaffolds with BLASTN (Qcov above 0.8 and identity above 50%). There is one triplet of rRNAs (25S, 18S, 5.8S) on non-anchored scaffolds. A 5.8S rRNA gene and a 18S rRNA gene are colocalized on chromosome 7, whereas a 25S rRNA gene is localized on chromosome 4. tRNAscan-SE[45] was used to identify tRNA genes on the scaffolds and 472 genes were predicted. The number of rRNA genes identified in the assembly is likely to be greatly underestimated because of the lack of sequencing and assembly efficiency for the part of the genome that includes the repeated sequences when using data mainly composed of 454 shotgun reads.

Theobroma cacao microRNA annotation

Sequences of mature plant miRNAs were retrieved from miRBase release 14[46] and used as queries to search the *T. cacao* genome assembly using BLASTN. Hits with no more than one mismatch from a query were expanded to 150 nt upstream and 150 nt downstream and examined by MIRcheck[47]. miRNA candidates that were on the same arm of the hairpin as the known family members and passed MIRcheck with the parameters "-mir_bulge",3, "-ass", 2,"-unpair" were collapsed to retain a single miRNA for a given hairpin if length variants or position variants are present. The decision of which variant to retain was made as follows: for length variants, if the miRNA family was expressed in *A. thaliana* (according to the data in Ma et al.[48]), then the miRNA variant with the length of the most abundantly expressed miRNA was kept; if not, a 21-mer was favored. For positional variants, the miRNA variant with the greatest number of similar miRNA sequences in miRBase was retained.

A total of 83 *T. cacao* microRNAs (miRNAs) from 25 families were computationally predicted based on sequence similarity with known miRNAs in miRBase release 14 (Supplementary Table 10). The miRNA population size is reasonable compared to the number of miRNAs in other plant genomes in miRBase (Supplementary Figure 8), although our tally of *T. cacao* miRNAs is certainly an underestimate, as we were limited to identification by homology.

Because 25 *T. cacao* miRNA families were encoded by 83 loci, the number of paralogous loci per family was examined. Compared with *A. thaliana*, the cumulative distribution of *T. cacao* miRNAs was similar to the more-conserved (MC) subset of the *A. thaliana* miRNAs

Nature Genetics: doi: 10.1038/ng.736

(annotated outside of the *Brassicaceae* family in miRBase release 14), but quite different from the miRNA population in *A. thaliana* (Supplementary Figure 9). This is expected, because the miRNA prediction method finds only miRNAs conserved between *T. cacao* and another species.

Theobroma cacao microRNA target prediction

miRNA targets were predicted with the PERL script "axtell_targetfinder.pl" from the CleaveLand 2 package[49]. Randomization using 8300 randomly-shuffled miRNAs (100 times of the total number of real miRNAs) was also done using the same script. An average of 0.7 targets were predicted at a complementarity score of three for each randomized miRNA; this noise estimation increased to over two at a score of four. Therefore, a cutoff score of three was used for target prediction (higher target prediction scores indicate less complementarity).

Predicted targets were used as queries to search *A. thaliana* (TAIR9) and *Oryza sativa* coding sequences, and the top BLASTX hit with an E value <= $10E^{-4}$ to indicate the potential function of the predicted targets. 89 targets of 19 miRNA families were predicted, and 85 top hits (68 unique proteins) in *A. thaliana* and 83 top hits (66 unique) in *O. sativa* were identified by BLASTX search.

Because GO annotation was not yet available for *T. cacao*, GO annotation for the 68 unique *A. thaliana* genes homologous to the predicted *T. cacao* targets was used to search for GO term enrichment[50]. The terms "nucleus", "transcription factor activity" and "developmental processes" were the most significantly enriched terms in each GO category (Supplementary Table 11) for *T. cacao* miRNA target homologs in *A. thaliana* compared to the entire *A. thaliana* genome. These results are consistent with previous findings[51] that many conserved miRNA targets are transcription factors involved in developmental processes.

Identification of LRR-LRK genes in the *T. cacao* genome

The LRR-RLK receptors contain extracellular domains consisting of 1 to about 30 LRRs flanked by two cysteine pairs, a single-pass transmembrane domain (TM) and an intracellular serine/threonine protein kinase domain (KD)[52]. Based on this structural profile, we retrieved *T. cacao* (Tc) and *A. thaliana* (At) LRR-RLK genes in 3 steps using the proteome of *A. thaliana* (TAIR release 9: 33,200 sequences) from the TAIR website[53] and our predicted cacao proteome. First, we ran the hmmsearch program[54] to search for the LRR Hidden Markov Model (HMM) profile (PF00560)[55] in the 28,798 Tc protein sequences (23,529 mapped and 5,269 unmapped). On this set of sequences, we again used the hmmsearch program, this time seeking the kinase HMM profile (PF00069.16). Second, we extracted the kinase domain sequences of these proteins and aligned them with clustalw2 (default parameters)[56]. Finally, based on this alignment, we generated a phylogenetic tree by the maximum likelihood method with 100 bootstrap replicates (default parameters)[57]. We annotated tree leaves according to previous studies in order to classify the Tc sequences into one of the 19 subfamilies of LRR-LRK[58]. All manipulations on phylogenetic trees were performed with the treedyn[59] and treeview[60] programs. Some were performed on the "phylogeny.fr" web site[61].

Nature Genetics: doi: 10.1038/ng.736

Characterization of *T. cacao* genes orthologous to NBS-encoding genes

Genes coding for nucleotide-binding site (NBS) proteins play a determining role in resistance to pathogens and in the progression of the cell cycle[62,63]. This NBS-encoding gene family is rather abundant in plant genomes[64-66,62]. The NBS R-gene family is subdivided into different groups based on the structure of the N-terminal and C-terminal domain of the protein. The N-terminal domain either has a Coiled-Coil (CC) motif, a TIR (Toll Interleukin Receptor) motif or a sequence without obvious CC or TIR motifs. The C-terminal domain is either with or without a Leucine Rich-Repeat (LRR) motif[67]. We have identified and characterized the set of cocoa genes orthologous to R-gene-related NBS-encoding genes with (1) a description of the conserved domains of all the NBS proteins which were used to conduct phylogenetic analyses and (2) a distribution of NBS-encoding orthologous genes across pseudomolecules.

Classification of the predicted genes encoding NBS domains

Cocoa protein sequences, mapped and unmapped, were screened using hidden Markov models (HMM) to search for the Pfam NBS (NB-ARC) family PF00931 domain (E value cutoff of 1.0)[68] using hmmsearch version 3 [54]. Of 369 sequences that were manually cleaned, we kept 297 sequences. To detect TIR domains, the 297 predicted NBS-encoding amino acid sequences were screened using the HMM model Pfam TIR PF01582 (E value cutoff of 1.0). To detect LRR motifs, Pfam HMM searches using models for LRR_1 (PF00560), LRR_2 (PF07723) and LRR_3 (PF07725)[68] were used to screen predicted cocoa NBS-encoding amino acid sequences. CC motifs were detected using Paircoil2 with a P score cutoff of 0.025 [69].

NBS domain motif description

The NBS domain is characterized at least by five conserved motifs[70]. A specific consensus was deduced from the aligned sequences of all cocoa genes orthologous to NBS-encoding genes (Supplementary Figure 10). The consensus sequence of the P-loop motif, xGxGGxGKT(T/A)Lxx, was found in the majority of the cocoa genes orthologous to NBS-encoding genes except for eight predicted genes.

The consensus sequence of the kinase-2 motif, KxxLLVLDDVWxx, was found in 86% of all genes orthologous to NBS-encoding genes with a tryptophane (W) which is most often associated with CC-NBS proteins. The kinase-3 consensus sequence (xGsKxxxTTRxxx), the putative membrane-spanning motif (xCxGLPLAxxxx) with the consensus GLPL and the consensus MHDL motif, (xxxMHDLxxD), were found respectively in 90%, 85% and 70% of all NBS-encoding cocoa orthologous genes.

Total number and organization of T. cacao genes orthologous to NBS-encoding genes

Among 28,798 automatically annotated genes, a total of 297 non redundant genes orthologous to NBS-encoding genes were identified and manually verified (Supplementary Figure 10). The cocoa NBS-encoding orthologous gene family accounts for approximately 0.9% of total predicted genes. This value is similar to that of other eudicot plants: 0.7% for *A. thaliana*, 1% for *P. trichocarpa*, 1.2% for *M. truncatula* and 1.8% for *V. vinifera*[64-66]. The distribution of the number of genes according to the motifs framing the NBS domain for cocoa and these four other plant genomes is presented in Supplementary Table 14[71,64,66]. The TIR-NBS-LRR and TIR-NBS orthologous genes are under represented in the cocoa genome compared to other plants.

Nature Genetics: doi: 10.1038/ng.736

Phylogenetic analysis of NBS domains

The 297 non redundant NBS orthologous genes from *T. cacao* and eight representative *A. thaliana* NBS-encoding genes[72] were further studied by comparison to At-NBS (AT3G44670, AT4G26090, AT1G12220, AT3G07040, AT3G46530, AT5G43470, AT4G33300, AT5G45510) sequences obtained from the NIBLRRS Project website[73]. The sequences were aligned using MAFFT[74]. We applied a masking procedure to the optimized alignment to detect and remove amino acid columns/positions containing either no or a low phylogenetic signal. Our workflow uses a modified version of the AL2CO software for calculation of positional conservation[75]. The amino acid positions retained for phylogenetic constructions share a minimum conservation index of 2 together with a percentage of gaps below 50%. A phylogenetic tree was constructed using the PHYML[76] with bootstrap multiple alignment resampling set at 400. PHYML first constructs a BioNJ tree using the Neighbor-Joining tree algorithm and then optimizes this tree to improve its likelihood by successive iteration. All manipulations on phylogenetic trees were performed with the treedyn[59] program.

The phylogenetic distribution of non-TIR-NBS-encoding orthologous genes indicates that one-third of cocoa genes are structured around those of *A. thaliana* while the other two-thirds are divided into five cocoa-specific expansions, (Supplementary Fig. 11) including two major subgroups. According to the classification of *Arabidopsis* NBS-At genes by Meyers et al. 2003[72], the cocoa genes organized around class CNL-B (AT1G12220 and AT4G26090) belong mainly to the cluster of pseudomolecule 6. Those that are organized around class CNL-A/C/D belong mainly to the two first clusters of the pseudomolecule 10. A group that contained four NBS-LRR orthologous genes from the first cluster of the pseudomolecule 2 was identified from the remainder of cocoa NBS-encoding orthologous genes.

Identification of NPR1 genes in the cocoa genome

Plants have evolved a complex network of defense responses, often associated with a response local to the site of infection[77]. In addition, defenses are also systemically induced in remote parts of the plant in a process known as systemic acquired resistance (SAR). Multiple studies in both monocots and dicots have shown that salicylic acid (SA) plays a central role as a signaling molecule in SAR. NPR1 (Nonexpressor of pathogenesis-related 1), a central mediator of the plant defense response, was originally identified by screening for *npr1* mutants that were insensitive to SA[78]. It is believed that NPR1 also plays a role in the jasmonic acid (JA) signaling pathway and mediates the crosstalk between the SA and JA defense pathways to fine tune defense responses[79]. *NPR1* encodes a protein containing ankyrin repeats and a BTB/POZ domain, both of which mediate protein-protein interactions in animals[80]. *NPR1* is constitutively expressed, and NPR1 protein is present as inactive oligomers in the cytoplasm of the cell. Upon SAR induction, the redox state of the cell is altered, resulting in the reduction of NPR1 to its active monomeric form, which moves into the nucleus where it can regulate defense gene transcription via interactions with TGA transcription factors[81,82].

The *NPR* gene family of *Arabidopsis* consists of *NPR1* and five *NPR1*-like genes encoding proteins with significant similarity to NPR1, named *NPR1-like 2* (*NPR2*), *NPR3, NPR4, BLADE-ON-PETIOLE2 (BOP2*; also named *NPR5*), and *BOP1* (also named *NPR6*)[83]. These can be grouped into three subfamilies based on phylogenetic analysis (NPR1/2, NPR3/4 and

Nature Genetics: doi: 10.1038/ng.736

BOP1/2). BOP2 and BOP1 have functionally redundant roles in the regulation of symmetry during leaf morphogenesis and abscission[84]. The functions of NPR3 and NPR4 have been suggested to involve repression and/or activation of the plant defense response pathway[85,83]. Recent work in the Guiltinan laboratory strongly suggests that *Arabidopsis* NPR3 acts as a repressor of defense responses during flower development (M. Guiltinan, unpublished data). A growing body of evidence has revealed that the salicylic acid-dependent, NPR1-mediated defense pathway is also conserved in other plant species across wide phylogenetic distances including orthologs in grapevine[86], tomato[87], apple[88] wheat[89] and rice[89]. Thus, it appears that the mechanisms of SA-dependent, NPR1-mediated defense responses likely evolved very early in the emergence of the plant kingdom.

As key regulators of the defense pathway in plants, the *NPR1* gene family is of potential importance for breeding of disease resistant varieties. To characterize the *NPR1* orthologous family in *T. cacao*, the cocoa genome sequence was searched using BLAST with each of the 6 *Arabidopsis NPR1* family member gene sequences as queries. Full length protein sequences of all six *Arabidopsis NPR1* gene family members were used to search the *T. cacao* genome assembly V1.0 database using the TBLASTN[27] program with an E-value cutoff of 1E-40. Four cacao genes were identified with e-values below 5E-41. The next closest hit had an e-value of 1E-15 and was not considered a bona fide NPR1 family member. Using the TBLASTN program, a full length protein sequence of *Arabidopsis* NPR1 was used to search the Phytozome database (http://www.phytozome.net/) to obtain NPR-like genes from poplar, Medicago and grape (e-value cutoff of 1E-20). Six NPR1 family member genes were identified in *P. trichocarpa* (Poplar), four in *M. truncatula* (medicago) and three in *V. vinifera* (grape). A phylogenetic tree was constructed (Supplementary Fig. 12) using coding sequences from each gene to evaluate the genetic relatedness of the cocoa NPR1 orthologous family members to those from other species. The cocoa genome contains a single orthologous gene in the NPR1/2 and BOP subfamilies and two orthologous genes in the NPR3/4 subfamily.

Genome distribution of *T. cacao* genes orthologous to NBS, LRR-LRK and NPR1-like genes and comparative mapping with QTLs related to disease resistance in *T. cacao*

The distribution of all defense-related genes orthologous to NBS, LRR-LRK and NPR1-like genes is represented in Supplementary Fig. 13. Among the 297 unique NBS-encoding orthologous genes, 237 were mapped on the cocoa pseudomolecules. The genes orthologous to NBS-encoding genes were distributed across the ten chromosomes in 46 singletons and 41 clusters comprising between 2 and 17 tightly linked genes. Similar results were observed for sunflower[90], cucumber[91] and poplar[92] in which 75% of the NBS genes are located within clusters, indicating that they have evolved through tandem duplications, similar to the situation in other known plant genomes. Of the four NPR1-like genes in cacao, three were mapped on the pseudomolecules (PM 5, 6 and 9) and one is associated with the unassembled sequences.

A meta-analysis of 76 QTLs related to disease resistance identified in *T. cacao* in 16 different experiments was recently made by Lanaud et al.[93]. Their genetic localization was compared with the distribution of NBS, LRR-LRK and NPR1-like orthologous cocoa genes on the pseudomolecules using the Spidermap Software (JF Rami non published data). (Supplementary Fig. 13).

Nature Genetics: doi: 10.1038/ng.736

Considering an average confidence interval of about 20 cM for the QTLs identified[93] for nearly all QTLs it is possible to find a corresponding genome region containing NBS, or LRR-LRK genes. This comparative mapping provided a large number of candidate resistance genes, potentially underlying the QTLs of resistance already identified, and which have to be confirmed by complementary functional approaches. Of the NPR1 orthologous genes, only a single co-localization was found between a gene most closely related to NPR1 and a QTL for resistance to witches' broom disease (Tc09t007660 on PM9).

Genome distribution of lipid and flavonoid orthologous genes and comparative mapping with QTLs for traits related to fat and flavonoids

QTLs for butter fat content and hardness were detected in three studies conducted by Lanaud et al.[93], Araujo et al.[94] and Alvarez et al.[95]. QTLs for flavonoid content (epicatechin and procyanidins) were identified by Alvarez et al.[95] in a Venezuelian cocoa population corresponding to hybrid Criollo types. QTLs for cotyledon, leaves, staminode, sepal and fruit colors were identified by Marcano et al.[96]. Astringency is known to be related to certain procyanidin compounds such as tannins. Several QTLs linked to chocolate astringency were previously identified by Lanaud et al.[93] and are also reported in this analysis.

The QTLs projection on a same consensus map, suitable for QTL comparisons was made using a similar strategy, and with the same consensus map as those used by Lanaud et al.[93] for the meta-analysis of QTLs of resistance (Supplementary Fig. 15). Of the ten QTLs for cocoa butter quality, seven are located close to genes in the pathway. For example, the strongest QTL for fat content localized to the bottom arm of LG9 shows a close localization to a cluster of genes orthologous to KCS, KASIII and FAD3 genes on PM9 (Supplementary Table 14).

Of the 18 QTLs for flavonoid traits, 11 showed co-localization to genes orthologous to key genes in the pathway. A QTL for astringency located near the top of LG1 is closely associated with three orthologous genes in the pathway located on PM1. QTLs for epicatechin and procyanidin dimers reside on LG3, very close to a gene for LAR, a key enzyme leading to formation of the flavan-3-ols, including catechin (Supplementary Table 15). On LG4, a QTL for organ coloration (purple anthocyanins) is localized very close to an othologous gene encoding OMT, the first committed enzyme in the anthocyanin pathway. A similar pigment QTL on LG6 is co-localized with a gene for ANR on PM6. This gene product acts one step after the branch point to anthocyanins, and reduced expression of this gene would be expected to correlate with darker coloration.

Cacao genome synteny, duplication, evolution and paleohistory.

Arabidopsis, grape, poplar, soybean, papaya sequence databases.

Genome sequences of *Arabidopsis* (5 chromosomes - 33,198 genes - 119 Mb), grape (19 chromosomes - 21,189 genes - 302 Mb), poplar (19 chromosomes - 30,260 genes - 294 Mb), soybean (20 chromosomes - 46,194 genes - 949 Mb), and papaya (9 chromosomes - 19,205 genes – 234 Mb) were used as described in Salse *et al*.[97].

Nature Genetics: doi: 10.1038/ng.736

Synteny and duplication analysis.

Three new parameters were recently defined in Salse *et al.*[97] to increase the stringency and significance of BLAST sequence alignment by parsing BLASTP results and rebuilding HSPs (High Scoring Pairs) or pairwise sequence alignments to identify accurate paralogous and orthologous relationships.

Distribution of K_S distances (MYA scale) for paralogous and orthologous gene pairs

We performed sequence divergence and speciation event datation analysis based on the rate of non synonymous (*Ka*) vs. synonymous (*Ks*) substitutions calculated with PAML (Phylogenetic Analysis by Maximum Likelihood)[98]. We used average substitution rate (r) of 6.5×10^{-9} substitutions per synonymous site per year for grasses in order to calibrate the ages of the gene under consideration[99,100]. The time (*T*) since gene insertion was then estimated using the formula $T = Ks$ /r.

Nature Genetics: doi: 10.1038/ng.736

Supplementary Tables

Supplementary Table 1. Characteristics and quality of BAC libraries of *Theobroma cacao* var. Criollo

Supplementary Table 1. Characteristics of BAC libraries from *T. cacao* var. Criollo

Species	DNA fragmentation	Total number of clones	Total number of plates	% non insert containing clones	% chloroplast clones	Avg insert size (kb)	Estimated genome coverage
T. cacao	*Hind*III	18,432	48	0	0	135	5.14x
T. cacao	*Eco*RI	25,344	66	0	0	137	8.04x

Supplementary Table 2. Raw sequencing data overview.

	Number of reads	Number of bases	Coverage	Insert size (bp)
Roche/454 Single reads	17,615,336	5,665,734,388	13.2x	NA
Roche/454 Mate-pairs reads	8,819,944	1,398,260,416	3.3x	8,000
Sanger BAC ends	84,547	71,705,251	0.2x	136,000
Illumina Paired-end reads	397,959,108	19,102,037,184	44.4x	200

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 3. Nuclear DNA content (2C) and genome size of 27 cocoa clones (2n = 2x = 20) or related taxa.

