Pangenomic exploration of *Theobroma cacao*: New Insights into Gene Content Diversity and Selection During Domestication

Xavier Argout[1,2], Gaetan Droc[1,2,3], Olivier Fouet[1,2], Mathieu Rouard[2,3,4], Karine Labadie[5], Bénédicte Rhoné[1,2,6], Gaston Rey Loor[7], Claire Lanaud[1,2]

[1.] CIRAD, UMR AGAP Institut, F-34398 Montpellier, France

[2.] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France

[3.] French Institute of Bioinformatics (IFB) - South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, F-34398 Montpellier, France

[4.] Bioversity International, Parc Scientifique Agropolis II, Montpellier, France

[5.] Génomique Métabolique, Genoscope, Institut de Biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, 91000, Evry, France.

[6.] CATIE – Centro Agronómico Tropical de Investigación y Enseñanza, Turrialba, Cartago, 30501, Costa Rica

[7.] INIAP, EET-Pichilingue. CP 24 Km 5 vía Quevedo El Empalme, Quevedo, Los Ríos, Ecuador

Abstract

The Cacao tree (*Theobroma cacao L.)* is a major cash crop and an important source of income for millions of farmers across Africa, Asia and Latin America. However, cacao farmers and producing countries are facing multiple challenges including pest and disease management, sustainable production under threat of climate changes and diversification of cocoa flavor profiles. Addressing these challenging requires a deeper understanding of the existing genetic diversity of the species. Yet, very little is known about the intraspecific gene content variation.

In this study, we used the genome of 216 accessions of *T. cacao* (including 185 newly re-sequenced) covering a broad genetic diversity of the species to construct the first pan-genome of the cacao tree. As a result, we predicted a total of 30,489 pan-genes, enriching the reference Criollo genome with 1,407 new genes.

Our analysis revealed that only a small fraction of these genes (9.2%) is dispensable, being absent in some individuals. However, these genes could represent a valuable resource for breeding efforts aimed at improving disease resistance in the species.

We used this new pangenome resource to gain insights into how diversification and domestication processes have influenced genomic variability within the species. Notably, we observed a significant loss of genes within the domesticated Criollo genetic group. Furthermore, we found evidences suggesting that domestication processes have had an impact on the vigor and disease tolerance of Criollo accessions. In summary, our research has contributed to a better understanding of the cacao tree's genome diversity. These findings offer new avenues for biological discovery and breeding, ultimately addressing the challenges faced by cacao farmers and producing countries.

Introduction

Theobroma cacao L., commonly known as the cocoa tree, is a perennial diploid species native to the upper Amazon basin and is a pivotal cash crop for millions of smallholder farmers in countries across the humid tropical regions of the world[1]. The cocoa sector is currently facing many challenges which include low productivity[2], significant production losses due to pests and diseases[3], contamination by heavy metals such as Cadmium[4], impacts of climate change[5] and deforestation concerns[6]. Furthermore, in recent years, consumer preferences have shifted towards high-quality chocolate products using sustainable practices. This trend has prompted farmers and manufacturers to search for diverse and unique cocoa flavor profiles, and to produce fine and flavored cacao beans. Comprehensive exploration of the *T. cacao* genetic diversity at genomic level provides resources that can help in addressing these challenges. Effectively utilizing this information should support breeding programs to develop varieties adapted to these different environmental contexts[7].

Over the past decade, significant progresses have been made on cocoa genetic improvement through the generation and utilization of genomic data. To date, annotated chromosome-scale genomes of two *T. cacao* accessions are available: the B97-61/B2 Criollo reference genome produced in 2011[8], further improved in 2017[9] and the genome of Matina1-6, an Amelonado cultivar sequenced in 2013[10]. These two genomes have significantly advanced scientific discoveries in *Theobroma cacao*, allowing for instance the development of molecular markers and the candidate genes identification through GWAS or gene expression analysis using NGS[11]. They also enabled substantial progress in modern cocoa breeding approaches, using marker assisted or genomic selection. Nevertheless, recent studies in plant genomics indicate that one or few reference genomes cannot capture the full range of genomic diversity of plant species, including the presence/absence of protein coding genes sequences, which may lead to a bias associated with the use of a single reference[12]. The development of the pan-genome, defined here as a representation of all DNA sequences within a species, can address this issue.

Plant pan-genomes have been increasing dramatically in recent years[13]. The concept of pan-genome includes core genes, found in all individuals, as well as variable genes (or dispensable genes), which are absent in some individuals or genetic groups. Gene presence and absence variations (PAVs) are an important source of genetic divergence at a population level and their understanding can support applications for crop improvement because they can contribute to trait variation. Functional analysis of variable genes in multiple crops have revealed their enrichment in genes related to disease resistance, abiotic stresses, environmental adaptation, and quality when compared to gene core fraction[14–17].

Very little is known about the complete genomic content across *T. cacao* species. In comparison to other plant species, the cocoa tree genome size is relatively small (430Mb) but can vary by about 10% depending on the genotype[8]. The genetic diversity of the species has been widely studied, particularly using microsatellites and SNPs as molecular markers[18–21]. The current genetic diversity of the species includes several genetic groups that could have diversified through isolation in refugia during the last glaciation followed by human selection and domestication[18,22]. Among the different genetic groups observed, the Criollo, highly regarded by the global cocoa industry, stands out due to its nearly unique and highly homozygous genotype, resulting from domestication processes along with genetic drift after transport from South to Central America[23].

Recently, analyses of the genetic diversity among native cocoa trees collected during surveys in the Ecuadorian Amazon have broadened our understanding of the diversity of the species and improved the knowledge of the global genetic structure of *Theobroma cacao*. In this work, we utilized a subset of these genetic resources, along with representative samples from all other genetic groups, to construct the first *Theobroma cacao* pan-genome. This pan-genome was based on deep whole genome resequencing data generated from 216 accessions, encompassing the known diversity of the species up to the present day. We comprehensively investigated gene PAVs inside the diversity of the species and captured new protein coding genes not included in the reference genomes. We also examined how the processes of diversification and domestication of the *Theobroma cacao* genetic groups have shaped their gene content and searched for genes under selection during the domestication process of the Criollo group. Through this work, we aim to provide insights and tools that can guide future efforts in cocoa breeding and conservation strategies.

Results

De novo assembly of *Theobroma cacao* accessions and pan-genome construction

We selected a total of 216 *T. cacao* accessions (Figure 1a, Supplementary Table 2) covering a broad genetic diversity of the species[21,24,25]. Among these accessions, we newly sequenced the genome of 185 cocoa accessions by Illumina technology (2x150bp reads) that we combined with Illumina genome sequences of 31 accessions retrieved from public databases (Bioproject PRJNA558793). After low quality and sequencing adapters removal, read coverage per accession ranged from 40x to 75x with a mean of 58x.

We performed *de novo* assembly for each of the 216 accessions and generated a total of 76.24 Gbp of contig sequences longer than 500 bp with a N50 value of 13,357 bp. The size of the assembly ranged from 309.0 Mbp to 505.5 Mbp per accession with a mean of 354.6 Mbp (Supplementary Table 2). All accessions showed assembly completeness assessed by BUSCO higher than 92.4%, with a mean for all *de novo* assemblies of 95.2%, a value similar to that reported for the B97-61/B2 reference genome[9] (97.3%).

The process of identification of sequences not included in the B97-61/B2 reference genome[9] resulted in a total of 1.59 Gbp of cleaned non-reference sequences. The final non reference genome dataset was consolidated as 48.2 Mbp of sequences longer than 1,000 bp after redundancy removal.

A total of 1,407 candidate protein-coding gene models were predicted in the newly assembled contigs. From them, 99.2% exhibited significant similarities (evalue > 1e-5) with proteins of the NR database[26] and 93.2% were annotated with gene ontology terms. Gene ontology analysis revealed that the largest groups of genes included in the non-reference genome were involved in defense response, hydrolysis of proteins, protein phosphorylation, intracellular signal transduction and production of siRNA (Supplementary Figure 1).

The final *Theobroma cacao* pan-genome, including the B97-61/B2 reference and non-reference genome dataset, had a total size of 372.9 Mbp and contained 30,489 protein-coding gene models. The pan-genome saturation curve (Figure 1b) shows that the total number of pan-genes increases as additional genomes are added, until a plateau is reached at about 50 accessions. It indicates that our sampling led to a closed pan-genome and that the diversity of *Theobroma cacao* was largely captured by the selected set of accessions.

We also observed a continuously decrease of the core-genome size with the increase of population size, until a substantial decrease of the slope of the curve detected from 100 accessions, suggesting that the large number of accessions included in our dataset was large enough for an accurate characterization of the core-genome.

Core and dispensable genome

We then categorized the presence and absence of each protein-coding gene model according to their presence frequencies among the 216 *T. cacao* accessions (Figure 1c). A vast majority of genes, 27,687 (90,8%), were defined as core genes while 2,802 were identified as variable. Variable genes were composed of 806 (2,6%) softcore genes, absent in at least one accession but present in more than 95% of the accessions, 1,924 shell genes (6,3%) present in 5-95% of the accessions, and 72 cloud genes (0,3%) present in less than 5% of the accessions. Interestingly, none of the genes were specific to a single accession, suggesting the high level of representativeness of the *T. cacao* pan-genome provided by the 216 accessions.

