Supplementary Materials for

Genomic insights into the domestication of the chocolate tree, *Theobroma cacao* L.

Omar E. Cornejo[#], Muh-Ching Yee, Victor Dominguez, Mary Andrews, Alexandra Adams, Erika Strandberg, Donald Livingstone III, Conrad Stack, Pathmanathan Umaharan, Alberto Romero, Stefan Royaert, Nilesh R. Tawari, Pauline Ng, Ray Schnell, Wilberth Phillips, Keithanne Mockaitis, Carlos D. Bustamante[#], Juan C. Motamayor[#].

*Correspondence to: juan.motamayor@effem.com, cdbustam@stanford.edu

Supplementary Materials:

Materials and Methods

Figures S1-S41

Tables S1-S15

References (##-##)

1 800 386 1215 (5088)

Collection and DNA preparation for sequencing

We have constructed the first multi-population genomic variability dataset for cacao by analyzing the genomes of 200 accessions, selected from major collections of cacao. These samples represent a comprehensive collection from various geographical origins (genetic cluster), domestic accessions of common use in worldwide crops, and wild admixed individuals. Table S1 contains the Accession IDs and the Tree ID of collected trees (when available) as well as the sample source where the leaf samples were obtained. USDA approval was given for the import of leaf material for DNA extraction and library preparation.

Table S1 | List of accessions sequenced in this study. Sample source (collection of origin), Tree ID (when available), and research group are included.

Accession ID	*Sample source*	*Tree ID*	*Group*
2076	Ecuador		Indiana
2126	Ecuador		Indiana
2367	Ecuador		Indiana
2416	Ecuador		Indiana
2462	Ecuador		Indiana
2699	Ecuador		Indiana
2748	Ecuador		Indiana
AGU_3339_12	Trinidad	18276	Stanford
AMAZ-11_G18_A1	Ecuador		Indiana
AMAZ-11_G21_A10	Ecuador		Indiana
AMAZ-14_G23_A1	Ecuador		Indiana
AMAZ-14_G24_A5	Ecuador		Indiana
AMAZ_12	Trinidad	17443	Stanford
AMAZ_15_15	Trinidad	17421	Stanford
AM_1_54	Costa Rica		Stanford
AM_2_18	Costa Rica		Stanford
BE_10	Costa Rica	17422	Stanford
BR25	Indonesia (gDNA USDA)		Stanford

B_6_3	Trinidad	TC17477	Stanford
B_6_8	Trinidad	18277	Stanford
Brisas-1	(Ecuador) gDNA USDA	TC02491	Stanford
CAB_71_PL3	gDNA Miami	TC02565	Stanford
CAB_76_PL3	gDNA Miami	TC02567	Stanford
CAB_77_PL5	gDNA Miami	TC02568	Stanford
CATIE_1000	Costa Rica	17423	Stanford
CCAT1119_EET544_A1	Ecuador		Indiana
CCAT1858_EET547	Ecuador		Indiana
CCAT4675_EET575	Ecuador		Indiana
CCAT4688_EET576_A1	Ecuador		Indiana
CCAT4998_EET577	Ecuador		Indiana
CCN10	Costa Rica	17446	Stanford
CCN51	Brazil, MARS		Indiana
CC_71	Costa Rica		Stanford
CL_27_50	Trinidad	TC17482	Stanford
COCA_3370_5	Costa Rica	TC04611	Stanford
CRUZ_7_14	Costa Rica		Stanford
CRU_101	Trinidad	TC17492	Stanford
CRU_59	Trinidad		Stanford
CRU_89	Trinidad	TC17488	Stanford
CS_146	Costa Rica	11453	Stanford
CUR3_G35_A1	Ecuador		Indiana
CUR3_G37_A6	Ecuador		Indiana
CUR3_G38_A8	Ecuador		Indiana
CUR3_G39_A10	Ecuador		Indiana
Catongo	Costa Rica	17445	Stanford
EB-19-1S	(Ecuador) gDNA USDA		Stanford
EB2237_A1	Ecuador		Indiana

EET103_Borde	Ecuador	17424?	Indiana
EET397	Costa Rica	17447	Stanford
EET400_Arbol-122	Ecuador	Tc21291	Indiana
EET446_A1	Ecuador		Indiana
EET451_A1	Ecuador		Indiana
EET462_A1	Ecuador		Indiana
EET58_A1	Ecuador		Indiana
EET95_A2	Ecuador		Indiana
EET_103	Ecuador	17424	Indiana
EET_395	Trinidad	18186	Stanford
EET_400	Costa Rica	TC17439	Stanford
EET_58	Trinidad	TC17727	Stanford
EET_59	Trinidad	18302	Stanford
ELP_20_A	Costa Rica	TC02198	Stanford
FSC_13	Costa Rica	18315	Stanford
GU255_V	Costa Rica	17425	Stanford
GU_114_P	Trinidad	Tc17459	Stanford
GU_175_P	Trinidad	Tc17460	Stanford
GU_291_F	Trinidad		Stanford
GU_222	Trinidad		Stanford
GU_300_P	Trinidad	Tc17463	Stanford
GU_307	Costa Rica	17426	Stanford
GU_308A	Costa Rica	TC02175	Stanford
ICA_70	Trinidad	18321	Stanford
ICS39	Costa Rica	17448	Stanford
ICS40	Costa Rica	17449	Stanford
ICS_1	leaf from Miami		Stanford
ICS_43	Trinidad	18331	Stanford
ICS_6	Trinidad	TC00551	Stanford

ICS_60	Trinidad	Tc18242	Stanford
ICS_61	Trinidad		Stanford
ICS_95	gDNA Miami	Tc16546	Stanford
IMC67	Trinidad	TC17736	Stanford
IMC_12	Trinidad	18339	Stanford
IMC_14	Trinidad	17733	Stanford
IMC_20	Trinidad	TC00707	Stanford
IMC_36	Trinidad	TC00709	Stanford
IMC_47	Costa Rica	Tc16547	Stanford
IMC_50	Trinidad	TC00560	Stanford
IMC_51	Trinidad	TC00753	Stanford
JA_5_35	Trinidad		Stanford
JA_5_36	Trinidad		Stanford
JA_5_5	Trinidad		Stanford
K82	Papua New Guinea		Indiana
KA2	Papua New Guinea		Indiana
LCT46_A1	Ecuador		Indiana
LCTEEN_141	gDNA Miami	Tc01453	Stanford
LCT_EEN_46	Trinidad	TC01431	Stanford
LCT_EEN_83_S-8	Trinidad		Stanford
LP_3_40	Trinidad	18348	Stanford
LP_4_48	Trinidad		Stanford
LX_32	Costa Rica		Stanford
LX_43	Costa Rica		Stanford
M01	Indonesia (gDNA USDA)		Stanford
M02	Indonesia (gDNA USDA)		Stanford
M04	Indonesia (gDNA USDA)		Stanford
M05	Indonesia (gDNA USDA)		Stanford
M06	Indonesia (gDNA USDA)		Stanford

