nature genetics

The genome of *Theobroma cacao*

Xavier Argout[1,24], Jerome Salse[2,24], Jean-Marc Aury[3–5,24], Mark J Guiltinan[6,7,24], Gaetan Droc[1], Jerome Gouzy[8], Mathilde Allegre[1], Cristian Chaparro[9], Thierry Legavre[1], Siela N Maximova[6], Michael Abrouk[2], Florent Murat[2], Olivier Fouet[1], Julie Poulain[3–5], Manuel Ruiz[1], Yolande Roguet[1], Maguy Rodier-Goud[1], Jose Fernandes Barbosa-Neto[9], Francois Sabot[9], Dave Kudrna[10], Jetty Siva S Ammiraju[10], Stephan C Schuster[11], John E Carlson[12,13], Erika Sallet[8], Thomas Schiex[14], Anne Dievart[1], Melissa Kramer[15], Laura Gelley[15], Zi Shi[7], 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[15,7], Spencer Brown[18], Mickael Bourge[18], Wolfgang Golser[10], Xiang Song[10], 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[10], W Richard McCombie[15], Emmanuel Guiderdoni[1], Francis Quetier[23], Olivier Panaud[9], Patrick Wincker[3–5], Stephanie Bocs[1] & Claire Lanaud[1]

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

Theobroma cacao L. is a diploid tree fruit species ($2n = 2x = 20$ (ref. 1)) endemic to the South American rainforests. Cocoa was domesticated approximately 3,000 years ago[2] in Central America[3]. The Criollo cocoa variety, having a nearly unique and homozygous genotype, was among the first to be cultivated[4]. Criollo is now one of the two cocoa varieties providing fine flavor chocolate.

However, due to its poor agronomic performance and disease susceptibility, more vigorous hybrids created with foreign (Forastero) genotypes have been introduced. These hybrids, named Trinitario, are now widely cultivated[5]. Here we report the sequence of a Belizean Criollo plant[6].

Consumers have shown an increased interest for high-quality chocolate, and for dark chocolate, containing a higher percentage of cocoa[7]. Fine-cocoa production is nevertheless estimated to be less than 5% of the world cocoa production due to the low productivity and disease susceptibility of the traditional fine-flavor cocoa varieties. Therefore, breeding of improved Criollo varieties is important for sustainable production of fine-flavor cocoa.

3.7 million tons of cocoa are produced annually (see URLs). However, fungal, oomycete and viral diseases, as well as insect pests, are responsible for an estimated 30% of harvest losses (see URLs). Like many other tropical crops, knowledge of *T. cacao* genetics and genomics is limited. To accelerate progress in cocoa breeding and the understanding of its biochemistry, we sequenced and analyzed the genome of a Belizean Criollo genotype (B97-61/B2). This genotype is suitable

[1]Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)-Biological Systems Department-Unité Mixte de Recherche Développement et Amélioration des Plantes (UMR DAP) TA A 96/03-34398, Montpellier, France. [2]Institut National de la Recherché Agronomique UMR 1095, Clermont-Ferrand, France. [3]Commissariat à l'Energie Antomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France. [4]Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France. [5]Université d'Evry, Evry, France. [6]Penn State University, Department of Horticulture and the Huck Institutes of the Life Sciences, University Park, Pennsylvania, USA. [7]Penn State University, Plant Biology Graduate Program and the Huck Institutes of the Life Sciences, University Park, Pennsylvania, USA. [8]Institut National de la Recherche Agronomique (INRA)-CNRS Laboratoire des Interactions Plantes Micro-organismes (LIPM), Castanet Tolosan Cedex, France. [9]UMR 5096 CNRS-Institut de Recherche pour le Développement (IRD)-Université de Perpignan Via Domitia (UPVD), Laboratoire Génome et Développement des Plantes, Perpignan Cedex, France. [10]Arizona Genomics Institute and School of Plant Sciences, University of Arizona, Tucson, Arizona, USA. [11]Penn State University, Department of Biochemistry and Molecular Biology, University Park, Pennsylvania, USA. [12]Penn State University, the School of Forest Resources and the Huck Institutes of the Life Sciences, University Park, Pennsylvania, USA. [13]The Department of Bioenergy Science and Technology (WCU), Chonnam National University, Buk-Gu, Gwangju, Korea. [14]Unité de Biométrie et d'Intelligence Artificielle (UBIA), UR875 INRA, Castanet Tolosan, France. [15]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [16]INRA, UR 1279 Etude du Polymorphisme des Génomes Végétaux, CEA Institut de Génomique, Centre National de Génotypage, CP5724, Evry, France. [17]Penn State University, Bioinformatics and Genomics PhD Program and Department of Biology, University Park, Pennsylvania, USA. [18]Institut des Sciences du Végétal, UPR 2355, CNRS, Gif-Sur-Ivette, France. [19]Centre National de la Recherche Agronomique (CNRA), Divo, Côte d'Ivoire. [20]Comissão Executiva de Planejamento da Lavoura Cacaueira (CEPLAC), Itabuna Bahia, Brazil. [21]Centro Nacional de Biotecnología Agrícola, Instituto de Estudios Avanzados (IDEA), Caracas, Venezuela. [22]Chocolaterie VALRHONA, Tain l'Hermitage, France. [23]Département de Biologie, Université d'Evry Val d'Essonne, Evry, France. [24]These authors contributed equally to this work. Correspondence should be addressed to X.A. (xavier.argout@cirad.fr).

Received 10 August; accepted 1 December; published online 26 December 2010; doi:10.1038/ng.736

Table 1 Global statistics of the genome assembly and annotation of *Theobroma cacao*

Assembly		Number	N50 (kb)	Longest (kb)	Size (Mb)	Percentage of the assembly
Contigs	All	25,912	19.8	190	291.4	–
Scaffolds	All	4,792	473.8	3,415	326.9	100
	Anchored on chromosomes	385		3,415	218.4	66.8
	Anchored on chromosomes and oriented	206		3,415	162.8	49.8
Annotation		Number				
Genes	Protein coding	28,798			96.4	29.4
	rRNA	6[a]			<0.03	<0.01
	tRNA	473			<0.03	<0.01
	miRNA	83			<0.03	<0.01
	Transposable Element	17,342			52.6	16.1
Transposable elements		67,575			84.0	25.7

[a]The rRNA number is greatly underestimated due to the sequencing method (**Supplementary Note**).

for a high-quality genome sequence assembly because it is highly homozygous as a result of the many generations of self-fertilization that occurred during the domestication process.

RESULTS

Sequencing and assembly

We used a genome-wide shotgun strategy incorporating Roche/454, Illumina and Sanger sequencing technologies. The International Cocoa Genome Sequencing consortium (ICGS) produced a total of 17.6 million 454 single reads, 8.8 million 454 paired end reads, 398.0 million Illumina paired end reads and about 88,000 Sanger BAC end reads, corresponding to 26 Gb of raw data (**Supplementary Note**, **Supplementary Tables 1** and **2** and **Supplementary Figs. 1** and **2**). We used the Roche/454 and Sanger raw data to produce the assembly. This represented ×16.7 coverage of the 430-Mb genome of B97-61/B2, whose size was estimated by flow cytometry (**Supplementary Note** and **Supplementary Table 3**).

