Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process

Omar E. Cornejo1,2, Muh-Ching Yee2,3, Victor Dominguez4, Mary Andrews4, Alexandra Sockell2, Erika Strandberg2,5, Donald Livingstone III6,7, Conrad Stack6, Alberto Romero6, Pathmanathan Umaharan8, Stefan Royaert6, Nilesh R. Tawari9, Pauline Ng9, Osman Gutierrez10, Wilbert Phillips11, Keithanne Mockaitis4,12, Carlos D. Bustamante2 & Juan C. Motamayor6

Domestication has had a strong impact on the development of modern societies. We sequenced 200 genomes of the chocolate plant *Theobroma cacao* L. to show for the first time to our knowledge that a single population, the Criollo population, underwent strong domestication ~3600 years ago (95% CI: 2481–13,806 years ago). We also show that during the process of domestication, there was strong selection for genes involved in the metabolism of the colored protectants anthocyanins and the stimulant theobromine, as well as disease resistance genes. Our analyses show that domesticated populations of *T. cacao* (Criollo) maintain a higher proportion of high-frequency deleterious mutations. We also show for the first time the negative consequences of the increased accumulation of deleterious mutations during domestication on the fitness of individuals (significant reduction in kilograms of beans per hectare per year as Criollo ancestry increases, as estimated from a GLM, \( P = 0.000425 \)).
organized-state societies were only possible after the development of agriculture, which involved the domestication of numerous plants and animals\textsuperscript{3,4}. We are just starting to understand this process more from a genetic perspective, thanks to the expanding genomic technologies. Of particular interest is identifying the demographic scenario, timeline for domestication, and genomic consequences for species that have been critical in the development of societies\textsuperscript{2,3}. Among all species, there is a special place in the general culture for the domestication of \textit{Theobroma cacao} L., the plant from which chocolate is made. The chocolate tree has played a fundamental role in the development of Mesoamerican civilizations\textsuperscript{7} and has been a topic of research for over 100 years, but its domestication history has remained controversial\textsuperscript{8–10}. While the domestication history of cacao has sparked the interest across diverse disciplines, our knowledge of the process is incomplete and often involves partial information focused on a few genetic groups, a few geographical regions, fragmented archeological information, or a limited number of genetic markers\textsuperscript{8–10}. In this work, we report and analyze whole-genome variation of 200 \textit{T. cacao} individuals (Supplementary Table 1) to investigate the evolutionary origin of Criollo, the cacao tree domesticated in Mesoamerica. We also examine the consequences of the domestication process in the genomic architecture of the accumulation of deleterious mutations along the genome, which in turn allowed us to understand critical limits of Criollo (domesticated) cacao productivity.

The process of domestication of cacao has sparked the interest of a diverse set of disciplines and yet our knowledge of the process is incomplete and often involves partial information focused either on only a few genetic groups, a few geographical regions, fragmented archeological information, or a limited number of genetic markers\textsuperscript{8–12}. Current dogma suggests that cacao was introduced to Mesoamerica in Olmec times from the cacao varieties present in the Upper Amazon (Northern South America), the hotbed of diversity for the species\textsuperscript{6,8}. Another line of evidence suggests that the route of domestication of the chocolate tree could have dispersed throughout the Amazon Basin along two routes: one leading north and another leading west\textsuperscript{13}. According to this hypothesis, domestication of cacao would have occurred in South America and then spread to Central America and Mexico, carried during trade by native Americans\textsuperscript{14}. Anthropological research supports the view of a domestication event occurring in Mesoamerica\textsuperscript{6,9,12}. In addition, the continuous intermixing of farmed and wild cacao trees has likely continued to shape both gene pools in recent times\textsuperscript{11,15,16}. Both the impact of ancient domestication processes and modern hybridization on the genetic variation in the species are largely unknown in \textit{T. cacao}.

Traditional classifications of cacao have recognized the Criollo and Forastero groups or cultivars, and an additional hybrid of Criollo × Forastero called Trinitario\textsuperscript{7}. Biological characteristics of these groups have been described. More detailed genetic analyses using microsatellite markers have uncovered a large number of genetic groups as well as a clear differentiation between the trees found in the Amazon Basin and the Criollo varieties found in Central America\textsuperscript{10}. This work helped characterize the cacao germplasm into ten major genetically differentiated groups: Amelonado, Contamana, Criollo, Curaray, Guiana, Iquitos, Marañon, Nacional, Nanay, and Purús\textsuperscript{10}. Additional analyses performed with microsatellites suggested that Criollo, the most likely representative of the cacao domesticated in Mesoamerica, is more closely related to trees from the Colombia–Ecuador border than trees from other South American groups\textsuperscript{10}. Yet, there is a huge gap in information about the extent of genomic variation in the species, which makes it difficult to propose clear scenarios for the evolution of natural populations and the domestication of \textit{T. cacao}. Although it is widely recognized that some of these ten populations have contributed in recent times to the genetic makeup of crops, the majority of them remain as wild populations\textsuperscript{10}. The most closely related species to \textit{T. cacao} is putatively \textit{Theobroma grandiflorum}\textsuperscript{7}, but the biological characteristics of the trees and the fruits are dramatically different to those found in \textit{T. cacao}\textsuperscript{18}.