M07	Indonesia (gDNA USDA)		Stanford
MAN_15_2	Costa Rica	17428	Stanford
MATINA_Tica2	Ecuador		Indiana
MO_109	Trinidad	Tc18213	Stanford
MO_4	Trinidad	18362	Stanford
MO_9	Trinidad	18370	Stanford
MO_99	Trinidad	TC17508	Stanford
MXC_67	Trinidad	18391	Stanford
M_8	Trinidad		Stanford
Matina	Costa Rica		Indiana
Mocorongo	Costa Rica	17429	Stanford
NA45	Costa Rica	TC00602	Stanford
NA702	Trinidad	TC0930	Stanford
NA_286	Trinidad	18395	Stanford
NA_331	Trinidad	Tc00923	Stanford
NA_33_T2	Trinidad	TC00797	Stanford
NA_712	Trinidad	TC0629	Stanford
NA_92	Trinidad	TC00657	Stanford
NH_40	Bolivia, gDNA from Miami	6451	Stanford
NH_53	Bolivia, gDNA from Miami	6464	Stanford
OC_61	Trinidad	Tc17475	Stanford
PA107_A1	Trinidad		Stanford
PA289	Trinidad	TC00511	Stanford
PA_107	Trinidad	18407	Stanford
PA_120	Trinidad	Tc18221	Stanford
PA_121	Costa Rica, gDNA from Miami	TC00955	Stanford
PA_13	gDNA Miami	Tc18218	Stanford
PA_137	Trinidad	TC15958	Stanford
PA_150	Trinidad	TC00501	Stanford

PA_169	Trinidad	TC15974	Stanford
PA_218	Trinidad	18192	Stanford
PA_51	Miami	Tc00686	Stanford
PA_56	Trinidad	18396	Stanford
PA_70	gDNA Miami	TC15954	Stanford
PA_88	gDNA Miami	TC00983	Stanford
PBC123	Indonesia (gDNA USDA)		Stanford
PMF_20	gDNA Miami	Tc11280	Stanford
PMF_27	gDNA Miami	11287	Stanford
POUND_10_B	Trinidad	TC00858	Stanford
POUND_7_B	Trinidad	18419	Stanford
PlayaAlta1	Trinidad	Tc16545	Stanford
Pound_7	Costa Rica		Stanford
RB39PL1	Costa Rica	TC02522	Stanford
RB_40	Costa Rica	TC00449	Stanford
RB_47_PL3	Costa Rica	TC02518	Stanford
REDAMEL_1_27	Trinidad	18180	Stanford
REDAMEL_1_31	Trinidad		Stanford
*SCA6***	gDNA Miami	Tc16548	Stanford
SCA_10	Trinidad	TC00984	Stanford
SCA_11	Trinidad	TC00882	Stanford
SCA_19	Costa Rica	TC00522	Stanford
SCA_24.2	Costa Rica	TC00523	Stanford
SCA_5	gDNA Miami	TC00884	Stanford
SC_1	Costa Rica		Stanford
SIAL169	Costa Rica	TC00179	Stanford
SIAL70	Costa Rica	TC00185	Stanford
SIAL84	Costa Rica	TC00186	Stanford
SIC806	Costa Rica	TC00200	Stanford

SIL-1_G56_A6	Costa Rica		Stanford
SJ_2_22	Trinidad	18442	Stanford
SLC_4	Trinidad	TC17513	Stanford
SNA0707	Costa Rica		Stanford
SPA_7	Costa Rica		Stanford
SPEC_194_75	Trinidad	Tc18235	Stanford
SPEC_54_1	Costa Rica	TC05194	Stanford
T675_A645_A1	Ecuador		Stanford
T678_B60_A1	Ecuador		Stanford
T680_A1	Ecuador		Stanford
T682_A1_D147	Ecuador		Stanford
T684_EET233_A1	Ecuador		Stanford
T685_EET387_A1	Ecuador		Stanford
T686_LCT-368_A1	Ecuador		Stanford
T695_SCA-6_A1	Ecuador		Stanford
TAP10_G12_A1	Ecuador		Stanford
TAP3_G70_A2	Ecuador		Stanford
TAP6_G12_A1	Ecuador		Stanford
TIP-1_G41_A1	Ecuador		Stanford
TRD86	Trinidad		Stanford
TRD_45	Trinidad	18443	Stanford
TSA654Zymo	Costa Rica	17457	Stanford
TSH1188	Brazil, MARS		Stanford
TSH516	Costa Rica	17455	Stanford
UF12	Costa Rica	17434	Stanford
UF273_T1	Costa Rica		Stanford
UF273_T2	Costa Rica		Stanford
UF_11	Costa Rica	18446	Stanford
UF_668	Costa Rica	17456	Stanford

UF_676	Costa Rica	TC13037	Stanford
UNAP2_G78_A2	Ecuador		Stanford
criollo	Costa Rica		Indiana
mvP30	Indonesia		Indiana
mvT85	Indonesia		Indiana
sp1	Venezuela (gDNA USDA)		Stanford
sp3	Venezuela (gDNA USDA)		Stanford
sp9	Venezuela (gDNA USDA)		Stanford

Sample marked with a ** is an offset from what SCA6 should be. It resulted in an admixed individual and researchers interested in looking at SCA6 should not use this accession as a representative sample from SCA6 (admixture analysis showed this is a hybrid).

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 grams of leaf material per extractions per sample were cut and placed in homogenization tubes with ceramic pearls and lysis buffer. Samples where homogenized in a FastPrep-24™ (MP Biomedicals, LLC) placed in a cold room at 4 C for 60 seconds at a speed of 4.5 m/sec. If the tissue was not homogenized throuroughly, the tissues were homogenized for an additional 20 – 40 seconds 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 the extraction was assessed with 2% E-Gel (Invitrogen, Carlsbad, CA). Most of the samples were prepared using Nextera DNA Sample Preparation Kits (Epicentre, Chicago, IL, USA) and NEBnext® Ultra DNA Library Prep Kit for Illumina (New England BioLabs, Inc). The remaining samples were prepared by first shearing 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 (HMW) DNA from cacao leaves. Approximately 450 milligrams of leaf material per sample was ground to powder under liquid N_2 using mortar and pestle. Tissue powder was homogenized and washed twice by vortexing in 3 ml of ice cold 100 mM HEPES, 0.1% PVP-40, 4% b-mercaptoethanol followed by centrifugation at 7000 rpm in an Eppendorf F35-6-30 rotor. Nuclei were extracted from tissue pellets on ice in 50 mM