Species	Accession name	Number of measures	2C DNA content (pg)	SD (pg)	Bonferroni's grouping (P = 0.05) [a]	Genome size (Mbp/1C)
Theobroma cacao (ranked by size of genome)	B97 61/B-2	5	0.88	0.023	*abc*	430
	CATONGO	3	0.93	0.001	*bd*	455
	ICS 100	4	0.89	0.016	*bc*	435
	ICS 95	3	0.91	0.031	*b*	445
	IMC 78	3	0.97	0.013	*e*	474
	LAF1	2	0.84	0.007	*a*	411
	LCTEEN 162/S 10-10	3	0.99	0.021	*ef*	484
	LCTEEN 255	5	0.97	0.025	*ef*	474
	LCTEEN 413	4	1.01	0.057	*efg*	494
	LCTEEN 82	3	0.94	0.007	*be*	460
	LCTEEN 83	3	0.95	0.005	*be*	465
	LCTEEN 85	4	0.94	0.01	*bd*	460
	Matina 1-7	3	0.93	0.002	*bd*	455
	Na 226	2	0.94	0.0001	*be*	460
	OC 77	2	0.91	0.02	*bcdh*	445
	SPA 5	2	0.92	0.012	*bcdg*	450
Theobroma grandiflora		3	1.02	0.06	*efgh*	499
Theobroma kanukensis		5	0.76	0.019	*j*	372
Theobroma microcarpa		5	0.73	0.005	*k*	357
Theobroma speciosa		3	1.02	0.022	*e*	499
Herrania albiflora		3	0.69	0.009	*i*	337
Herrania balaensis		3	0.82	0.012	*a*	401
Herrania breviligulata		3	0.8	0.027	*a*	391
Herrania camargoana		3	1.05	0.015	*g*	513
Herrania nitida		4	0.7	0.009	*i*	342
Cola nitida		3	4.86	0.039	*l*	2377

[a] 2C content is not significantly different within each class identified by the same letter following Bonferroni's system.

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 4. Cocoa genome assembly overview.
N50, N80, N90 refer to the size (or number) above which 50%, 80% and 90% of the total length of the sequence assembly, respectively, can be found

	Contigs	Scaffolds
Number	25,912	4,792
Cumulative size (Mb)	291.4	326.9
Average Size (Kb)	11.2	68.2
N50 size (Kb)	19.8	473.8
N50 number	4,097	178
N80 size (Kb)	8.0	143.9
N80 number	11,043	542
N90 size (Kb)	4.8	75.5
N90 number	15,723	854
Largest size (Kb)	190	3,415

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 5: Overview of the anchoring of the assembly on the cocoa linkage groups.

Linkage group (LG) or pseudomolecule (PM)	Size (cM/Mbp)	Number of markers with positive blast	Number of scaffolds	Number of anchored and oriented scaffolds	Gene number per pseudomolecule
LG1 **PM1**	94.1 31.27	191	46	34	3588
LG2 **PM2**	101.1 27.75	146	55	31	2879
LG3 **PM3**	76.9 25.47	158	49	23	2664
LG4 **PM4**	64.2 23.50	126	39	21	2627
LG5 **PM5**	78.1 25.66	141	46	19	2623
LG6 **PM6**	64 15.48	77	29	18	1872
LG7 **PM7**	52.6 14.17	57	27	12	1417
LG8 **PM8**	59.2 11.53	70	16	12	1488
LG9 **PM9**	100.9 28.46	171	49	25	3017
LG10 **PM10**	59.5 15.16	55	29	11	1354
total number		**1192**	**385**	**206**	**23529**
total length	**750.6 cM / 218.45 Mbp**		**218.4 Mbp**	**162.8 Mbp**	
Total not anchored	**108.89 Mbp**		**4407**		**5269**

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 6: Transposable elements detected in the *T. cacao* genome.

Element	Number of families	Number of elements
Class I		
Copia	290	18,060
Gypsy	159	12,622
non classified	198	19,260
Class II		
tranposons	36	7,284
MITEs	1353	14,598

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 7: Relative copy number of *ThCen* and *Gaucho* repeated sequences in the cocoa genome.

The relative copy number of *TcCen* and *Gaucho* repeated sequences was evaluated, in comparison with B97-61/B-2, in a panel of *T. cacao* genotypes differing in genome size. The copy number was estimated by the hybridization signal intensities on Southern blots on which *Hind* III-restricted DNA from each cocoa accession was hybridized using *TcCen* and *Gaucho* probes. The relative signal intensities were evaluated by Image Quant in comparison with B97-61/B-2 hybridization signals after normalization by DNA concentrations. The relative DNA amounts compared to those of B97-61/B-2 were also estimated by Image Quant in an agarose gel image after electrophoresis of DNA restricted by *Hind*III restriction enzyme and BET staining.

Genotypes	**relative DNA amount**	**genome size (Mb)**	**relative *TcCen* copies**	**relative *Gaucho* copies**
Theobroma cacao genotypes				
LAN2	1.08	411	1.46	0.61
LAF1	0.78	418	1.51	1.17
B 97 61/B-2	1.00	430	1.00	1.00
UF676	2.20	434	1.88	0.64
ICS 100	1.31	435	1.57	0.89
ICS 95	1.24	445	1.83	0.71
CATONGO	1.62	455	1.95	0.60
MATINA 1/7	1.37	455	1.97	0.49
LCTEEN 82	1.13	460	2.31	0.67
LCTEEN 85	1.09	460	2.33	0.71
UPA 402	1.12	461	1.96	0.53
LCTEEN 83	0.86	465	2.44	0.57
IMC 78	0.96	474	1.98	0.62
LCTEEN 162/S 10-10	0.90	484	2.50	0.48
LCTEEN 413	1.26	494	1.52	0.69
minimum *T. cacao* value		411	1.00	0.48
maximum *T. cacao* value		494	2.50	1.17

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 8: Features of *Theobroma cacao* genes in comparison with *Arabidopsis thaliana* and *Vitis vinifera*.

	Theobroma cacao	*Arabidopsis thaliana* (TAIR9)	*Vitis vinifera* (8X)
Assembled chromosomes (bp)	218,466,233	119,146,348	303,085,820
Unanchored assembled sequences (bp)	108,886,888	-	194,422,951
GC content (%)	34.38		
Protein coding genes	28,798	27,379[1]	30,434
Mean gene size with UTR (bp)	3,346	2,183	7,413
Median gene size with UTR (bp)	2,582	1,898	3,398
Mean gene density per 100kb	10.19	22.64	6.39
Coding exons	144,998	138,883	149,351
Mean coding exons per gene	5.03	5.07	4.90
Mean coding exon size (bp)	231	237	224
Median coding exon size (bp)	133	133	129
Mean intergenic region (bp)	6,319	2,187	7,918
Median intergenic region (bp)	2,130	888	3,136

[1] For gene comprising predicted alternative splice variants, the first (.1) representative has been selected.

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 9: Summary of gene family clustering. Description of clusters of orthologous (or paralogous) genes obtained after applying OrthoMCL. For BBMH (Best BLAST Mutual Hits), we report also the number of cocoa genes with a reciprocal best-hit relationship with the organism. Numbers in parentheses indicate the percent of BBMH with cocoa.

Species	# genes	# genes after cdhit	# genes in families	# groups	# groups specific	# BBMH / cocoa
T. cacao	28,801	28,219	18,595	12,954	682	
A. thaliana	33,200	26,809	20,672	12,103	1047	13,086 (46,4%)
V. vinifera	21,189	19,919	12,565	9,363	247	11,245 (39,8%)
P. trichocarpa	45,778	36,348	24,374	13,419	960	15,156 (53,7%)
G. max	55,787	36,112	27,374	13,152	1148	14,051 (49,8%)
Total	184,755	147,407	103,580	18,154		

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 10: miRNA families found in *Theobroma cacao*.

miRNA family	Number of paralogous loci	Number of plant species where also found	List of species
156	7	17	aqc, ath, bdi, bna, ghr, gma, mtr, osa, ppt, pta, ptc, sbi, sly, smo, sof, vvi, zma
160	3	15	aqc, ath, bdi, bra, gma, mtr, osa, ppt, ptc, sbi, sly, smo, tae, vvi, zma
162	1	10	ath, cpa, ghr, gma, mtr, osa, ptc, sly, vvi, zma
164	3	11	ath, bna, bra, gma, mtr, osa, ptc, sbi, tae, vvi, zma
166	4	17	aqc, ath, bdi, bna, ghr, gma, mtr, osa, ppt, pta, ptc, pvu, sbi, sly, smo, vvi, zma
167	3	17	aqc, ath, bdi, bna, bra, gma, lja, mtr, osa, ppt, ptc, sbi, sly, sof, tae, vvi, zma
168	1	11	aqc, ath, bna, gma, mtr, osa, ptc, sbi, sof, vvi, zma
169	14	13	aqc, ath, bdi, bna, ghb, gma, mtr, osa, ptc, sbi, sly, vvi, zma
171	8	18	aqc, ath, bdi, bna, bol, bra, gma, mtr, osa, ppt, pta, ptc, sbi, sly, smo, tae, vvi, zma
172	5	13	aqc, ath, bdi, bol, bra, gma, mtr, osa, ptc, sbi, sly, vvi, zma
319	1	14	aqc, ath, gma, mtr, osa, ppt, pta, ptc, pvu, sbi, sly, smo, vvi, zma
390	2	11	ath, bna, ghr, gma, mtr, osa, ppt, pta, ptc, sbi, vvi
393	2	9	ath, bna, gma, mtr, osa, ptc, sbi, vvi, zma
394	2	6	ath, osa, ptc, sbi, vvi, zma
395	2	10	aqc, ath, mtr, osa, ppt, ptc, sbi, sly, vvi, zma
396	5	15	aqc, ath, bna, ghr, gma, lja, mtr, osa, pta, ptc, sbi, smo, sof, vvi, zma
397	1	8	ath, bdi, bna, osa, ptc, sbi, sly, vvi
398	2	9	aqc, ath, bol, gma, mtr, osa, pta, ptc, vvi
399	9	14	aqc, ath, bdi, bna, ghr, mtr, osa, ptc, pvu, sbi, sly, tae, vvi, zma
403	2	3	ath, ptc, vvi
529	1	4	aqc, osa, ppt, sbi
530	2	3	aqc, osa, ptc
535	1	4	aqc, osa, ppt, vvi
827	1	3	ath, osa, ptc
2111	1	2	ath, bna

aqc: *Aquilegia coerulea,* ath: *Arabidopsis thaliana* , bdi: *Brachypodium distachyon* , bna: *Brassica napus* , bol: *Brassica oleracea* , bra: *Brassica rapa* , cpa: *Carica papaya* , ghb: *Gossypium herbecium* , ghr: *Gossypium hirsutum* , gma: *Glycine max* , lja: *Lotus japonicus* , mtr: *Medicago truncatula* , osa: *Oryza sativa* , ppt: *Physcomitrella patens* , pta: *Pinus taeda* , ptc: *Populus trichocarpa* , pvu: *Phaseolus vulgaris* , sbi: *Sorghum bicolor* , sly: *Solanum lycopersicum* , smo: *Selaginella moellendorffii* , sof: *Saccharum officinarum* , tae: *Triticum aestivum* , vvi: *Vitis vinifera* , zma:

Supplementary Table 11. Gene ontology (GO) annotation of cacao miRNA target homologs in *A. thaliana*. Terms in bold indicates the most significant enrichment in each GO category

term	ontology category	gene number of cacao miRNA target homologs in *A. thaliana* (total 67)	gene number in *A. thaliana* genome (total 34278)	p-value
Nucleus	cellular component	24	2609	6,97E-12
Extracellular	cellular component	7	441	2,37E-06
Transcription factor activity	molecular function	23	1679	3,93E-15
Other enzyme activity	molecular function	17	3345	5,30E-05
DNA or RNA binding	molecular function	14	2714	1,93E-04
Transporter activity	molecular function	7	1242	2,85E-03
Developmental processes	biological process	24	2006	2,01E-14
Transcription	biological process	20	1709	5,67E-12
Other cellular processes	biological process	43	10140	1,08E-09
Other metabolic processes	biological process	41	9410	1,77E-09
Other biological processes	biological process	12	1913	7,32E-05

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 12. Number of LRR-RLK genes (or orthologous genes) in each of the 19 subfamilies in *Arabidopsis thaliana* (At), *Theobroma cacao* (Tc) and *Populus trichocarpa* (Pt).

Subfamily	At	Tc	Pt*
LRR-I	44	12	19
LRR-II	14	10	20
LRR-III	46	37	63
LRR-IV	3	3	8
LRR-V	9	6	11
LRR-VI-1	5	5	7
LRR-VI-2	5	4	10
LRR-VII	8	6	12
LRR-VIII-1	8	7	15
LRR-VIII-2	12	19	50
LRR-IX	4	5	12
LRR-Xa	7	7	25
LRR-Xb	6	5	22
LRR-XI	32	54	54
LRR-XII	8	63	90
LRR-XIIIa	3	2	4
LRR-XIIIb	3	2	4
LRR-XIV	2	2	6
LRR-XV	2	4	4
Total	221	253	436

*From Lehti-Shiu et al. 2009

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 13. Classification of *Theobroma cacao* orthologous genes into one of the 19 LRR-RLK subfamilies. If available, one representative member of *Arabidopsis thaliana* gene is cited per subfamily.

Subfamily	***T. cacao* accession numbers**	***Arabidopsis* gene names of one representative member per subfamily**
LRR-I	Tc00g002950 Tc00g003020 Tc00g042320 Tc00g002900 Tc06g007080* Tc06g007100* Tc06g007020* Tc09g008480 Tc02g025930 Tc05g005850 Tc01g007500 Tc01g022480	LRRPK (light-repressible receptor protein kinase)
LRR-II	Tc00g050290 Tc02g012140 Tc01g008780 Tc01g013050 Tc02g030940 Tc04g015680 Tc02g030920 Tc09g014280 Tc02g024160 Tc06g014110	AtSERK1 (SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE 1)
LRR-III	Tc01g005680 Tc02g009430 Tc05g018150 Tc02g026770 Tc04g013850 Tc02g019380 Tc02g020570 Tc09g004910 Tc05g008010 Tc09g010460 Tc00g057690 Tc03g006900 Tc04g016010 Tc09g035300 Tc10g016710 Tc03g028420 Tc03g017680 Tc09g006650 Tc06g010180 Tc06g005910 Tc00g034880 Tc09g000010 Tc02g011910 Tc03g000770 Tc04g000610 Tc07g002050 Tc02g030240 Tc08g000060 Tc01g005520 Tc01g022920 Tc05g001620 Tc10g000820 Tc06g000470 Tc09g033140 Tc01g009920 Tc08g003050 Tc01g002030	
LRR-IV	Tc01g037760 Tc03g018540 Tc03g025890	
LRR-V	Tc06g010890 Tc00g016770 Tc06g019780 Tc00g090980 Tc06g016520 Tc00g028150	SCM-SUB
LRR-VI-1	Tc04g002970 Tc03g014610 Tc02g033910 Tc05g023650 Tc10g010660	
LRR-VI-2	Tc03g027220 Tc06g014220 Tc09g034340 Tc10g001090	
LRR-VII	Tc09g002810 Tc06g014170 Tc03g026200 Tc04g018460 Tc00g061270 Tc04g016210	
LRR-VIII-1	Tc04g015880 Tc09g035330 Tc04g015890 Tc00g020540 Tc00g006550 Tc00g031660 Tc00g031610	
LRR-VIII-2	Tc06g013680 Tc06g013690 Tc06g013660 Tc06g013650 Tc06g013700 Tc06g013710 Tc01g014070 Tc07g010630 Tc07g010680 Tc07g010700 Tc07g010770 Tc07g010730 Tc06g013640 Tc07g014090 Tc06g011190 Tc06g011240 Tc06g011260 Tc06g011210 Tc06g011180	
LRR-IX	Tc02g029860 Tc04g005810 Tc04g020290 Tc00g005590 Tc00g012460	
LRR-Xa	Tc00g075310 Tc04g008190 Tc00g054780 Tc01g020480	BIR1 (BAK1-interacting like kinase 1)
LRR-Xb	Tc03g010530 Tc07g000200 Tc01g010390 Tc02g029320 Tc03g019480 Tc00g055300 Tc02g030270 Tc07g008390	BRI1 (BRASSINOSTEROID INSENSITIVE 1)

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 13. (continued)

LRR-XI	Tc02g000140 Tc01g026140 Tc05g025870 Tc03g017040 Tc08g003840 Tc01g007220 Tc02g000040 Tc02g001640 Tc07g000640 Tc03g002110 Tc04g021800 Tc01g022970 Tc00g052790 Tc06g017670 Tc05g025560 Tc04g021990 Tc02g002570 Tc08g006630 Tc08g006640 Tc08g015570 Tc06g015260 Tc02g016020 Tc01g007200 Tc09g008880 Tc09g008890 Tc09g004330 Tc05g018030 Tc08g014220 Tc05g016360 Tc06g002650 Tc03g021590 Tc01g025700 Tc01g025760 Tc10g014070 Tc10g014090 Tc10g014230 Tc00g040550 Tc08g012170 Tc00g061190 Tc00g040450 Tc00g040530 Tc00g040440 Tc00g040490 Tc09g002700 Tc09g002720 Tc08g009140 Tc04g026560 Tc08g009130 Tc03g017130 Tc00g007810 Tc01g038750 Tc01g038760 Tc01g038730 Tc01g038770	CLV1 (CLAVATA 1)
LRR-XII	Tc06g013030 Tc10g001680* Tc05g004050* Tc08g007960* Tc09g014550* Tc06g004870* Tc06g004910* Tc03g016440* Tc03g016460* Tc07g009880* Tc00g035980* Tc00g075630* Tc00g001540* Tc00g004260* Tc06g000790* Tc06g013970* Tc06g014130* Tc00g058390* Tc00g081090* Tc07g010620* Tc07g010550* Tc07g010600* Tc07g010590* Tc07g010520* Tc07g011460* Tc00g080960* Tc00g081000* Tc00g062690* Tc00g058340* Tc00g050640* Tc00g081210* Tc00g081080* Tc00g081100* Tc00g081190* Tc00g050580* Tc00g062650* Tc00g081070* Tc00g058300* Tc00g050570* Tc10g008830* Tc07g004700* Tc04g022010 Tc04g022030 Tc04g022000 Tc10g001940 Tc10g001600 Tc10g001950 Tc10g001590 Tc10g001980 Tc10g001970 Tc10g001930 Tc10g001610 Tc05g026830 Tc10g001660 Tc10g011300 Tc05g016100 Tc05g022420 Tc10g001670 Tc04g015500 Tc05g002710 Tc09g030110 Tc05g002730 Tc10g001630	FLS2 (FLAGELLIN-SENSITIVE 2)
LRR-XIIIa	Tc03g031220 Tc03g009500	FEI1 (named for the Chinese word for fat)
LRR-XIIIb	Tc09g005700 Tc03g018800	ER (ERECTA)
LRR-XIV	Tc02g006980 Tc08g002130	
LRR-XV	Tc02g032000 Tc04g010660 Tc01g033210 Tc04g005310	TOAD2-RPK2
unclassified	Tc09g003290 Tc09g004160 Tc03g013850 Tc01g040010 Tc04g017660 Tc00g086930	EVR (EVERSHED)

* putative member (unsupported by bootstrap value)

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 14: Numbers of orthologous genes found in *Theobroma cacao* (Tc), *Populus trichocarpa* (Pt), *Vitis vinifera* (Vv), *Medicago truncatula* (Mt), and number of genes in *Arabidopsis thaliana* (At) that encode NBS domains similar to those in plant R proteins.

	Tc	Pt	Vv	Mt	At
TIR-NBS-LRR	8	78	97	118	93
CC-NBS-LRR	82	120	203	152	51
NBS-LRR	104	132	159	-	3
NBS	53	62	36	328	1
CC-NBS	46	14	26	25	5
TIR-NBS	4	10	14	38	21
Total NBS-LRRgenes	194	330	459	-	147
Total NBS genes	297	416	535	661	174

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 15. List of *Theobroma cacao* genes orthologous to encoding key enzymes in the storage lipid biosynthesis pathway. Gene copy numbers and full length protein sequences for *Arabidopsis* were obtained from the Arabidopsis Lipid Gene Database (Mekhedov) (http://lipids.plantbiology.msu.edu/). Full length *Arabidopsis* protein sequences from all 67 *Arabidopsis* genes in the database were used to query the *T. cacao* assembly V1.0 genome database using the TBLASTN program. An E-value cutoff of $1*e^{-24}$ was used for all genes except for the acyl carrier protein gene family, for which an e-value cutoff of $1*e^{-12}$ was used because of its short length (137 amino acids). For enzymes with multiple gene copies in *Arabidopsis*, full length protein sequences of each copy were used to query the cacao genome and a non-redundant set of all hits was listed. Standard gene designation (Gene), enzyme activity (Enzyme), Gene copy numbers and locus numbers for each predicted cacao gene in the *T. cacao* assembly V1.0 genome database are indicated.