In this work, we explore the extent of whole-genome variation in \textit{T. cacao} L. and investigate the evolutionary origin of Criollo, the cacao tree domesticated in Mesoamerica. We provide a detailed analysis of the population genetic structure in the species and analyze the evolutionary history of the populations, with an emphasis on domestication and the process of selection during domestication. Finally, we show how the reproductive system in cacao and the relatively recent process of domestication have strongly influenced the accumulation of deleterious mutations in domesticated cacao, with measurable consequences on the fitness of the individuals. This last result is a strong validation of the cost of domestication hypothesis which proposes that the process of domestication of a species will result in the increase in number, frequency, or proportion of deleterious genetic variants, hindering genetic improvement of domesticated species\textsuperscript{19,20}. We show not only that indeed, there is an increase in the higher-frequency deleterious variants, but also show that this increase is associated with a reduction of individual fitness in domesticated cacao.

### Results

#### Genetic variation in \textit{Theobroma cacao} L.

Resequencing of 200 accessions (see Supplementary Table 1 for details on accessions selected for the study) at high coverage (average coverage 22×) generated ~4.52 trillion base pairs. After aligning the reads to the cacao reference (Matina-v1.12\textsuperscript{1}), we identified 7,412,507 single-nucleotide polymorphisms (SNPs). We found that cacao presents a high genetic variability of ~5 SNPs per kilobase per individual, similar to what has been observed in Arabidopsis\textsuperscript{22,23} (Supplementary Figure 1). Although the large majority of identified variants are noncoding, we identified 322,275 (4.35% of the total SNPs) missense variants and 220,043 (2.97% of the total SNPs) synonymous variants in 29,408 genes. We also identified 10,062 variants predicted to change the splice donor sites, which could be responsible for polymorphism, impacting the number of transcripts produced\textsuperscript{24}. Among the potential changes that alter the length of the transcripts, we identified 8470 stop loss (0.114% of total SNPs), 16,956 stop gains (0.229% of total SNPs), and 8588 stop losses (0.116% of total SNPs). Overall, this SNP dataset represents a new resource for cacao biology that we hope will accelerate breeding programs (for a catalog of SNPs contextualized with respect to gene annotation, see Supplementary Table 2 and Fig. 1).

#### Population genetic structure and signature of domestication

Model-based clustering analysis using ADMIXTURE\textsuperscript{24} enabled us to identify ten genetic clusters, consistent with previous analyses\textsuperscript{10}, and allowed us to correctly assign overall ancestry to previously uncharacterized accessions (Fig. 2a). We also present, for the first time, estimates of global ancestry for natural admixed individuals and man-made hybrids, revealing the relative contribution of the ten main populations to individual ancestry (Fig. 2a and Supplementary Figure 2). Our results show that there is an overrepresentation of genetic material from Amelonado, Criollo, and Nacional ancestry in the majority of admixed individuals. There is a concomitant underutilization of other genetic groups in current agronomist practice, which provides Blue Ocean opportunities for crop improvement (see Supplementary Information). Analyses of genetic diversity show significant differences in π across populations (Supplementary Table 2 and Fig. 1).
Table 3, $p < 2e^{-16}$) and along the genomes within populations, a pattern that is consistent with what has been observed in other species (see Supplementary Information for more details and Supplementary Figures 4 and 5).

Further analysis of the population structure shows that the Criollo group is clearly differentiated from the rest of the genetic clusters along the first axis of a multidimensional scaling (MDS) analysis (Fig. 2b). The second component of the MDS analysis presents a gradient separating genetic clusters roughly from Pacific to Atlantic (bottom to top of the second component), consistent with a natural process of differentiation from the higher diversity groups on the Pacific side of the Amazonian Basin to those of lower diversity on the Atlantic side (See Fig. 2c, excluding domesticated Criollo, $Y = 0.202 - 72.71 \times, p = 0.02$, Supplementary Table 4). It has been proposed that the center of origin of the species is located in the Western Amazon7,25. Our observation of the significant decline ($p = 0.02$ as per the model above) in genetic diversity from Pacific to Atlantic is consistent with the suggested center of origin for the species. The spread of admixed individuals in the MDS space is consistent with our admixture analysis in which individuals fall into two general categories: one which presents admixture between Criollo and Amelonado with a minor contribution from other groups and other hybrids presenting admixture along the Atlantic–Pacific gradient. There is a pattern of strong differentiation between all genetic clusters (Fig. 2d, $F_{ST}$ values range between 0.16 and 0.65), with a larger differentiation between Criollo and any other group, consistent with either a scenario of strong drift during a recent process of domestication or an old diversification of the Criollo population from the rest of the genetic clusters. Given previous anthropological and genetic evidence6,12,14, it is more likely that this pattern is the result of domestication from a small pool of seeds that were used to create the Criollo group (drift), however, it is also possible that the strong divergence of Criollo from all the other groups is the result of a combination of both scenarios (diversification of the Criollo ancestral population and human-mediated genetic drift).