This assembly, performed with Newbler software (Roche, Inc.), consists of 25,912 contigs and 4,792 scaffolds (**Table 1**, **Supplementary Note** and **Supplementary Table 4**). Eighty percent of the assembly is in 542 scaffolds, and the largest scaffold measures 3.4 Mb. We determined the N50 (the scaffold size above which 50% of the total length of the sequence assembly can be found) to be 473.8 kb. The total length of the assembly was 326.9 Mb, which represents 76% of the estimated genome size of the *T. cacao* genotype B97-61/B2 (430 Mb). In addition, we used a high coverage of Illumina data (×44 coverage of the genome), which has a different error profile than 454 data, to improve accuracy of the *T. cacao* genome sequence (Online Methods).

The resulting assembly appears to cover a very large proportion of the euchromatin of the *T. cacao* genome. We confirmed the high genome coverage of this assembly by comparing it to the unigene resource (38,737 unigenes assembled from 715,457 expressed sequence tag (EST) sequences from the B97-61/B2 genotype). We recovered 97.8% of the unigene resource in the *T. cacao* genome assembly.

Using a set of 1,259 molecular markers from a consensus genetic linkage map established from two mapping populations, 94% of the markers (1,192) were unambiguously located on the assembly, allowing us to anchor 67% of the assembled 326 Mb along the ten *T. cacao* linkage groups. The remaining 33% of the genome assembly was in 2,207 scaffolds. Fifty percent of the assembly could be anchored and oriented (**Supplementary Note**, **Supplementary Table 5** and **Supplementary Fig. 3**). The average ratio of genetic-to-physical distance was 1 cM per 444 kb in centromeric regions and 1 cM per 146 kb in distal chromosomal regions.

Gene content and repeated sequences annotation

We performed the identification and annotation of transposable elements using a two step approach: the first approach was based on the *de novo* identification of transposable elements from the assembled scaffolds and the second one was based on the search for transposable elements from the unassembled reads (**Supplementary Note**).

This *de novo* search lead to the identification of a total of 67,575 transposable element–related sequences in the assembled cocoa sequences (**Table 1**). The second step led to the identification of three highly repeated transposable element families from the unassembled reads dataset. The most common transposable element was a long terminal repeat (LTR) retrotransposon that we named Gaucho. It is a Copia-like element 11,297 bp in length that is repeated approximately 1,100 times, based upon its occurrences in the 454 unassembled sequences. Fluorescence *in situ* hybridization (FISH) analysis revealed that Gaucho is distributed on all chromosome arms but is found mainly in their median region (as opposed to the centromeric and telomeric regions), a classical feature shared by many LTR retrotransposons in plants (**Fig. 1**). The other two highly repeated families were another Copia-like LTR retrotransposon and a Mu-type transposon.

Class I elements were the most abundant, representing 69% of the total transposable elements in the cocoa genome, with a total of 290 Gypsy-like and 159 Copia-like families. In addition, we identified 36 transposons and 1,353 miniature inverted-repeat transposon element (MITE) families (class II) (**Supplementary Table 6**). The most highly repeated transposon families were Mutator and Vandal (Mu type), with copy numbers of 994 and 1,978, respectively. Transposable elements were particularly abundant in centromeric regions, as illustrated in **Figure 2**; this feature was already observed in other sequenced genomes[8].

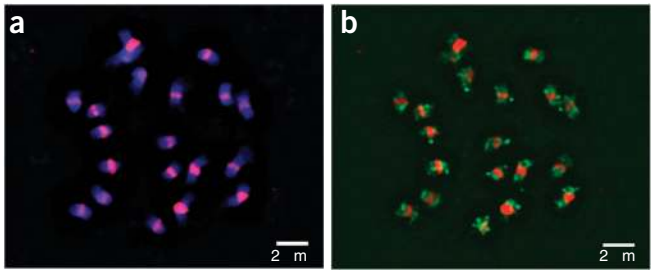

Figure 1 FISH analysis of *T. cacao* chromosomes. (**a**) *In situ* hybridization of *T. cacao* chromosomes stained with DAPI (blue) using a ThCen repeat probe (red). (**b**) *In situ* hybridization using Gaucho LTR retrotransposon (green) and ThCen repeat (red) probes.

Altogether, the transposable elements identified in both assembled (84 Mb) and un-assembled (20.3 Mb) reads represent about 24% of the *T. cacao* genome. This value is substantially lower than that for other sequenced genomes of similar size, for example, rice (35%, for 380 Mb)[9] and grape (41.4%, for 475 Mb)[10]. However, sequencing and assembling of highly repeated sequences can be expected to be the major limitation of *de novo* sequencing of a complex genome using next-generation sequencing. This is particularly true for transposon element families that have undergone very recent amplification, like in the case of Gaucho. Therefore, we conclude that the total contribution of repetitive elements to the whole cocoa genome may be underestimated.

In addition, we identified a tandem repeat sequence (that we named ThCen) from the 454 repeat reads. This tandem repeat is 212 bp long and, when used as probe in a FISH experiment, was found to be located in the centromeres of all cocoa chromosomes (**Fig. 1**). Tandem repeats and retrotransposons are the major components of plant centromeres[11]. The copy number of Gaucho and ThCen repeat sequences varied up to 2.5-fold among *T. cacao* genotypes from various genetic origins. We observed a positive Pearson correlation (r = 0.56) between genome size and ThCen repeat copy number, suggesting a possible contribution of ThCen repeats in genome size variation (**Supplementary Note**, **Supplementary Fig. 4** and **Supplementary Table 7**).

We annotated the genome sequence using the integrative gene prediction package EUGene[12] following specific training for *T. cacao* (**Supplementary Note**). Homology searching and functional annotation (**Supplementary Fig. 5**) led to the identification of 28,798 *T. cacao* protein-coding genes (**Table 1**), with an average gene size of 3,346 bp and a mean of 5.03 exons per gene (**Supplementary Table 8**). Compared to the smaller *Arabidopsis thaliana* genome, the *T. cacao* genome has a higher gene number, a similar exon number per gene and a lower mean gene density per 100 kb (**Supplementary Table 8**). The genes in *T. cacao* were more abundant in subtelomeric regions (**Fig. 2**), as previously observed in other sequenced plant genomes[13].

The comparison of cocoa, *A. thaliana*, grape, soybean and poplar proteomes revealed 6,362 clusters of genes (totaling 52,176 genes) distributed among all five eudicot genomes and 682 gene families (totaling 2,053 genes) specific to the cocoa genome (**Fig. 3**, **Supplementary Note** and **Supplementary Table 9**). Most of these 682 gene families encode hypothetical proteins, as supported at the transcript level. The functional analysis of these five proteomes using gene ontology terms revealed a similar pattern among them (**Supplementary Note** and **Supplementary Fig. 6**). A specific feature of the *T. cacao* clusters common to the other species was the relatively high level of metabolic and cellular processes (**Supplementary Note** and **Supplementary Figs. 6** and **7**). On the other hand, grape- and cocoa-specific clusters showed the highest unknown-function percentages (**Supplementary Note** and **Supplementary Fig. 7**).