The comparison of variable genes detected in this pan-genome analysis with the 2 annotated genomes[9,10] indicated that the pan-genome comprised 583 *Theobroma cacao* protein coding gene models not previously cataloged (Figure 1d). We also observed that a substantial part of the variable genes is absent in either Matina1-6 (34%) or B97-61/B2 (43%) genomes, highlighting the relevance of this new resource for the understanding of the genomic basis of traits in the *Theobroma cacao* species.

Gene PAVs population genetic structure

We first examined the genetic structure of the population with the protein-coding gene presence and absence variations (PAVs) matrix build from the 2,802 variable genes found in the 216 *T. cacao* accessions. From admixture cross entropy criterion, we identified a minimum at K=15 (Figure 2a) indicating that the studied population comprised 15 genetic clusters. Out of these 15 genetically differentiated groups, 11 were identified and consistently named in accordance with Motamayor *et al.,* 2008[18] and Fouet at al., 2022[21] : Amelonado, Caqueta, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay and Purús. The remaining four genetic clusters newly identified were named based on their respective geographical locations:, Apaporis, Pangui, Napo and Tiwinza.

Notably, Caquetá and Pangui genetic groups dissociated early from other accessions at a low K value (respectively K=4 and K=8) (Supplementary Figure 2), suggesting that these genetic groups comprise uncommon gene space among the genetic diversity of *Theobroma cacao*. Conversely, the accessions belonging to the Tiwinza genetic group dissociated from the Curaray genetic group only at K=15.

We then estimated the individual admixture coefficients and identified 95 accessions exhibiting ancestry higher than 85% to their respective identified genetic cluster (Figure 2a). These accessions were retained to investigate the genomic specificity of the genetic groups while the remaining 121 accessions with a coefficient of membership lower than 85% were considered as admixed.

The first axis of the PCA analysis (Figure 2b) clearly differentiated the domesticated Criollo group from the other genetic groups while the second axis separated roughly the genetic groups from East to West according to their geographical coordinates. Furthermore, the third component of the PCA analysis revealed a distinct pattern of separation among genetic groups based on their South to North geographic coordinates, notably pronounced in the accessions collected in Ecuador. The PCA analysis also indicated high similarity between the Curaray and Tiwinza genetic groups as the accessions from these 2 genetic groups overlap in the projection of the 3 main principal axes.

Finally, phylogenetic analysis of the 216 *T. cacao* accessions was consistent with admixture and PCA analyses (Figure 2c). The Neighbor-joining tree supported 15 clades, grouping accessions similarly to the 15 groups predicted by the admixture analysis.

Selection of gene PAVs during *T. cacao* diversification

We further investigated whether the processes of diversification of the *Theobroma cacao* genetic groups have shaped their gene content by exanimating the distribution of gene PAVs into the 15 groups. We found that accessions of the Iquitos and Napo genetic groups had the highest gene content (gene mean number of 29835 and 29806 respectively) (Figure 3A), suggesting that the differentiation processes leading to these two genetic groups had the least impact on their gene content. At the opposite, the domesticated Criollo genetic group encoded significantly (Newman-Keuls test at $p<0.05$) less genes than other genetic groups (29298), suggesting that the domestication process of *T. Cacao* has led to gene loss.

Since some admixture was observed in several accessions of the 15 genetic groups (Figure 2a), we assigned a gene to a genetic group if it occurred in at least 25% of its comprised accessions. We then compared the intersection between the representative gene space of the 15 genetic groups (Figure 3b). We observed that very few genes were specific to a genetic group (Supplementary Figure 3), highlighting the fact that the boundaries between genetic groups are porous and multiple instances of hybridization among the genetic groups may have occurred during their natural process of differentiation along the Amazon Basin. The gene space of the Iquitos genetic group intersected with a very high intensity (>=92%) with almost all genetic groups, revealing the high diversity of the gene space of this genetic group. Interestingly, intersection between the Criollo gene space and gene space of other genetic groups indicates that Caquetá genetic group encodes the highest percentage (94%) of the genes found in the Criollo genetic group, while the Curaray genetic group, a closely related group of the Criollo group[27], only covered 88% of the Criollo gene content.

Functional analysis of variable genes

Gene Ontology (GO) enrichment of protein-coding genes of the variable genome identified significant enriched GO terms (p-value < 0.01) associated to biological processes involved in the plant immunity, gene transcription regulation, plant growth and cellular processes, protein metabolic processes and biosynthesis of aromatic compounds (Table 1) which is consistent with previous findings in other plants (Bayer et al., 2020). Among the most interesting GO terms, 282 variable protein-coding genes were associated with the term "defense response" (GO :006952), suggesting that the variable part of the *T. cacao* genome could play an important role in pathogen resistance processes. Most of them (193) were predicted in the non- reference assembled contigs and were thus not included in the reference genome. We also identified 31 variable genes annotated with "meristem maintenance" (GO:0010073) and 8 genes annotated with "lignin biosynthetic process" (GO:0009809), two key biological processes involved in plant growth and potentially involved in *T. cacao* vigor. Finally, 12 and 11 genes are annotated with "diterpenoid and triterpenoid biosynthetic process" (respectively GO:0016102 and G0:0016104 respectively) indicating that the variable genome may be involved in cocoa flavor compounds.

Pan-genome-wide analysis of resistance gene analogues (RGAs)

Since the dispensable genome of *T. cacao* seemed particularly enriched in genes involved in plant immunity, we identified at a pangenome-wide scale, RGAs, an important gene family involved in disease resistance in plants. A total of 1385 RGAs were identified in the *T. cacao* pangenome, comprising 654 receptor-like kinase proteins (RLK), 350 nucleotide binding site proteins (NBS), 201 receptor-like proteins (RLP) and 180 transmembrane coiled-coil proteins (TM-CC) (Table 2). Among them, 210 RGAs were variable, indicating that the variable genome may play important roles in determining the level of tolerance to certain diseases observed in particular cocoa accessions. The largest class of variable RGAs corresponded to NBS (30%) followed by RLP (26%) while only 6% and 7% of the RLK and TM-CC resistance gene respectively were found to be variable, indicating that for *T. cacao*, NBS and RLP genes show a greater variability than RLK and TM-CC genes. We then analyzed the distribution of RGAs among the 95 accessions representative of the 15 genetic cluster (with ancestry >0.85). We observed that accessions belonging to the Pangui genetic groups contained the highest number of RGAs (1364), indicating that the Pangui genetic group comprises almost the entire repertoire of the T. cacao RGAs (90% of the T. cacao variable RGAs). This result suggests that germplasm originating from the Pangui genetic group may contain valuable resistance genes that would be worth exploiting. Conversely, 35% of the *T. cacao* variable RGAs were absent in the domesticated Criollo genetic group which comprised the lowest number of RGAs (1312), showing that the domestication process of *T. Cacao* has led to RGA gene loss.

Protein-coding genes under selection during the Criollo domestication

The origin of the domesticated Criollo has been hypothesized to be located in the Upper Amazon region, which includes parts of Ecuador and Colombia[21,28]. To investigate the putative protein-coding genes that might have been under selection during the domestication process of the Criollo group, we compared the gene frequencies between Criollo accessions and accessions native from the putative center of origin of the domesticated Criollo.

We identified in our georeferenced data, a subset of 35 native accessions with Caquetá, Curaray and Napo ancestries, collected by Allen et al. in 1988[29] in North Ecuador and South Colombia (Figure 4a), hereafter referred to as “native accessions”. Analysis of gene frequencies between accessions from the Criollo group and native accessions (Figure 4b, Supplementary Table 3, Supplementary Table 4) identified 71 genes which present significant higher frequencies in the domesticated Criollo group and therefore would have been positively selected during Criollo domestication and were considered as favorable genes. We also found 624 genes with significantly higher frequencies in native accessions that we regarded as unfavorable or unselected genes during Criollo domestication. No significant enriched Gene Ontology terms was associated with the favorable Criollo genes while defense response was the most enriched term found in the unfavorable genes (Figure

4c). KEGG functional annotation of the unselected genes (Supplementary Table 4) identified 87 genes involved in plant-pathogen interaction pathway (ko04626). Another noteworthy finding was the identification of genes related to the Phenylpropanoid biosynthesis and Flavonoid biosynthesis pathways (ko00940 and ko00941) within the unselected genes. Notably, this included four genes encoding shikimate O-hydroxycinnamoyltransferase (HCT), a central enzyme involved in both pathways. These pathways provide a rich source of metabolites in plants, necessary for the biosynthesis of lignin and are the starting point for the production of many other important compounds, such as anthocyanins[30,31].