Gene	Enzyme	Gene Copy Number		Cacao Locus Numbers			
		Arabidopsis	Cacao				
ACC2	Homomeric Acetyl-CoA Carboxylase	1	1	Tc08g009450			
CAC2	Heteromeric acetyl-CoA carboxylase BC subunit	1	1	Tc00g000210			
BCCP (CAC1A)	Heteromeric acetyl-CoA carboxylase BCCP subunit	2	3	Tc05g019490	Tc04g010240	Tc03g011680	
CAC3	Heteromeric acetyl-CoA carboxylase alpha-CT subunit	1	2	Tc01g036350	Tc05g003050		
ACCD	Heteromeric acetyl-CoA carboxylase beta-CT subunit	1	1	Tc04g003850			
MAT	Plastidial Malonyl-CoA : ACP Malonyltransferase	1	1	Tc09g034600			
KAS I	Ketoacyl-ACP synthase I	1	2	Tc06g012880	Tc03g029270		
KAS II	Ketoacyl-ACP synthase II	1	3	Tc09g006480	Tc06g010360	Tc06g010370	
KAS III	Ketoacyl-ACP synthase III	1	1	Tc03g025440			
KAR	Plastidial Ketoacyl-ACP Reductase	5	3	Tc02g026480	Tc00g069670	Tc04g004190	
	Plastidial Hydroxyacyl-ACP Dehydrase	2	1	Tc01g009360			
ENR1	Plastidial Enoyl-ACP Reductase	1	2	Tc01g015430	Tc05g018860		
FAB2	Stearoyl-ACP Desaturase	7	8	Tc04g017510	Tc04g017520	Tc05g012840	Tc04g017540
				Tc04g005590	Tc09g024040	Tc08g012550	Tc01g009910
ACP	Plastidial Acyl Carrier Protein	5	3	Tc05g024590	Tc04g027340	Tc09g005970	
ACP	Mitochondrial Acyl Carrier Protein	3	4	Tc01g039970	Tc05g019890	Tc06g020860	Tc05g008650
FATA	Acyl-ACP Thioesterase Fat A	2	1	Tc01g022130			
FATB	Acyl-ACP Thioesterase Fat B	1	5	Tc09g010360	Tc03g015170	Tc02g017580	Tc03g003720
				Tc00g060380			
FAD2	ER Oleate Desaturase	1	2	Tc05g018800	Tc01g015280		
FAD3	ER Linoleate Desaturase	1	1	Tc09g029750			
FAD4	Phosphatidylglycerol Desaturase	1	1	Tc07g004100			
FAD5	Monogalactosyldiacylglycerol Desaturase	1	3	Tc06g001950	Tc00g079390	Tc06g001920	
FAD6	Plastidial Oleate Desaturase	1	1	Tc09g002840			
FAD7/FAD8	Plastidial Linoleate Desaturase	2	2	Tc05g002310	Tc10g001360		
MCCA	3-Methylcrotonyl-CoA Carboxylase (biotinylated subunit)	1	1	Tc01g028230			
MCCB	3-Methylcrotonyl-CoA Carboxylase (non-biotinylated subunit)	1	1	Tc02g008870			
KCS	β-Ketoacyl-CoA Synthase	21	20	Tc02g006950	Tc04g024460	Tc09g005920	Tc00g000460
				Tc08g002160	Tc02g022050	Tc00g092330	Tc00g015810
				Tc09g010640	Tc00g053150	Tc00g092340	Tc00g053110
				Tc09g013510	Tc00g092320	Tc03g018920	Tc09g032420
				Tc04g024470	Tc01g033850	Tc00g015820	Tc00g053140
KCR	Ketoacyl-CoA Reductase	2	2	Tc02g025790	Tc02g025850		
ECR	Enoyl-CoA Reductase	1	1	Tc10g003320			
LACS	Long Chain Acyl-CoA Synthetase	2	7	Tc01g032080	Tc02g002110	Tc03g028900	Tc04g023400
				Tc05g029290	Tc01g019380	Tc08g011430	
Totals		71	84				

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 16: List of *Theobroma cacao* genes orthologous to genes encoding key enzymes of the flavonoid biosynthesis pathway. Gene copy numbers were determined using the BLASTP program with full-length protein sequences obtained from the loci listed in Supplementary Table 15 as queries. For *V. vinifera* (grape), *P. trichocarpa* (poplar) and *A. thaliana* the Phytozome database (http://www.phytozome.net) was queried. For cacao the *T. cacao* assembly V1.0 database was queried. Standard gene designation (Gene), locus used to obtain protein sequences for Blast searches (Query), enzyme activity (Enzyme), gene copy numbers followed by e-value cutoff (highest e-value accepted), and locus numbers for each predicted cacao gene in *T. cacao* assembly V1.0 database are indicated. For the chalcone and stilbene synthase enzymes, it was not possible to distinguish gene function using sequence data alone, so the genes were grouped into one family.

Gene	Query	Enzyme	Gene Copy Number (Highest Blast e Value)				Cacao Locus Numbers			
			Grapevine	Poplar	Arabidopsis	Cacao				
PAL	AT2G37040	PHENYLALANINE AMMONIA-LYASE	15 (0)	5 (0)	4 (0)	3 (0)	Tc05g006550	Tc10g005730	Tc00g004780	
C4H	AT2G30490	CINNAMATE 4-HYDROXYLASE	3 (1e-169)	3 (0)	1 (0)	3 (1e-164)	Tc09g035230	Tc10g016540	Tc09g021750	
4CL	AT1G51680	COUMARATE-4-CoA LIGASE	13 (1e-73)	17 (1e-78)	12 (1e-72)	14 (1e-81)	Tc00g013150	Tc06g010940	Tc03g014500	Tc08g004630
							Tc05g022370	Tc09g033930	Tc03g030380	Tc10g007060
							Tc00g026030	Tc02g026580	Tc04g002850	
							Tc00g081720	Tc03g014490	Tc08g004620	
CHS / STSY	AT5G13930	CHALCONE AND STILBENE SYNTHASE	35 (1e-130)	14 (1e-76)	4 (1e-71)	7 (1e-73)	Tc00g003270	Tc04g021630	Tc05g016950	Tc05g016940
							Tc05g016920	Tc02g005640	Tc01g038610	
CHI	AT3G55120	CHALCONE ISOMERASE	1 (1e-84)	1 (1e-82)	2 (1e-58)	3 (1e-43)	Tc00g015870	Tc10g004370	Tc00t015860	
F3H	AT3G51240	FLAVANONE 3-HYDROXYLASE	3 (1e-142)	2 (1e-171)	1 (0)	4 (1e-51)	Tc01t001700	Tc01g001680	Tc03g022510	Tc01g033570
F3'H	AT5G07990	FLAVONOID 3'-HYDROXYLASE	0	0	1 (0)	3 (1e-99)	Tc03g011930	Tc00g013390	Tc09g006880	
F3'5'H	POPTR_0009s07320	FLAVONOID 3',5'-HYDROXYLASE	10 (0)	2 (0)	0	1 (0)	Tc09g02420			
FLS	AT5G08640	FLAVONOL SYNTHASE	2 (1e-124)	3 (1e-128)	6 (1e-78)	3 (1e-95)	Tc05g019560	Tc08g010270	Tc08g010240	
DFR	AT5G42800	DIHYDROFLAVONOL-4-REDUCTASE	1 (1e-123)	3 (1e-148)	1 (0)	18 (1e-42)	Tc08g002380	Tc01g002920	Tc08g002390	Tc01g037650
							Tc01g037660	Tc07g005240	Tc01g037700	Tc01g037670
							Tc02g012780	Tc06g004070	Tc00g058870	Tc05g021970
							Tc01g040340	Tc05g021930	Tc09g034440	Tc01g040310
							Tc08g008380	Tc01g034710		
LDOX (ANS)	AT4G22880	LEUCOANTHOCYANIDIN DIOXYGENASE	1 (1e-154)	2 (1e-149)	1 (0)	2 (1e-51)	Tc03g026420	Tc05g002410		
LAR	POPTR_0008s11540	LEUCOANTHOCYANIDIN REDUCTASE	2 (1e-126)	3 (1e-125)	0	2 (1e-123)	Tc03g002450	Tc02g034610		
ANR	AT1G61720	ANTHOCYANIDIN REDUCTASE	1 (1e-120)	2 (1e-102)	1 (0)	1 (1e-111)	Tc06g018030			
UFGT	AT5G54060	UDP-GLUCOSE: FLAVANOID 3-O-GLUCOSYLTRANSFERASE	1 (1e-153)	2 (1e-114)	1 (0)	11 (1e-42)	Tc07g008830	Tc06g015080	Tc06g015090	Tc07g009090
							Tc07g009020	Tc07g009080	Tc07g009000	Tc07g009050
							Tc08g006110	Tc01g007890	Tc09g007510	
AS	AB044884.1	AUREUSIDIN SYNTHASE	5 (1e-120)	13 (1e-106)	0	7 (1e-108)	Tc06g016250	Tc06g016260	Tc06g016300	Tc06g016270
							Tc06g016280	Tc02g006800	Tc04g018040	
OMT	AT5G54160	O-METHYLTRANSFERASE	7 (1e-107)	2 (1e-166)	1 (0)	14 (1e-47)	Tc03g021850	Tc08g011220	Tc08g011250	Tc00g058650
							Tc04g009850	Tc00g018080	Tc05g023810	Tc05g023880
							Tc03g022520	Tc01g033580	Tc01g033590	
							Tc05g023890	Tc05g023830	Tc05g023860	
Totals			100	74	36	96				

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 17. Classification of *Theobroma cacao* orthologous genes into the 13 terpenoïd-encoding gene subfamilies.
FPPS, farnesyl diphosphate synthase; **GGPPS**, geranylgeranyl diphosphate synthase; **GPPS**, geranyl diphosphate synthase; **SQS** squalene synthase; **PSY**, phytoene synthase; **HMGS**, HMG-CoA synthase;

SUBFAMILY	***T. CACAO* ACCESSION NUMBERS**	**Compound class**
GPPS	Tc07g003160, Tc09g029280	Monoterpene
limonene synthase	Tc00g044690, Tc06g017940, Tc07g016360, Tc07g016390	Monoterpene
linalool synthase	Tc06g016360, Tc06g016370, Tc06g016390, Tc06g017500, Tc06g017520, Tc06g017530, Tc06g017930	Monoterpene
myrcene synthase	Tc00g016370, Tc00g033730, Tc03g027710, Tc03g027720, Tc06g017980	Monoterpene
ocimene synthase	Tc06g017510, Tc06g017920	Monoterpene
pinene synthase	Tc00g007730, Tc00g012940, Tc00g033760	Monoterpene
FPPS	Tc06g021060, Tc07g000710	Sesquiterpene
Germacrene-D synthase	Tc00g085410, Tc07g005070, Tc07g005280, Tc07g005390, Tc07g016300	Sesquiterpene
Cadinene synthase	Tc00g033820, Tc00g067420, Tc04g011420, Tc07g005080, Tc07g005290, Tc07g005310, Tc07g005320, Tc07g005310, Tc07g005330, Tc07g005340, Tc07g005350	Sesquiterpene
GGPPS	Tc00g085980, Tc01g003740, Tc02g004290, Tc04g014990, Tc04g015000, Tc06g012170, Tc06g012180	Diterpene, Phytoene,
Casbene synthase	Tc00g044710, Tc00g054380	Diterpene
PSY	Tc00g029810, Tc00g062310, Tc01g015090, Tc03g025560	Phytoene
SQS	Tc02g007320, Tc02g007330, Tc02g007380	Triterpene

Nature Genetics: doi: 10.1038/ng.736

Supplementary Table 18. ***T. cacao genome synteny,*** The table illustrates the synteny relationships (lines) identified between the *T. cacao* (first column, number of genes in parenthesis) and *Arabidopsis*, poplar, grape, soybean and papaya chromosomes (second column). The number of orthologous genes per chromosomes is shown in parenthesis.

Cacao chromosome	Arabidopsis chromosome
C1(1321)	A1(63)-A2(85)-A3(69)-A4(116)-A5(64)
C2(935)	A1(89)-A2(30)-A4(61)-A5(10)
C3(1018)	A1(70)-A3(17)-A4(63)-A5(130)
C4(837)	A1(21)-A2(9)-A3(94)-A4(36)-A5(44)
C5(820)	A1(24)-A2(51)-A3(78)-A5(37)
C6(616)	A1(59)-A2(15)-A4(66)-A5(26)
C7(287)	A1(5)-A5(32)
C8(605)	A1(67)-A4(10)-A5(5)
C9(1137)	A1(17)-A2(95)-A3(22)-A4(72)-A5(76)
C10(290)	A2(23)-A3(14)

Cacao chromosome	Poplar chromosome
C1(1321)	P14(164)-P19(6)-P1(6)-P2(120)-P3(18)-P4(6)-P5(66)-P7(124)-P11(5)-P9(8)
C2(935)	P1(20)-P3(13)-P4(24)-P5(9)-P7(27)-P8(48)-P9(77)-P10(88)-P13(21)-P17(11)
C3(1018)	P3(120)-P4(13)-P11(3)-P12(109)-P15(87)-P1(31)
C4(837)	P1(103)-P2(19)-P4(22)-P7(33)-P8(48)-P10(94)-P11(2)-P14(14)-P16(14)-P17(36)-P19(5)
C5(820)	P4(3)-P6(77)-P13(142)-P16(88)-P19(25)
C6(616)	P1(27)-P3(34)-P4(92)-P6(15)-P8(3)-P11(79)-P13(1)-P15(11)
C7(287)	P1(59)-P11(60)-P13(13)
C8(605)	P2(147)-P5(64)-P6(2)-P13(1)-P14(9)-P19(34)
C9(1137)	P1(49)-P3(7)-P4(5)-P6(129)-P9(115)-P13(6)-P18(150)-P19(6)
C10(290)	P10(54)-P19(10)-P1(8)-P8(34)

Cacao chromosome	Grape chromosome
C1(1321)	G12(99)-G13(32)-G2(12)-G3(1)-G15(119)-G4(118)-G5(2)-G6(6)-G7(325)-G18(10)
C2(935)	G10(16)-G12(69)-G1(218)-G3(151)
C3(1018)	G2(244)-G16(54)-G17(256)-G8(1)
C4(837)	G1(1)-G5(328)-G14(181)-G8(23)
C5(820)	G4(18)-G5(7)-G14(111)-G8(336)-G16(18)
C6(616)	G1(11)-G10(109)-G12(11)-G9(120)-G19(6)
C7(287)	G18(1)-G19(183)
C8(605)	G18(381)
C9(1137)	G3(21)-G4(186)-G6(268)-G11(235)-G13(4)
C10(290)	G11(2)-G13(178)

Cacao chromosome	Soybean chromosome
C1(1321)	S10(12)-S11(5)-S13(5)-S14(13)-S16(12)-S18(7)-S19(14)-S1(16)-S2(19)-S3(32)-S4(6)-S5(13)-S6(12)-S7(8)-S8(19)-S9(3)
C2(935)	S20(8)-S1(7)-S2(20)-S4(5)-S8(29)-S10(6)-S11(9)-S12(12)-S13(6)-S14(4)
C3(1018)	S11(15)-S17(17)-S20(10)-S1(12)-S4(12)-S5(21)-S6(13)-S7(6)-S8(7)-S9(17)
C4(837)	S20(11)-S1(3)-S2(2)-S5(11)-S7(7)-S8(10)-S9(10)-S10(6)-S11(1)-S13(8)-S15(10)-S16(17)-S17(7)-S18(6)-S19(11)
C5(820)	S10(24)-S11(1)-S13(8)-S16(1)-S19(26)-S20(13)-S3(25)-S7(8)
C6(616)	S13(23)-S15(20)-S1(1)-S5(6)-S7(23)-S8(16)-S9(5)
C7(287)	S13(15)-S14(3)-S2(5)-S6(6)-S8(5)
C8(605)	S14(17)-S17(9)-S4(16)-S6(39)
C9(1137)	S4(10)-S6(14)-S11(11)-S12(10)-S13(16)-S14(3)-S15(5)-S17(25)-S18(9)
C10(290)	S20(11)-S10(12)-S13(5)

Cacao chromosome	Papaya chromosome
C1(1321)	Py2(202)-Py3(49)-Py4(33)-Py5(176)-Py6(171)-Py8(1)
C2(935)	Py3(214)-Py4(54)-Py5(23)-Py6(24)-Py8(8)-Py9(169)
C3(1018)	Py2(127)-Py3(11)-Py5(1)-Py6(16)-Py8(366)
C4(837)	Py1(64)-Py3(13)-Py5(65)-Py6(175)
C5(820)	Py3(77)-Py4(233)-Py6(78)-Py9(1)
C6(616)	Py1(3)-Py2(42)-Py5(26)-Py7(150)-Py8(54)-Py9(24)
C7(287)	Py2(162)-Py5(5)-Py9(7)
C8(605)	Py1(316)-Py2(8)
C9(1137)	Py9(136)-Py2(7)-Py3(36)-Py5(32)-Py6(43)-Py7(280)
C10(290)	Py4(15)-Py9(139)

Nature Genetics: doi: 10.1038/ng.736

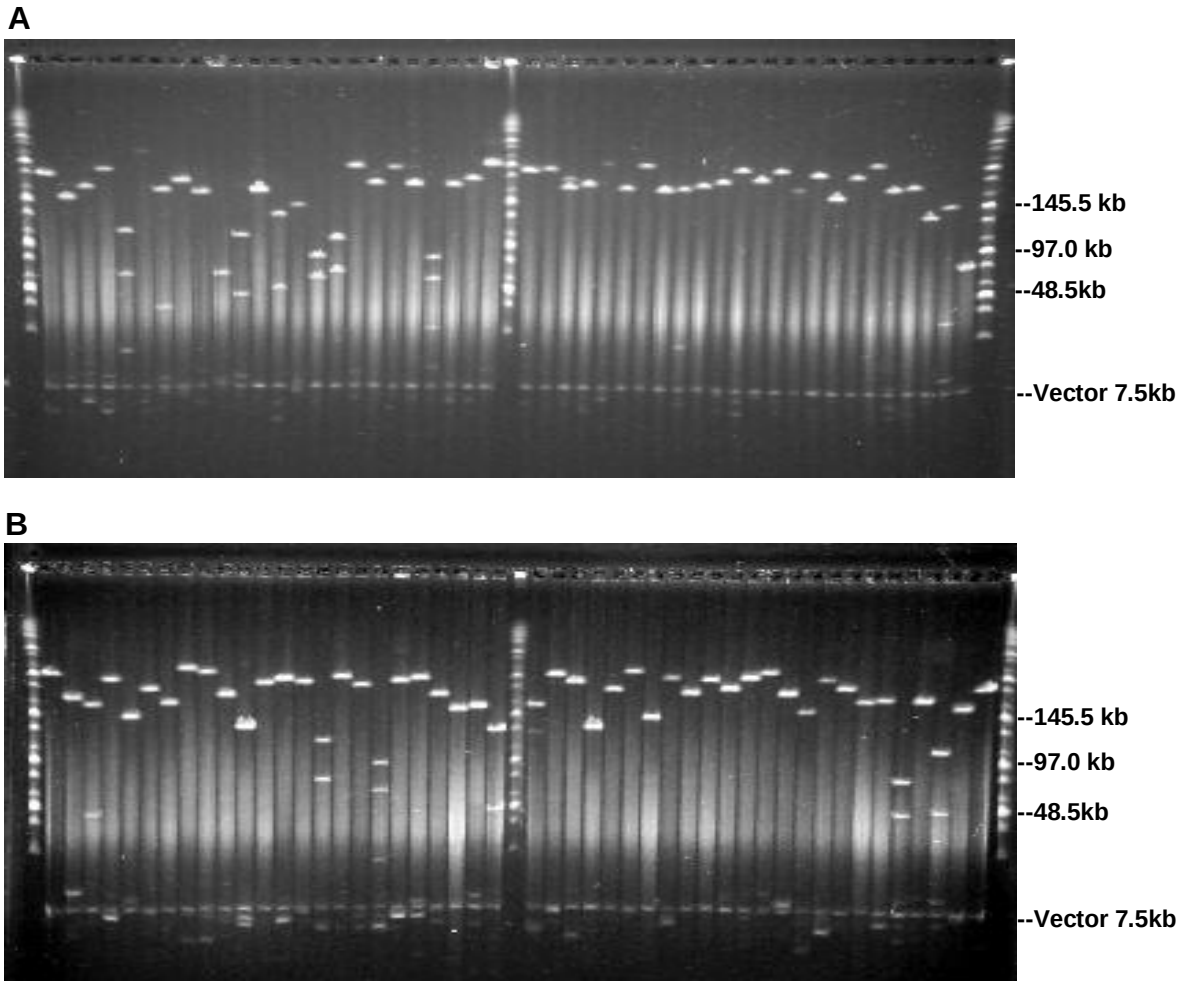
Supplementary Table 19. ***T. cacao genome duplication.*** The table illustrates seven ancestral duplications identified in the *T. cacao* genome. The duplicated blocks (1 to 3) are mentioned in columns and the start/end position on the corresponding chromosomes