MicroRNAs (miRNAs) are short noncoding RNAs that regulate target genes transcriptionally or post-transcriptionally. Many of them play important roles in development and stress responses[14]. A total of 83 *T. cacao* miRNAs from 25 families were computationally predicted based on sequence similarity to known plant miRNAs in

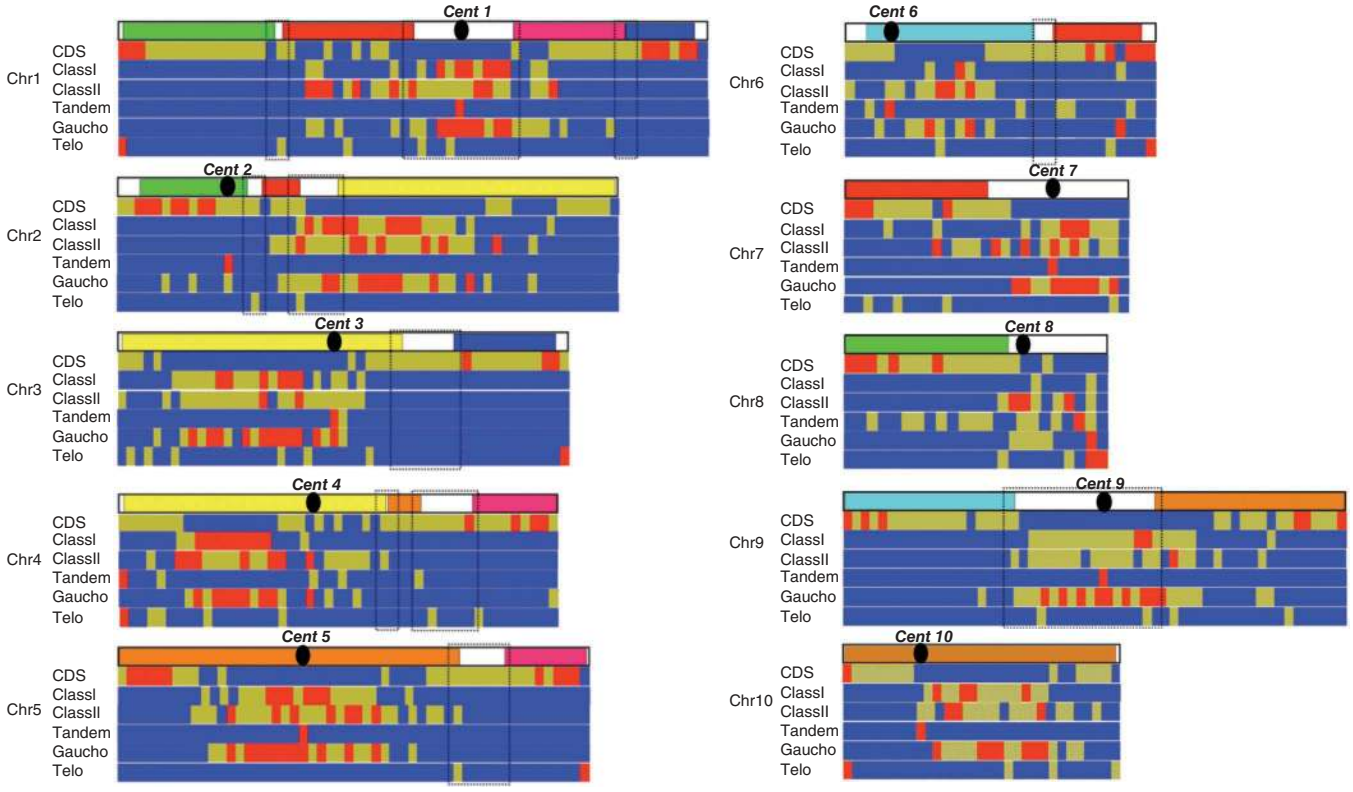

Figure 2 *T. cacao* genome heat map. The ten *T. cacao* chromosomes harboring 11 chromosome fusions (in black dotted boxes) identified in these genomes are illustrated according to their ancestral chromosomal origin (see paleo-chromosome color code in **Fig. 4**). Centromeres are marked 'Cent'. For the ten chromosomes, heat maps are provided for the CDS (blue <60%, yellow 60%–90% and red >90%), class I and II transposable elements (blue <80%, yellow >80% and red ~100%), ThCen and Gaucho elements (blue <50% of maximum, yellow ≥50% of maximum and red = maximum) and telomeric repeats (blue = 0, yellow <40% and red >40%). Only the elements present in the assembled part of the genome are represented. Therefore, the genome distribution of the repeated sequences represented in this figure could be biased due to the major limitations of *de novo* sequencing of complex genomes using next-generation sequencing (NGS), which is limited in its ability to assemble highly repeated sequences.

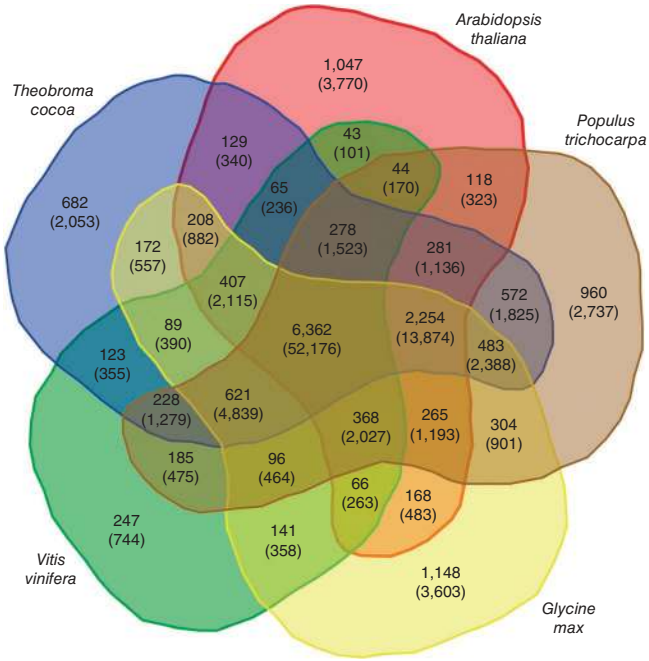

Figure 3 Venn diagram showing the distribution of shared gene families among *Theobroma cacao*, *Arabidopsis thaliana, Populus trichocarpa, Glycine max* and *Vitis vinifera*. Numbers in parentheses indicate the number of genes in each cluster. The Venn diagram was created with web tools provided by the Bioinformatics and Systems Biology of Gent (see URLs).

miRBase 14 (ref. 15) (**Supplementary Note**, **Supplementary Table 10** and **Supplementary Figs. 8** and **9**). Ninety-one *T. cacao* miRNA targets were predicted. Most predicted targets were homologous to known miRNA targets in other plant species, but there was a profound bias toward putative transcription factors compared to the other species (**Supplementary Table 11**), suggesting that miRNAs are major regulators of gene expression in *T. cacao*.

Disease resistance–related genes

Fungal and oomycete diseases are a major constraint to world cocoa production, and the search for natural disease resistance is one of the main objectives of all *T. cacao* breeding programs. Resistance genes (*R* genes) are divided into 2 classes: *NBS-LRR*, the nucleotide-binding site leucine-rich repeat class of genes, and *RPK*, the receptor protein kinase class of genes[16].