Tris-Cl pH 8.0, 50 mM EDTA, 50 mM NaCl with 15% sucrose, and centrifuged at 3600 rpm to pellet trace 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 NH_4OAc 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 HMW fraction (24 to >/=60 Kb) was assessed by migration on Genomic DNA Screen Tape, Agilent TapeStation 2200 Software (A.01.04) (Agilent) and secondarily quantified by fluorimetry 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), and minimally-amplified libraries were prepared using the TruSeq DNA Sample Prep Kit (Illumina) with 4 cycles of PCR, at the Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign (UIUC). All library preparation steps were according 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), quantitated by qPCR, pooled in sets of 12 at equimolar concentration, and sequenced as paired 2 x 161 nt reads on a UIUC HiSeq2500 instrument using HiSeq SBS sequencing kit version 4. Fastq files were generated with Casava1.8.2

Read processing and single nucleotide polymorphism identification

The Illumina data was basecalled using Illumina software CASAVA 1.8.2 and sequences were de-multiplexed with a requirement of full match of the 6 nucleotide index that was used for library preparation. Samples prepared using Nextera were hard clipped 13 nucleotides from the 5' end. Following demultiplexing, raw sequenced data was analyzed for quality using FastQC[1]. We performed adaptive quality trimming (setting a quality threshold of 25), and additional hard trimming of the reads based on stabilization of the base composition on the 5' end of the sequences using TrimGalore! and cutadapt [2,3]. Sets of reads from individual samples where mapped to the Matina-v1.1 reference genome[4], using the burrow-wheeler aligner bwa[5], 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 preprocessed prior to performing SNP identification with samtools/Picard Tools and Bamtools[6-8], to mark duplicates, fix mate pair information, correct unmapped reads flags and obtain overall mapping statistics. We followed recommendations of the Genome Analysis Toolkit to perform base quality recalibration and local realignment to minimize false positives during the SNP calling procedure[9]. Finally, we performed genotype calling using Real Time Genomics population analysis tool to speed the process of SNP identification[10]. Calls were also called with GATK and a suitable subset of single nucleotide polymorphisms were kept after a combination of Variant Quality Score Recalibration (VQSR) and hard filters that included thresholds in coverage (maximum coverage = 200*50X), quality by depth (QD 2) estimated from the division of variant confidence by unfiltered depth of non-reference samples, fisher strand test (FS 50), and the root mean

square of the mapping quality across samples (MQ 30). Variants identified were phased using shapeit v2.12 [11,12]. The phasing was performed per chromosome for the 10 main chromosomes using only bi-allelic sites.

Identified single nucleotide polymorphisms were annotated using SNPEff[13]. For this, we used the current gene annotation from the Matina-v1.1 reference genome[4] to construct a new database for *Theobroma cacao*. This database was used to annotate the observed polymorphisms following their potential effect on gene expression according to their position with respect to the coding regions. The number of changes belonging to the main 16 categories is presented in Table S2 and Figure S1. As discussed in the main text, synonymous and especially missense variants are relevant because they were used to annotate the potential impact of these mutations in protein function. Splice acceptor (Splice_acceptor) and splice donor (Splice_donor) variants are relevant to potentially understand the underlying polymorphism in the variation for the number of differentially spliced transcripts in the species and among populations. The small overall number of start lost variants is also indicative of high conservation in the expression among genes or the changes in the use of potential codons for the start of translation of protein products. Perhaps even more interesting is the relatively higher number of stop gains observed in the analysis. Most of the stop gains (> 60%) seem to be located towards the end of the genes, which suggest that their effect in preventing the appropriate generation of proteins is rather limited and their negative effects will be minimal on average.

It is not surprising that most polymorphism is found in intergenic regions or generally annotated as upstream or downstrem of genes. After non-coding variants found in intergenic regions, SNPs found in intronic regions are the most prevalent. Changes in intronic regions can often be neutral, from a functional perspective, but could contribute to differential lifespan of mRNAs, differential processing of the immature RNA including regulation of nonsense-mediated decay[14]. There is an increase intron variant research because it has long been recognized that introns might be involved in mRNA transport or chromatin assembly[15,16]; and we expect that the variants indentified in the genome will facilitate further work involving post-transcriptional regulation of expression.

Table S2 | Annotation of polymorphic sites in the cacao genome.

Type of change	*scaffold_1*	*scaffold_2*	*scaffold_3*	*scaffold_4*	*scaffold_5*	*scaffold_6*	*scaffold_7*	*scaffold_8*	*scaffold_9*	*scaffold_10*	*Total*
splice_acceptor_variant	1137	1091	1077	1069	1155	1036	1041	941	1109	1006	10662
splice_donor_variant	1041	1065	1018	1011	1030	966	995	925	1034	977	10062
start_lost	881	857	848	870	870	819	836	806	862	821	8470
stop_gained	1852	1819	1676	1759	1950	1470	1727	1353	1704	1646	16956
stop_lost	888	865	870	862	888	853	828	829	871	834	8588
missense_variant	40756	36789	32090	35310	39172	27462	27189	21986	35056	26465	322275
*reg_region_ablation**	757	757	757	757	757	757	757	757	757	757	7570
splice_region_variant	6664	5877	5253	5278	5695	4573	3731	3707	5920	3863	50561
stop_retained_variant	826	813	800	807	815	806	795	789	804	794	8049

synonymous_variant	28976	25397	22197	23849	26200	19084	16520	15803	25544	16473	220043
3_prime_UTR_variant	35039	31120	27554	27252	29125	21817	17428	18657	31491	18345	257828
5_prime_UTR_variant	23013	19428	17491	17385	18449	14067	10813	11713	19689	11506	163554
*downstream_g_variant***	132861	132687	112431	112718	136761	91290	92960	72961	137247	96219	1118135
intergenic_region	301809	446024	298491	298152	434813	230026	234133	167223	406066	277742	3094479
intron_variant	92905	113223	105025	94243	119196	101995	122671	100143	116546	100909	1066856
upstream_gene_variant	321800	296212	243510	248306	286108	195094	170662	159877	297913	184225	2403707

* regulatory region ablation. ** downstream gene variant

A graphical representation of the number of SNPs per functional impact category and per chromosome (Figure S1).