The hypothesis of a single event of domestication (along with genetic drift after transport from South to Central America) predicts that Criollo would show a higher differentiation to other groups than that observed between any other pairwise comparison of the populations. Our population structure analyses are consistent with this prediction (Fig. 2b, d). Our model-based analysis of population differentiation with TreeMix26 provides evidence that Criollo was the result of a single domestication event, undergoing extreme drift after separating from its most closely related population (represented as a longer branch in Fig. 3a, b, similar structure was obtained in a Neighbor-Joining analysis presented in Supplementary Figure 3). This analysis also shows that Criollo is most closely related to Curaray, suggesting that the origin of domesticated cacao was a subset of ancient Curaray germplasm10, a genetic cluster that has been described for the North of Ecuador and South of Colombia10,27. After exploring multiple models of differentiation with admixture, we found no evidence supporting subsequent contributions of any group to the domesticated Criollo, with the exception of a potential recent contribution of Purus to Criollo (Fig. 2b). However, we learned from this analysis that multiple instances of admixture have potentially occurred among multiple groups during their natural process of differentiation along the Amazon Basin (see Supplementary Information).

Evolutionary genomics Theobroma cacao L. In addition to admixture and population differentiation analyses, we investigated the demographic history of the ten genetic clusters to understand the natural demographic process that has characterized the species historically. Given the relatively small number of accessions per group, we performed analysis with pairwise sequentially Markovian coalescent (PSMC) model28 and smc++,29 which allows evolutionary history inference by analyzing single diploid genomes. In general, the evolutionary history of T. cacao shows a common trend toward the reduction of population size/genetic diversity with time (Fig. 3c, d, and e). The median demographic history among accessions was used to show the overall trends for the evolutionary history of the groups. The process of reduction of effective population size started prior to the peopling of the Americas, which suggests that the overall reduction of genetic diversity in the species could be tied to environmental changes or historical changes in the distribution of pollinators and/or animals that are involved in the dispersion of the seeds during the Last Glacial Maximum. This result is
consistent with recent studies suggesting that most groups of *Theobroma cacao* could have diversified during the last glacia-
tion\(^2^7\). During the Last Glacial Maximum, it has been inferred
that the Amazon had dry seasons that lasted twice as long as
present day and rainfall could have dropped 25–35% from pre-
sent day records and presented remarkably lower CO\(_2\) con-
centration in the atmosphere\(^3^0,3^1\). Pockets of increased humidity
and more constant temperature during the year were constrained
to the vicinity of major river basins and the development of
refugia\(^3^0,3^1\). The fact that populations of cacao have been
decreasing historically is consistent with recent studies that have
analyzed the conservation status of more than 15,000 species of
Amazonian trees, predicting that *T. cacao* could suffer an addi-
tional 50% population decline in the near future\(^3^2\). Because of the
lack of reliability of these methods to resolve more recent
demographics, there is little that can be said about the apparent

**Fig. 2** Population genetic structure in *T. cacao*. **a** The ten main genetic clusters can be recovered (A.1), although further structure (11 clusters) seems to be meaningful given that a considerable number of admixed individuals present the ancestry from a subset of Amelonado ancestry (A. 2). Color bars on top of the admixed individuals show our suggested grouping for the hybrids. **b** Map of Central and South America showing the median coordinate locations for the origin of samples from each population sampled in this work (with the exception of Admixed). **c** MDS showing a gradient of differentiation form the West to the East side of the Amazon (PC2) and a major separation of the Criollo group that corresponds to the Mesoamerican domesticated group (PC1). **d** Significant decay of genetic diversity (\(\pi\)) for the species along PC2 is supportive of the origin of the species being in the western side of the Amazon Basin (Criollo is excluded, model: \(\pi \sim \text{group} + \epsilon, p < 2E^{-16}, r^2 = 0.19\)). **e** All ten population genetic groups that have been described for the species are highly
differentiated, with Criollo presenting a larger average F\(_{ST}\) when compared against all the other groups.
recent increase in population size. Additional analyses with a larger sample size per population and inference of recent effective population sizes with Identical By Descent (IBD)/linkage-based methods will be necessary to address this issue.