Within the *RPK* class, one family of plant-specific transmembrane receptors was shown to also possess *LRR*-class genes in its extracellular domain and have important roles in defense responses or in plant development[17]. The *LRR-RLK* family consists of more than 200 members in *A. thaliana* and more than 400 members in poplar[18]. Here we show that the *T. cacao* genome contains at least 253 *LRR-RLK* genes orthologous to *Arabidopsis LRR-RLK* genes (**Supplementary Note**). Reports indicate that the *LRR-RLK* family is divided into 19 subfamilies[18]. In Viridiplantae, some of these subfamilies have expanded dramatically. As in *Populus trichocarpa*, the *LRR-XII* subfamily has greatly expanded in the *T. cacao* genome (**Supplementary Note**, **Supplementary Table 12** and **Supplementary Table 13**), with approximately 36% of the *LRR* genes belonging to this subfamily.

The *NBS* genes, encoding nucleotide-binding site proteins, also play an important role in resistance to pathogens and in the cell cycle[19]. The *NBS* gene family is rather abundant in plant genomes, ranging from 0.6% to approximately 2% of the total gene number (**Supplementary Note**).

We identified a total of 297 non-redundant *NBS*-encoding orthologous genes in the *T. cacao* genome (**Supplementary Note**, **Supplementary Table 14** and **Supplementary Figs. 10** and **11**). Among these genes, one class, characterized by the *TIR* (encoding the toll interleukin receptor) motif, is markedly underrepresented in the *T. cacao* genome compared to other eudicot plants, with only 4% of *NBS* orthologous genes containing *TIR* motifs, in contrast to grape, poplar, Medicago (20%) or *Arabidopsis* (65%) (**Supplementary Table 14**). The *TIR* motifs have been shown to be present in basal angiosperms and eudicots, but are nearly absent in monocots[20]. It has been suggested that the *TIR-NBS-LRR* resistance genes are more ancient than the divergence of angiosperm and gymnosperms and that they have been lost in the cereal genomes[21]. Their lower level in the cocoa genome compared to other eudicots, and the close relatedness of cocoa to a common eudicot ancestor as shown by paleohistory studies (see below), suggests a divergent evolution of *NBS-TIR* orthologous cocoa genes from an ancestral locus, leading to a lower expansion of this gene family in *T. cacao*.

Another gene family that plays a major role in plant defense is the *NPR* gene family. *NPR1* is an *Arabidopsis* BTB/POZ domain protein that acts as a central mediator of the plant defense signal transduction pathway[22]. We surveyed the *T. cacao* genome sequence and found four related *T. cacao* orthologous genes corresponding to each of the *NPR1* subfamilies found in *Arabidopsis* (**Supplementary Note** and **Supplementary Fig. 12**). Recently, we showed that one of these genes (located on chromosome 9) is a functional ortholog of *Arabidopsis NPR1* by transgenic complementation[23].

We mapped the *NBS*, *LRR-LRK* and *NPR1* orthologous genes along the *T. cacao* pseudomolecules (**Supplementary Note** and **Supplementary Fig. 13**). They were distributed across the ten chromosomes, with a large number being organized in clusters, as is classically observed for these classes of genes[24].

A meta-analysis of quantitative trait loci (QTL) related to disease resistance previously identified in *T. cacao* was recently done[25]. We compared the QTL genetic localizations found in this previous study with the distribution of *NBS-LRR*, *LRR-LRK* and *NPR* orthologous cocoa genes (**Supplementary Fig. 13**). Considering an average confidence interval of about 20 cM for the 76 QTLs identified[25], most of the QTLs are located in genome regions containing candidate resistance genes. However, due to the fact that a large number of QTLs and candidate resistance genes are widespread across the genome, many colocalizations may have occurred at random. Therefore, the candidate genes potentially underlying QTLs need to be further studied by functional genomics approaches to confirm their potential roles in disease resistance in cocoa.

Genes potentially involved in cocoa qualities

T. cacao seeds are fermented, dried and then processed into cocoa mass (ground, dehusked seeds), cocoa butter (triacylglycerol storage lipids from cotyledonary endosperm cells) and cocoa powder (defatted mass consisting primarily of cell walls, endosperm storage proteins, starch and proanthocyanins, as well as other flavonoids, aromatic terpenes, theobromine and many other metabolites). In order to characterize the gene families involved in cocoa quality traits, we used a translational approach to survey the *T. cacao* genome using molecular and biochemical knowledge from *Arabidopsis*, poplar, grape and other model plant species.

Oils, proteins, starch and various secondary metabolites such as flavonoids, alkaloids and terpenoids comprise the principal molecular components of cocoa affecting flavor and quality. Storage lipids (triacylglycerols) provide carbon and energy reserves for germinating cocoa embryos (**Supplementary Fig. 14**). *T. cacao* seed storage

lipids (cocoa butter), representing 50% of the dry seed weight, are exceptional in their very high level of stearate (30–37%), which gives cocoa butter its relatively high melting point (34–38 °C)[26]. The unique fatty acid profile of cocoa butter enhances the olfactory qualities of chocolate and confectionaries and makes it valuable for cosmetic and pharmaceutical products. We discovered a total of 84 orthologous *T. cacao* genes potentially involved in lipid biosynthesis, which is 13 more than those discovered in *Arabidopsis*[27] (**Supplementary Table 15**). Consistent with the large amount of storage lipids produced in *T. cacao* seeds, the genome contained five additional genes encoding acyl-ACP thioesterase fat B (*FATB*) and three additional genes encoding ketoacyl-ACP synthase, the two key workhorse enzymes leading to the synthesis of triacylglycerols.

Flavonoids are a diverse group of plant secondary metabolites that play many important roles during plant development[28]. They are involved in plant defense against insects, pathogens and microbes, in absorption of free radicals and ultraviolet light, and in attraction of beneficial symbionts and pollinators. Proanthocyanidins are flavonoid polymers that are present in large amounts in *T. cacao* seeds (**Supplementary Fig. 15**). Recent evidence suggests that proanthocyanidins may be beneficial to human health by improving cardiovascular health, providing cancer chemopreventative effects and also through neuroprotective activities[29,30]. We identified 96 *T. cacao* genes orthologous to *Arabidopsis* genes that are involved in the flavonoid biosynthetic pathway (**Supplementary Table 16**), which is 60 more than are present in *Arabidopsis*. Of these genes, we evaluated the function of *TcANS*, *TcANR* and *TcLAR* in transgenic *Arabidopsis* and Tobacco, demonstrating that they are functional orthologs of *Arabidopsis* genes[31]. Notably, although *Arabidopsis* has only one gene encoding dihydroflavonol-4-reductase (*DFR*), the *T. cacao* genome contains 18 orthologous *DFR* genes. DFR catalyzes the reaction that produces the flavan-3,4-diols, the immediate precursors of the flavonoids catechin and epicatachin. These compounds can accumulate to as much as 8% of the dry *T. cacao* seed, making *T. cacao* one of the richest known sources of this phytonutrient[32].

Terpenoids constitute a large family of natural compounds and play diverse roles in plants as hormones, pigments and in plant-environment interactions and defense. They are major components of resins, essential oils and aromas[33]. Among them, two subclasses of terpenoids are particularly involved in aromas: monoterpenes (C10), which represent aromatic compounds that are the basis of floral essences and essential oils, and sesquiterpenes (C15), which may also constitute a defense response of plants toward microorganisms or insect aggression. Compared to bulk cocoa, a higher level of monoterpenes (such as linalool, an acyclic monoterpene alcohol found in the floral scent of *Clarkia breweri* and of many other plants species) has been observed in fine-flavored cocoa varieties like Criollo and Nacional, which are characterized by fruity and floral notes[34,35].