	Block1			Block2			Block3		
	chromosome	start	end	chromosome	start	end	chromosome	start	end
duplication1	c2	12716774	27462648	c3	208385	16091087	c4	349021	14314443
duplication2	c1	27207631	30674661	c3	16741484	24212437	c3	16741484	24212437
duplication3	c1	315357	7988483	c2	1350572	7237080	c8	43353	6712481
duplication4	c6	1071819	9467758	c9	739576	9589803	c9	739576	9589803
duplication5	c1	21083375	26683534	c4	18966341	23343107	c5	23329957	25395907
duplication6	c5	541440	5362779	c9	23851693	28019603	c10	333882	12953021
duplication7	c1	8722499	15224371	c6	10864133	14795052	c7	511932	6542889

Nature Genetics: doi: 10.1038/ng.736

Supplementary Figures

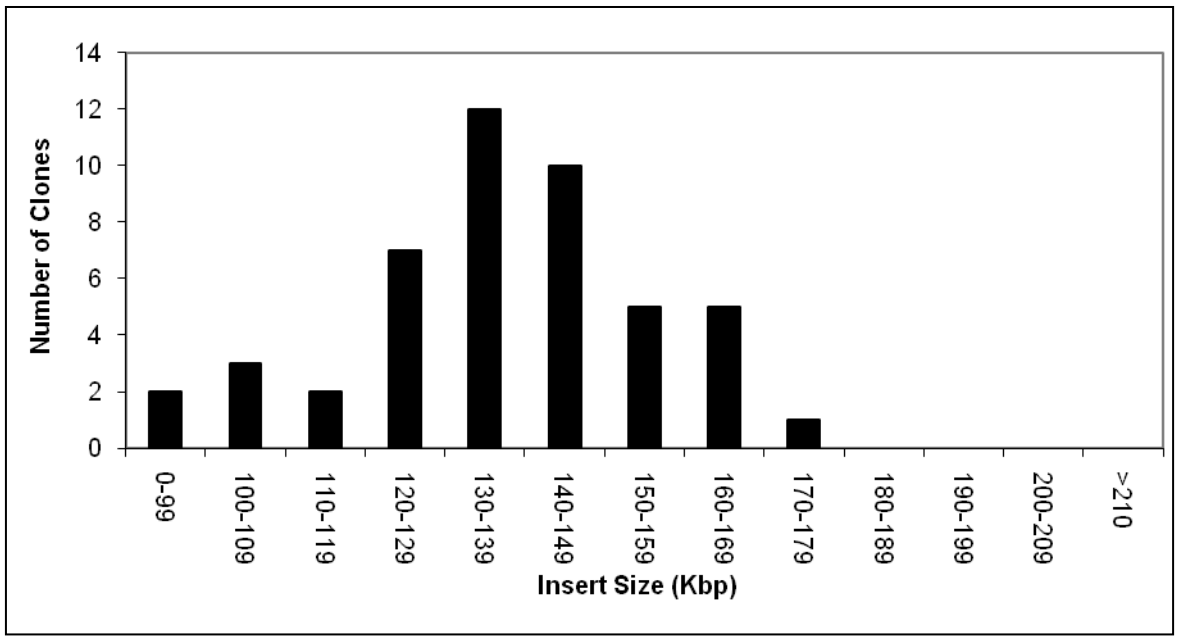
Supplementary Fig. 1. BAC clone inserts from two *Theobroma cacao* BAC libraries. A. BAC library TC_CBa composed of *Hind*III restricted fragments. B. BAC library TC_CBb composed of *Eco*RI restricted fragments. Plasmid BAC DNA was restricted with *Not*I enzyme and subjected to Pulsed Field Gel Electrophoresis (PFGE). Molecular size standards are indicated.



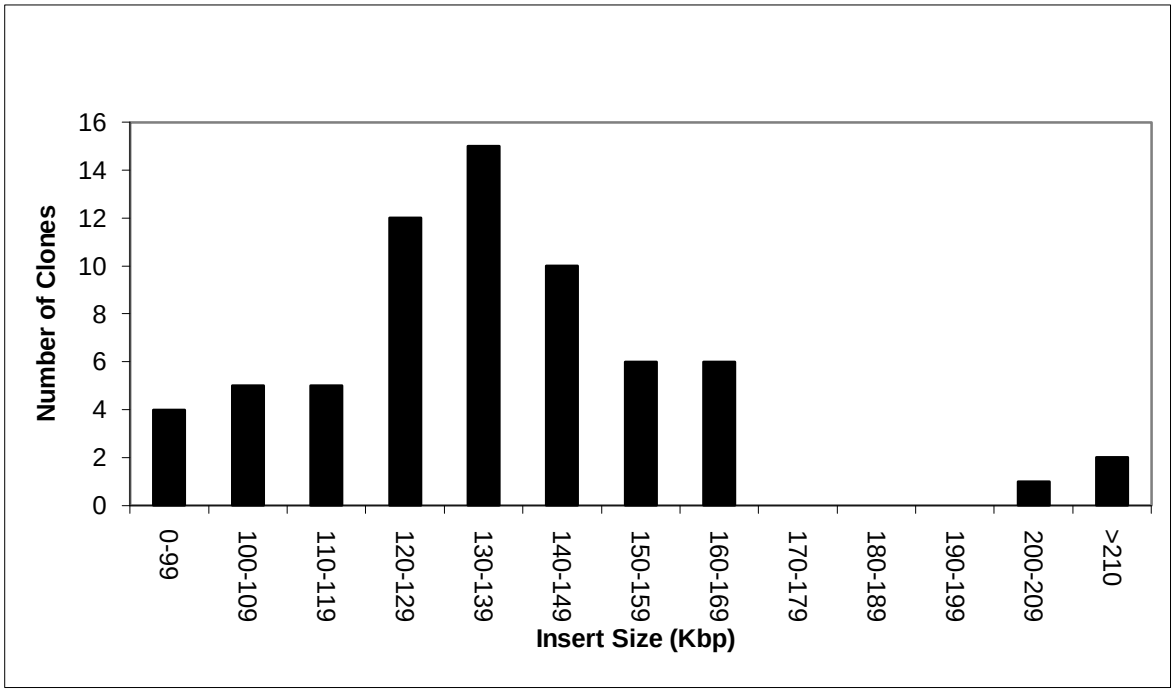
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 2. Histograms showing the frequency distribution of insert sizes in the two *Theobroma* BAC libraries. A. H*ind*III BAC library (TC_CBa); 135kbp average insert size. B. *Eco*RI BAC library (TC_CBb); 137kbp average insert size.

A



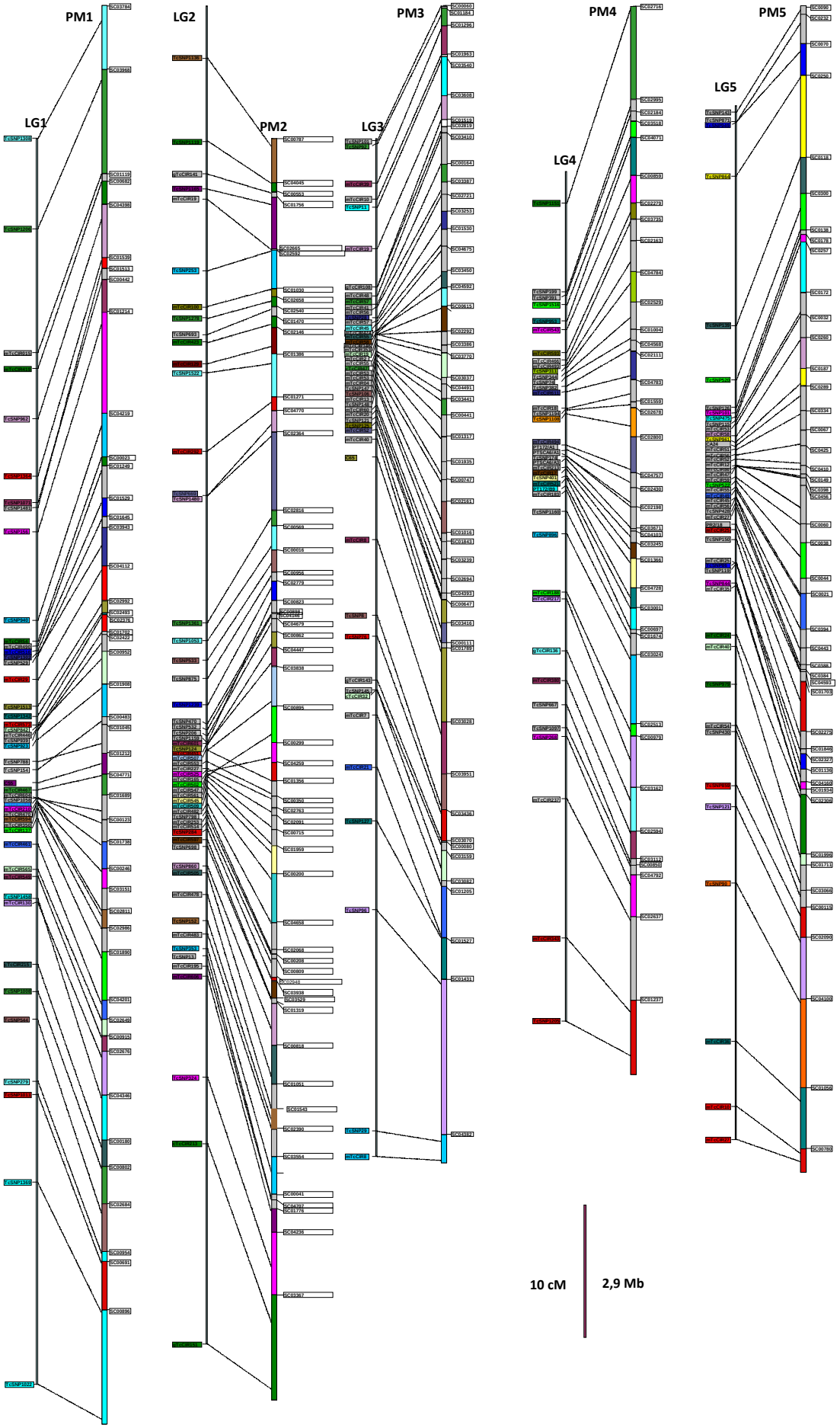
B



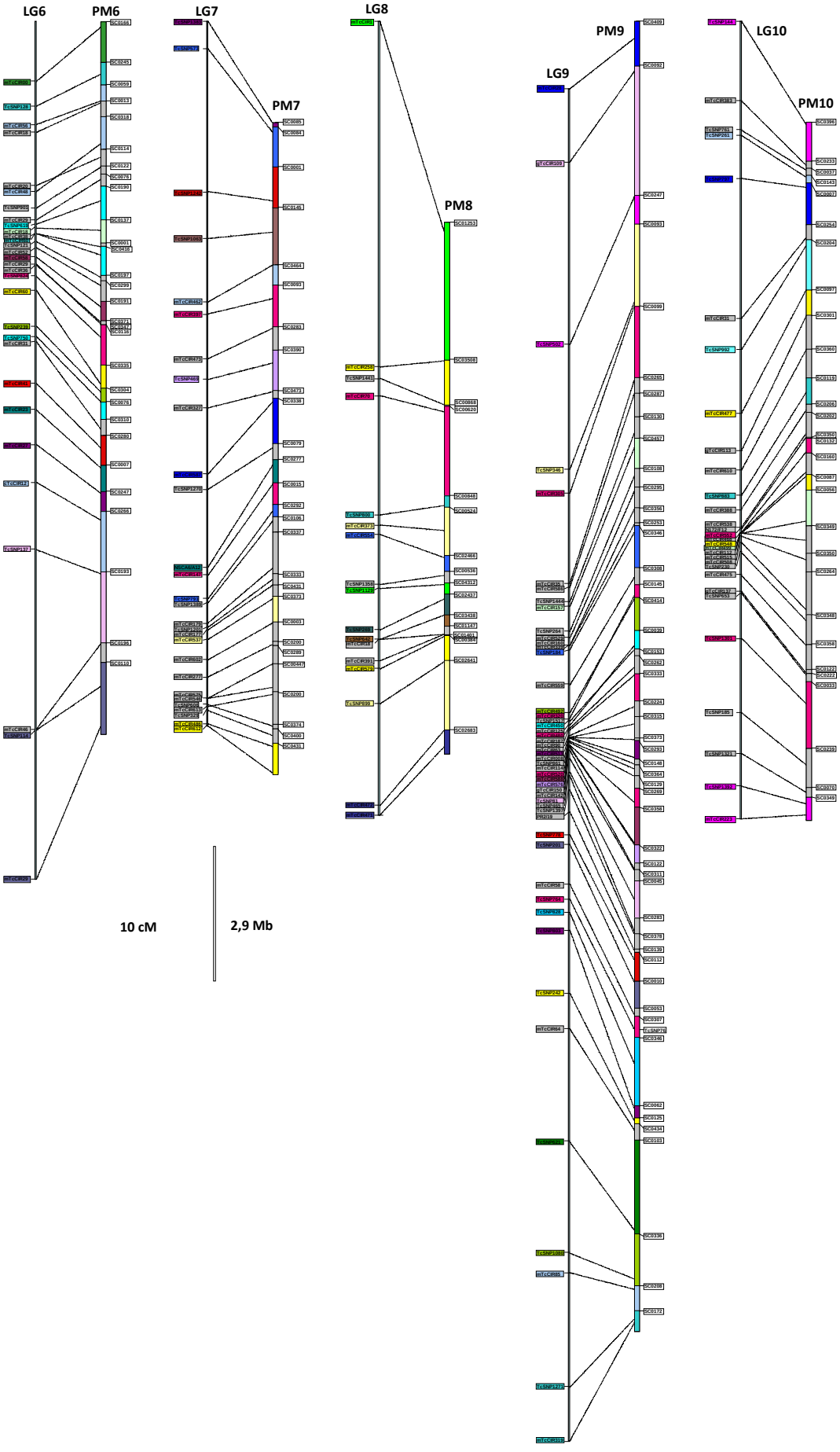
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 3. Map of the sequence scaffolds along the genetic map. Scaffolds that constitute the pseudomolecules (PM) are linked to the cocoa linkage groups (LG). At least one informative marker per scaffold has been represented in the figure. Scaffolds are represented on the right as colored bars (oriented) or as grey bars (random orientation).

Nature Genetics: doi: 10.1038/ng.736

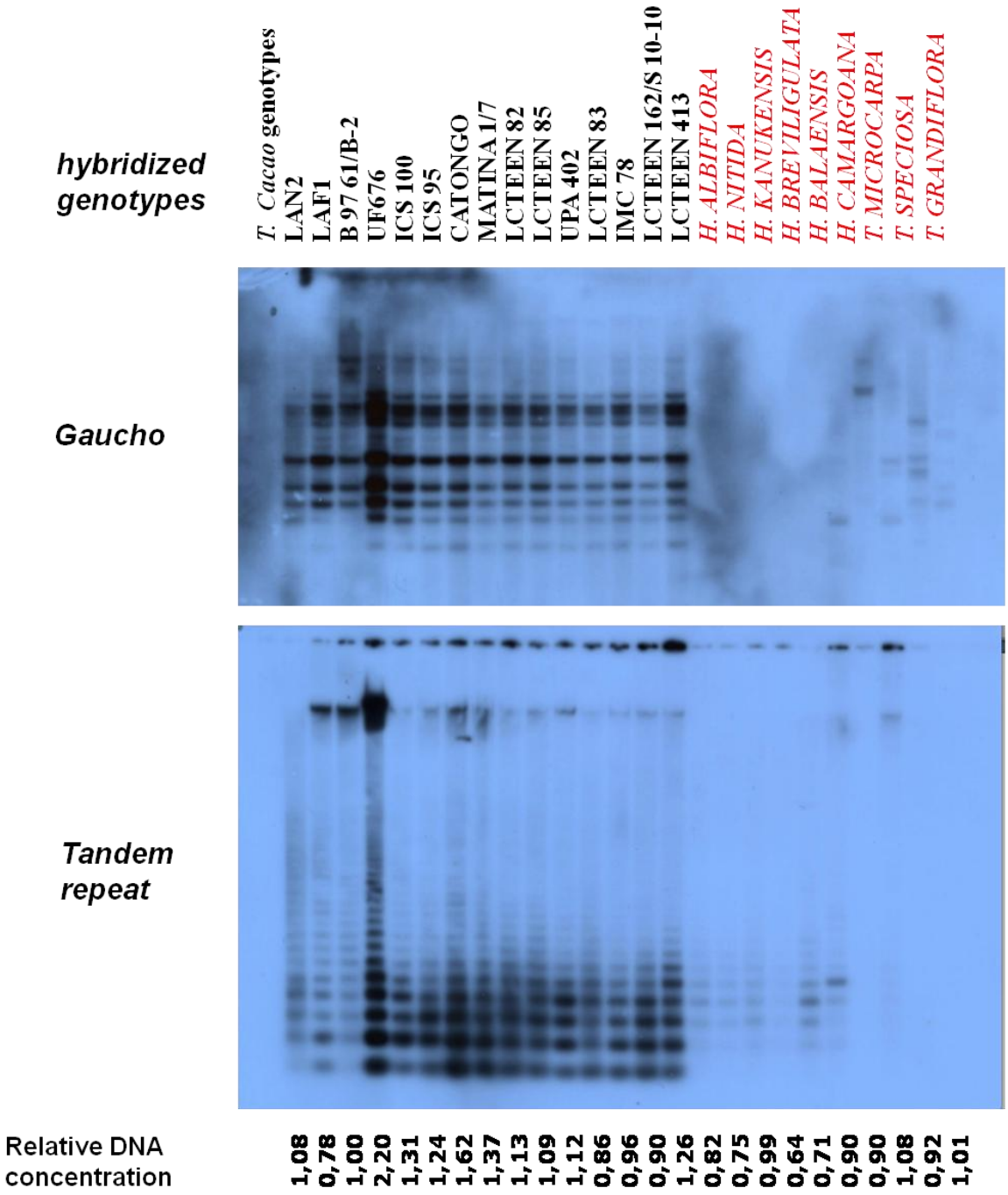


Nature Genetics: doi: 10.1038/ng.736



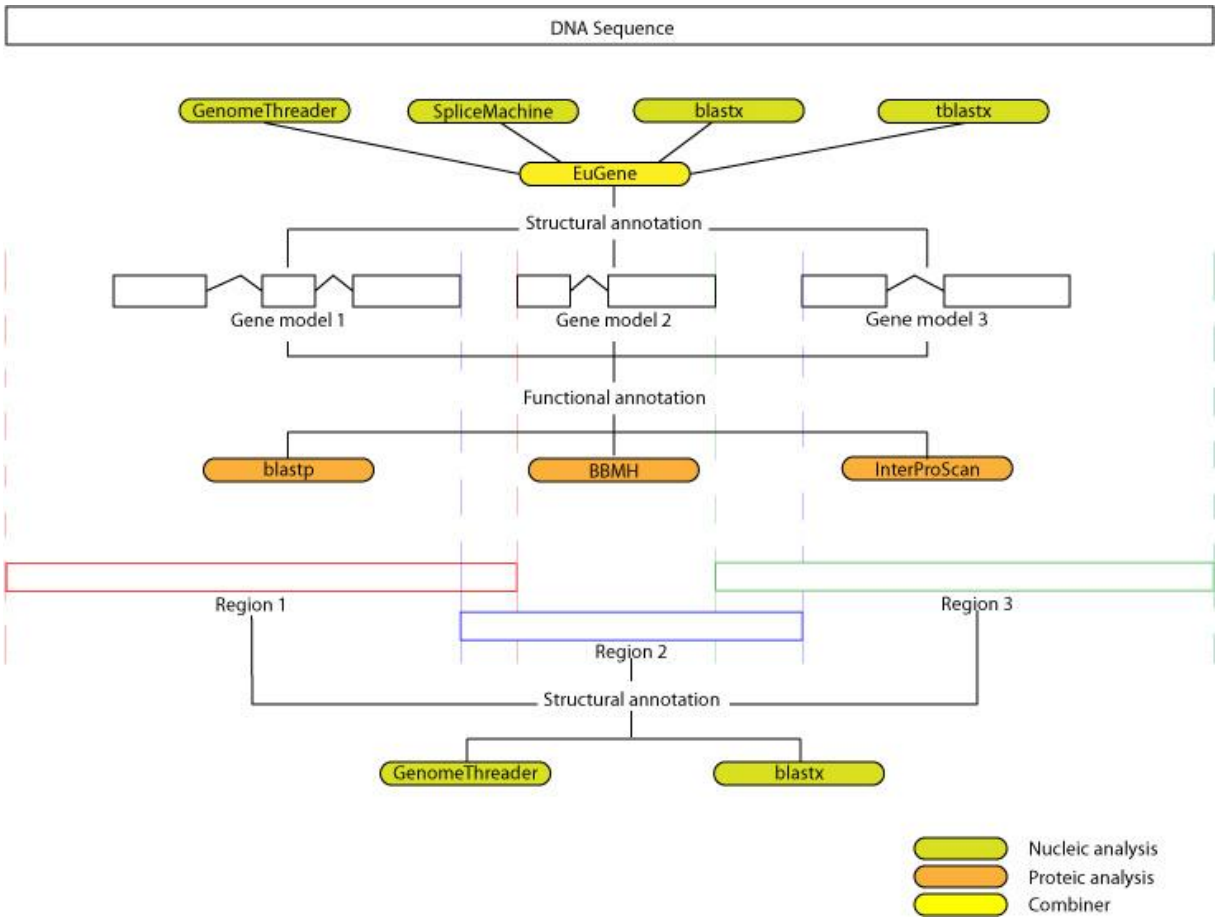
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 4. Southern blot hybridizations with probes from *Gaucho* and *ThCen* repeated sequences. Two micrograms of genomic DNA from cocoa genotypes (black type) and from representatives of related species or *Herrania* of *Theobroma* (red type) were digested with *Hind* III and separated on a 0.8% agarose gel. Each blot was probed individually with a *Gaucho* and *ThCen* repeat probe, as indicated in Supplementary Note.