Using what we learned from the admixture/principal components analysis (PCA) (Fig. 2a, b) and our overall assessment of the demographic history of the populations (Fig. 3c, d), we explored the evolutionary history of domestication of the Criollo...
group from a Curaray ancestor to answer two critical questions: (1) how long ago the ancestral Curaray populations gave rise to what is today known as the Criollo group and (2) the size of the founding population of Curaray ancestry that used to domesticate cacao Criollo in Central America. For this, we analyzed the frequency spectrum of variants under a model of isolation, with migration under a maximum likelihood framework with $\delta a_{6}^{33}$. Our analyses show that the fraction of the ancestral Curaray effective population used to domesticate Criollo in Mesoamerica was indeed very small and comprised $\sim 738/1476$ individuals (95% CI: $437/574$–$2647/3894$ individuals for mutation rates $7.1 \times 10^{-9}/3.1 \times 10^{-9}$, respectively; see Supplementary Figures 6 and 7). More importantly, we provide strong support from genomic data analysis that this process started 3600/7200 years before present (95% CI: 2481/4162–10,903/13,806 years before present for mutation rates $7.1 \times 10^{-9}/3.1 \times 10^{-9}$ mutations bp$^{-1}$ gen$^{-1}$, respectively). The observed distribution of shared variants for different minor allele frequency categories (Fig. 3f) fits well the predicted values under the best fitted model (Fig. 3g) with a distribution of residuals that show a good absolute fit (for details about the alternative demographic models tested see Supplementary Information). Our estimates for the time of separation of the Curaray and Criollo groups overlaps well with archeological evidence and what has been thought to be the onset of cultivation of Criollo in Mesoamerica$^{8,9,12,34}$. These results are consistent with findings of theobromine in Olmec pottery from the capital San Lorenzo as old as the Early Preclassic (1800–1600 BCE)$^{9,33}$. Our demographic analyses are also consistent with large-scale analyses of modern and ancient DNA, which pinpoint the colonization of the American continent by humans to roughly 13,000 years ago$^{36–38}$. In addition, recent analysis of the post-colonization demographic of human in South America is consistent with human populations staying in relative low numbers for the first 8000 years and then with the advent of agriculture and thus sedentism, experiencing a population expansion at $\sim 5000$ years ago, similar to that experienced during the Neolithic revolution elsewhere on the globe$^{39}$. In short, our understanding of human demographic history suggests that our inference of T. cacao domestication in Mesoamerica between 2481 and 13,806 years before present are strongly consistent with the history of human settling in the region, but our knowledge of human history suggests that times closer to the lower limit of the confidence interval or at least lower than 8000 years ago are more likely. A schematic of the best demographic model explaining the data is provided in Fig. 3h. Although we have been as rigorous as possible in the analysis (see Supplementary Figure 8), it will be important to validate the estimated age for divergence between Curaray and Criollo with methods that could better resolve recent demographics with a larger number of individuals from each population.

Our analyses also show that the patterns of linkage disequilibrium (LD) are consistent with the observed demographics, with Criollo populations showing a higher LD over longer stretches of the genome (Supplementary Figure 9 and Supplementary Information).

One of the greatly appreciated features of the domesticated Criollo cacao is the white cotyledon of the bean, which seems to be associated with desirable flavor qualities. Early work has suggested that decreased concentrations of polyphenols, methylxanthines, and anthocyanin precursors in the cotyledon are associated with this observation$^{40–42}$. Polyphenols and methylxanthines are responsible for the astringency and bitterness detected in cacao beans$^{43}$, and it is thought that the modification of these compounds during the process of fermentation contributes to the final flavor of a chocolate$^{43}$. In fact, during the process of fermentation, the concentration of polyphenols is reduced by up to 70%$^{44}$. Plants of the Criollo variety were likely selected during domestication to reduce this bitterness. We investigated the impact of artificial selection during domestication on the Criollo genome by looking for regions of increased differentiation between Criollo and its sister population Curaray, using XP-CLR, a method that seeks to identify changes in the distribution of allelic variation or changes in the 2D-site frequency spectrum along the chromosomes in sliding windows$^{45}$. We found several regions of the genome in which natural selection has produced higher differentiation between Curaray and Criollo than expected by demographics alone (Fig. 4, Supplementary Data 1 and 2). The most interesting result derives from the identification of genes encoding laccase 14, laccase/diphenol oxidase. Laccases are normally associated with the process of lignification, but it has recently shown that laccases are also involved in the metabolism of polyphenols$^{46–48}$; we hypothesize that selection on these genes likely results in the reduction of the concentration of polyphenols in cacao. We also identify signatures of selection in a region containing the gene encoding xanthine dehydrogenase 1, likely involved in the metabolism of methylxanthines (like theobromine) and also likely to have been the result of the process of selection for reduced bitterness$^{42}$. An additional list of genes in regions identified to be under selection is provided in Supplementary Table 3 and includes genes involved in genomic stability (structural maintenance of chromosomes), disease resistance, abiotic stress response (WRKY DNA-binding protein), transcriptional regulation (MYB domain), and signaling (cysteine-rich RLK receptor as well as S-domain-2 5 genes).