We identified 57 *T. cacao* genes that are orthologs of *Arabidopsis* genes that encode terpene synthase (*TPS*), which catalyze terpenoid synthesis[33], and nine pseudogenes (**Supplementary Table 17** and **Supplementary Figs. 16** and **17**). This number is higher than in *Arabidopsis* and poplar[36], which have 30 and 40 genes, respectively, and lower than in grape[10], which has 89 functional genes. The classification of *TPS*s in the different subclasses revealed that 34% of them correspond to monoterpenes and 31% correspond to sesquiterpenes. Two gene families are particularly expanded in *T. cacao*: the linalool synthase family (monoterpenes), which is represented by 7 genes clustered in a region of chromosome 6, and the cadinene synthase family (sesquiterpenes), comprising 10 members, among which 7 are localized in a same region of chromosome 7. Cadinene synthase is one of the key enzymes involved in the synthesis of gossypol, a toxic terpenoid produced in the seeds of cotton, a species belonging to Malvaceae, the same family as *T. cacao*. In cocoa, cadinene synthase has been found to be expressed in pod tissues[37] in response to attacks by mirids (*Sahlbergella singularis)*, a major insect pest of cocoa trees in Africa. Therefore, the cadinene synthase orthologous genes are candidates for elements of the cocoa insect resistance response.

Colocalization of quality related genes and QTL

Previous studies have reported QTLs associated with quality traits such as lipid and flavonoid content[38,39] (**Supplementary Note**). For most of these QTLs, genes encoding key enzymes of these biosynthetic pathways were found to be colocalized with most of these QTLs (**Supplementary Fig. 16**). For example, a major QTL for fat content is associated on chromosome 9 with a gene orthologous to *KCS*, encoding beta ketoacyl-CoA synthase, and is located very close to an ortholog of one member of the *FATB* gene group, which is specifically expanded in *T. cacao.* A strong QTL for cocoa butter hardness[39] was found localized in linkage group 7 near a gene orthologous to *FAD4*, which is involved in creating a bond between C2 and C3 of the lipid chain, resulting in lipids with a higher melting point, which, in terms of cocoa butter, represents greater hardness.

Similarly, we found each of the QTLs for astringent taste to be closely associated with genes potentially involved in the proanthocyanidins biosynthetic pathway (**Supplementary Fig. 15**). For example, two genes orthologous to that encoding flavonoid 3-hydroxylase (*F3H*) and one orthologous to that encoding dihydro-flavonol-4-reductase (*DFR*), which are specifically expanded in *T. cacao*, co-localize on linkage group 1 with a major astringency QTL.

T. cacao genome paleo-history

Angiosperms have been shown to evolve through rounds of paleo-polyploidy[10,40]. Two types of events have been reported in the literature for eudicots: an ancestral event (referenced as γ) and lineage-specific events (referenced as α and β). In order to investigate the paleo-history of the *T. cacao* genome, we characterized shared paleo-polyploidies based on the integration of orthologous relationships identified between *T. cacao* and five eudicot sequenced genomes (*Arabidopsis*[41], grape[10], poplar[36], soybean[42] and papaya[43]), as well as paralogous relationships identified between the ten *T. cacao* chromosomes.

Recently, we published a method for the identification of orthologous regions between plant genomes as well as for the detection of duplicated blocks within genomes based on integrative sequence alignment criteria combined with statistical validations[44]. This approach has been recently applied to available monocot and eudicot genomes and has allowed us to propose a common ancestor with five (core gene set of 9,138) and seven proto-chromosomes (core gene set of 9,731) for monocots and eudicots, respectively[45]. We have integrated the *T. cacao* genome sequence information (23,529 gene models anchored) into our previous paleo-genomics analysis in order to investigate the *T. cacao* evolutionary paleo-history (Online Methods).

Using the alignment parameters and statistical tests reported previously[44], 7,866 orthologous relationships covering 80% of the *T. cacao* genome were identified between the *T. cacao* and the *Arabidopsis*, poplar, grape, soybean and papaya genomes. The chromosome-to-chromosome orthologous relationships that were established between the *T. cacao* and the five sequenced eudicot genomes are illustrated in **Figure 4a** and are available in **Supplementary Table 18**.