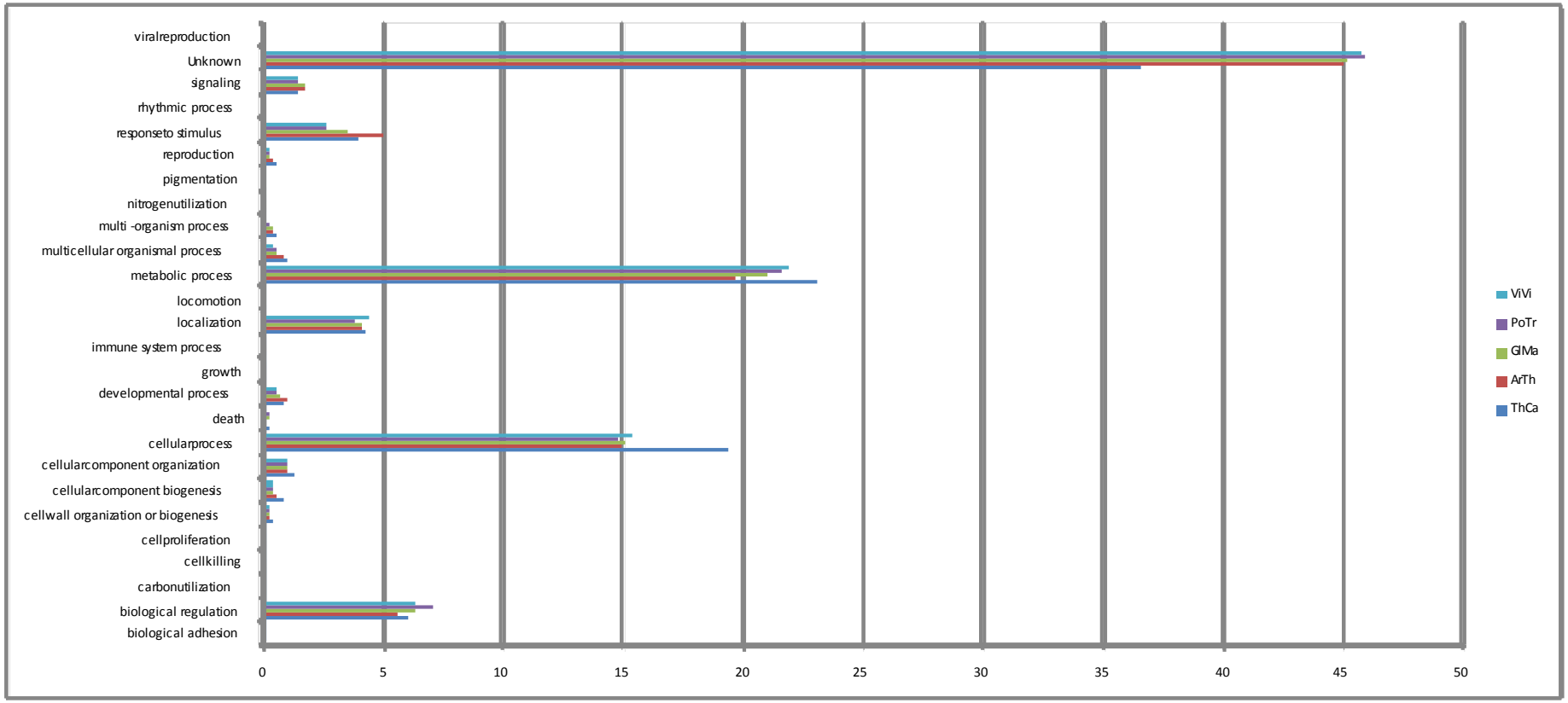
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 5. Structural and functional annotation workflow. Gene model predictions were produced using the integrative gene prediction platform EuGene[33] with statistical models trained for *T. cacao*. Translation starts and splice sites were predicted by SpliceMachine[101]. Available *T. cacao* ESTs were aligned on the genome using GenomeThreader[34]. Similarities to proteins from several datasets were searched using NCBI BLASTX[27] Similarities to *A. thaliana*, *Gossypium*, *V. vinifera*, *Citrus* and *T. cacao* ESTs were searched using NCBI TBLASTX. For each predicted coding sequence, we performed several analyses (BLASTP, InterProScan, BBMH) to transfer functional annotations. Then, we extracted for each gene model (n) a genomic region between the end of the gene preceding (n–1) and the beginning of the next gene (n+1) and we ran GenomeThreader and BLASTX to improve the structure of the gene if necessary.



Nature Genetics: doi: 10.1038/ng.736

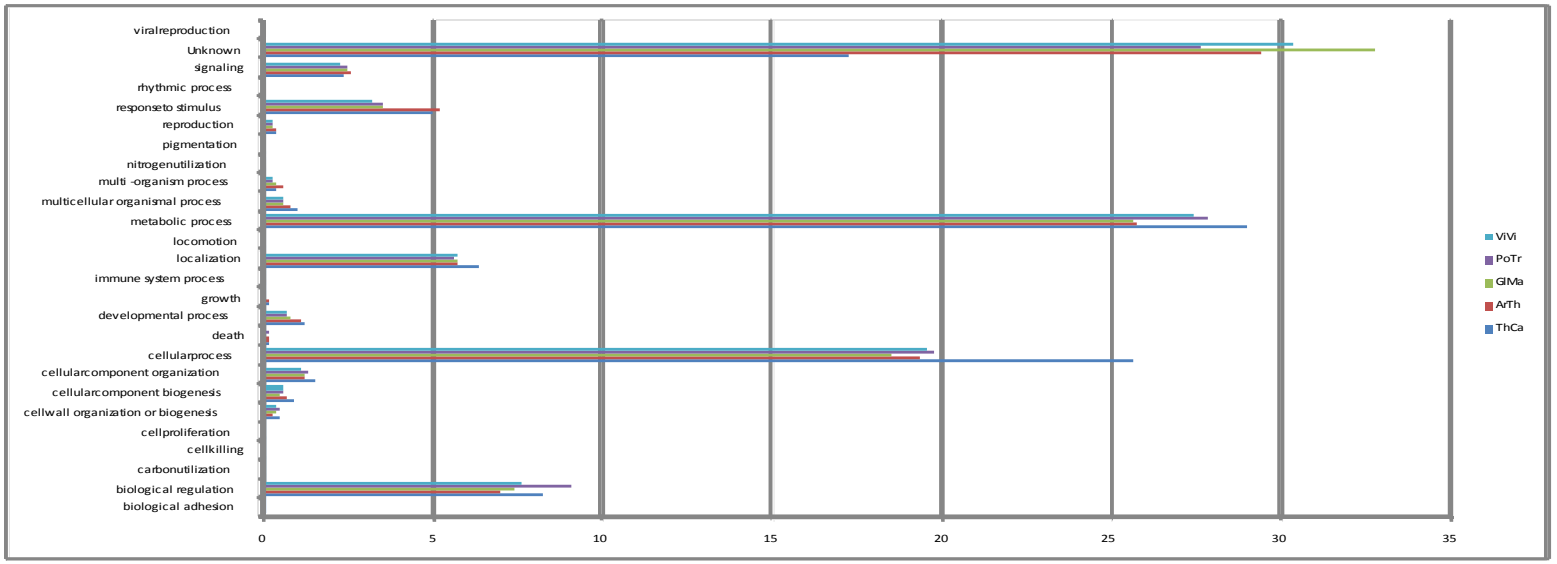
Supplementary Fig. 6. Distribution of Gene Ontology terms for *Theobroma cacao, Vitis vinifera, Arabidopsis thaliana, Populus trichocarpa* and *Glycine max* genes.



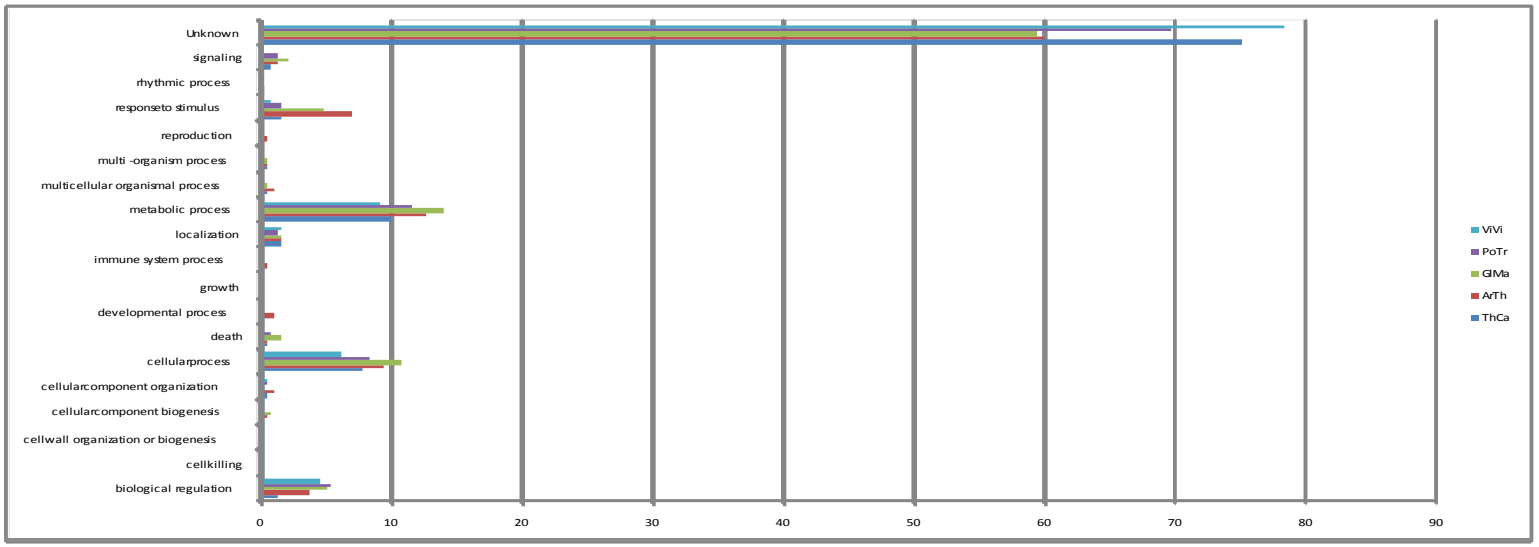
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 7. Distribution of Gene Ontology terms for genes contained in (A) common and (B) specific cluster families

A

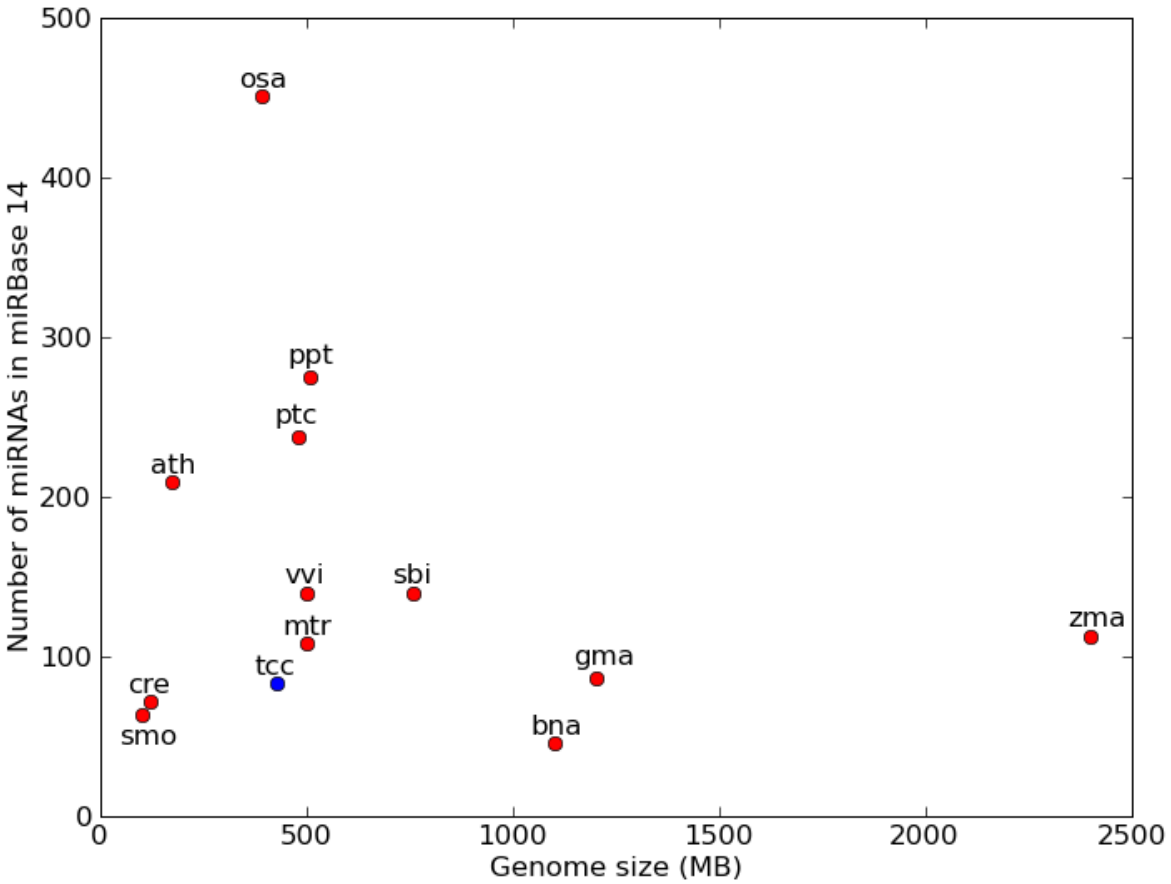


B



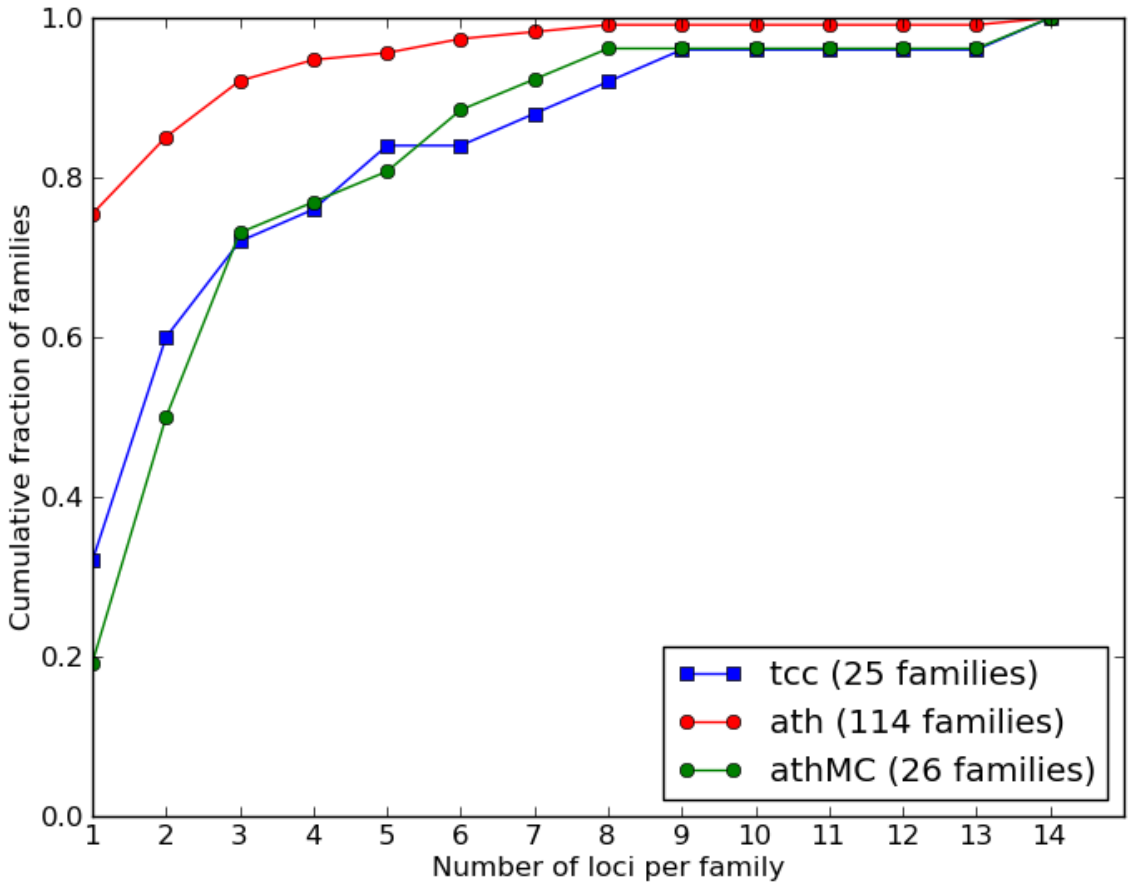
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 8. Number of miRNAs in each plant species in miRBase 14 and *Theobroma cacao* in relation to genome size. All plant species with more than 45 miRNAs in miRBase 14 and a known genome size (NCBI http://www.ncbi.nlm.nih.gov/genomeprj/?term=txid33090[Organism%3Aexp]) are plotted. No obvious correlation of miRNA population and genome size is observed. The high variation in the number of miRNAs among the different species is mostly due to different discovery methods (prediction only versus experimental confirmation) or different levels of stringency of the prediction.
ath: *Arabidopsis thaliana*, bna: *Brassica napus*, cre: *Chlamydomonas reinhardtii*, gma: *Glycine max*, mtr: *Medicago truncatula*, osa: *Oryza sativa*, ppt: *Physcomitrella patens*, ptc: *Populus trichocarpa*, sbi: *Sorghum bicolor*, smo: *Selaginella moellendorffii*, tcc: *Theobroma cacao*, vvi: *Vitis vinifera*, zma: *Zea mays*.



Nature Genetics: doi: 10.1038/ng.736

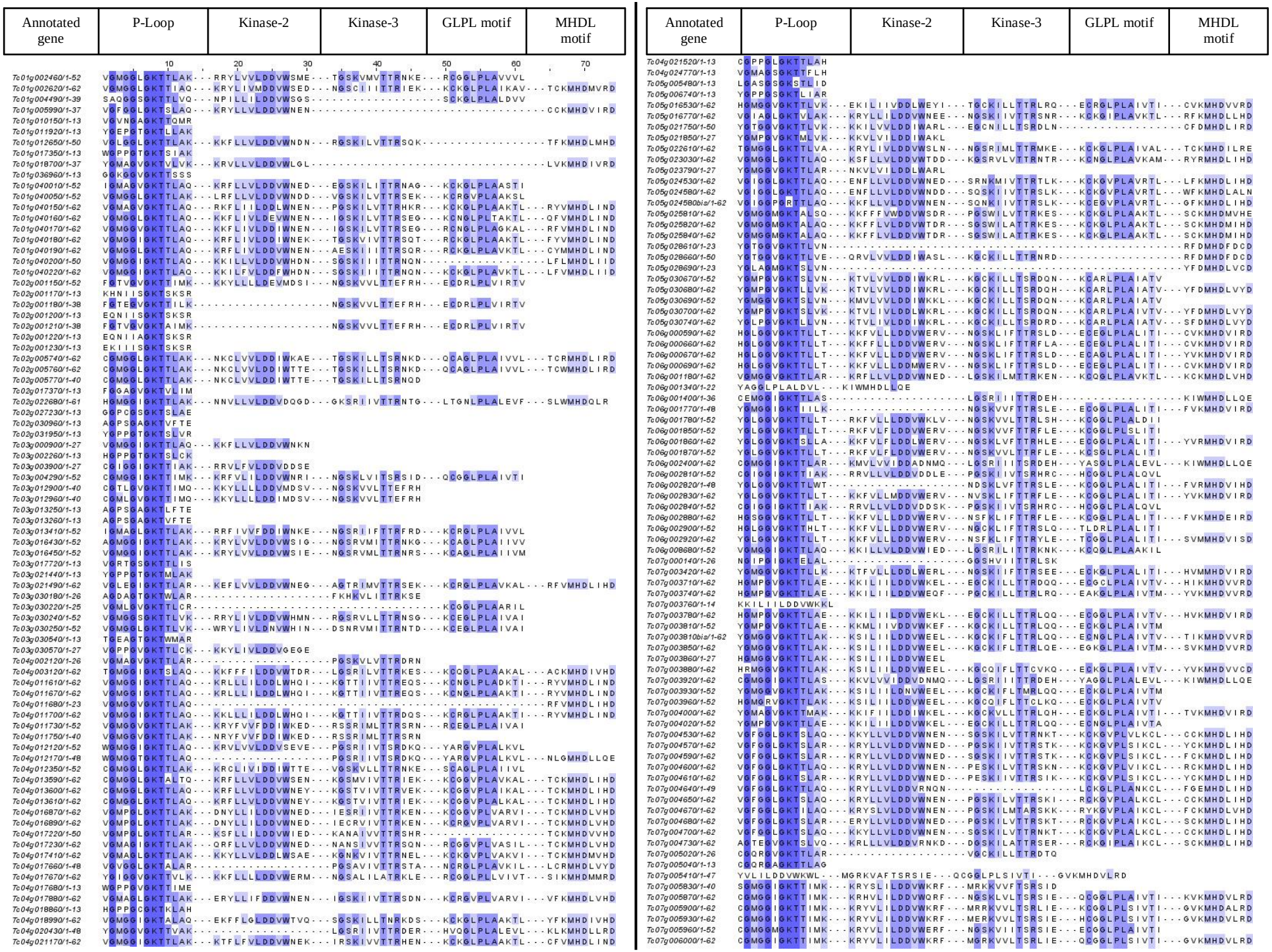
Supplementary Fig. 9. Cumulative distributions of the number of loci per miRNA family in *T. cacao* and *A. thaliana*. tcc: *T. cacao*, ath: *A. thaliana*, athMC: *A. thaliana* more conserved families.