Most mutations that appear in the genome are deleterious and have the potential for reducing reproductive success$^{49–51}$. The fate of these mutations and their transit time in a population strongly depends on the intensity of genetic drift, purifying selection, and the degree of dominance of the mutations. Mathematical
population geneticists were long concerned with the impact of the accumulation of deleterious mutations in a population\textsuperscript{52}. The process of domestication in animals and plants has been used as a framework to study how intense selection of some desirable traits affects the accumulation of deleterious mutations in the population\textsuperscript{3,53}. Yet, thus far, we have little evidence of how the process of accumulation of deleterious mutation affects traits associated with fitness or, in the case of crops, productivity.

**Test of the cost of domestication hypothesis in** *T. cacao* **L.**

Populations of cacao have been declining over time and a natural consequence of the reduction in population size is increasing in inbreeding. Because all ten populations of cacao are experiencing reductions in population size, it is expected that this process will have a similar effect across populations, and the differences in magnitude of inbreeding will reflect differences in population size. We observe an increase in the amount of inbreeding (estimated as F-statistics\textsuperscript{54}) when the Admixed cluster of individuals (expected to have low inbreeding) is compared with the ten populations defined in Fig. 1 (Fig. 5a, Kruskal–Wallis test chi-squared = 803.45, df = 10, p-value < 2.2e\textsuperscript{−16}, significant Nemenyi’s post-hoc tests between Admixed and all populations, except Iquitos and Nacional). These differences between the coefficients of inbreeding can be partially explained as a function of the differences in historical population size among genetic clusters (Fig. 5b, see Supplementary Information).

Population genetics theory predicts that selfing increases the efficiency in the elimination of recessive deleterious mutations, when compared with outcrossing populations, because variants otherwise hidden in heterozygous individuals will be exposed to the action of natural selection\textsuperscript{55,56}. In contrast, domestication is a process that has been shown to contribute to the maintenance of deleterious mutations in higher frequency in populations\textsuperscript{3,53}. The impact of domestication on arboreal crops is not well understood, and it is even less understood in a plant-like cacao that in other organisms, including dogs and humans\textsuperscript{53,60}, it has not been shown what the impact of the accumulation of deleterious mutations is on fitness. We tested the hypothesis that the accumulation of deleterious mutations due to domestication would decrease fitness by examining the relationship between Criollo ancestry and a measure of performance in cacao using an independent dataset. Individuals were genotyped with a SNP array that was developed using a subset of genotypes inferred as part of this work and published elsewhere\textsuperscript{61}. We measured bean (seed) productivity (yield in kilograms of bean per hectare per year for each plant) as a measure of fitness. We inferred proportional ancestry to a new set of admixed individuals for which productivity had been assessed (Fig. 5d and Supplementary Figure 11) and showed that there is a significant negative relationship between Criollo ancestry and fitness (Fig. 5e, with Criollo ancestry decreasing yield per hectare per year in \sim 319.9 units per percent unit of ancestry, p = 0.000425, additional details for the model are available in the Supplementary Information). We also demonstrate that despite the decrease in fitness in domesticated populations, there is no loss in quality and ability to prepare chocolate from its beans (Supplementary Figure 12).
In summary, we provide the first general view of how natural diversification has shaped genetic variation in *Theobroma cacao*. We provide the first comprehensive view of the demographic scenario involved in the domestication of the Criollo variety and identify genes that could function in the desirable taste attributes of the Criollo variety. More importantly, we show how genomic resources can be successfully used to evaluate how the process of domestication has shaped the pattern of accumulation of deleterious mutations in arboreal crops, impacting fitness in a considerable way, validating for this species the cost-of-domestication hypothesis.²⁰
Conclusions
Our study has provided a closer look at the evolutionary history of Theobroma cacao. We have developed a great resource for breeders and researchers, which include over 7 M SNPs, and corresponding genomic annotation for those variants. The results of the work presented in this manuscript shed light on a diverse array of questions that range from a deeper characterization of the genetic population structure in cacao populations to increasing our understanding of the evolutionary history of domestication in cacao. Most importantly, our work has provided strong genomic evidence supporting the cost-of-domestication hypothesis, stating that the process of improvement and selection for desirable traits is hindered by the undesired accelerated accumulation of deleterious mutations.

Methods
Sampling
We sampled leaves from accessions at the Cacao Research Unit at the University of West Indies and CATIE in Costa Rica (Supplementary Table 1).