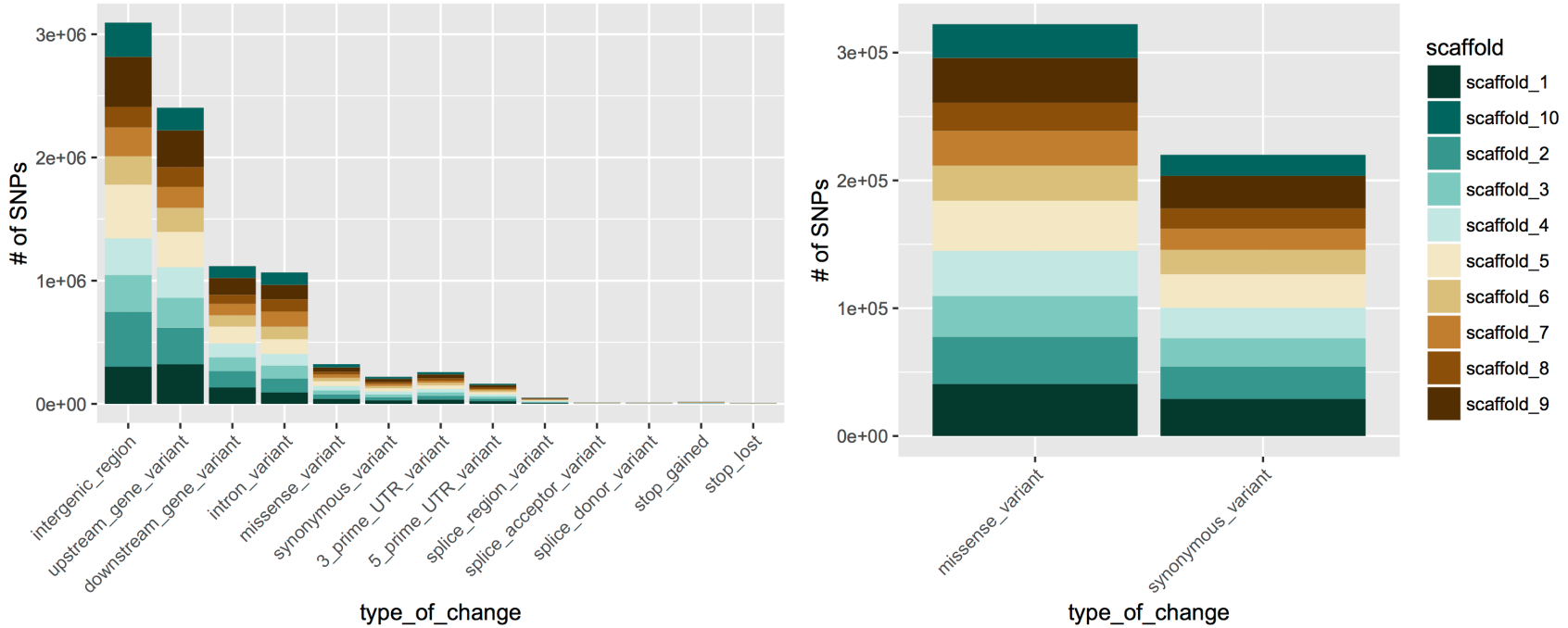

Figure S1 | Number of single nucleotide polymorphisms categorized by functional impact in transcript variation per chromosome. Right panel presents detail of the comparative number of synonymous and non synonymous mutations.

Distribution of genetic variation among genetic groups.

We characterize the distribution of genetic variation in the populations estimating variation using two approximations for the inference of genetic variation: Watterson's theta (θ_w)[17] and the number of pairwise differences per site (π)[18]. We used vcftools[19] to estimate both statistics in windows of 1 kilobase (Kb). Under a standard neutral model (with no changes in effective population size, drift and mutation balance) both statistics should converge to the same measurement (the basis for Tajima's D [20]). Overall, there is agreement between the two measurements of genetic diversity, with a slight underestimation of genetic polymorphism under θ_w (Figure S2). The orthogonal comparison between our SNP data and a 6K SNPs chip specifically designed for cacao on selected accessions (CCN51, TSH1188, Pound7 and UF273_T1/T2) was 99.9% concordant.

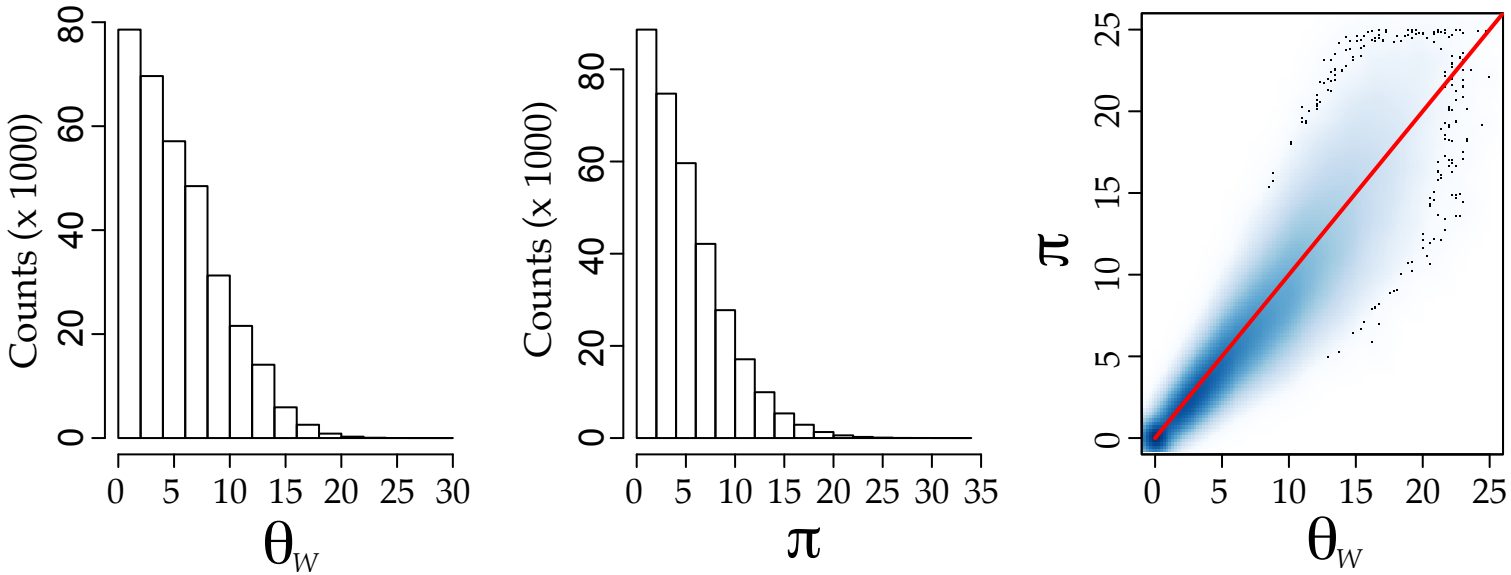

Figure S2 | Left and center panels present the distribution of Watteron's θ_w per Kb, and π per KB respectively. Right hand panel presents the scatterplot of Watterson's θ_w, and π; ther red line corresponds to the 1:1 relationship between the two.