Discussion

In this study, we analyzed the genetic diversity of the *Theobroma cacao* species through the prism of its pangenome, define here as all the genes contained in the species, whether present in all individuals (core genome) or absent in one or more individuals (variable genome). For this purpose, we analyzed a population of 216 cacao trees covering a wide genetic diversity and including, for the majority of the studied accessions, subspontaneous genetic material, collected during surveys in the main geographical areas of origin of *Theobroma cacao*[21,29,32–34]. The construction of the pangenome has increased the genome of the *Theobroma cacao* species, with 48.2 Mb of new genomic sequences and 1,407 new candidate protein-coding genes absent from the B97-61/B2 reference genome[9]. Analysis of the cocoa pangenome composition showed, unusually, a very high content of core genes (90.8%), in contrast to what has been observed in other plant species such as melon[35] (73.5%), white lupin[36] (78.5%), tomato[37] (74.2%) and cotton[38] (74.1%). We also highlighted that very few genes were specific to the genetic groups identified and that none of the genes were specific to a single accession. It could therefore be hypothesized that the evolutionary processes behind the diversification of the cocoa tree through genetic drift or domestication have had little impact on the gene content between accessions, and that the species has not yet suffered any major gene loss.

Recent research[39] has also suggested that genetic mixing among various *T. cacao* genetic groups might have occurred several millennia ago due to human influence. This gene flow between ancestral genetic groups would thus have limited the effects of their original differentiation and led to today's *T. cacao* populations. Clement et al. (2015)[40] proposed that Amazonia played a pivotal role as a major world center for plant domestication. Human societies in the Amazonian regions have been significantly altering species composition in various ecosystems since the Pleistocene to the Early Holocene, simultaneously amassing valuable crop genetic resources as their societies expanded. Intensive trading interactions, including the exchange of plants, facilitated by the extensive river networks across Amazonia, occurred between distant human populations thus favorizing gene flow between contrasted genetic groups. For instance, an important center of plant genetic resources was identified in the Iquitos region of Peru[40], where multiple *T. cacao* genetic groups have been documented[22]. The *T. cacao* Iquitos genetic group, characterized by its exceptionally diverse gene content that extensively overlaps with nearly all other genetic groups, may have originated through the hybridization of multiple ancestral genetic groups. This hypothesis is reinforced by the high rate of heterozygosity found in Iquitos accessions (49.3%), as reported by Fouet et al. (2022)[21].

Another noteworthy finding related to the Guiana genetic group, which, despite its remote location in the far eastern region of the Amazon basin (French Guiana), shares a close genetic relationship with the Peruvian Nanay, Marañon, Amelonado and Iquitos accessions, the latest comprising 95% of the gene space of the Guiana group. The Guiana genetic group is also one of the group with the fewest gene content and was found to be highly homozygous (Observed Heterozygosity of 7.9%)[21]. Archaeological research evidenced the presence of human populations settle in French Guiana dating from 4500 to 2000 BC[41]. The distribution of linguistic groups in contact in Amazonia[42] shows that the Tupi language family extended from Iquitos until to the mouth of the Amazon, and on either side including French Guiana, thus supporting possible contact between these two regions. All these results suggest that accessions from the Iquitos region could have been transported by humans to French Guiana through Amazonian rivers, initiating a domestication process that may have involved gene losses.

Evolutionary origin of the Criollo genomes and Impact on gene content

Examination of the distribution of gene PAVs among the 15 genetic groups revealed that accessions belonging to the early cultivated Criollo genetic group exhibited the most restricted gene catalogs (29,298). Gene losses have been particularly observed in plants and animals in populations subjected to domestication or genetic improvement[37,38,43,44]. The gene losses observed in the Criollo genetic group thus provide further evidence of domestication processes that led to its genetic differentiation, such as those already reported in previous studies with SSR and SNP markers[27,28]. While there is a scientific consensus suggesting that Criollo was domesticated from cacao populations originating in the Upper Amazon[21,27,28], the genetic origin of the Criollo group remains controversial. Some authors[20,27] argued that Criollo could have been domesticated from ancient Curaray germplasm, a genetic group described from North Ecuador to South Colombia. A recent study based on molecular analysis of cocoa trees native to the upper Amazon suggests that Criollo is most closely related to the Caqueta genetic material, collected in southern Colombia[21,29]. Using PAVs, the results presented in our study confirmed the closer genetic similarity of Criollo accessions to accessions from the Caqueta group. We found that the Caqueta group contains 94% of the Criollo gene space, in contrast to the Curaray germplasm, which encodes only 88% of the genes found in Criollo. This greatest similarity is also reinforced in the Neighbor-Joining analysis where Criollo accessions are closely related to Caqueta. All these results suggest that the domestication of Criollo could have been initiated from individuals close to the Caqueta genetic group.

We further investigated putative protein-coding genes under selection during the domestication process of the Criollo group by identifying significant different frequencies between accessions from the Criollo group and subspontaneous accessions collected in the putative area of origin of the Criollo group. We identified 71 positively selected genes and 624 negatively selected genes during Criollo domestication. The domestication process led to a significant loss of genes involved in defense response, which could partially explain the high disease susceptibility of Criollo observed in the fields[45,46]. Nowadays, Criollo cocoa trees are grown in very restricted areas and has been gradually replaced by Criollo hybrids (Trinitario) which provide more disease resistance and productivity[23,28,47].

The analysis of the unselected genes during Criollo domestication also revealed the negative selection of genes involved in the metabolic pathways of phenolic compounds, such as Phenylpropanoid and Flavonoid biosynthesis pathways. The flavonoids constitute the largest and most diverse group of phenolic compounds in cocoa beans, among which anthocyanins, responsible for the purple color of cacao beans, are generally present in large quantities[48]. One of the main desirable criteria of the Criollo varieties is their beans specificities, which are white or very slightly colored and recognize for their high-quality flavor properties with low astringency and low bitterness due to low polyphenolic compounds content such as anthocyanins[49]. These particular features of the Criollo beans let us hypothesized that early selection of Criollo varieties during domestication has been done on their bean's qualities. Furthermore, several authors have observed that Criollo exhibits very low vigor and is particularly challenging to propagate vegetatively[23,47,50,51]. As phenolic compounds are also involved in lignin production processes and in regulating cell growth, the selection of Criollo on the criterion of white bean color and low bitterness might also have been associated with a counter-selection of genes involved in plant vigor.

Potential impact of PAVs on disease resistance breeding

Pest and disease resistance is one of the major targets of most cocoa breeding programs[52]. To investigate the resistance gene resources available in *T. cacao*, we used the pan-genome to identify resistance gene analogues (RGAs) and analyzed their distributions and PAVs among the different genetic groups. We found that a relatively substantial fraction of resistance genes resides in dispensable loci (15%), underscoring the importance of conserving and utilizing *T. cacao* genetic diversity for disease control.

The most common sources of disease resistance used presently by breeding programs belong to accessions from Marañon, Nanay, Contamana and Iquitos ancestries[3]. These accessions were collected in Peru by Pound in 1938 and 1943[33,53] to search for *T. cacao* accessions resistant to witches' broom disease. This collected genetic material also had the advantage of being resistant or tolerant to other widespread diseases, such as black pod disease caused by several *Phytophthora* species. Other sources of resistance to *Phytophthora* were more recently observed in subspontaneous *T. cacao* accessions from Guiana ancestry[54–56]. Surprisingly, no significant expansion of RGAs was found in accessions from these genetic groups. Interestingly we found that accessions from Pangui and Tiwinza ancestries, collected since 2010 in the Ecuadorian provinces of Zamora-Chinchipe, Morona- Santiago, and Pastaza[21] contained the largest number of RGAs. This result suggests that new sources of disease resistance could be available in this germplasm, currently being evaluated at INIAP research stations in Ecuador.

Materials and methods

Plant material

In this study, leaf tissues were collected leaf from 185 accessions maximizing the available of *Theobroma cacao* as described by Fouet *et al.* (2022)[21], and from 3 accessions from *Theobroma cacao* related species, *Theobroma bicolor*, *Theobroma grandiflorum* and *Theobroma speciosum,* used as outgroups (Supplementary Table 1).

Genome sequencing and reads cleaning

Genomic DNA was extracted and purified for each accession using the protocol published by Risterucci *et al.* (2000)[57] and modify by Allegre *et al. (* 2012)[58]. Sequencing libraries were prepared as previously described in Alberti *et al.* (2017)[59] using an 'on beads' protocol. Each library was sequenced using 151 base-length read chemistry on a paired-end flow cell on the Illumina NovaSeq6000 (Illumina, USA). After Illumina sequencing, an in-house quality control process was applied to the reads that passed Illumina quality filters. The first step discarded low quality nucleotides (Q < 20) from both ends of the reads. Next, Illumina sequencing adapters and primer sequences were removed from the reads. Then, reads shorter than 30 nucleotides after trimming were discarded. Trimming and removal steps were achieved using custom-developed software based on the FastX package (https://github.com/institut-de-genomique/fastxtend). Finally, using SOAP aligner[60], read pairs that aligned to the phage phiX genome and the Enterobacteria phage PhiX174 reference sequence (GenBank: NC_001422.1) were identified and discarded.