DNA extraction and sequencing libraries preparation
Samples processed at Stanford University were prepared as follows:

DNA was extracted using ZR Plant/Seed DNA MiniPrep™ (Zymo Research Inc). Approximately 3 g of leaf material per extraction per sample was cut and placed in homogenization tubes with ceramic pearls and lysis buffer. Samples were homogenized in a FastPrep-24™ (MP Biomedicals, LLC) placed in a cold room at 4°C for 60 s at a speed of 4.5 m·s⁻¹. If the tissue was not homogenized thoroughly, the tissues were homogenized for an additional 20–40 at the same speed. DNA was quantified using a Qubit™ 3.0 fluorometer (ThermoFisher Scientific), using a dsDNA HS Assay Kit. Additionally, overall quality of extracted DNA was assessed with 2% 1x Gel (Invitrogen, Carlsbad, CA). Most of the samples were prepared using Nextera DNA Sample Preparation Kits (Epizentec, Chicago, IL, USA) and NEBNext® Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Inc). The remaining samples were prepared by first shearing the genomic DNA using a M220 Focused-ultrasonicator™ ( Covaris Inc) and NEBNext® Ultra DNA Library Prep Kit for Illumina (New England Biolabs, Inc). Libraries were quantified on Agilent 2100 Bioanalyzer High Sensitivity DNA chip for concentration and size distribution, pooled in sets of 3–4 per batch, and sequenced on the HiSeq 2000/2500 platform at the Stanford Sequencing Service Center (100 cycles, paired read mode).

Samples processed at Indiana University were prepared as follows:

DNA was extracted using a protocol customized for enrichment of high molecular weight DNA from cacao leaves. Approximately 450 mg of leaf material per sample was ground to powder under liquid N₂ using mortar and pestle. Tissue powder was homogenized and washed twice by vortexing in 3 ml of 10 mM HEPES, 0.1% PVP-40, 4% b-mercaptoethanol, followed by centrifugation at 7000 rpm in an Eppendorf 5415 R at 4°C. Nuclei were extracted from tissue-pellets on ice in 50 mM Tris-Cl pH 8.0, 50 mM EDTA, and 50 mM NaCl with 15% sucrose, and centrifuged at 3600 rpm to pellet the cellular debris. Nuclei were lysed at 70°C for 15 min in 20 mM Tris-Cl pH 8.0, 10 mM EDTA with the addition of SDS to a final concentration of 1.5%. Protein was precipitated on ice with the addition of NH₄OAc to a final concentration of 2.7 M, pelleted twice by centrifugation at 7000 rpm. DNA was precipitated using gentle inversion in an equal volume of cold isopropanol, followed by centrifugation at 7000 rpm. DNA pellets were washed in 70% ethanol and resuspended in 10 mM Tris-Cl, 1 mM EDTA using wide bore pipette tips, DNA quality and quantity in the high molecular weight fraction (24 to 40 kb) was assessed by migration on Genomic DNA Screen Tape, Agilent TapeStation 2200 Software (A01.01.04) (Agilent) and secondarily quantified by fluorometry using the dsDNA HS Assay Kit (Invitrogen) with a Qubit™ 2.0 fluorometer (ThermoFisher). Sequencing libraries were prepared either as unamplified NGS libraries, using the PCR-free DNA library kit (KAPPA) or minimally amplified libraries were prepared using the TrueSeq DNA Sample Prep Kit (Illumina) with four cycles of PCR at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana–Champaign (UIUC). All library preparation steps were according to the manufacturer with the exception that after shearing for minimally amplified libraries, DNA was cleaned through a Zymo column and size selected to retain only 400–600 bp fragments. All libraries were evaluated for quality using an Agilent 2100 Bioanalyzer High Sensitivity DNA Assay (Agilent), quantified by qPCR, pooled in sets of 12 at equimolar concentration, and sequenced as paired 2 × 151nt reads on a UHCC HiSeq2500 instrument using HiSeq SBS sequencing kit version 4. Fastq files were generated with CASAVA 1.8.2.

Read processing and SNP identification
The Illumina data were basecalled using Illumina software CASAVA 1.8.2, and sequences were demultiplexed with a requirement of full match of the six nucleotide index that was used for library preparation. Samples prepared using Nextera were hard clipped 13 nt from the 5’ end. Following demultiplexing, raw sequenced data was analyzed for quality using FastQC®. We performed adaptive quality trimming (setting a quality threshold of 23) and additional hard trimming of the reads based on the stabilization of the base calling (quality) in the 5’ end of the sequences using TrimGalore! and cutadapt. Sets of reads from individual samples were analyzed using the ADMIXTURE genome24, using the bowe-warelier aligner BWA® with relaxed conditions for the editing distance (0.06), as it was expected that T. cacao has a high genetic diversity. Aligned sam files were processed prior to performing SNP identification with Samtools/Picard Tools to mark duplicates, and Bamtools® to generate mate pair information, correct unmapped reads flags, and obtain overall mapping statistics. We followed the recommendations of the Genome Analysis Toolkit to perform base quality recalibration and local realignment to minimize false positives during the SNP calling procedure35. Finally, we performed genotype calling using Real Time PCR libraries to get a final population annotation. SNPs were filtered in regards of our filtering procedure28. As we used the current common annotation from the Matina-v1.1 reference genome23 to construct a new database for Theobroma cacao. This database was used to annotate the observed polymorphisms following their potential effect on gene expression and functionality according to their position with respect to the coding regions.