For convenience, we present the assessment of the distribution of genetic diversity genome-wide per population and across the genome per population, using the number of pairwise differences per site (π). Our analyses reveal that there are remarkable differences in the magnitude of genetic variation among populations of cacao. We used a generalized linear model compare the genetic diversity among groups. Our Generalized linear model assumed a Gaussian family against the log value of the genetic diversity using a model of the form: $\log(Y) = \beta_0 + \beta_i + \epsilon$ where *i* corresponds to the population or genetic group (Amelonado, Admixed, Contamanta… etc).

log.glm <- glm(log(PI) ~ Group, family=gaussian, data=data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-5.8607	-0.6223	0.0838	0.6870	3.7276

Our results clearly show that belonging to a genetic group modifies considerably the expectation of the levels of genetic diversity observed in a sample, when compared to the Amelonado population (Table S3). We used Amelonado to compare against as it presents small levels of variation overall and because the reference genome employed (Matina) is an individual of Amelonado ancestry. It is remarkable that Criollo presents the largest impact towards reducing the expected genetic diversity. This is consistent with our results that Criollo populations present a very reduced effective population size, probably the result of a very strong and relatively recent domestication event. Not surprisingly, admixed individuals present the largest positive effect towards the increase on genetic diversity, but this effect is very like that observed of the estimates for the Contamana group which, to the effects of this work, is considered a wild population.

Table S3 | Coefficients GLM model adjusted to explain the differences in genetic diversity by group.

Coefficients	**Estimate**	**Std._Error**	**t_value**	**Pr(>|t|)**
(Intercept)	-7.457655	0.002082	-3582.4	<2e-16***
Contamana	1.544563	0.002707	570.6	<2e-16***
Criollo	0.189763	0.003874	48.99	<2e-16***
Curaray	1.427195	0.002825	505.26	<2e-16***
Guianna	0.572881	0.002931	195.5	<2e-16***
Iquitos	1.317511	0.002728	482.9	<2e-16***
Maranon	1.0866	0.002732	397.8	<2e-16***

Nacional	1.28207	0.002754	465.6	<2e-16***
Nanay	0.537165	0.002784	192.9	<2e-16***
Purus	1.210039	0.002731	443	<2e-16***
Admixed	1.614775	0.002774	582.1	<2e-16***

The relative impact that overall ancestry has on genetic diversity is partially explained by the differences in the demographic history inferred for each population (see main text and below). Because *T. cacao* presents a self-incompatibility system, and some individuals can self-fertilize and others are obligate outcrossers, then the differential proportion of self-compatible individuals among populations will strongly impact the magnitude of genetic diversity maintained in the population[21,22].

Our analyses also reveal that the distribution of genetic diversity along the genome of *T. cacao* presents a heavy tail, similar to what has been observed in other organisms[23,24] (Figure S3). Differences in the distribution of genetic polymorphism along the genome have been interpreted, under population genetic premises and appropriate models, as corresponding differences in the effective population size along the genome[23-26], which could then affect the rate of adaptive molecular evolution in eukaryotes[25]. Areas of the genome with a larger effective population size could be more prone to fast adaptive evolution from standing variation than other regions of the genome and the differential distribution of genetic variation along the chromosomes in different genetic groups suggests that *T. cacao* harbors a large potential for adaptation from standing variation (Figure S3, S4). It can be seen in Figure S4 that there are regions of the genome with considerably more variation than others, a pattern a requires further investigation. In the context of management of a domesticated species, this implies that potentially different regions of the genome are more amenable of artificial selection than others in different populations of *T. cacao*. Our analysis reveals that the process of domestication of the Criollo variety in Mesoamerica was the result of a single event, with no evidence of recent gene flow to this group from other genetic clusters. Also, we find no additional signatures for domestication as strong as the one found for the Criollo group in the rest of the genetic clusters. We show for the first time that the process of differentiation of the genetic clusters presents a complex pattern of historical admixture. Our analysis of genetic variation on the admixed individuals reveals that, despite the large number of well-differentiated populations in cacao, only a few ancestry components can be found in admixed individuals, suggesting there is a large amount of untapped genetic variation in the species. We expect that the resources we have generated will help improve cacao crops and that this contribution will have important repercussions on the economy of producing countries. Our analysis also shows that genetic diversity is not uniform across loci. Genetic diversity presents a distribution with a long tail suggestive of areas of the genome with unusually high polymorphism, as has been described for other species (see supplementary figure S1[25,26]). Moreover, genetic diversity is not uniformly distributed across genetic groups (see supplementary Figure S2. The difference in overall genetic diversity across groups is likely due to the combined contribution of differences in effective population size and demographic histories, as well as differences in selfing rates across groups, as *T. cacao* presents both self-compatible and self-incompatible individuals within the species [27,28].

The difference in overall genetic diversity across groups is likely due to the combined contribution of differences in effective population size and demographic histories, with Criollos showing the lowest genetic diversity (π=0.27%) and Contamanas, Nacional, and Admixed individuals presenting the highest diversity (π=0.32%, π=0.31%, π=0.37%, respectively, supplementary Figure S2). We also identify a clear pattern of high heterogeneity in the distribution of genetic diversity along the genome suggesting differences in effective population size along the genome potentially driven by artificial and natural selection (supplementary figures S3, S4).