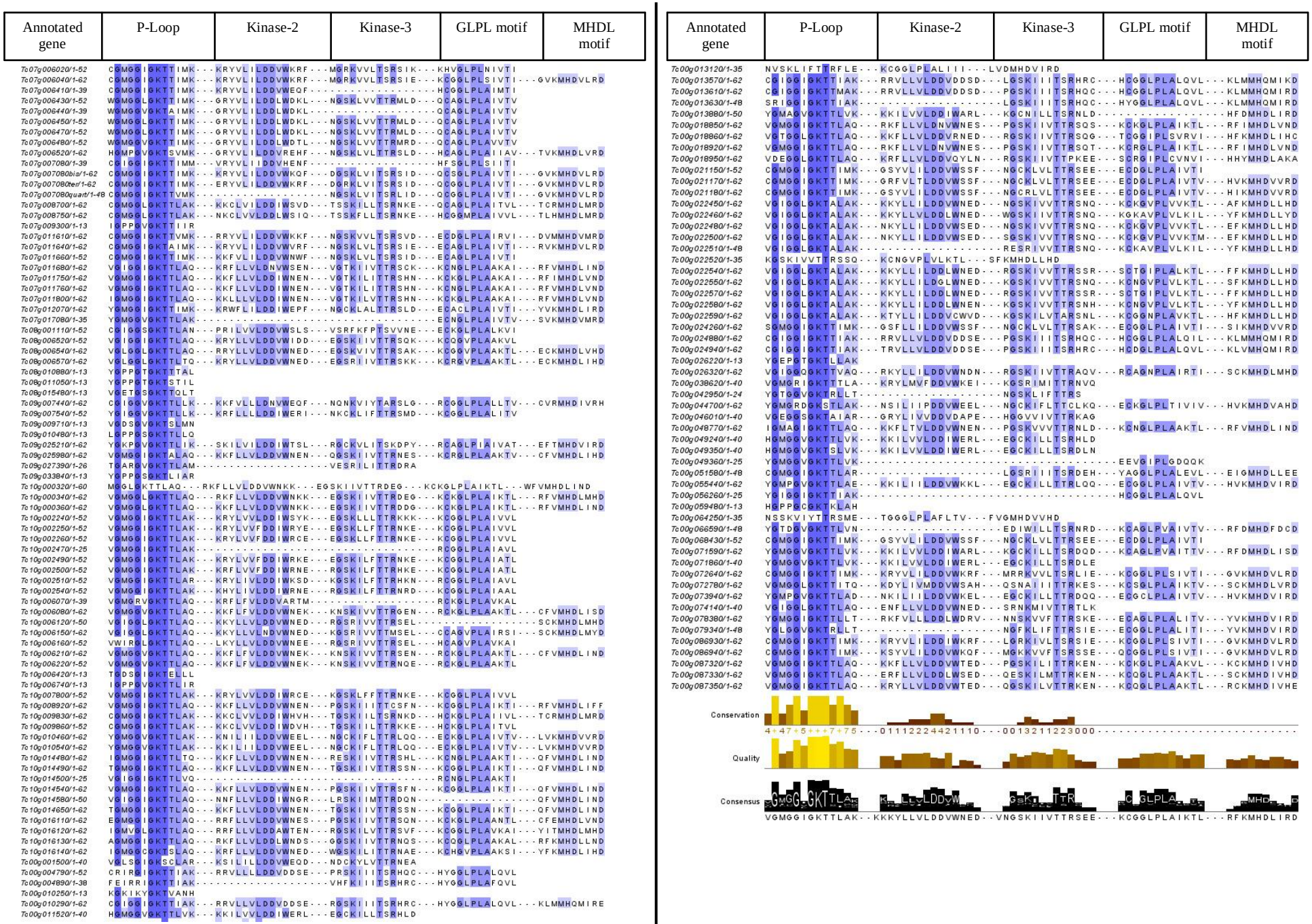


Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 10: NBS motifs of predicted cocoa genes orthologous to NBS-encoding genes. The visualization of multiple alignments of NBS motifs was monitored with Jalview software (http://www.jaview.org/). Conserved consensus sequences are highlighted in blue, with the blue intensity proportional to the % identity.

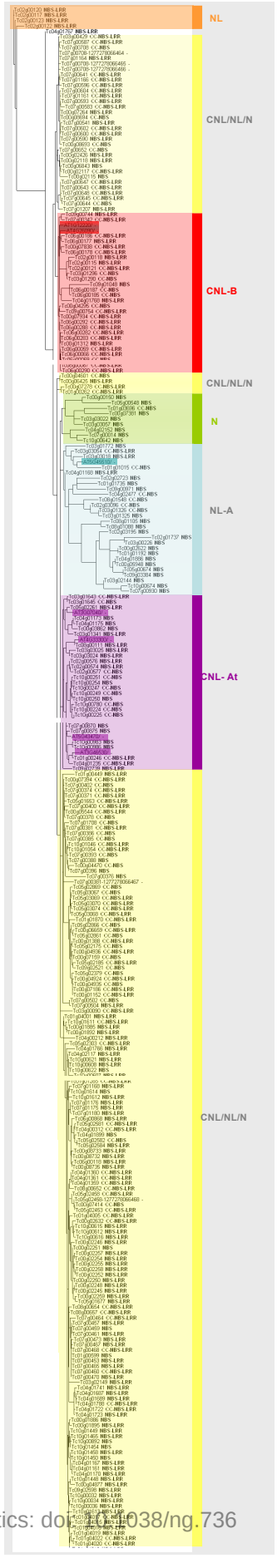
Nature Genetics: doi: 10.1038/ng.736


Annotated gene
P-Loop
Kinase-2
Kinase-3
GLPL motif
MHDL motif
Annotated gene
P-Loop
Kinase-2
Kinase-3
GLPL motif
MHDL motif


Annotated gene
P-Loop
Kinase-2
Kinase-3
GLPL motif
MHDL motif
Conservation
Quality
Consensus

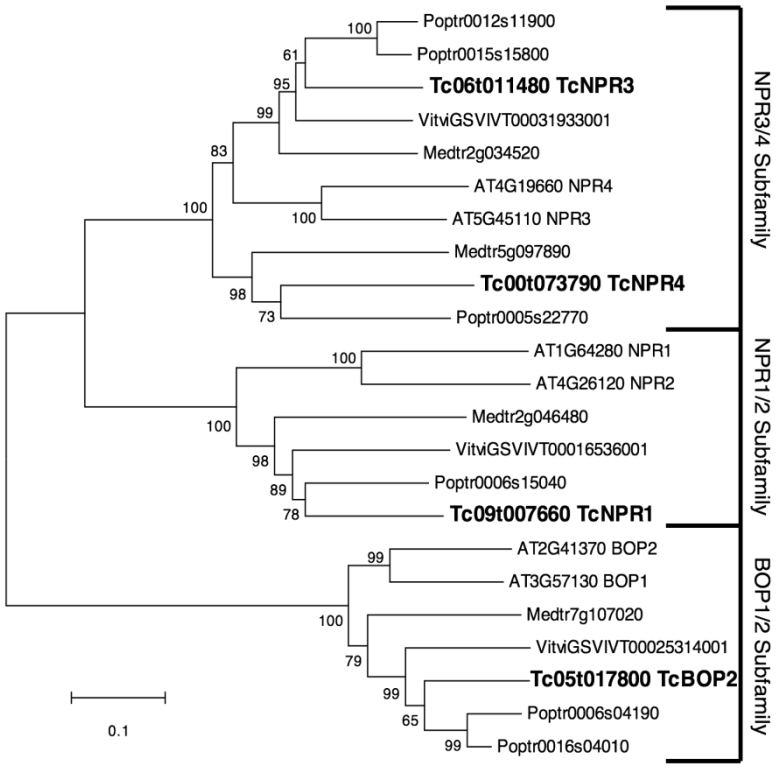
Supplementary Fig. 11: Phylogenetic tree based on NBS domains from non-TIR-NBS orthologous genes. According to Meyers et al, 2003[72]: the red box represents the CNL-B class (AT4G26090, AT1G12220); purple box, CNL-A/C/D class (AT3G07040, AT3G46530, AT5G43470, AT4G33300); blue box, NL-A class (AT5G45510). Three subclasses are specific to *T. cacao*: yellow box, CNL/NL/N class; green box, N class; and orange box, NL class.

Nature Genetics: doi: 10.1038/ng.736



Nature Genetics: doi:10.1038/ng.736

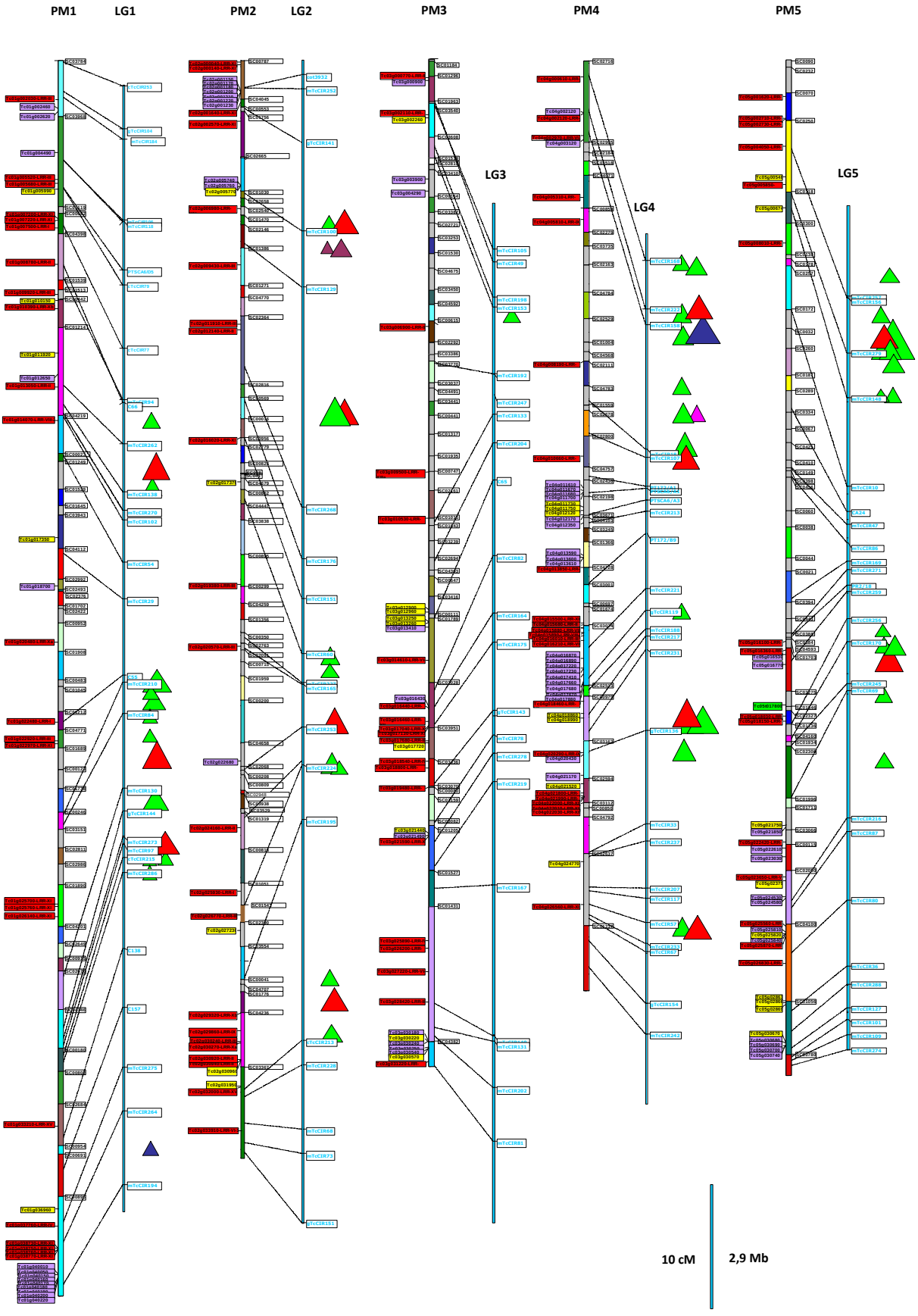
Supplementary Fig. 12. Phylogenetic analysis of the *T. cacao NON-EXPRESSOR OF PATHOGENESIS-RELATED 1* (*NPR1*) gene orthologous family. Full length protein sequences of all six *Arabidopsis NPR1* gene family members were used to search the *T. cacao* genome assembly V1.0 database using the TBLASTN[27] program with an E-value cutoff of 1*e-^{40}. Four cacao genes were identified with e-values below 5*e-^{41}. The next closest hit had an e-value of 1*e^{-15} and was not considered a bona fide NPR1 family member. Using the TBLASTN program, a full length protein sequence of *Arabidopsis* NPR1 was used to search the Phytozome database (http://www.phytozome.net/) to obtain NPR-like genes from poplar, Medicago and grape (e-value cutoff of 1*e^{-20}). Six NPR1 family member genes were identified in *P. trichocarpa* (Poplar), four in *M. truncatula* (medicago) and three in *V. vinifera* (grape). Multiple DNA alignment of 23 NPR genes from five species was carried out using MUSCLE[102] software. The phylogenetic tree was constructed with MEGA 4.0[103] software using the neighbor-joining method with the option of pairwise deletion. Gene locus IDs are included; bootstrap values are indicated next to each node and were obtained from 2000 replicates. A scale bar indicating a rate of 0.1 base pair substitutions per site is indicated at the bottom. Three subfamilies of *NPR1* genes are designated with brackets on the right.



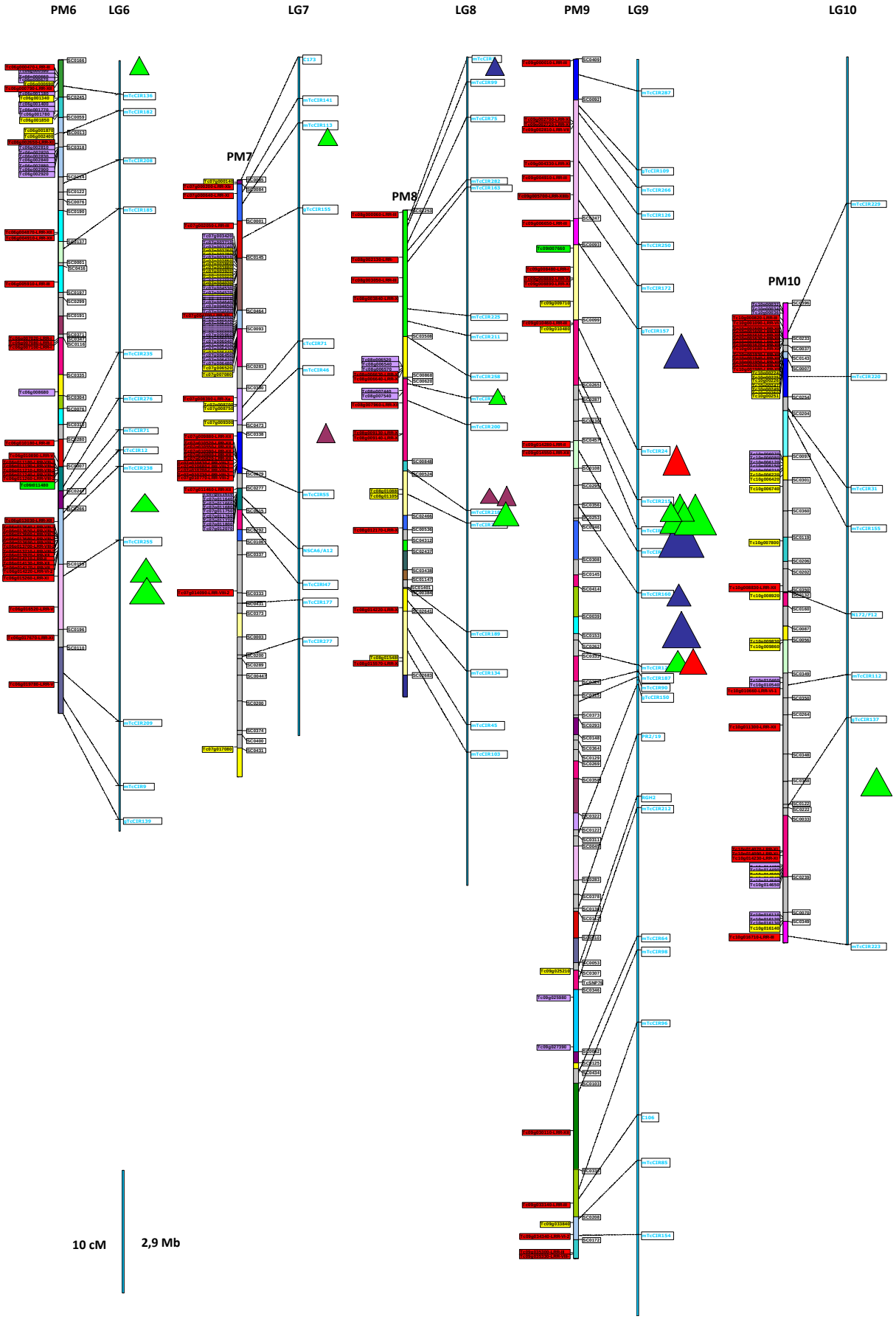
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 13. Mapping of genes orthologous to NBS, LRR-LRK and NPR1-like genes on pseudomolecules (PM) and comparative genome localisations with QTLs related to disease resistance identified on *T. cacao*. Orthologs of NBS-LRR (purple bars), NBS not LRR (yellow bars), LRR-LRK (red bars) and NPR1-like (green bars) genes are represented to the left of the pseudomolecules. QTLs, represented by triangles, are positioned to the right of the linkage groups (LG) as described by Lanaud et al., (2009)[93]: green triangles correspond to *Phytophthora* resistance, red triangles correspond to consensus QTLs correspond to *Phytophthora* resistance identified by meta-analyses, blue triangles correspond to Witches'broom disease due to *Moniliphthora perniciosa*, and purple triangles correspond to QTLs related to frosty pod due to *Moniliphthora roreri.*