To complement the genetic diversity of the dataset, Illumina genome sequencing data (150bp paired end reads) of 31 accessions[24] were retrieved from NCBI SRA database (BioProject PRJNA558793). Quality control of sequencing reads was performed using the same protocol mentioned above, and sequencing depth of these accessions was normalized to 64X with fastq-sample (https://github.com/fplaza/fastq-sample), corresponding to the third quartile of the sequencing coverage of the newly sequenced 185 *T. cacao* accessions.

Pangenome assembly and annotation

We applied an assemble-then-map approach. For each accession, cleaned Illumina reads were *de novo* assembled using Megahit v1.2.9[61] with a minimum length of contigs of 500bp (option --min-contig-len=500). The completeness and accuracy of the assembly were assessed using BUSCO v5.0.0[62] in genome mode with the lineage dataset eudicots_odb10.

Contigs were then aligned to a reference dataset comprising the "B97-61/B2" reference genome[9] and the complete *Theobroma cacao* chloroplast genome (GenBank : HQ244500.2) using NCBI Megablast v2.9.0[63] with e-value threshold of 1e-5. Sequences ≥ 500bp and below 90% of identity with alignment with the reference dataset were considered novel (non-reference) and retained for further analysis. The non-reference sequences of each genotype were then searched against the GenBank nucleotide collection NT using Megablast (parameter -evalue=1e-5). Subsequently, we used the BASTA taxonomic classification tool[64] with default parameters was employed to identify and eliminate sequences not belonging to the green plants clade or those related to known plant mitochondrial or chloroplast genomes. The cleaned nonreference sequences of all accessions were then combined to reduce sequence redundancy with cd-hit-est of CD-HIT package v4.8.1[65] using a sequence identity threshold of 90%. The nonreferenced, cleaned and nonredundant dataset was then combined to the "B97-61/B2" reference genome to form the *Theobroma cacao* pan-genome.

Automatic gene prediction was performed on the pan-genome with the integrative EuGene Eukaryotic Pipeline (EGNEP) v1.5.1 including EuGene v4.2a[66]. A custom statistical model for splice-site detection was applied to improve the quality of prediction based on the published reference genome[9]. Three protein databases were aligned to contribute to translated regions detection, (i) the proteome of *T. cacao* "B97-61/B2" reference genome, (ii), UniProt/SwissProt Viridiplantae, (iii) the proteome of *T. cacao* cultivar Matina 1-6[10]. Two transcriptome databases were used for transcriptional evidence, one representing the transcriptome of the "B97-61/B2" reference genome[8] and the second composed of Expressed Sequence Tags (ESTs) derived from diverse genotypes, tissues and experimental conditions[67]. Proteins hits from Repbase were filtered prior to functional annotation. Gene functions were assigned through InterProScan domain searches and similarity searches against the UniProt/Swissprot and UniProt/TrEMBL databases (BlastP). Gene Ontology (GO) annotation of the pan-genome protein-coding genes was implemented using OmicsBox/Blast2GO v2.1[68]. Prediction of plant resistance gene analogs was carried out with the RGAugury pipeline[69].

Analysis of Gene Presence/Absence Variation (PAV)

Sequencing reads from each accession were aligned to the *T. cacao* pan-genome using Bowtie2[70] v2.4.2 under the “very-sensitive-local” mode. Gene PAV was determined based on the cumulative coverage of exons of each protein-coding gene model in each accession using SGSGeneLoss[71] v0.1. A gene was considered present if a minimum of 20% of its exons were covered by at least two reads (options minCov = 2, lostCutoff = 0.2), otherwise it was considered absent.

Estimation of ancestral populations and ancestry coefficients was calculated from the PAV matrix of the 216 *Theobroma cacao* accessions using the admixture analysis provided by the LEA R package[72] v3.8.0. The sparsen nonnegative matrix factorization (snmf) function was first executed with K value ranging from 2 to 30 with 100 repetitions. Then, the lowest value of the cross-entropy criterion found between the different K values was used to select the number of ancestral populations that best explains the genotypic data. Accessions having ancestry proportion > 85% were considered representative of an ancestral population, otherwise were considered as admixed.

The population structure was then analyzed with principal component analysis (PCA) implemented in the LEA package. The number of significant components was computed with the Tracy-Widom test (pvalue<0.01).

Neighbor-Joining phylogenic analysis was conducted using Darwin Software[73] v6.0.21. The distance matrix was calculated from the PAV matrix of the 216 *T. cacao* accessions and 3 accessions from outgroups (*T. bicolor*, *T. grandiflorum* and *T. speciosum*). Distances were computed using the Sokal & Michener (simple matching) algorithm as implemented in Darwin software. The first topology was calculated using 95 accessions exhibiting ancestry > 0.85 with the weighted Neighbor-Joining method and 1000 bootstrap. The resulting NJ tree was then used as constraints to build the phylogenetic tree of the remaining 121 *T. cacao* accessions using the “Neighbor-Joining under topological constraints” method implemented in Darwin software. Finally, the 3 outgroup accessions were grafted a posteriori on the resulting NJ tree to avoid disturbance of these distant accessions on the resulting tree.

The identification of gene PAVs under selection during the process of domestication of the Criollo genetic group was done according to the methodology described in Gao et al., (2019)[37]. Basically, the frequency of each gene in the Criollo population and in the native population from the putative center of origin of the domesticated Criollo were first computed. Then, Fisher’s exact test was applied to test the significance of the difference in gene frequency between each group and p-values were adjusted using Benjamini and Hochberg False Discovery Rate (FDR) method. Genes exhibiting a FDR less than 0.001 and a frequency fold change exceeding 2 were considered to be at significantly different frequencies between the two groups and were identified either as favorable (selected in the Criollo population) or unfavorable (unselected) genes. GO enrichment was performed using TopGO R package v.2.48.0 with the Fisher's exact statistical test and the “elim” method. KEGG annotation was computed by eggNOG-mapper[74] v2.1.9.

Availability of data and material

Raw sequence reads were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) (BioProject: PRJNA880462). Pangenome assembly, Gene structure and functional and additional data sets are made available on Download section of the Cacao Genome Hub (https://cocoa-genome-hub.southgreen.fr/download).

Competing interests

The authors declare no competing financial interests.

Authors' contributions

X.A. and G.D performed bioinformatic and statistical analysis. O.F and K.L. performed wet-lab experiment and sequencing. O.F. and G.R.L. contributed to sample collection. M.R. and B.R. assisted with data analysis and manuscript revision. X.A. and C.L. conceived research designed analysis and wrote manuscript.

Acknowledgements

This work has been realized with the support of MESO@LR-Platform at the University of Montpellier, the financial support of CIRAD, and the technical support of the bioinformatics group of the UMR AGAP Institute, member of the French Institute of Bioinformatics (IFB) – South Green Bioinformatics Platform.

We thank CRC (Cocoa Research Center – Trinidad and Tobago), CATIE (Centro Agronómico Tropical de Investigación y Enseñanza- Costa Rica) and INIAP (Instituto Nacional de Investigaciones Agropecuarias – Ecuador) for providing leaf samples of *T. cacao* accessions studied in this work.

We thank Kelly Colonges for her help in DNA extraction.

We thank the I-Site MUSE and Valrhona for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" program with the reference ANR-16-IDEX-0006.

References

1. Fountain, A. & Hütz-Adams, F. 2015 Cocoa Barometer (USA edition). (2015).

2. Toledo-Hernández, M., Wanger, T. C. & Tscharntke, T. Neglected pollinators: Can enhanced pollination services improve cocoa yields? A review. *Agriculture, Ecosystems & Environment* **247**, 137–148 (2017).

3. Bailey, B. A. & Meinhardt, L. W. *Cacao Diseases*. (Springer International Publishing, 2016).

4. Vanderschueren, R. *et al.* Mitigating the level of cadmium in cacao products: Reviewing the transfer of cadmium from soil to chocolate bar. *Science of The Total Environment* **781**, 146779 (2021).

5. Läderach, P., Martinez-Valle, A., Schroth, G. & Castro, N. Predicting the future climatic suitability for cocoa farming of the world's leading producer countries, Ghana and Côte d'Ivoire. *Climatic Change* **119**, 841–854 (2013).

6. Ruf, F., Schroth, G. & Doffangui, K. Climate change, cocoa migrations and deforestation in West Africa: What does the past tell us about the future? *Sustain Sci* **10**, 101–111 (2015).

7. Laliberté, B. A Global Strategy for the Conservation and Use of Cacao Genetic Resources, as the Foundation for a Sustainable Cocoa Economy. *CacaoNet* (2012).

8. Argout, X. *et al.* The genome of Theobroma cacao. *Nature Genetics* **43**, 101–108 (2011).

9. Argout, X. *et al.* The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. *BMC Genomics* **18**, 730 (2017).

10. Motamayor, J. C. *et al.* The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* **14**, r53 (2013).

11. Daymond, A. & Bekele, F. Cacao. in *Cash Crops: Genetic Diversity, Erosion, Conservation and Utilization* (eds. Priyadarshan, P. M. & Jain, S. M.) 23–53 (Springer International Publishing, 2022). doi:10.1007/978-3-030-74926-2_2.