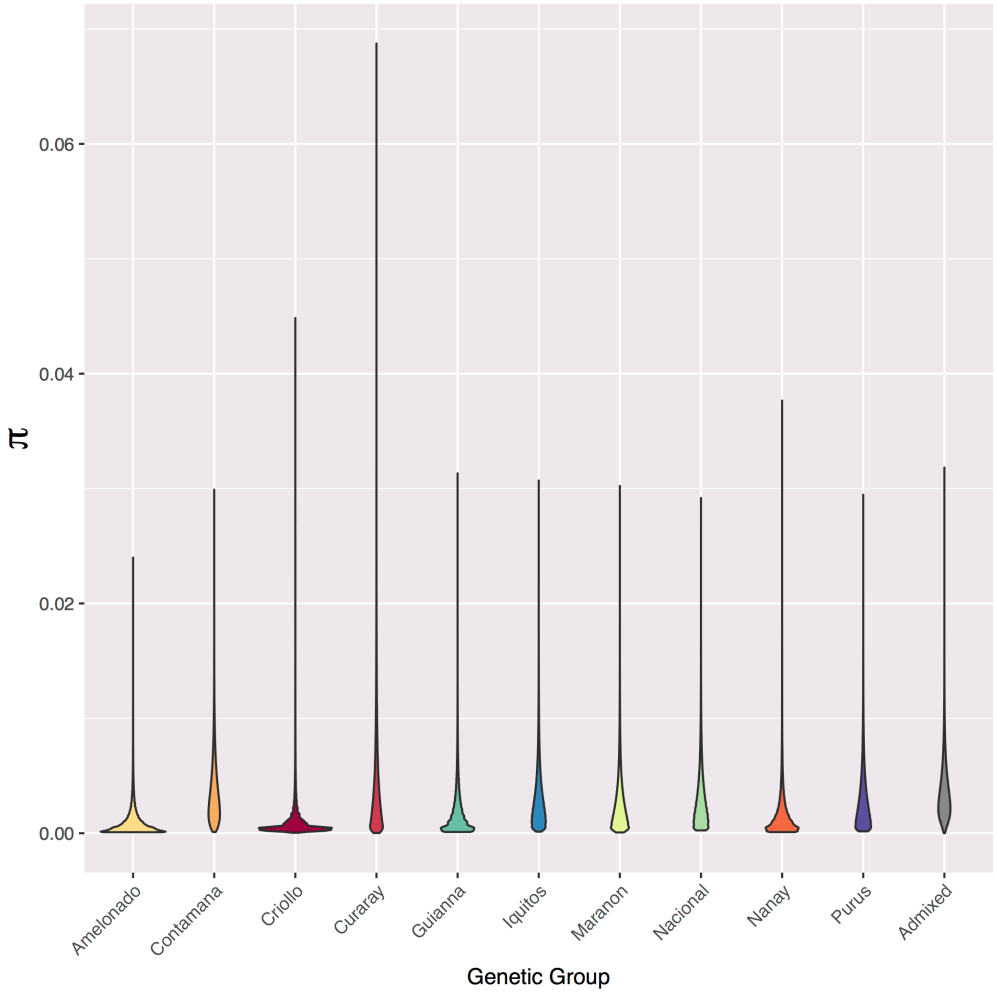

Figure S3 | Distribution of genetic diversity (measured as π), represented as violin plots. Differences in overall genetic diversity among groups is significant (see model fitting in text).

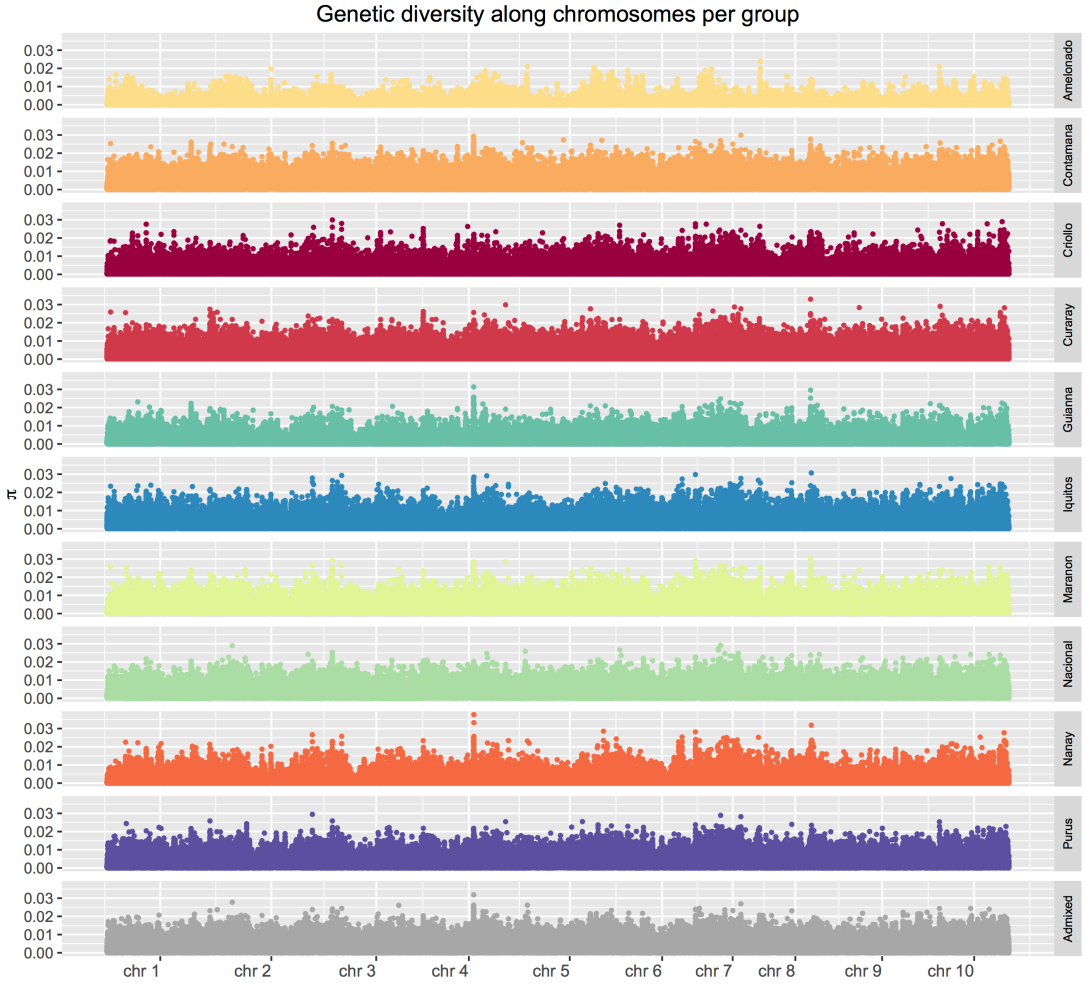

Figure S4 | Distribution of genetic diversity (measured as π per base) along the genome for each genetic group (including admixed individuals).

***Population Structure in* Theobroma cacao**

We used a *ADMIXTURE*[29], an implementation of an approach similar to well known *STRUCTURE*[30]. Based on an expectation-maximization algorithm, *ADMIXTURE* uses a maximum likelihood based approach to assign ancestry genome-wide and to 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 data[29]. We filtered our data and restricted our analysis to SNPs with minor allele frequency over 5% and we also pruned the data for LD as the approximations assumes unlinked loci. For this, we used vcftools[19] to estimate LD (r2) scores for each pair of SNPs in windows of 2000 SNPs and excluded one of the pair if r2 > 0.45. The windows were selected with 500 SNPs of overlap. The final dataset contained 63,374 SNPs. We analyzed this dataset using *ADMIXTURE* and set 2 to 18 "ancestral" populations (*K*=2 to *K*=18) in 100 replicates. We checked for convergence of individual *ADMIXTURE* runs at each *K* by evaluating the maximum difference in log likelihood (LL) scores in fractions of runs with the highest LL scores at each *K*. We assume that a global LL maximum was reached at a given *K* if at least 10% of the runs with the highest LL score show minimal variation in LL scores and present consistent assignment to the groups. It has been shown[31] that a threshold of 5 LL units is conservative enough to assure similar results to those obtained with *CLUMPP*[32]. Following this approximation, we concluded that the global LL maximum was reached in runs at *K*=2 to *K*=15 (at least). *ADMIXTURE* includes a cross-validation (CV) procedure to help choose the "best" *K*, which is defined as the *K* that has the best predictive accuracy. Our analysis suggests that using cross-validation it is not possible to distinguish between K=10, K=11 and K=12. Although K=10 and K=11 seem to be more likely (better likelihood scores examined using Akaike information criteria).