Nature Genetics: doi: 10.1038/ng.736



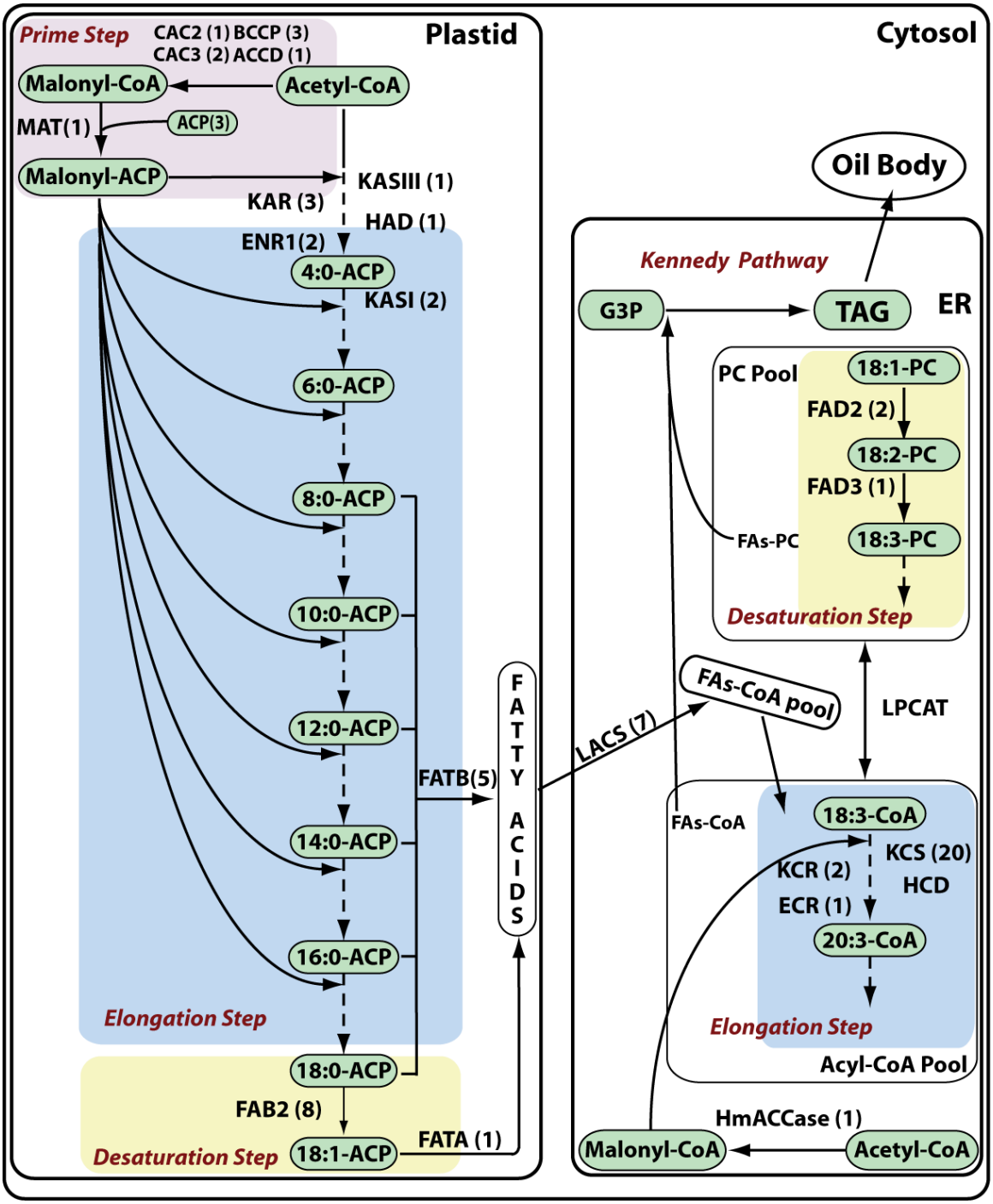
Nature Genetics: doi: 10.1038/ng.736



Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 14. Metabolic pathway for storage lipid biosynthesis adapted from Baud et al., 2010[104]. Orthologous gene copy number for each enzyme in *T. cacao* were determined as described in Supplementary Table 15. Enzymes involved the pathway are listed based on their sequential order and their compartmentalization in plastid and ER. Orthologous gene copy number in *T. cacao* are indicated in parentheses beside each enzyme abbreviation: CAC2, Heteromeric acetyl-CoA carboxylase BC subunit; BCCP, Heteromeric acetyl-CoA carboxylase BCCP subunit; CAC3, Heteromeric acetyl-CoA carboxylase alpha-CT subunit; ACCD, Heteromeric acetyl-CoA carboxylase beta-CT subunit; ACP: Acyl-carrier protein; CoA: coenzyme A; MAT, Plastidial malonyl-CoA : ACP malonyltransferase; KAS, Ketoacyl-ACP synthase; KAR, Plastidial ketoacyl-ACP reductase; HAD, Plastidial hydroxyacyl-ACP dehydrase; ENR1, Plastidial enoyl-ACP reductase; FAB2, Stearoyl-ACP desaturase; FATA, Acyl-ACP thioesterase; FATB, Acyl-ACP thioesterase;LACS, Long-chain acyl-CoA synthetase; FAD2, ER oleate desaturase; FAD3, ER linoleate desaturase; KCS, β-Ketoacyl-CoA synthase; KCR, Ketoacyl-CoA reductase; HCD, Hydroxyacyl-CoA dehydrase; ECR, Enoyl-CoA reductase; HmACCase, Homomeric acetyl-CoA carboxylase; LPCAT, Lysophosphatidylcholine acyltransferase (copy number was not determined for cacao); G3P, Glycerol-3-phosphate; TAG: Triacylglycerol. CAC2, BCCP, CAC3, and ACCD are the four subunits of ACCase in the plastid. Dashed arrows indicate the four-step elongation cycles catalyzed by KAS, KAR, HAD and ENR1, which is repeated multiple times during chain elongation.

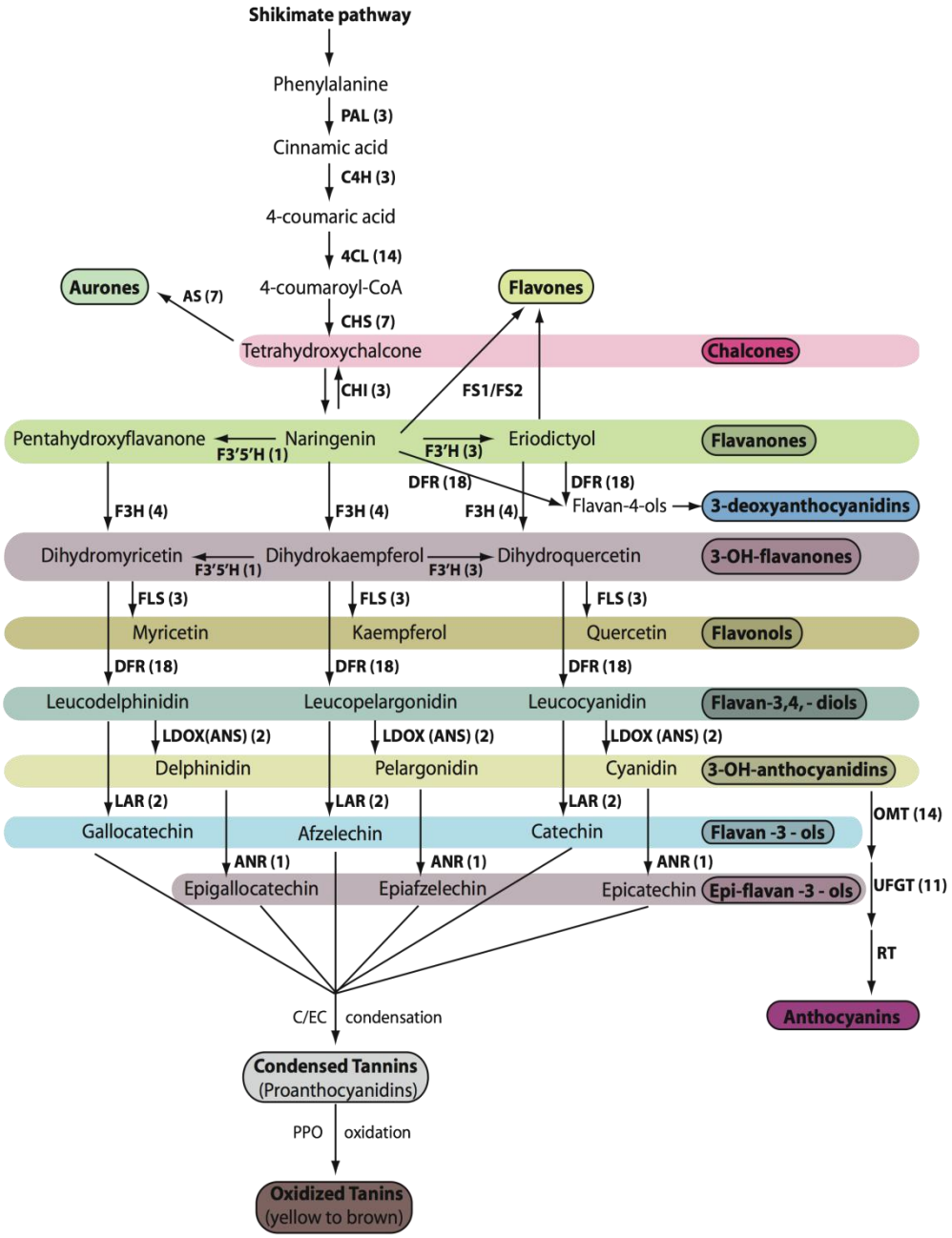
Nature Genetics: doi: 10.1038/ng.736



Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 15. **Metabolic pathway for flavonoid biosynthesis adapted from Lepiniec, L. *et al.*, 2006**[105]**.** *T. cacao* orthologous gene copy numbers for each enzyme were determined as described in Supplementary Table 16. Enzymes involved in the pathway are listed in sequential order (top to bottom): PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate-CoA ligase; CHS, chalcone synthase; AS, aureusidin synthase; CHI, chalcone isomerase; FS1/FS2, flavone synthase (copy number was not determined for cacao); F3H, flavanone 3-hydroxylase; F3'H, flavonoid 3′-hydroxylase; F3'5'H, flavonoid 3′,5′-hydroxylase; FLS, flavonol synthase; DFR, dihydroflavonol 4-reductase; LDOX (ANS), leucoanthocyanidin dioxygenase; LAR, leucoanthocyanidin reductase; ANR, anthocyanidin reductase; OMT, O-methyltransferase; UFGT, UDP-glucose:flavonoid 3-O-glucosyltransferase; RT, rhamnosyl transferase (copy number was not determined for cacao); C/EC refers to catechins/epicatechins, PPO refers to polyphenol oxydase.

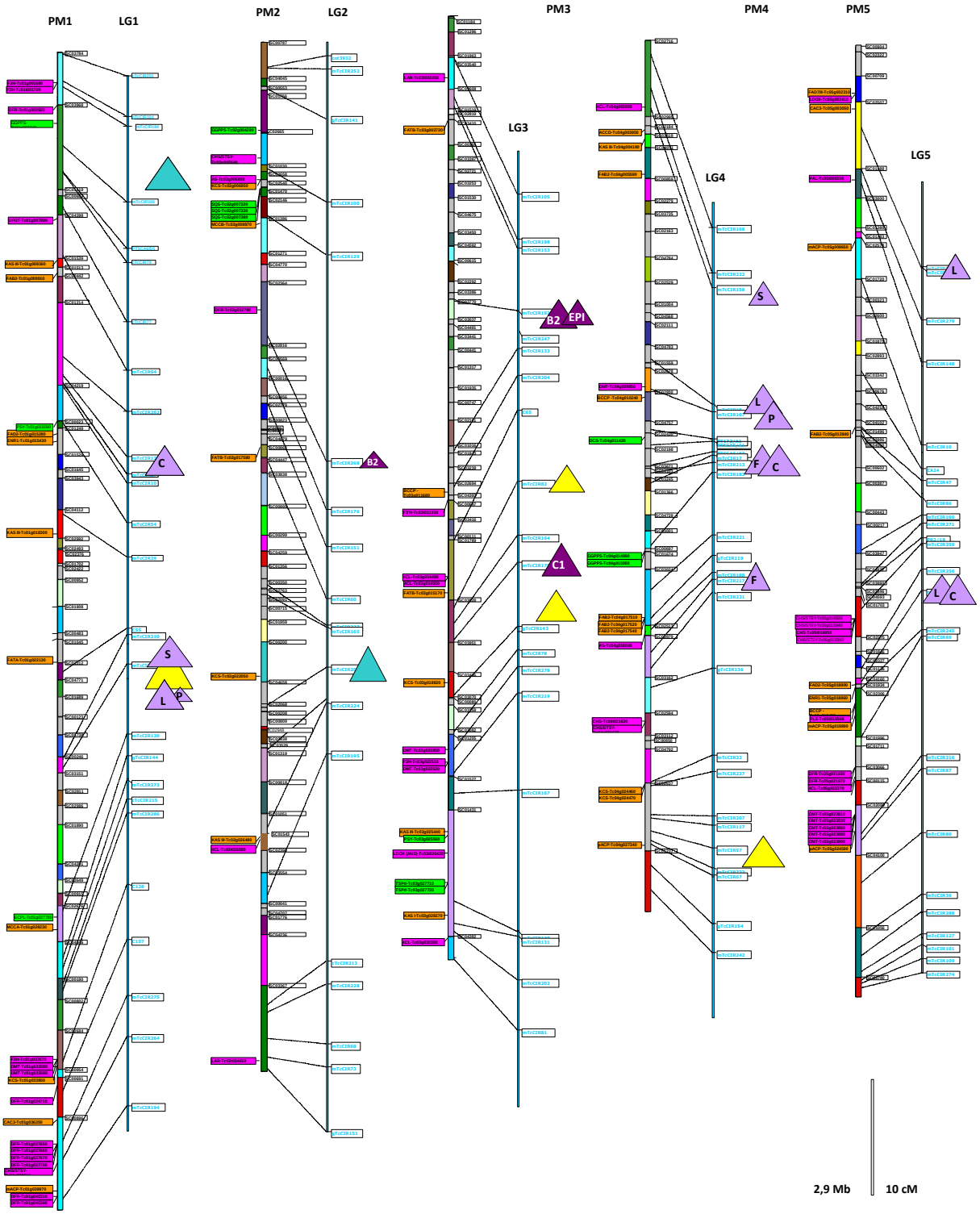
Nature Genetics: doi: 10.1038/ng.736



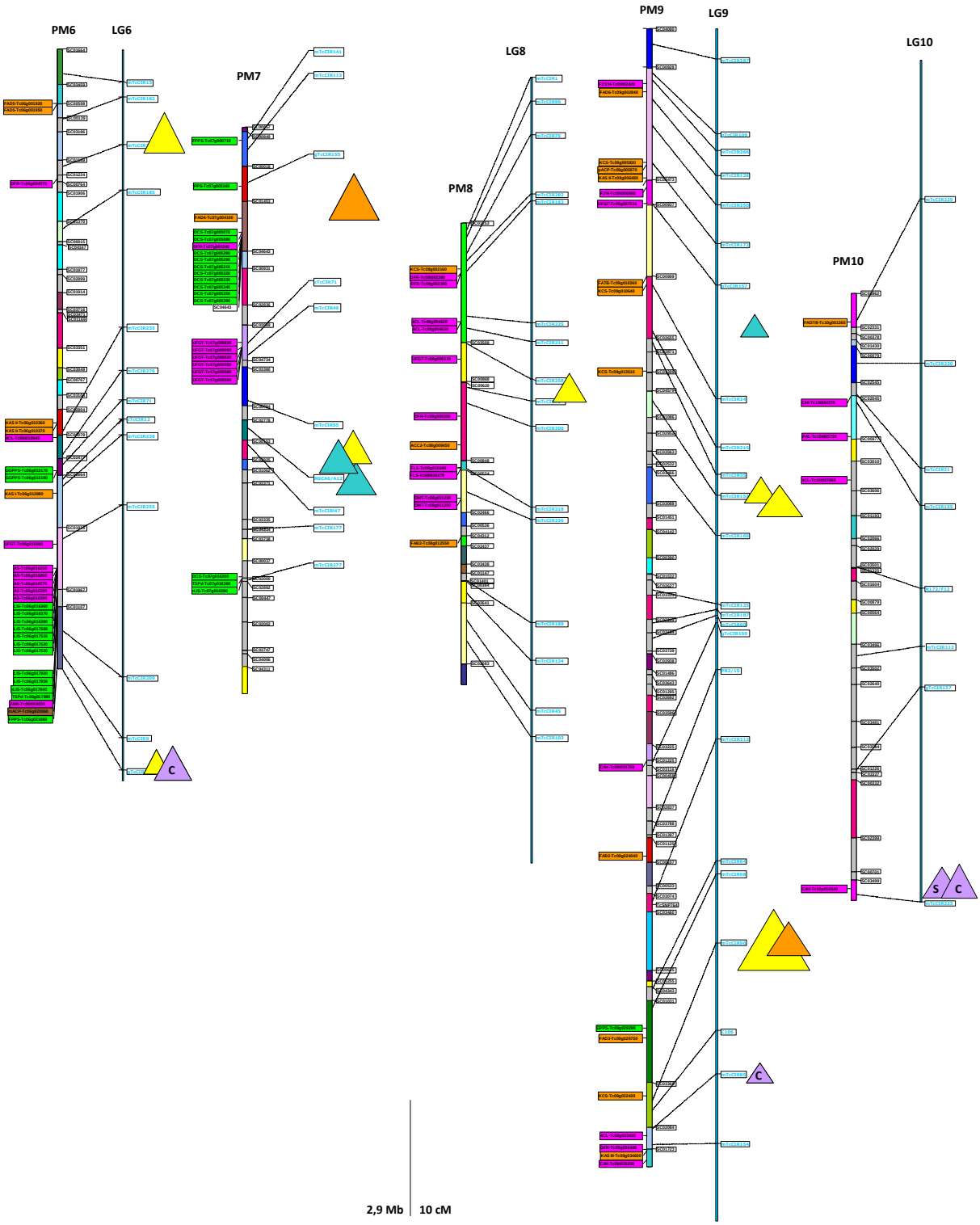
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 16. Mapping of lipid, flavonoid and terpenoid orthologs on pseudomolecules (PM) and comparative genome localizations with QTLs for related traits identified in *T. cacao*. QTLs, represented by triangles are positioned to the right of the linkage groups (LG) and correspond to butter fat content (yellow) and hardness (orange), procyanidin content (purple, EPI : epicathechin, B2, B5 : procyanidin dimer, C1 : procyanidin trimer), cocoa organs color (light purple, C : cotyledon, L : leaf, S : staminode, P : sepal, F :fruit) and to chocolate astringency (green). Lipid (orange bars), flavonol (purple bars) and terpene synthase (green bars) orthologs are represented to the left of the pseudomolecules.