12. Yang, X., Lee, W.-P., Ye, K. & Lee, C. One reference genome is not enough. *Genome Biol* **20**, 104 (2019).

13. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).

14. Tranchant□Dubreuil, C., Rouard, M. & Sabot, F. Plant Pangenome: Impacts on Phenotypes and Evolution. in *Annual Plant Reviews online* (ed. Roberts, J. A.) 453–478 (Wiley, 2019). doi:10.1002/9781119312994.apr0664.

15. Tao, Y., Zhao, X., Mace, E., Henry, R. & Jordan, D. Exploring and Exploiting Pan-genomics for Crop Improvement. *Molecular Plant* **12**, 156–169 (2019).

16. Gabur, I., Chawla, H. S., Snowdon, R. J. & Parkin, I. A. P. Connecting genome structural variation with complex traits in crop plants. *Theor Appl Genet* **132**, 733–750 (2019).

17. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. & Hirsch, C. N. How the pan-genome is

changing crop genomics and improvement. *Genome Biology* **22**, 3 (2021).

18. Motamayor, J. C. *et al.* Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma cacao L). *PLOS ONE* **3**, e3311 (2008).

19. Cosme, S. ., Cuevas, H. E., Zhang, D., Oleksyk, T. K. & Irish, B. M. Genetic diversity of naturalized cacao (Theobroma cacao L.) in Puerto Rico. *Tree Genetics & Genomes* **12**, (2016).

20. Osorio-Guarín, J. A. *et al.* Colombia a Source of Cacao Genetic Diversity As Revealed by the Population Structure Analysis of Germplasm Bank of Theobroma cacao L. *Front. Plant Sci.* **8**, 1994 (2017).

21. Fouet, O. *et al.* Collection of native Theobroma cacao L. accessions from the Ecuadorian Amazon highlights a hotspot of cocoa diversity. *PLANTS, PEOPLE, PLANET* **4**, 605–617 (2022).

22. Thomas, E. *et al.* Present Spatial Diversity Patterns of Theobroma cacao L. in the Neotropics Reflect Genetic Differentiation in Pleistocene Refugia Followed by Human-Influenced Dispersal. *PLoS ONE* **7**, e47676 (2012).

23. Lachenaud, P. & Motamayor, J. C. The Criollo cacao tree (Theobroma cacao L.): a review. *Genet Resour Crop Evol* **64**, 1807–1820 (2017).

24. Hämälä, T. *et al.* Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proceedings of the National Academy of Sciences* **118**, e2102914118 (2021).

25. Loor Solorzano, R. G. *et al.* Insight into the Wild Origin, Migration and Domestication History of the Fine Flavour Nacional Theobroma cacao L. Variety from Ecuador. *PLoS ONE* **7**, e48438 (2012).

26. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).

27. Cornejo, O. E. *et al.* Population genomic analyses of the chocolate tree, Theobroma cacao L., provide insights into its domestication process. *Communications Biology* **1**, 167 (2018).

28. Motamayor, J. C. *et al.* Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* **89**, 380–386 (2002).

29. Allen, J. B. Geographical variation and population biology in wild Theobroma cacao. (University of Edinburgh, 1988).

30. Fraser, C. M. & Chapple, C. The Phenylpropanoid Pathway in Arabidopsis. *The Arabidopsis Book* **9**, e0152 (2011).

31. Dias, M. C., Pinto, D. C. G. A. & Silva, A. M. S. Plant Flavonoids: Chemical Characteristics and Biological Activity. *Molecules* **26**, 5377 (2021).

32. Solorzano, R. G. L. *et al.* Rescue of cacao genetic resources related to the nacional variety: surveys in the Ecuadorian Amazon (2010-2013). *Revista ESPAMCIENCIA ISSN 1390-8103* **6**, 7–15 (2015).

33. Pound, F. J. Cacao and witches' broom disease (Marasmius perniciosus) of South America, with notes on other species of Theobroma. *Yuilles' Printery, Port of Spain, Trinidad and Tobago* **1**, 21–64 (1938).

34. Allen, J. & Lass, R. London cocoa trade Amazon project: final report, phase 1. *Cocoa Grower's Bull* **34**, 1–71 (1983).

35. Sun, Y. *et al.* Pan-Genome Analysis Reveals the Abundant Gene Presence/Absence Variations Among Different Varieties of Melon and Their Influence on Traits. *Frontiers in Plant Science* **13**, (2022).

36. Hufnagel, B. *et al.* Pangenome of white lupin provides insights into the diversity of the species. *Plant Biotechnology Journal* **n/a**, (2021).

37. Gao, L. *et al.* The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics* **51**, 1044–1051 (2019).

38. Li, J. *et al.* Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biology* **22**, 119 (2021).

39. Lanaud, C. *et al.* Past domestication of T. cacao in Latin America revealed by paleogenomics and analyses of methylxanthines. in (2022).

40. Clement, C. R. *et al.* The domestication of Amazonia before European conquest. *Proc. R. Soc. B.* **282**, 20150813 (2015).

41. Van Den Bel, M. M. *Archaeological Investigations between Cayenne Island and the Maroni River: A cultural sequence of western coastal French Guiana from 5000 BP to present.* (Sidestone Press, 2015).

42. Eriksen, Love. Nature and Culture in Prehistoric Amazonia□: Using G.I.S. to reconstruct ancient ethnogenetic processes from archaeology, linguistics, geography, and ethnohistory. vol. 12 (Lund University, 2011).

43. Wang, K. *et al.* The Chicken Pan-Genome Reveals Gene Content Variation and a Promoter Region Deletion in IGF2BP1 Affecting Body Size. *Molecular Biology and Evolution* **38**, 5066–5081 (2021).

44. Liu, Y. *et al.* Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162-176.e13 (2020).

45. Soria, V. Principal varieties of cocoa cultivated in tropical America. *Cocoa Growers' Bulletin* (1970).

46. Marita, J. M., Nienhuis, J., Pires, J. L. & Aitken, W. M. Analysis of Genetic Diversity in Theobroma cacao with Emphasis on Witches' Broom Disease Resistance. *Crop Science* **41**, 1305–1316 (2001).

47. Bekele, F. & Phillips-Mora, W. Cacao (Theobroma cacao L.) Breeding. in *Advances in Plant Breeding Strategies: Industrial and Food Crops* (eds. Al-Khayri, J. M., Jain, S. M. & Johnson, D. V.) 409–487 (Springer International Publishing, 2019). doi:10.1007/978-3-030-23265-8_12.

48. Oracz, J., Zyzelewicz, D. & Nebesny, E. The Content of Polyphenolic Compounds in Cocoa Beans (Theobroma cacao L.), Depending on Variety, Growing Region, and Processing Operations: A Review. *Critical Reviews in Food Science and Nutrition* **55**, 1176–1192 (2015).

49. Elwers, S., Zambrano, A., Rohsius, C. & Lieberei, R. Differences between the content of phenolic compounds in Criollo, Forastero and Trinitario cocoa seed (Theobroma cacao L.). *Eur Food Res Technol* **229**, 937–948 (2009).

50. Cheesman, E. E. Notes on the nomenclature, classification and possible relationships of cocoa populations. *Trop Agricult* **21**, 144–159 (1944).

51. Wood, G. A. R. & Lass, R. *Cocoa*. (John Wiley & Sons, 2008).

52. Monteiro, W. R., Lopes, U. V. & Clement, D. Genetic Improvement in Cocoa. in *Breeding Plantation Tree Crops: Tropical Species* (eds. Jain, S. M. & Priyadarshan, P. M.) 589–626 (Springer, 2009). doi:10.1007/978-0-387-71201-7_16.

53. Pound, F. Cacao and witches' broom disease (Marasmius perniciosus). Report on a recent visit to the Amazon territory of Peru, September, 1942-February, 1943. *Cacao and witches' broom disease (Marasmius perniciosus). Report on a recent visit to the Amazon territory of Peru, September, 1942-February, 1943.* (1943).

54. Lachenaud, Ph., Rossi, V., Thevenin, J.-M. & Doaré, F. The "Guiana" genetic group: A new source of resistance to cacao (Theobroma cacao L.) black pod rot caused by Phytophthora capsici. *Crop Protection* **67**, 91–95 (2015).

55. Paulin, D., Ducamp, M. & Lachenaud, P. New sources of resistance to Phytophthora megakarya identified in wild cocoa tree populations of French Guiana. *Crop Protection* **27**, 1143–1147 (2008).

56. Thevenin, J.-M. *et al.* Numerous Clones Resistant to Phytophthora palmivora in the "Guiana" Genetic Group of Theobroma cacao L. *PLOS ONE* **7**, e40915 (2012).

57. Risterucci, A. M. *et al.* A high-density linkage map of Theobroma cacao L. *Theor Appl Genet* **101**, 948–955 (2000).

58. Allegre, M. *et al.* Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of Theobroma cacao L. *DNA Res* **19**, 23–35 (2012).

59. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* **4**, 170093 (2017).

60. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).

61. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

62. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

63. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203–214 (2000).