Initial genetic analyses, with microsatellites markers, have uncovered a large number of genetic groups and clear differentiation between the trees found in the Amazon basin and the Criollo varieties found in Central America[33]. This work helped characterize cacao germplasm into 10 major genetically differentiated groups: Amelonado, Contamana, Criollo, Curaray, Guianna, Iquitos, Marañon, Nacional, Nanay and Purús[33]. 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[33]. Yet, there is a huge gap in our understanding of genomic variation in the species which makes it difficult to propose clear scenarios for the evolution of natural populations and the domestication of *T. cacao* and how this might be exploited by the agronomist for crop improvement and sustainability. The assignment of ancestries to K=10 is easy to interpret using previous work that has characterized the genetic variation of *T. cacao* with microsatellite markers and proposed 10 main populations or genetic groups to explain genetic differentiation in the species[33]. The assignment based on K=10 or K=11 produce overall similar results. Yet, K=11 reveals further population structure (which has been observed in previous analyses) which could be particulary important to understand the genetic ancestry of the hybrids (wild and domesticated). Based on our analyses, we designated five arbitrary groups of admixed individuals based on the major contribution of genetic clusters to mixed ancestry (Figure 5A, see also

Figure 1A of the main manuscript, left to right). The identification of individual samples and their assigned ancestry is provided in Figure S5. Their position of individuals in the plot is the same shown in the figure in the main document. Group I (horizontal red bar above plot) is characterized by a declining gradient in the contribution of Criollo ancestry, with a large majority of individuals presenting Criollo/Amelonado ancestry and a wide number of accessions presenting complex patterns of admixture that includes contributions from Nacional, Iquitos and Purus. Group II (horizontal yellow bar) is defined by a gradient of Curaray ancestry with major contributions from Nacional and Amelonado. Group III (horizontal green bar) is defined by a gradient of Nanay ancestry with major contributions from Iquito ancestry. Group IV (horizontal blue bar) is defined by a gradient in Contamana ancestry and presents major contribution from Iquitos and Nacional. Group V (horizontal purple bar) is defined by a relatively equivalent contribution of Amelonado and Nacional ancestry. There is evidence for additional substructure among these groups, not previously identified but consistent with observations of other genetic groups[34,35]. More specifically, we identify an additional component of ancestry that results from the decomposition of the Amelonado group into two clusters of ancestry. We believe that this second component of Amelonado ancestry is real, even though only a single individual can be fully assigned to it, because it is found to be the major contributor to the ancestry of Group V of admixed individuals. These results are noteworthy because most cultivated varieties seem to contain a large component of Criollo and Amelonado ancestry which means there ample genetic potential in this critically important crop for the plant breeder to exploit.

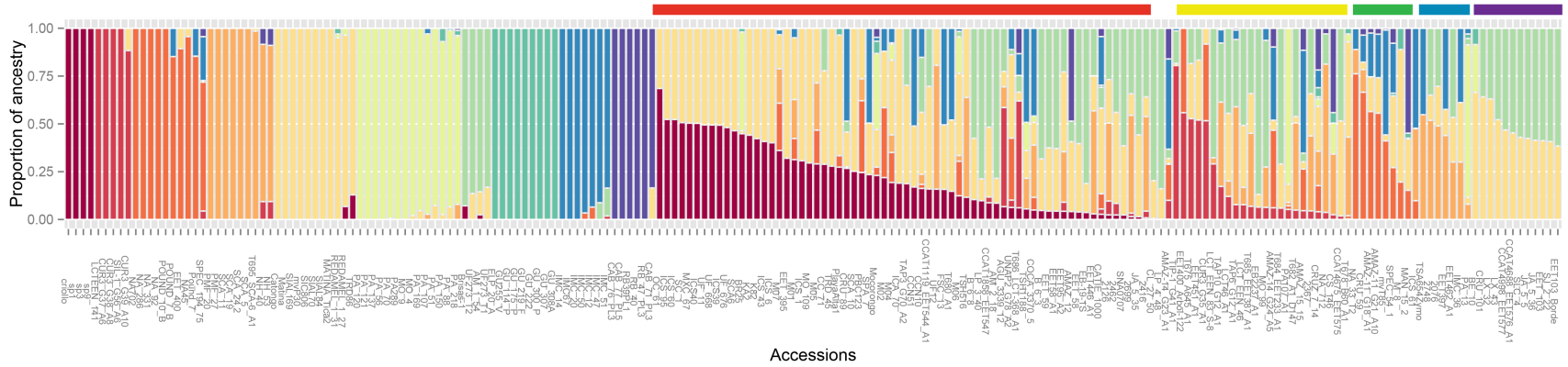

Figure S5 | Ancestry assignment (K=10) for all admixed individuals. The order of the groups corresponds to the assignment in Figure 1 of the main manuscript. From left to right, the colors correspond to the following groups: Criollo (dark red), Curaray (red), Nanay (dark orange), Contamana (orange), Amelonado (light orange), Marañon (light green), Nacional (green), Guianna (dark green), Iquitos (blue) and Purus (purple). The color bars on top of the admixed individuals correspond to each of the arbitratry groups defined for cacao on this work (see supplementary text) to help the work of breeders.

Although the discriminative analyses performed with *ADMIXTURE* provide a good approximation to identify the underlying number of popuations (or components of ancestry in a population of admixed individuals), it does not provide an intuitive way to interpret the relatedness among populations. In order to gain a better understanding of the population structure in *T. cacao* we performed a multimidensional scaling analysis on the same set of SNPs employed for the *ADMIXTURE* analysis. First we normalized the data (centered and standardize) following previous recommendations [36] and performed MDS analyses. We recapitulated the separation of groups observed in the *ADMIXTURE* analysis, and the MDS analysis added information about the relative differentiation between populations and the contributions of the different groups to the admixed individuals (See Figure 1 of main manuscript).

Admixture and MDS analyses provide a good graphical representation of the genetic structure in *T. cacao*. We measured population differentiation resulting from restrictions in gene flow between populations using Weir and Cockerham's F_{st} estimator[37] in windows of 5Kb, 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 comparisons (shown in Figure 1 of the manuscript). In addition to the overall analysis of population differentiation using F_{st}s, we show that there are regions of the genome that are more differentiated when all pairs of populations are compared to each other. This pattern of differentiation suggests that not only genetic diversity has distribution with long tails, but also the genetic divergence between populations. We exploit this feature in the analysis of selection where we show that some regions of the genome present significant differences in the local two-dimensional site frequency spectrum when compared to the genome-wide site frequency spectrum described by the demographics. In Figures S6 – S15 we represent the pairwise Fst along the genome for each population against the rest of the populations (S6: Amelonado, S7: Contamana, S8: Criollo, S9: Curaray, S10: Guianna, S11: Iquitos, S12: Maranon, S13: Nacional, S14: Nanay and S15: Purus). The grey line in each Figure represents the median F_{st} and the top red line the upper 95 % confidence interval. This is a pattern we examined in more detail between the separation of the Criollo and Curaray populations in the context of domestication.