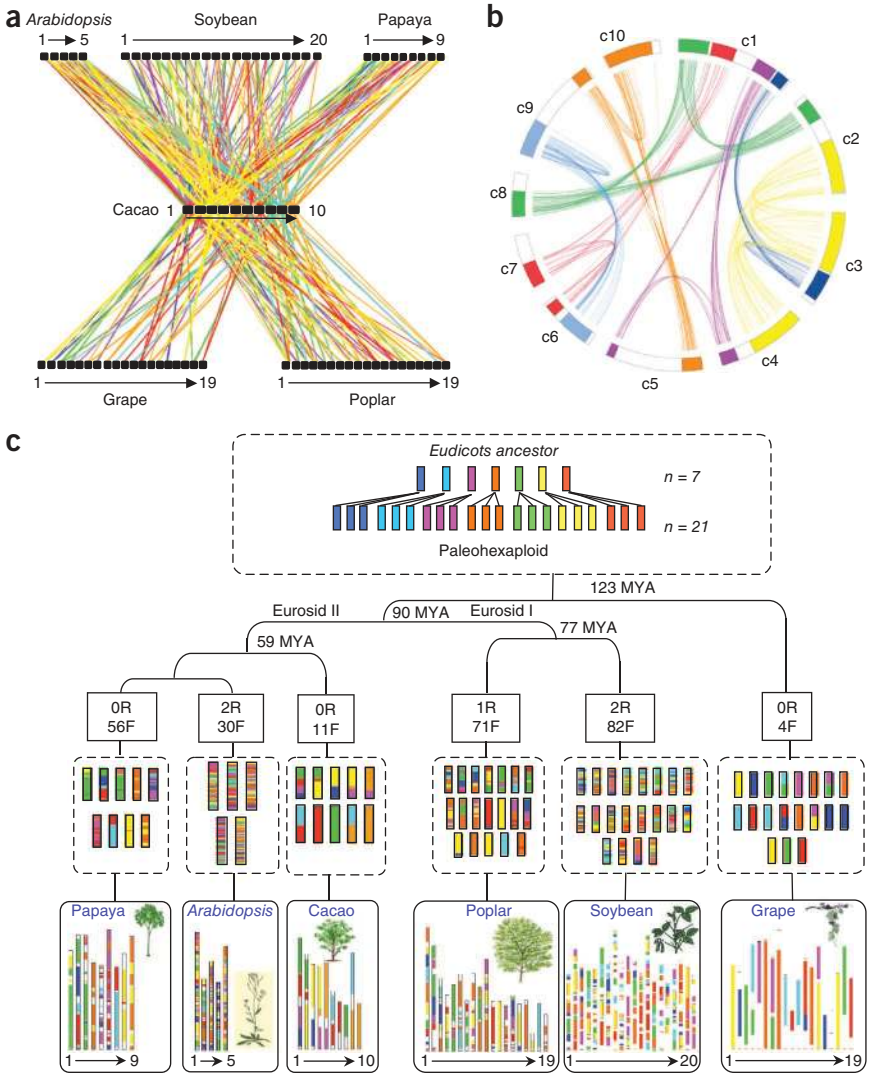

Figure 4 *T. cacao* genome paleohistory. (**a**) *T. cacao* genome synteny. A schematic representation of the orthologs identified between *cacao* chromosomes (c1 to c10) at the center and the grape (g1 to g19), *Arabidopsis* (a1 to a5), poplar (p1 to p19), soybean (s1 to s20) and papaya (p1 to p9) chromosomes. Each line represents an orthologous gene. The seven different colors used to represent the blocks reflect the origin from the seven ancestral eudicot linkage groups. (**b**) *T. cacao* genome duplication. The seven major triplicated chromosomes groups in *T. cacao* (c1 to c10) are illustrated (colored blocks) and related with paralogous gene pairs identified between the *T. cacao* chromosomes (colored lines). The seven different colors reflect the seven ancestral eudicot linkage groups. (**c**) *T. cacao* genome evolutionary model updated from Abrouk *et al.*[46]. The eudicot chromosomes are represented with a seven-color code to illustrate the evolution of segments from a common ancestor with seven protochromosomes (top). The different lineage-specific shuffling events that have shaped the structure of the six genomes during their evolution from the common paleo-hexaploid ancestor are indicated as R (for rounds of whole-genome duplication (WGD)) and F (for fusions of chromosomes). The current structure of the eudicot genomes is represented at the bottom of the figure.

Moreover, we aligned the 23,529 gene models from the *T. cacao* genome onto themselves. Seven blocks of duplicated genes (344 gene pairs) were identified and characterized in *T. cacao*, covering 64% of the genome and involving the following chromosome to chromosome (c) relationships: c2-c3-c4 (yellow), c1-c3 (blue), c1-c2-c8 (green), c6-c9 (light blue), c1-c4-c5 (purple), c5-c9-c10 (orange) and c1-c6-c7 (red) (**Fig. 4b** and **Supplementary Table 19**). We found this ancestral paleo-polyploidy event shared at orthologous positions between eudicot genomes on the following chromosome pair combinations in *T. cacao* compared to the seven ancestral triplicated chromosome groups reported in grape (g)[10]: g1-g14-g17/c2-c3-c4 (yellow), g2-g12-g15-g16/c1-c3 (blue), g3-g4-g7-g18/c1-c2-c8 (green), g4-g9-g11/c6-c9 (light blue), g5-g7-g14/c1-c4-c5 (purple), g6-g8-g13/c5-c9-c10 (orange), g10-g12-g19/c1-c6-c7 (red) (**Fig. 4c**). This result confirms the paleo-hexaploid origin of the eudicot species recently reported for the grape[10] and soybean[42] genomes. Moreover, we confirmed the known γ paleo-hexaploidization event in the *T. cacao* genome through classical Ks-based (synonymous substitution rate) data analysis between paralogous genes (**Supplementary Fig. 18**).

Based on the ancestral (γ) and lineage-specific duplications (α, β) reported for eudicots, it became possible to propose an evolutionary scenario that shaped the ten *T. cacao* chromosomes from the seven chromosomes of the eudicot ancestor and, more precisely, to the 21 chromosomes of the paleo-hexaploid ancestor (**Fig. 4c**). We suggest, from the 21 chromosomes intermediate ancestor, at least 11 major chromosome fusions (referenced as 'F' in **Fig. 4c**) to reach the actual ten-chromosome structure (compared to 30, 4, 71, 82 and 56 reported, respectively, for *Arabidopsis*, grape, poplar, soybean and papaya genomes)[46].

Finally, in order to gain insight into our understanding of the molecular mechanisms driving the chromosome number reduction from the 21 chromosome ancestor intermediate to the actual 10 chromosome structure of the *T. cacao* genome, we produced heat maps scoring particular features such as CoDing Sequences (CDS), transposable element repeats for class I and II, tandem-Gaucho elements and telomeric repeats (**Fig. 2**). We observed a classical distribution pattern of CDS and transposon elements that were more abundant at the telomeric and centomeric regions of the chromosomes, respectively. Moreover, we identified a clear correlation between the position of chromosome fusion points (dotted rectangles) and the occurrence of telomeric repeats (telomeric remnants collocating with eight out of the eleven sites) consistent with a telomere-telomere recombination process leading to the chromosome fusion events reported previously in eudicots[47].

DISCUSSION

We sequenced the genome of *T. cacao*, resulting in the assembly of 76% of its genome and identification of 28,798 protein-coding genes, among which 23,529 (82%) were anchored onto the ten cocoa chromosomes. A large proportion of the euchromatin of the *T. cacao* genome is likely covered by this assembly, allowing for the recovery of 97.8% of the *T. cacao* unigene resource. We found that 682 gene

families are specific to *T. cacao*, as compared to *A. thaliana*, grape, soybean and poplar proteomes. Only 24% of the *T. cacao* genome consists of transposable elements, a lower percentage than in other genomes of similar size. The analysis of specific gene families that are potentially linked to cocoa qualities and disease resistance showed that particular expansion or reduction of some gene families appears to have occurred. The mapping of these gene families along the cocoa chromosomes and comparison with the genome regions involved in trait variation (QTLs) constitutes an invaluable source of candidate genes for further functional studies that aim to discover the specific genes directly involved in trait variation. This draft genome sequence will facilitate a better understanding of trait variation and will accelerate the genetic improvement of *T. cacao* through efficient marker-assisted selection and exploitation of genetic resources. Using an updated version of the Newbler software (released by Roche on 8/17/2010), we performed a second-generation assembly of the cacao genome data (ICGS Assembly 1.2). The new assembly covers 84.3% of the *T. cacao* B97-61/B2 genome, with a N50 scaffold size of 5.624 Mb and the largest scaffold of 18.20 Mb. This enhanced assembly enabled us to improve its' anchorage onto the genetic map, which now includes approximately 87% of the assembled sequences on ten pseudochromosomes. Additional details of this and further assembly improvements will be available on the ICGS website (see URLs).

This study has highlighted the close evolutionary relationship of the *T. cacao* genome to the eudicot putative ancestor, showing a limited number of recombinations between ancestral chromosomes, as has also been observed in grape[10]. *T. cacao*, which has only ten pairs of chromosomes, is easily propagated by both sexual and vegetative methods and can be transformed[48], and therefore, it represents a new and simple model to study the evolutionary processes, gene function, genetics and biochemistry of tree fruit crops.