64. Kahlke, T. & Ralph, P. J. BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods in Ecology and Evolution* **10**, 100–103 (2019).

65. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

66. Sallet, E., Gouzy, J. & Schiex, T. EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. *Methods Mol Biol* **1962**, 97–120 (2019).

67. Argout, X. *et al.* Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of Theobroma cacao L. generated from various tissues and under various conditions. *BMC Genomics* **9**, 512 (2008).

68. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**, 3420–3435 (2008).

69. Li, P. *et al.* RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 852 (2016).

70. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357–359 (2012).

71. Tay Fernandez, C. G. *et al.* An SGSGeneloss-Based Method for Constructing a Gene Presence-Absence Table Using Mosdepth. *Methods Mol Biol* **2512**, 73–80 (2022).

72. Frichot, E. & François, O. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6**, 925–929 (2015).

73. Perrier, X. & Jacquemoud-Collet, J. P. DARwin software. (2006).

74. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).

Table 1. Gene ontology (GO) term enrichment of biological processes found in variable genes

GO ID	GO Term	Annotated	Significant	Expected	P-value
Immunity					
GO:0006952	defense response	1065	282	83.78	< 1e-30
GO:0042742	defense response to bacterium	129	20	10.15	2.6e-3
Gene transcription regulation					
GO:0070919	production of siRNA involved in gene silencing by small RNA	118	57	9.28	< 1e-30
GO:0030422	production of siRNA involved in RNA interference	135	57	10.62	5.6e-28
GO:0001172	transcription, RNA-templated	68	26	5.35	2.7e-12
GO:0070918	production of small RNA involved in gene silencing by RNA	189	74	14.87	2.7e-7
GO:0015074	DNA integration	354	55	27.85	8.5e-7
GO:0090502	RNA phosphodiester bond hydrolysis, endonucleolytic	364	49	28.64	1.5e-4
GO:0035556	intracellular signal transduction	420	52	33.04	7.4e-4
GO:0031047	gene silencing by RNA	291	91	22.89	1.5e-3
GO:0010495	long-distance posttranscriptional gene silencing	35	8	2.75	5.0e-3
GO:0045899	positive regulation of RNA polymerase II transcription preinitiation complex assembly	6	3	0.47	8.1e-3
Protein metabolic processes					
GO:0006508	proteolysis	1524	200	119.89	9.3e-14
GO:0006468	protein phosphorylation	1523	184	119.82	1.2e-9
GO:0006438	valyl-tRNA aminoacylation	6	4	0.47	5.0e-4
GO:0006480	N-terminal protein amino acid methylation	10	4	0.79	5.5e-3
Plant growth and cellular processes					
GO:0010073	meristem maintenance	107	31	8.42	9.5e-11
GO:0051923	sulfation	44	19	3.46	1.9e-10
GO:0009809	lignin biosynthetic process	24	8	1.89	3.4e-4

GO:0006207	'de novo' pyrimidine nucleobase biosynthetic process	42	10	3.3	1.3e-3
GO:0070682	proteasome regulatory particle assembly	4	3	0.31	1.8e-3
GO:0009220	pyrimidine ribonucleotide biosynthetic process	62	11	4.88	8.4e-3
GO:0120029	proton export across plasma membrane	31	7	2.44	9.0e-3
Biosynthesis of aromatic compounds					
GO:0016104	triterpenoid biosynthetic process	35	11	2.75	4.9e-5
GO:0016102	diterpenoid biosynthetic process	62	12	4.88	2.9e-3

Table 2: The number of different RGA candidates and subfamilies found on the *T. cacao* pangenome and across the 15 genetic groups. The numbers in parentheses represent the number of variable genes. AME=Amelonado, CAQ=Caqueta, CRI=Criollo, CUR=Curaray, APA=Apaporis, GUI=Guiana, IQU=Iquitos, NAP=Napo, MAR=Marañon, NAC=Nacional, NAN=Nanay, PAN=Purus, TIW=Tiwinza

Individuals	**RGA class**				
	NBS	RLK	RLP	TM-CC	Total
Pangenome	350 (105)	654 (43)	201 (52)	180 (10)	**1385 (210)**
AME	321 (76)	644 (33)	184 (35)	178 (8)	**1327 (152)**
APA	329 (84)	648 (37)	187 (38)	178 (8)	**1342 (167)**
CAQ	337 (92)	648 (37)	195 (46)	177 (7)	**1357 (182)**
CON	329 (84)	647 (36)	183 (34)	179 (9)	**1338 (163)**
CRI	303 (58)	649 (38)	184 (35)	176 (6)	**1312 (137)**
CUR	334 (89)	649 (38)	193 (44)	179 (9)	**1355 (180)**
GUI	336 (91)	648 (37)	189 (40)	178 (8)	**1351 (176)**
IQU	340 (95)	650 (39)	186 (37)	178 (8)	**1354 (179)**
MAR	333 (88)	648 (37)	189 (40)	179 (9)	**1349 (174)**
NAC	330 (85)	652 (41)	189 (40)	179 (9)	**1350 (175)**
NAN	327 (82)	646 (35)	184 (35)	178 (8)	**1335 (160)**
NAP	335 (90)	650 (39)	197 (48)	179 (9)	**1361 (186)**
PAN	341 (96)	651 (40)	193 (44)	179 (9)	**1364 (189)**
PUR	323 (78)	648 (37)	185 (36)	178 (8)	**1334 (159)**
TIW	338 (93)	649 (38)	196 (47)	179 (9)	**1362 (187)**

Figure legends

Figure 1. Pan-genome of Theobroma cacao. (a) Geographical distribution of the 216 T. cacao accessions used for pan-genome construction. (b) Pan-genome saturation curve indicating the distribution of the number of pan and core genes along with different numbers of sequenced individuals. (b) Pan-genome gene classification. (d) Comparison of variable genes detected in the T. cacao pan-genome with the Criollo (pink) and Amelonado (red) annotated T. cacao genomes.

Figure 2. Population structure and phylogeny of the 216 T. cacao accessions based on gene PAVs. (a) Model-based clustering analysis with ancestral kinship (K = 15). The y-axis quantifies cluster membership and the x-axis lists the different accessions. Cluster AME=Amelonado, CAQ=Caqueta, CRI=Criollo, CUR=Curaray, APA=Apaporis, GUI=Guiana, IQU=Iquitos, NAP=Napo, MAR=Marañon, NAC=Nacional, NAN=Nanay, PAN=Purus, TIW=Tiwinza. (b) 3D-PCA scatter plot of the first three principal components. (c) Neighbor-joining tree was calculated from PAV distance matrix computed using the Sokal & Michener algorithm. Numbers indicate branch length and colors genetic group. The phylogenetic tree was edited using iTOL v6 (https://itol.embl.de)

Figure 3. Gene number and gene space intersection among the T. cacao genetic groups. (a) Violin plots showing gene abundance for the 95 accessions exhibiting ancestry higher than 85% to their respective identified genetic cluster**.** Neighbor-joining tree was calculated from PAV distance matrix between the 95 accessions. Letters indicate comparable means according to the test of Newman-Keuls (probability p = 0.05). (b) Intersection plot of the gene space of the 15 genetic groups. “Very Low” indicates intersection <0.7, “Low” >=0.7 and <=1st quartile (0.85), “Medium” >1st quartile and <=2nd quartile (0.88), “High” >2nd quartile and <=3rd quartile (0.91), “Very High” > 3rd quartile.

Figure 4. Protein-coding genes under selection during the Criollo domestication

(a) Geographic location of the 35 *T. cacao* accessions native from the putative center of origin of the domesticated Criollo. (b) Scatter plot showing gene occurrence frequencies in the population of 35 accessions native from the putative center of origin of the domesticated Criollo and in the 13 accessions of the Criollo group. (c) Enriched GO terms in unfavorable genes during Criollo domestication.