Nature Genetics: doi: 10.1038/ng.736



Nature Genetics: doi: 10.1038/ng.736



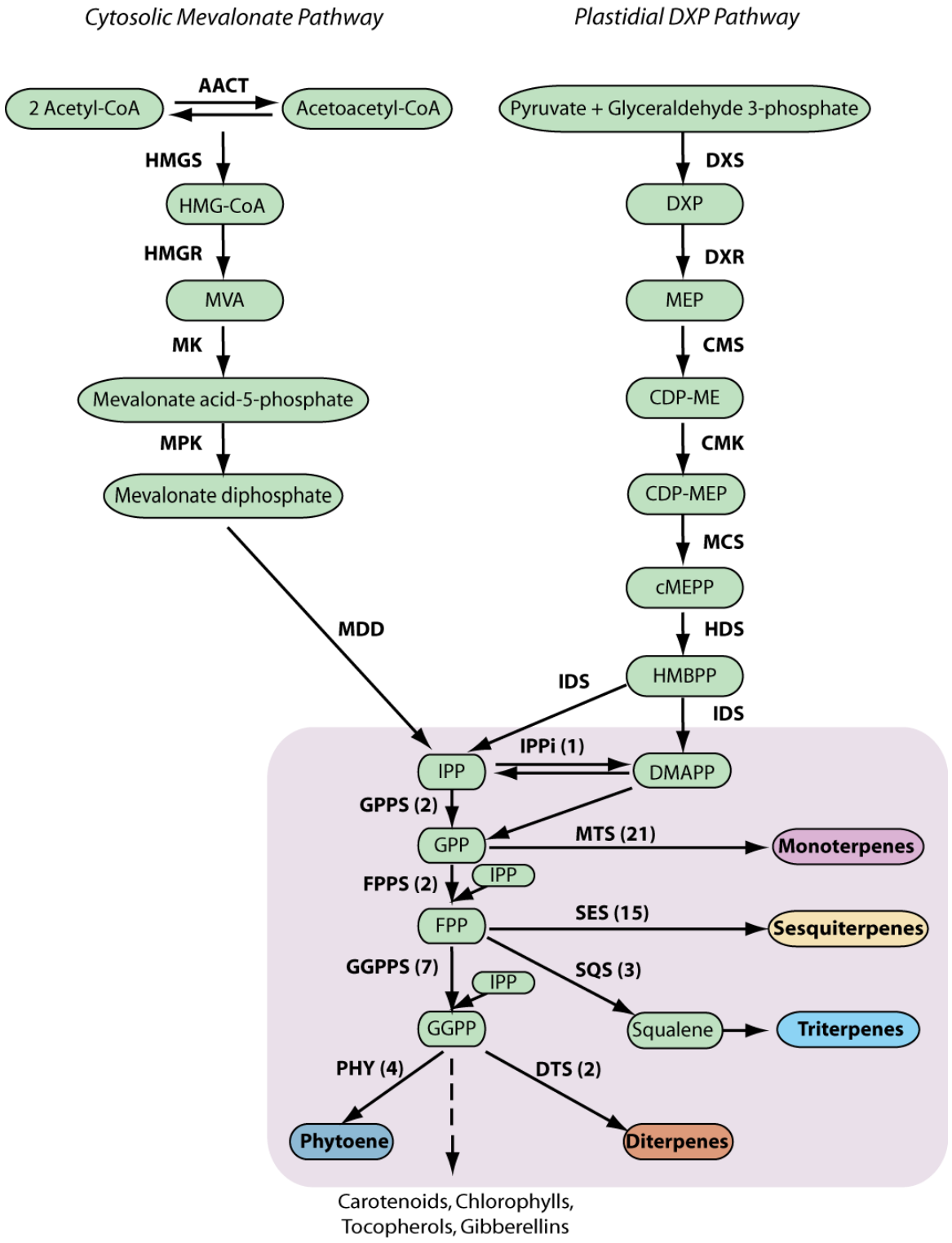
Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 17. **Metabolic pathway for isoprenoid biosynthesis adapted from Liu *et al.*, (2005)[106].** *T. cacao* orthologous gene copy numbers for each enzyme that was determined, as described in Supplementary Table 17, are shown in parentheses beside each enzyme abbreviations:

AACT, acetoacetyl-coenzyme A (CoA) thiolase; **CMS**, 2-*C*-methyl-*D*-erythritol 4-phosphate cytidyl transferase; **DTS**, diterpene synthase; **DXR**, 1-deoxy-*D*-xylulose 5-phosphate reductoisomerase; **DXS**, 1-deoxy-*D*-xylulose 5-phosphate synthase; **FPPS**, farnesyl diphosphate synthase; **GGPPS**, geranylgeranyl diphosphate synthase; **GPPS**, geranyl diphosphate synthase; **HMGR**, 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase; **IPPi**, isopentenyl diphosphate isomerase; **MTS**, monoterpene synthase; **SES**, sesquiterpene synthase; **SQS** squalene synthase; **MK**, mevalonate kinase; **MPK**, mevalonate-5-phosphate kinase; **CMK**, 4-(cytidine 5'-diphospho)-2-*C*-methyl-*D*-erythritol kinase; **MDD**, mevalonate diphosphate decarboxylase; **IDS**, isopentenyl diphosphate/dimethylallyl diphosphate synthase; **MCS**, 2-*C*-methyl-*D*-erythritol 2,4-cyclodiphosphate synthase; **HDS**, 1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate synthase; **PSY**, phytoene synthase; **HMGS**, HMG-CoA synthase;

HMG-CoA, 3S-hydroxy-3-methylglutaryl coenzyme A; **DXP**, 1-deoxy-*D*-xylulose 5-phosphate; **MVA**, 3R-mevalonic acid; **MEP**, 2-*C*-methyl-*D*-erythritol 4-phosphate; **CDP-ME**, 4-(cytidine 5'-diphospho)-2*C*-methyl-*D*-erythritol; **CDP-MEP**, 4-(cytidine 5'-diphospho)-2*C*-methyl-*D*-erythritol 2-phosphate; **cMEPP**, 2*C*-methyl-*D*-erythritol 2,4-cyclodiphosphate; **DMAPP**, Dimethylallyl diphosphate; **HMBPP**, 1-hydroxy-2-methyl-2-(*E*)-butenyl 4-diphosphate; **IPP**, isopentenyl diphosphate; **GPP**, geranyl diphosphate; **FPP**, farnesyl diphosphate; **GGPP**, geranylgeranyl diphosphate.

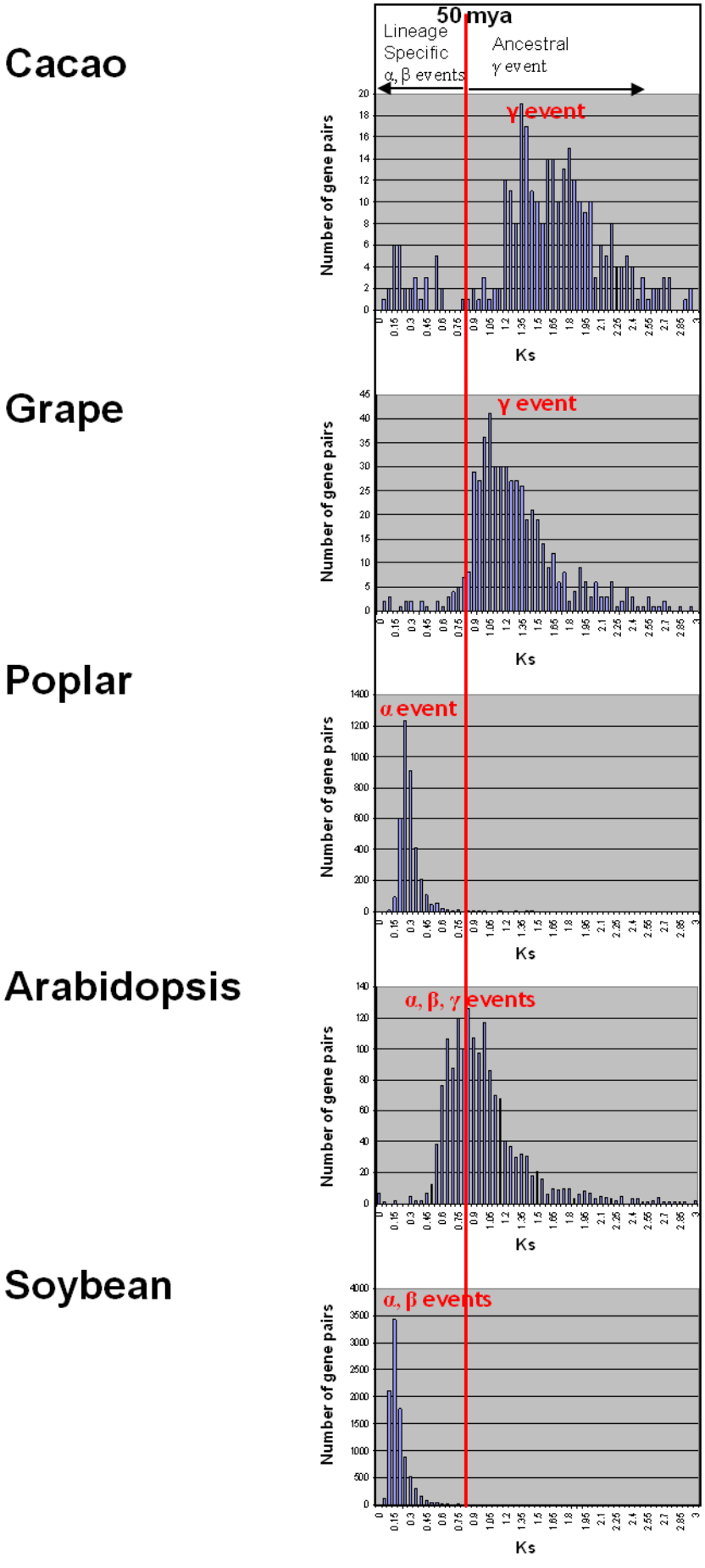
Nature Genetics: doi: 10.1038/ng.736



Nature Genetics: doi: 10.1038/ng.736

Supplementary Fig. 18. Dating of the *T. cacao* genome duplications. The distribution of Ks distance values observed for the paralogous gene pairs identified for the *T.cacao*, grape, poplar, Arabidopsis, soybean genomes are illustrated with bars as number of duplicated gene pairs (y-axis) per Ks values (x-axis) intervals from 0 to 3. The distinct rounds of whole genome duplication (α, β, γ,) reported for the eudicot genome paleohistory are highlighted in red. The red vertical line represent the separation between lineage specific WGD (left) and shared paleo-WGD (right).

Nature Genetics: doi: 10.1038/ng.736



Nature Genetics: doi: 10.1038/ng.736

Supplementary References

1. Mooleedhar, V., Maharaj, W. & O'Brien, H. The collection of Criollo cocoa germplasm in Belize. *Cocoa Grower's Bull.* **49**, 26-40 (1995).
2. Motilal, L.A. et al. The relic Criollo cacao in Belize–genetic diversity and relationship with Trinitario and other cacao clones held in the International Cocoa Genebank, Trinidad. *Plant Genetic Resources* **8**, 106-115 (2010).
3. Lagunes Gálvez, S., Loiseau, G., Paredes, J.L., Barel, M. & Guiraud, J.P. Study on the microflora and biochemistry of cocoa fermentation in the Dominican Republic. *International journal of food microbiology* **114**, 124–130 (2007).
4. Schwan, R.F. & Wheals, A.E. The microbiology of cocoa fermentation and its role in chocolate quality. *Critical reviews in food science and nutrition* **44**, 205–221 (2004).
5. de Brito, E.S. et al. Structural and chemical changes in cocoa (Theobroma cacao L) during fermentation, drying and roasting. *Journal of the Science of Food and Agriculture* **81**, 281–288 (2001).
6. Stark, T., Bareuther, S. & Hofmann, T. Sensory-guided decomposition of roasted cocoa nibs (Theobroma cacao) and structure determination of taste-active polyphenols. *J Agric Food Chem* **53**, 5407-5418 (2005).
7. Luo, M. & Wing, R.A. An improved method for plant BAC library construction. *Methods in Molecular Biology (Clifton, N.J.)* **236**, 3-20 (2003).
8. Ammiraju, J.S.S. et al. The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. *Genome Res.* **16**, 140-147 (2006).
9. Risterucci, A.M. et al. A high-density linkage map of Theobroma cacao L. *TAG Theoretical and Applied Genetics* **101**, 948–955 (2000).
10. http://www.genome.arizona.edu/orders/.
11. Kim, H. et al. Comparative physical mapping between Oryza sativa (AA genome type) and O. punctata (BB genome type). *Genetics* **176**, 379-390 (2007).
12. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-85 (1998).
13. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Research* **12**, 656-664 (2002).
14. Aury, J. et al. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008).
15. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)* **24**, 713-714 (2008).
16. Marie, D. & Brown, S.C. A cytometric exercise in plant DNA histograms, with 2C values for 70 species. *Biology of the Cell / Under the Auspices of the European Cell Biology Organization* **78**, 41-51 (1993).
17. Galbraith, D.W. et al. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science (New York, N.Y.)* **220**, 1049-1051 (1983).
18. Greilhuber, J., Dolezel, J., Lysák, M.A. & Bennett, M.D. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. *Ann. Bot. (Lond.)* **95**, 255-260 (2005).
19. Dolezel, J., Bartos, J., Voglmayr, H. & Greilhuber, J. Nuclear DNA content and genome size of trout and human. *Cytometry. Part A: The Journal of the International Society for Analytical Cytology* **51**, 127-128; author reply 129 (2003).
20. Lanaud, C. et al. A genetic linkage map of Theobroma cacao L. *Theor.Appl. Genet.* **91**, 987–993 (1995).

Nature Genetics: doi: 10.1038/ng.736

21. Pugh, T. et al. A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. *Theor.Appl. Genet.* **108**, 1151-1161 (2004).
22. Fouet, O. et al. Structural characterization and mapping of functional EST-SSR markers in Theobroma cacao. *non published data*
23. Allegre, M. et al. A high-density consensus genetic map for Theobroma cacao L. *non published data*
24. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265-268 (2007).
25. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
26. Jurka, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res* **110**, 462-467 (2005).
27. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
28. D'hont, A. et al. Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Molec. Gen. Genet.* **250**, 405–413 (1996).
29. Argout, X. et al. Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of Theobroma cacao generated from various tissues and under various conditions. *BMC Genomics* **9**, 512 (2008).
30. Argout, X. et al. ESTtik : a semi-automatic cDNA sequence analysis and annotation pipeline including SSR and SNP search tools. *15th International Cocoa Research Conference* **1**, 501-506 (2007).
31. Pertea, G. et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-2 (2003).
32. Mituyama, T. et al. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Research* **37**, D89-92 (2009).
33. Foissac, S. et al. Genome annotation in plants and fungi: EuGène as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
34. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965–978 (2005).
35. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods in Molecular Biology (Clifton, N.J.)* **406**, 89-112 (2007).
36. Poole, R.L. The TAIR database. *Methods in Molecular Biology (Clifton, N.J.)* **406**, 179-212 (2007).
37. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank. *Nucleic Acids Res.* **34**, D16-20 (2006).
38. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178-183 (2010).
39. Wasmuth, J.D. & Blaxter, M.L. prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* **5**, 187 (2004).
40. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res* **37**, D211-215 (2009).
41. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
42. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **13**, 2178-2189 (2003).

Nature Genetics: doi: 10.1038/ng.736

43. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
44. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-6 (2005).
45. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
46. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154-158 (2008).
47. Jones-Rhoades, M.W. & Bartel, D.P. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Molecular Cell* **14**, 787-799 (2004).
48. Ma, Z., Coruh, C. & Axtell, M.J. Arabidopsis lyrata small RNAs: Transient MIRNA and small interfering RNA loci within the Arabidopsis genus. *The Plant Cell Online* **22**, 1090 (2010).
49. http://homes.bio.psu.edu/people/faculty/Axtell/AxtellLab/Software.html.
50. Berardini, T.Z. et al. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol* **135**, 745-755 (2004).
51. Axtell, M.J. & Bowman, J.L. Evolution of plant microRNAs and their targets. *Trends Plant Sci* **13**, 343-349 (2008).
52. Diévart, A. & Clark, S.E. Using mutant alleles to determine the structure and function of leucine-rich repeat receptor-like kinases. *Curr Opin Plant Biol* **6**, 507-16 (2003).
53. http://www.arabidopsis.org.
54. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics* **23**, 205-211 (2009).
55. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320-2 (1998).
56. Larkin, M.A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-8 (2007).
57. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696-704 (2003).
58. Lehti-Shiu, M.D., Zou, C., Hanada, K. & Shiu, S.H. Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol* **150**, 12-26 (2009).
59. Chevenet, F., Brun, C., Banuls, A.L., Jacq, B. & Christen, R. TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).
60. Page, R.D. Visualizing phylogenetic trees using TreeView. *Curr Protoc Bioinformatics* **6**, Unit 6 2 (2002).
61. Dereeper, A. et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**, W465-9 (2008).
62. DeYoung, B.J. & Innes, R.W. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* **7**, 1243-9 (2006).
63. van der Biezen, E.A. & Jones, J.D. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Current Biology: CB* **8**, R226-227 (1998).
64. Mun, J., Yu, H., Park, S. & Park, B. Genome-wide identification of NBS-encoding resistance genes in Brassica rapa. *Molecular Genetics and Genomics: MGG* **282**, 617-631 (2009).
65. Porter, B.W. et al. Genome-wide analysis of Carica papaya reveals a small NBS resistance gene family. *Molecular Genetics and Genomics: MGG* **281**, 609-626 (2009).

Nature Genetics: doi: 10.1038/ng.736

66. Ameline-Torregrosa, C. et al. Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant Medicago truncatula. *Plant Physiology* **146**, 5-21 (2008).
67. McHale, L., Tan, X., Koehl, P. & Michelmore, R.W. Plant NBS-LRR proteins: adaptable guards. *Genome Biology* **7**, 212 (2006).
68. Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Research* **36**, D281-288 (2008).
69. McDonnell, A.V., Jiang, T., Keating, A.E. & Berger, B. Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics (Oxford, England)* **22**, 356-358 (2006).
70. Meyers, B.C. et al. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *The Plant Journal: For Cell and Molecular Biology* **20**, 317-332 (1999).
71. Yang, S., Zhang, X., Yue, J., Tian, D. & Chen, J. Recent duplications dominate NBS-encoding gene expansion in two woody species. *Molecular Genetics and Genomics: MGG* **280**, 187-198 (2008).
72. Meyers, B.C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R.W. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell* **15**, 809-834 (2003).
73. http://niblrrs.ucdavis.edu/At_RGenes/.
74. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002).
75. Pei, J. & Grishin, N.V. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics (Oxford, England)* **17**, 700-712 (2001).
76. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).
77. Vlot, A.C., Klessig, D.F. & Park, S. Systemic acquired resistance: the elusive signal(s). *Current Opinion in Plant Biology* **11**, 436-442 (2008).
78. Dong, X. NPR1, all things considered. *Current Opinion in Plant Biology* **7**, 547-552 (2004).
79. Koornneef, A. & Pieterse, C.M.J. Cross talk in defense signaling. *Plant Physiology* **146**, 839-844 (2008).
80. Stogios, P.J., Downs, G.S., Jauhal, J.J.S., Nandra, S.K. & Privé, G.G. Sequence and structural analysis of BTB domain proteins. *Genome Biology* **6**, R82 (2005).
81. Boyle, P. et al. The BTB/POZ domain of the Arabidopsis disease resistance protein NPR1 interacts with the repression domain of TGA2 to negate its function. *The Plant Cell* **21**, 3700-3713 (2009).
82. Johnson, C., Mhatre, A. & Arias, J. NPR1 preferentially binds to the DNA-inactive form of Arabidopsis TGA2. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1779**, 583–589 (2008).
83. Liu, G., Holub, E.B., Alonso, J.M., Ecker, J.R. & Fobert, P.R. An Arabidopsis NPR1-like gene, NPR4, is required for disease resistance. *Plant J.* **41**, 304-318 (2005).
84. McKim, S.M. et al. The BLADE-ON-PETIOLE genes are essential for abscission zone formation in Arabidopsis. *Development* **135**, 1537-1546 (2008).
85. Zhang, Y. et al. Negative regulation of defense responses in Arabidopsis by two NPR1 paralogs. *The Plant Journal: For Cell and Molecular Biology* **48**, 647-656 (2006).
86. Le Henanff, G. et al. Characterization of Vitis vinifera NPR1 homologs involved in the regulation of pathogenesis-related gene expression. *BMC Plant Biology* **9**, 54 (2009).

Nature Genetics: doi: 10.1038/ng.736

87. Chen, Y.Y. et al. Virus-induced gene silencing reveals the involvement of ethylene-, salicylic acid-and mitogen-activated protein kinase-related defense pathways in the resistance of tomato to bacterial wilt. *Physiologia Plantarum* **136**, 324–335 (2009).
88. Malnoy, M., Jin, Q., Borejsza-Wysocka, E.E., He, S.Y. & Aldwinckle, H.S. Overexpression of the apple MpNPR1 gene confers increased disease resistance in Malus x domestica. *Molecular Plant-Microbe Interactions: MPMI* **20**, 1568-1580 (2007).
89. Chern, M., Fitzgerald, H.A., Canlas, P.E., Navarre, D.A. & Ronald, P.C. Overexpression of a rice NPR1 homolog leads to constitutive activation of defense response and hypersensitivity to light. *Molecular Plant-Microbe Interactions: MPMI* **18**, 511-520 (2005).
90. Radwan, O. et al. Genetic diversity and genomic distribution of homologs encoding NBS-LRR disease resistance proteins in sunflower. *Molecular Genetics and Genomics: MGG* **280**, 111-125 (2008).
91. Huang, S. et al. The genome of the cucumber, Cucumis sativus L. *Nature Genetics* **41**, 1275-1281 (2009).
92. Kohler, A. et al. Genome-wide identification of NBS resistance genes in Populus trichocarpa. *Plant Molecular Biology* **66**, 619-636 (2008).
93. Lanaud, C. et al. A meta–QTL analysis of disease resistance traits of Theobroma cacao L. *Mol. Breed.* **24**, 361–374 (2009).
94. , I.S. et al. Mapping of Quantitative Trait Loci for Butter Content and Hardness in Cocoa Beans (Theobroma cacao L.). *Plant Mol. Bio. Rep.* **27**, 177–183 (2009).
95. Alvarez, C. et al. Association mapping studies of biochemical compounds related to quality traits in Theobroma. *non published data*
96. Marcano, M. et al. A genomewide admixture mapping study for yield factors and morphological traits in a cultivated cocoa (Theobroma cacao L.) population. *Tree Genetics & Genomes* **5**, 329–337 (2009).
97. Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. & Feuillet, C. Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Briefings Bioinf.* **10**, 619-630 (2009).
98. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725-736 (1994).
99. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nature Genetics* **20**, 43-45 (1998).
100. Gaut, B.S., Morton, B.R., McCaig, B.C. & Clegg, M.T. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 10274-10279 (1996).
101. Degroeve, S., Saeys, Y., De Baets, B., Rouzé, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics (Oxford, England)* **21**, 1332-1338 (2005).
102. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).
103. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599 (2007).
104. Baud, S. & Lepiniec, L. Physiological and developmental regulation of seed oil production. *Prog Lipid Res* **49**, 235-49 (2010).
105. Lepiniec, L. et al. Genetics and biochemistry of seed flavonoids. *Annual Review of Plant Biology* **57**, 405-430 (2006).

Nature Genetics: doi: 10.1038/ng.736

106. Liu, Y., Wang, H., Ye, H. & Li, G. Advances in the plant isoprenoid biosynthesis pathway and its metabolic engineering. *J. Integr. Plant. Biol.* **47**, 769–782 (2005).

Nature Genetics: doi: 10.1038/ng.736