URLs. Cocoa statistics, http://www.icco.org/economics/market.aspx, http://www.dropdata.org/cocoa/cocoa_prob.htm; ICGS website, http://cocoagendb.cirad.fr/gbrowse/cgi-bin/gbrowse/theobroma/; web tools provided by the Bioinformatics and Systems Biology of Gent, http://bioinformatics.psb.ugent.be/webtools/Venn/; MUST, http://csbl1.bmb.uga.edu/ffzhou/MUST/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The *Theobroma cacao* whole-genome sequences are deposited in the EMBL, GenBank and DDBJ databases under accession numbers CACC01000001–CACC01025912. A genome browser and further information on the project are available from http://cocoagendb.cirad.fr/gbrowse and http://cocoagendb.cirad.fr.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank CIRAD, the Agropolis foundation, the Région Languedoc Roussillon, Agence Nationale de la Recherche (ANR), Valrhona and the Venezuelan Ministry of Science, Technology and Industry for their financial contribution to this project. We thank the Toulouse Midi-Pyrénées bioinformatic platform for providing us with computational resources. Activities at Pennsylvania State University were supported by a gift from the Hershey Corp. and through support from the Schatz Center for Tree Molecular Genetics in the School of Forest Resources. Acquisition of an Illumina sequencer by the Cold Spring Harbor Laboratory was supported by the National Science Foundation grant DBI0923128 to W.R.M. We would like to thank F.C. Baurens, Y. Jiao, O. Garsmeur and C. dePamphilis for helpful advice and assistance with bioinformatics. We would like to express our special appreciation to V. Mooleedhar, who collected the cacao accession in Belize and made it available to us for our analysis.

AUTHOR CONTRIBUTIONS

X.A., J.S., J.-M.A., M.J.G., J.G., D.K., M.J.A., S. Brown, K.G., A. D'Hont, A. Dievart, D.B., D.I., P.C., R.W., W.R.M., E.G., F.Q., O.P., P.W., S. Bocs and C.L. designed the analyses.

X.A., J.S., J.-M.A., M.J.G., J.G., M.R., D.K., M.J.A., S. Brown, A. D'Hont, D.B., W.R.M., O.P., P.W., S. Bocs and C.L. managed the several components of the project.

X.A., M.A., O.F., Y.R., A.B., M. Bocca, D.C., R.R., M.T., J.M.A., K.G., I.K., J.-M.A. and C.L. performed material preparation and multiplication, DNA and RNA extractions, genotyping, genetic mapping and anchoring of the assembly.

D.K., J.S.S.A., W.G. and X.S. performed BAC libraries.

J.-M.A., J.P., S.C.S., J.E.C., M.K., L.G. and W.R.M. performed sequencing and assembly.

X.A., G.D., J.G., M. Allegre, T.L., S.N.M., E.S., T.S., Z.S., C.V., V.G., Y.Z., B.P. and S. Bocs performed automatic and manual gene annotations and database management.

C.C., J.F.B.-N., F.S., A.M.R., M.J.A., Z.M., O.P. and S. Brown performed repeated elements and miRNA analyses.

M.R.-G., M. Bourge, S. Brown and A. D'Hont performed in situ hybridizations and genome-size evaluations.

M.J.G., G.D., T.L., S.N.M., M.R., A. Dievart, Z.S., X.S. and Y.Z. performed gene family analyses.

J.S., M. Abrouk and F.M. performed evolution analyses.

X.A., J.S., J.-M.A., M.J.G., G.D., J.G., C.C., T.L., S.N.M., M.R., M.R.-G., D.K., S.C.S., A. D'Hont, A. Dievart, X.S., M.J.A., S. Brown, P.C., F.Q., O.P., S. Bocs and C.L. wrote and/or revised the paper.

C.L. initiated and coordinated the whole project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Davie, J.H. Chromosome studies in the Malvaceae and certain related families. II. *Genetica* **17**, 487–498 (1935).
2. Henderson, J.S., Joyce, R.A., Hall, G.R., Hurst, W.J. & McGovern, P.E. Chemical and archaeological evidence for the earliest cacao beverages. *Proc. Natl. Acad. Sci. USA* **104**, 18937–18940 (2007).
3. Coe, S.D. & Coe, M.D. *The True History of Chocolate*. (Thames and Hudson Ltd., London, England, 1996).
4. Motamayor, J.C. *et al.* Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* **89**, 380–386 (2002).
5. Motamayor, J.C., Risterucci, A.M., Heath, M. & Lanaud, C. Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. *Heredity* **91**, 322–330 (2003).
6. Mooleedhar, V., Maharaj, W. & O'Brien, H. The collection of Criollo cocoa germplasm in Belize. *Cocoa Grower's Bull.* **49**, 26–40 (1995).
7. Cocoa Resources in consuming Countries–ICCO Market Committee, 10th meeting. *EBRD Offices London, MC* **10**, 16 (2007).
8. Paterson, A.H. *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
9. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
10. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
11. Wolfgruber, T.K. *et al.* Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
12. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGène as a model platform. *Curr. Bioinform.* **3**, 87–97 (2008).
13. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
14. Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell* **136**, 669–687 (2009).
15. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
16. Afzal, A.J., Wood, A.J. & Lightfoot, D.A. Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol. Plant Microbe Interact.* **21**, 507–517 (2008).

17. Diévart, A. & Clark, S.E. LRR-containing receptors regulating plant development and defense. *Development* **131**, 251–261 (2004).
18. 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).
19. DeYoung, B.J. & Innes, R.W. Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* **7**, 1243–1249 (2006).
20. Tarr, D.E.K. & Alexander, H.M. TIR-NBS-LRR genes are rare in monocots: evidence from diverse monocot orders. *BMC Res. Notes* **2**, 197 (2009).
21. Pan, Q., Wendel, J. & Fluhr, R. Divergent evolution of plant NBS-LRR resistance gene homologues in dicot and cereal genomes. *J. Mol. Evol.* **50**, 203–213 (2000).
22. Mukhtar, M.S., Nishimura, M.T. & Dangl, J. NPR1 in plant pefense: it's not over 'til it's turned over. *Cell* **137**, 804–806 (2009).
23. Shi, Z., Maximova, S., Lui, Y., Verica, J. & Guiltinan, M.J. Functional analysis of the *Theobroma cacao NPR1* Gene in *Arabidopsis*. *BMC Plant Biol.* **10**, 248 (2010).
24. Lehmann, P. Structure and evolution of plant disease resistance genes. *J. Appl. Genet.* **43**, 403–414 (2002).
25. Lanaud, C. *et al.* A meta–QTL analysis of disease resistance traits of *Theobroma cacao* L. *Mol. Breed.* **24**, 361–374 (2009).
26. Griffiths, G. & Harwood, J.L. The regulation of triacylglycerol biosynthesis in cocoa (*Theobroma cacao*) L. *Planta* **184**, 279–284 (1991).
27. Beisson, F. *et al. Arabidopsis* genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. *Plant Physiol.* **132**, 681 (2003).
28. Pourcel, L., Routaboul, J., Cheynier, V., Lepiniec, L. & Debeaujon, I. Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* **12**, 29–36 (2007).
29. Spencer, J.P. Flavonoids and brain health: multiple effects underpinned by common mechanisms. *Genes Nutr.* **4**, 243–250 (2009).
30. Rimbach, G., Melchin, M., Moehring, J. & Wagner, A.E. Polyphenols from cocoa and vascular health-a critical review. *Int. J. Mol. Sci.* **10**, 4290–4309 (2009).
31. Liu, Y. Molecular analysis of genes involved in the synthesis of proanthocyanidins in theobroma cacao. *Thesis* 1–146 (2010).
32. Tomas-Barberan, F.A. *et al.* A new process to develop a cocoa powder with higher flavonoid monomer content and enhanced bioavailability in healthy humans. *J. Agric. Food Chem.* **55**, 3926–3935 (2007).
33. 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).
34. Ziegleder, G. Linalol contents as characteristics of some flavour grade cocoas. *Z. Lebensm. Unters. Forsch.* **191**, 306–309 (1990).
35. Chanliau, S. & Cros, E. Influence du traitement post-récolte et de la torréfaction sur le développement de l'arôme cacao. *12th Int. Cocoa Res. Conf., Salvador de Bahia (Brazil)* 959–964 (1996).
36. Tuskan, G.A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596–1604 (2006).
37. 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).
38. Lanaud, C. *et al.* Identification of QTLs related to fat content, seed size and sensorial traits in *Theobroma cacao* L. *Proc. 14th Int. Cocoa Res. Conf.* 13–18 (2003).
39. Araújo, 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).
40. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
41. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. *Nature* **408**, 796–815 (2000).
42. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
43. Ming, R. *et al.* The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**, 991–996 (2008).
44. 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).
45. Salse, J. *et al.* Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. USA* **106**, 14908–14913 (2009).
46. Abrouk, M. *et al.* Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**, 479–487 (2010).
47. Murat, F. *et al.* Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **11**, 1545–1547 (2010).
48. Maximova, S.N. *et al.* Over-expression of a cacao class I chitinase gene in *Theobroma cacao* L. enhances resistance against the pathogen, *Colletotrichum gloeosporioides*. *Planta* **224**, 740–749 (2006).