Supplementary Figure 1

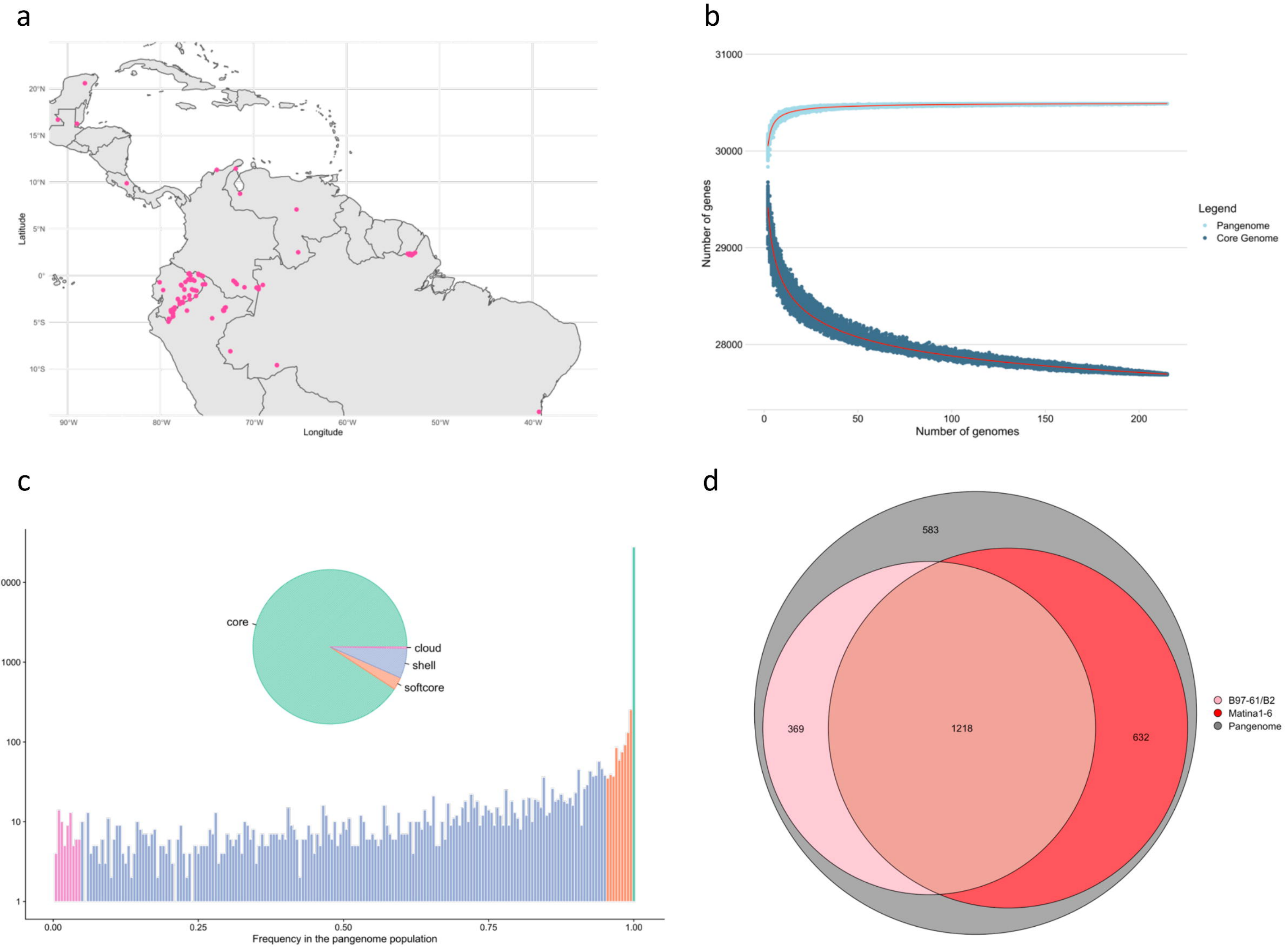
Word clouds of Gene ontology terms found in the non-reference genes belonging to different biological processes.

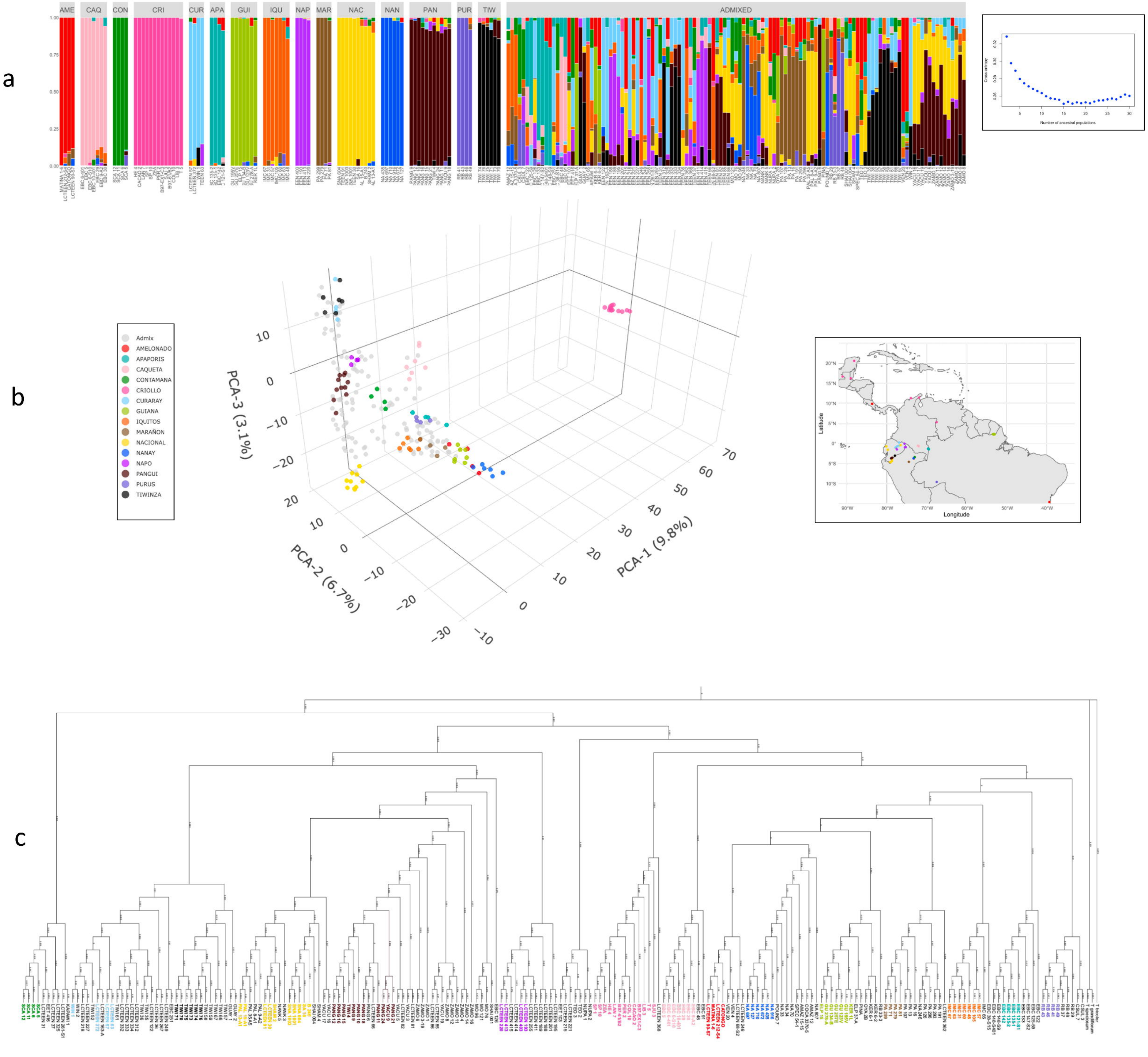
Supplementary Figure 2

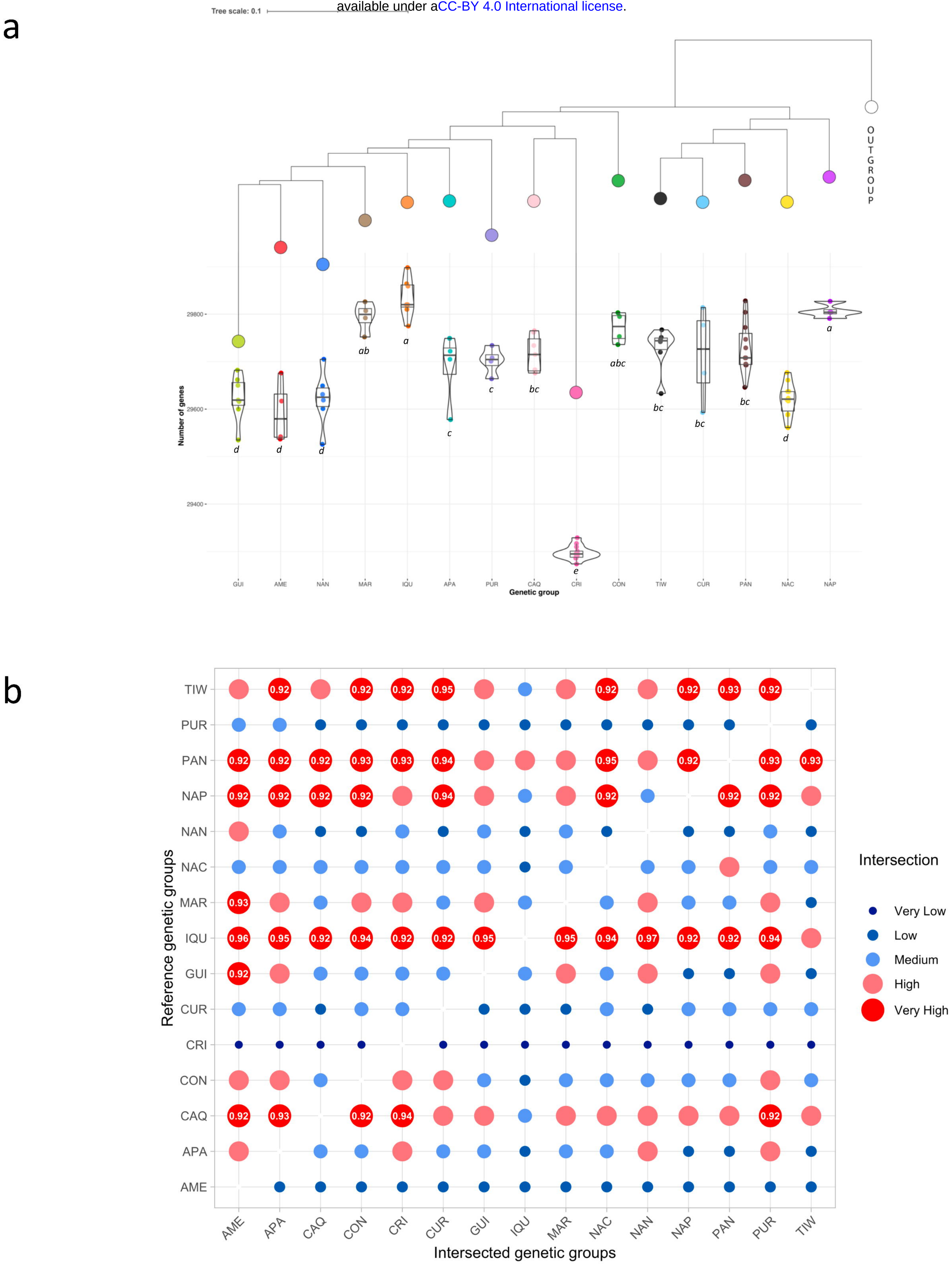
Model-based clustering (from K=2 to K=15) of the 216 *T. cacao* accessions based on the 2,802 gene PAVs.