Population genetic analyses
We characterize the distribution of genetic variation in the populations, estimating variation using two approximations for the inference of genetic variation: Watterson’s theta (θw)24 and the number of pairwise differences per site (ns)73. We used vcftools76 to estimate both statistics in windows of 1 kb. Generalized Linear models explore the differences in diversity among populations are explained in the section Distribution of Genetic variation among genetic groups in the Supplementary text. We used ADMIXTURE™ as an implementation of an approach similar to well-known STRUCTURE™. Based on an expectation–maximization algorithm, ADMIXTURE uses a maximum likelihood-based approach to assign ancestry genome wide and visualize the genetic structure of the T. cacao populations. A cross-validation procedure is employed to select the most likely number of clusters that explains the structure of the data24. We filtered our data and restricted our analysis to SNPs with minor allele frequency over 5%, and we also pruned the data sets of reads from individual samples, fisher strand test (FS 50), and the root mean square of the mapping quality across samples (MQ 30). Variants identified were phased, per population, using shapeit v2.12 on a subset of variants in which the minor allele frequency (MAF) > 0.057,2. The phasing was performed per chromosome for the ten main chromosomes using only biallelic sites.

Identification of SNPs were annotated using SNPinfo®. For this, we used the current gene annotation from the Matina-v1.1 reference genome23 to mark SNPs with pair information, correct unmapped reads flags, and obtain overall mapping statistics. We followed the recommendations of the Genome Analysis Toolkit to perform base quality recalibration and local realignment to minimize false positives during the SNP calling procedure35. Finally, we performed genotype calling using Real Time PCR libraries to get a final population annotation. SNPs were filtered in regards of our filtering procedure28. As we used the current common annotation from the Matina-v1.1 reference genome23 to construct a new database for Theobroma cacao. This database was used to annotate the observed polymorphisms following their potential effect on gene expression and functionality according to their position with respect to the coding regions.

Population genetic analyses
We characterize the distribution of genetic variation in the populations, estimating variation using two approximations for the inference of genetic variation: Watterson’s theta (θw)24 and the number of pairwise differences per site (ns)73. We used vcftools76 to estimate both statistics in windows of 1 kb. Generalized Linear models explore the differences in diversity among populations are explained in the section Distribution of Genetic variation among genetic groups in the Supplementary text. We used ADMIXTURE™ as an implementation of an approach similar to well-known STRUCTURE™. Based on an expectation–maximization algorithm, ADMIXTURE uses a maximum likelihood-based approach to assign ancestry genome wide and visualize the genetic structure of the T. cacao populations. A cross-validation procedure is employed to select the most likely number of clusters that explains the structure of the data24. We filtered our data and restricted our analysis to SNPs with minor allele frequency over 5%, and we also pruned the data sets of reads from individual samples, fisher strand test (FS 50), and the root mean square of the mapping quality across samples (MQ 30). Variants identified were phased, per population, using shapeit v2.12 on a subset of variants in which the minor allele frequency (MAF) > 0.057,2. The phasing was performed per chromosome for the ten main chromosomes using only biallelic sites.

Identification of SNPs were annotated using SNPinfo®. For this, we used the current gene annotation from the Matina-v1.1 reference genome23 to construct a new database for Theobroma cacao. This database was used to annotate the observed polymorphisms following their potential effect on gene expression and functionality according to their position with respect to the coding regions.
likelihood units is conservative enough to assure similar results to those obtained with CLUMPP. In addition to the admixture analysis, we performed a multidimensional scaling analysis on the same set of SNPs employed for ADMIXTURE. First, we normalized the data (centered and standardized) following previous recommendations and performed MDS analyses using Singular Value Decomposition on the normalized data using the cmd scale function in R.

We inferred genetic relationships among the distinct genotypes from restrictions in gene flow between populations using Weir and Cockermah's $F_{ST}$ estimator in windows of 5 kb, after filtering out low-frequency alleles. To summarize the genome-wide differentiation among populations, we estimated the mean of $F_{ST}$ estimators across windows and standard error for every pair of populations.

The population structure of populations in South America was created using ggpmaps in R. The maps used in ggpmaps are obtained from Google maps (open access source) and the diamonds used for the positioning of the populations were modified to increase the size in the Illustrator.

We fitted a generalized linear model to explain the differences in genetic diversity along the Pacific and Atlantic axes of genetic differentiation captured in the second component of a multidimensional scaling. For this, we estimated the centroids for PC1 and PC2 of the data presented in Fig. 1b. These centroids were used as predictors ($\beta_i$) to explain the differences in mean genetic diversity per population (measured as $\pi$, $Y$ in the following model) under a simple linear model with a Gaussian family $Y = \beta_0 + \beta_1 + \epsilon$. Admixed individuals were excluded from the analysis.

We used a model-based approach to infer the population relationships between the ten main groups as implemented in TreeMix to identify the relationships between populations and identify signatures of domestication.