ONLINE METHODS

High molecular weight DNA preparation. High molecular weight DNA was prepared from nuclei of B97-61/B2 cocoa leaves according to previously described protocols[49], except that the steps of filtration were replaced by five successive filtrations with nylon filters (SEFAR NITEX) having a decreasing mesh diameter: 250 µM, 100 µM, 50 µM and two times 11 µM. (**Supplementary Note**).

Construction of BAC libraries. Two BAC libraries were constructed from cocoa leaves. DNA was isolated from nuclei collected in agarose plugs and DNA digestions were performed with HindIII or EcoRI, followed by ligation to the pAGIBAC1 vector (a modified pIndigoBAC536Blue with an additional *Swa*I site[49]). Ligation products were transformed into DH10B T1 phage-resistant *Escherichia coli* cells (Invitrogen) and plated on Lysogeny broth agar that contained chloramphenicol (12.5 µg ml^{-1}), X-gal (20 mg ml^{-1}) and Isopropyl β-D-1-thiogalactopyranoside (0.1 M). For characteristics, quality assessment and estimated genome coverage see the **Supplementary Note** and **Supplementary Table 1**.

Genome sequencing. The genome was sequenced using a genome-wide shotgun strategy. All data were generated using next-generation sequencers: Roche/454 GSFLX (×16.5 coverage) and Illumina GAIIx (×44 coverage), except for data from BAC ends (×0.2 coverage), which were produced by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers (**Supplementary Note** and **Supplementary Table 2**).

Genome assembly and automatic error corrections with Solexa/Illumina reads. 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 by Newbler. The 454 assembly was improved by automatic error corrections with Solexa/Illumina reads, which have a different bias in error type, as described previously[50]. Short-read sequences were aligned on the cocoa genome assembly using the SOAP software (with a seed size of 12 bp and a maximum gap size allowed on a read of 3 bp). Only uniquely mapped reads were retained (**Supplementary Note**).

Estimation of nuclear DNA content by flow cytometry. The genome sizes of B97-61/B2 and a panel of diverse cocoa clones were estimated by flow cytometry[51]. Leaves of studied samples and internal standards were chopped with a razor blade and then stained using propidium iodide. DNA content of 5,000–10,000 stained nuclei was determined using a CyFlow SL3 flow cytometer with a 532-nm green solid state laser (100 mWatt). The monoploid C-value (1C) was calculated and expressed in Mb using the conversion factor 1 pg DNA = 978 Mb (**Supplementary Note**).

Anchoring of the assembly in the genetic map. A consensus map was established from two progenies: an F1 progeny of 256 individuals (UPA 402 × UF676) and an F2 progeny of 136 individuals recently produced (Scavina 6 × ICS1). The F1 progeny was previously used to established the cocoa reference map, which includes 600 markers[52,53]. New SSR and SNP markers were mapped in these two progenies, and a consensus map including 1,259 markers was established[54]. BLAT software was used to align the markers of the genetic map with the scaffolds (**Supplementary Note**).

Prediction of transposable elements. The annotation of transposable elements in the cocoa genome was achieved in two stages. First, a combination of *de novo* analyses (for example, LTR_finder, LTRharvest, MUST (see URLs)) and extrinsic comparisons (BLAST) was conducted. Then, a *de novo* approach was used to construct highly repeated elements from the 3,220,522 unassembled reads. A total of 67,575 transposable elements were annotated (**Supplementary Note**).

***In situ* hybridization of transposable element probes.** FISH was performed on mitotic metaphase spreads prepared from meristem root tip cells. The probes were labeled with Alexa-488 dUTP and Alexa-594 dUTP by random priming (Fisher Bioblock Scientific) and the *in situ* hybridizations were performed according to D'Hont *et al.*[55] (**Supplementary Note**).

Prediction of protein-coding genes. Gene structures were predicted using EUGene[12]. Translation start sites and RNA splice sites were predicted by SpliceMachine[56]. Available *T. cacao* ESTs were aligned onto the scaffolds using GenomeThreader[57]. Similarities to proteins from Swiss-Prot, TAIR, Malvacae GenBank extraction, *Glycine max* high confidence gene models[58] and translated *T. cacao* EST contigs were searched using BLASTX. Similarities to *A. thaliana*, *Gossypium*, *V. vinifera*, *Citrus* and *T. cacao* ESTs were searched using TBLASTX. A total of 50,582 genes were predicted, giving a final count of 28,798 *T. cacao* genes after filtering (**Supplementary Note**).

Gene family analysis. Protein domains were searched using InterProscan against the InterPro database. Cocoa, *Arabidopsis*, grape, poplar and soybean Best Blast Mutual Hit (BBMH) were computed, and protein clustering was done using OrthoMCL[59].

Synteny and duplication analysis. Arabidopsis, grape, poplar, soybean and papaya proteomes were aligned using an approach based on BLASTP in order to identify accurate paralogous and orthologous relationships[44,45]. K_S divergence (million year ago (MYA) scale) for paralogous and orthologous gene pairs as well as speciation events dating were calculated with PAML[60].

49. 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).
50. Aury, J.M. *et al.* High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008).
51. 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).
52. 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).
53. Fouet, O. *et al.* Structural characterization and mapping of functional EST-SSR markers in *Theobroma cacao*, in the press.
54. Allegre, M. *et al.* A high-density consensus genetic map for *Theobroma cacao* L., in the press.
55. D'hont, A. *et al.* Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Gen. Genet.* **250**, 405–413 (1996).
56. Degroeve, S., Saeys, Y., De Baets, B., Rouzé, P. & Van de Peer, Y. SpliceMachine: predicting splice sites from high-dimensional local context representations. *Bioinformatics* **21**, 1332–1338 (2005).
57. Gremme, G., Brendel, V., Sparks, M.E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
58. Schmutz, J. *et al.* Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
59. Li, L., Stoeckert, C.J. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
60. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).