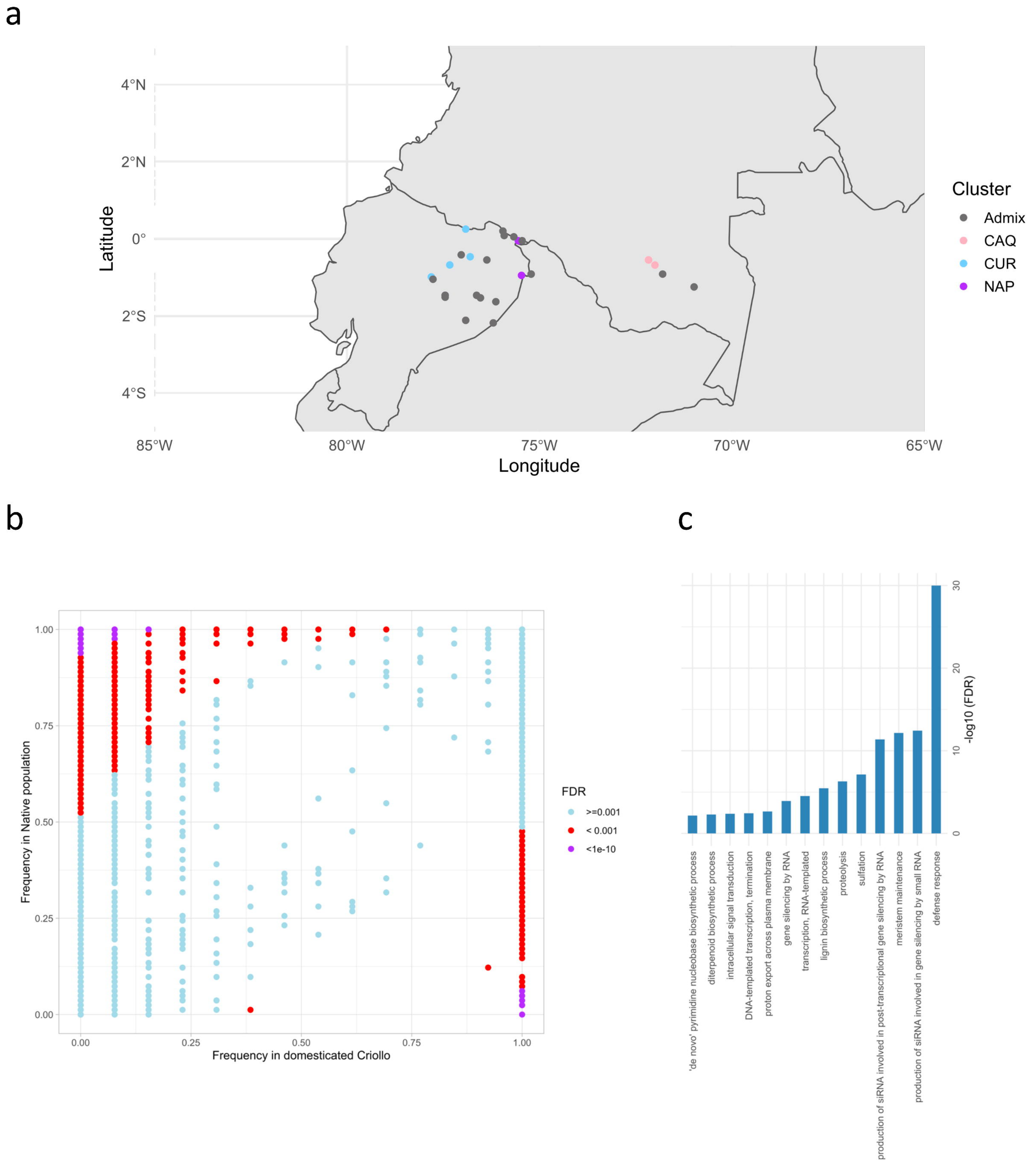
Supplementary Figure 3

Principal component analysis of the 216 *T. cacao* accessions based on the 2,802 gene PAVs.


a
20°N
15°N
10°N
5°N
0°
5°S
10°S
Latitude
90°W
80°W
70°W
60°W
50°W
40°W
Longitude
b
31000
30000
29000
28000
Number of genes
0
50
100
150
200
Number of genomes
Legend
● Pangenome
● Core Genome
c
10000
1000
100
10
1
Gene count
0.00
0.25
0.50
0.75
1.00
Frequency in the pangenome population
core
cloud
shell
softcore
d
583
369
1218
632
● B97-61/B2
● Matina1-6
● Pangenome


a
AME
CAQ
CON
CRI
CUR
APA
GUI
IQU
NAP
MAR
NAC
NAN
PAN
PUR
TIW
ADMIXED
1.00
0.75
0.50
0.25
0.00
Cross-entropy
Number of ancestral populations
5
10
15
20
25
30
b
Admix
AMELONADO
APAPORIS
CAQUETA
CONTAMANA
CRIOLLO
CURARAY
GUIANA
IQUITOS
MARAÑON
NACIONAL
NANAY
NAPO
PANGUI
PURUS
TIWINZA
PCA-3 (3.1%)
PCA-2 (6.7%)
PCA-1 (9.8%)
10
0
−10
−20
20
10
0
−10
−20
−30
−10
0
10
20
30
40
50
60
70
20°N
15°N
10°N
5°N
0°
5°S
10°S
90°W
80°W
70°W
60°W
50°W
40°W
Latitude
Longitude
c


a
Tree scale: 0.1
OUTGROUP
Number of genes
29800
29600
29400
d
d
d
ab
a
c
c
bc
e
abc
bc
bc
bc
d
a
GUI
AME
NAN
MAR
IQU
APA
PUR
CAQ
CRI
CON
TIW
CUR
PAN
NAC
NAP
Genetic group
b
Reference genetic groups
TIW
0.92 0.92 0.92 0.95 0.92 0.92 0.93 0.92
PUR
PAN
0.92 0.92 0.92 0.93 0.93 0.94 0.95 0.92 0.93 0.93
NAP
0.92 0.92 0.92 0.92 0.94 0.92 0.92 0.92
NAN
NAC
MAR
0.93
IQU
0.96 0.95 0.92 0.94 0.92 0.92 0.95 0.95 0.94 0.97 0.92 0.92 0.94
GUI
0.92
CUR
CRI
CON
CAQ
0.92 0.93 0.92 0.94 0.92
APA
AME
AME
APA
CAQ
CON
CRI
CUR
GUI
IQU
MAR
NAC
NAN
NAP
PAN
PUR
TIW
Intersected genetic groups
Intersection
Very Low
Low
Medium
High
Very High


a
4°N
2°N
0°
2°S
4°S
Latitude
85°W
80°W
75°W
70°W
65°W
Longitude
Cluster
● Admix
● CAQ
● CUR
● NAP
b
1.00
0.75
0.50
0.25
0.00
Frequency in Native population
0.00
0.25
0.50
0.75
1.00
Frequency in domesticated Criollo
FDR
● >=0.001
● < 0.001
● <1e-10
c
30
20
10
0
-log10 (FDR)
'de novo' pyrimidine nucleobase biosynthetic process
diterpenoid biosynthetic process
intracellular signal transduction
DNA-templated transcription, termination
proton export across plasma membrane
gene silencing by RNA
transcription, RNA-templated
lignin biosynthetic process
proteolysis
sulfation
production of siRNA involved in post-transcriptional gene silencing by RNA
meristem maintenance
production of siRNA involved in gene silencing by small RNA
defense response