We used admixture and multidimensional scaling to infer the demographic history of populations using individual genomes and small sets of individuals per population. First, we used the pairwise sequentially Markovian coalescent as implemented in PSMC, second, we used SMC+++, a likelihood-free method that can leverage information from multiple individuals from the population (as opposed to PSMC) to infer population size changes in the past. We assumed a mutation rate $\mu = 7.1 \times 10^{-9}$ mutations $/bp/\text{gen}$.

We also examined the effect of uncertainty in mutation rates by including analysis following recent work suggesting that mutation rates could be half of that estimated previously on the order of $3.1 \times 10^{-9}$ to go from seed to seed in cacao. The generation time of 5 years, based on the observation that it takes 5 years on average for a cacao tree to produce its first flowers, could be used to infer the population history. The true history of chocolate, however, is much more complex.

We measured population differentiation resulting from restrictions in gene flow using the $F_{ST}$ statistic. To summarize the genome-wide site frequency spectrum, for this, we set fixed windows of 0.05 cM for 200 SNPs and grid size of 2 kb. For these analyses, we used the geno2d chip of 15 K SNPs specified for cacao that was developed in parallel to this work.

We also examined the effect of uncertainty in mutation rates by including analysis following recent work suggesting that mutation rates could be half of that estimated previously on the order of $3.1 \times 10^{-9}$ to go from seed to seed in cacao. The generation time of 5 years, based on the observation that it takes 5 years on average for a cacao tree to produce its first flowers, could be used to infer the population history. The true history of chocolate, however, is much more complex.

We measured population differentiation resulting from restrictions in gene flow using the $F_{ST}$ statistic. To summarize the genome-wide site frequency spectrum, for this, we set fixed windows of 0.05 cM for 200 SNPs and grid size of 2 kb. For these analyses, we used the geno2d chip of 15 K SNPs specified for cacao that was developed in parallel to this work.

We also examined the effect of uncertainty in mutation rates by including analysis following recent work suggesting that mutation rates could be half of that estimated previously on the order of $3.1 \times 10^{-9}$ to go from seed to seed in cacao. The generation time of 5 years, based on the observation that it takes 5 years on average for a cacao tree to produce its first flowers, could be used to infer the population history. The true history of chocolate, however, is much more complex.

We measured population differentiation resulting from restrictions in gene flow using the $F_{ST}$ statistic. To summarize the genome-wide site frequency spectrum, for this, we set fixed windows of 0.05 cM for 200 SNPs and grid size of 2 kb. For these analyses, we used the geno2d chip of 15 K SNPs specified for cacao that was developed in parallel to this work.
73. Cingolani, P. et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* 3, 35 (2012).

74. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276 (1975).

75. Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *P Natl. Acad. Sci. USA* 76, 5269–5273 (1979).

76. Danecck, P. et al. The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158 (2011).

77. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959 (2000).

78. Behar, D. M. et al. The genome-wide structure of the Jewish people. *Nature* 466, 238–242 (2010).

79. Jakobsson, M. & Rosenberg, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806 (2007).

80. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PloS Genet.* 2, e190 (2006).

81. Lynch, M. Evolution of the mutation rate. *Trends Genet* 26, 345–352, https://doi.org/10.1016/j.tig.2010.05.003 (2010).

82. Ossowski, S. et al. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* 327, 92–94 (2010).

83. Crow, J. F. & Kimura, M. *An introduction to population genetics theory*. 591 (Blackwell Press, Caldwell, 1970).

**Acknowledgements**

All accessions of cacao sequenced in this work were obtained from the Cocoa Research Center (University of West Indies, Trinidad), the Centro Agronómico Tropical de Investigación y Educación (CATIE, Costa Rica); the samples obtained from NARS in Cameroon and the Ivory Coast. MARS provided with all the necessary funds to sequence accessions via a contract with CDB, and OEC was supported via this grant. This research used resources from the Center for Institutional Research Computing at Washington State University and Stanford Computing.

**Author contributions**

O.E.C., C.D.B., and J.C.M. designed the study. O.E.C., M.C.Y., A.S., E.S., V.D., M.A., and K.M. contributed to the wet lab work (library preparation and sequencing of the samples). D.L., C.S., A.R., P.U., W.P., S.R., R.S., and J.C.M. contributed to the selection of accessions to be sequenced and contributed along with the expertise to facilitate the analysis of the biological features of the plants. O.E.C. contributed analyses. N.R.T. and P. N. contributed to the development of the SIFT database for annotation. O.E.C., C.D.B., and J.C.M. wrote the paper. All authors reviewed the paper.

**Additional information**

Supplementary information accompanies this paper at https://doi.org/10.1038/s42003-018-0168-6.

**Competing interests:** Co-authors Donald Livingstone III, Conrad Stack, Alberto Romero, Stefan Royaert, Osman Gutierrez, and Juan C. Motamayor are or have been employees of the funding agency MARS Inc. The authors declare no other competing financial or non-financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.