Theobroma cacao *differentiation along the West to East axis in the Amazon basin.*

We fitted a model (described in Figure 1C of the main manuscript) to explain the differences in genetic diversity along the Pacific/Atlantic axis 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 Figure 1B (main manuscript). These centroids were used as predictors (β_i) to explain the differences in mean genetic diversity per population (measured as π, Y in the folllwing model) under a simple linear model with a Gaussian family ($Y = \beta_0 + \beta_i + \epsilon$). Admixed individuals were excluded from the analysis.

Our analysis shows a significant association between geographic location (as described by genetic differentiation, table S4) and genetic diversity, with larger genetic diversity available in groups closer to the Pacific end of the Amazon Basin (negative PC2 values) and a progressive reduction in genetic diversity towards the Atlantic.

Coefficients:
(Intercept) β_i
0.2015 -72.7140

Table S4 | Analysis of Variance Table

Response:- PC2	**Df**	**Sum_Sq**	**Mean_Sq**	**F_value**	**Pr(>F)**
pi	1	0.036303	0.036303	9.3134	0.01578*
Residuals	8	0.031183	0.003898		

Model-based analysis of population differentiation

We used a model-based approach to infer the population relationships between the 10 main groups as implemented in TreeMix[38]. This program allows to estimate the evolutionary history of populations by modelling how the share genetic variation and drift plays a role in determining the genetic relationships between populations. It allows to explicitly model how genetic variants along the genome drift and extend those models to explicitly include migration and how it contributes to the drift genetic components. For the analyses with TreeMix we used only intergenic regions. We used our annotation of the reference Matina genome to create bed files with intervals corresponding to the intergenic regions of the genome and extracted SNPs in these regions for our estimations. Bed files can be made available upon request. Two important results from this analysis are: i) that the domesticated Criollo populations have undergone a large amount of drift (larger than any other population) and ii) that all of the evidence of migration and admixture suggest that no additional contribution of any group has occurred after the domestication of *T. cacao* in Mesoamerica. The strongest evidence indicates a recent contribution of Iquitos to Nanay (red arrow in Figure 2B), which is consistent with the partial ancestry of Iquitos identified in some of the individuals belonging to the Nanay group in the admixture analysis (Figure 1A). Ancient admixture analyses, in the form of pairwise f_3 statistics [39], confirm that Criollo and Curaray are significantly closer to one another than to any other group and no evidence of admixture can be found.

***5. Demographic history in* Theobroma cacao**

We can model the distribution of variation within genomes to provide insights about the history and demography of ancestral populations[40]. We used a method developed to infer the demographic history of populations using individual genomes, the pairwise sequentially Markovian coalescent (PSMC)[40], to characterize the changes in effective population size (N_e) of the ancestral populations. PSMC uses the distribution of heterozygote sites throughout the genome to estimate the time to the most recent common ancestor of a segment of sequence. For this, we first phased the genetic information of genomes belonging to each one of the 10 populations characterized in cacao using ShapeIT[11,12]. Then, individual genomes were used to infer changes in demographic history for each population. We then combined the inferred history from multiple individuals from the same population and estimated smoothed PSMC curves per populations (shown in main manuscript). The results across individuals were highly similar and smooth spline regression showed that all populations of cacao seemed to have undergone a population decline since Last Glacial Maximum. The Criollo populations have a much smaller population size but we have detected a similar trend in population reduction. For our estimation of the population sizes, we assumed that mutation rate for *Theobroma cacao* followed typical mutation rates estimated in other plants (Arabidopsis) [41,42] of 7.1 x 10^{-9} mutations per base pair per generation. We also assumed a generation time (the time that it takes to go from seed to seed) as 5 years[27]. Although the general trend is towards the loss of genetic diversity, two different dynamics are evident. First, the Curaray populations and, to a lesser extent the Iquitos and Purus populations, show signatures of an initial increase in their population size followed by a decline. This pattern that could be explained by admixture as it has been observed in other organisms[43]; as well as real population increases and decreases in time. Second, we observe a much more recent and far smaller overall population size for the Criollo group which is consistent with the idea of a strong domestication event in recent times from a relatively small pool of individuals (see Figure 2D).

Effects of historical Population Size on Inbreeding

We observe an increase in the amount of inbreeding (estimated as F statistics [44]) when the admixed cluster of individuals is compared to the naturally defined genetic groups (Figure 4A). The general trend shows an increase in inbreeding from Iquitos, Nacional, Curaray, Contamana, Marañon and Purus and even higher levels of inbreeding in Guianna, Criollo, Nanay, and Amelonado (Figure 4A). Yet, there are striking patterns. Amelonado presents a much higher level of inbreeding, given what would be expected under its historical demographics. *T. cacao* shows a unique self-incompatibility mating system where some individuals in the species are self-incompatible (SI) while some other are self-compatible (SC) [28]. A reduced number of accessions in the species have been characterized for SC/SI, and there has not been a thorough assessment of the distribution of SC/SI in most genetic groups, so that the overall frequency of SI/SC in the species is largely unknown. Despite this, there is field evidence suggesting that the Amelonado population presents a high frequency of SC individuals. Similarly, most plants in the Criollo group have also been described to be SC[27].

Detailed demographic analysis of Cacao dosmestication

To further understand the population demographic history during the process of domestication we used a different approximation and build a demographic model based on the observations from the results of the PSMC and TreeMix analysis. The results from TreeMix suggested that Criollo and Curaray are the most related populations and the results from PSMC showed that both populations have been declining over time. We use an approximation based on the comparison of the observed site frequency spectrum and simulations in a maximum likelihood framework to decide which model better explained the data, as implemented in the program δaδi[45]. We examined three alternative models: i) a simple model of isolation without migration (model A); ii) a model of isolation with migration (model B); and iii) a modified model of isolation with migration in which we allow ancestral population prior to the split to be changing in time and the populations post split to change in time (model C). For each model, we estimated the corresponding likelihood and compared the relative fit of the models using Akaike information criteria. The fitting to model A, isolation with no migration, was the worst model explain the data (LL = -15818.1, AIC =31642.2). The fitting to Model B, representing a simple isolation with migration model, was the second best fit (LL = -1251.61, AIC = 2513.22). The best fitted model is Model C, a change in the ancestral population size prior to split and also after the split (LL= -664.88, AIC=1345.76). The AIC values support Model C as the best fitting model. All estimations were performed masking the rare variants present as singletons in either populations or as doubletons in either or both populations. The reason for this, is that Criollo and Curary populations have a significant number of individuals able to self and selfing affects the coalescence by increasing the coalescence rate at the top of the genealogies[46,47] and we have observed via simulations that can strongly impact the frequency of rare variants (namely singletons and doubletons). The figures for fitted model C in the main manuscript have blank boxes in the 2D-sfs representing this.
Current dogma suggests cacao was introduced to Mesoamerica in Omec times from cacao varieties present in the Upper Amazon (Northern South America), the hotbed of diversity for the species[27,48]. Anthropological research, in particular, supports this view [27,49,50]. 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. According to this hypothesis, domestication of cacao would have occurred in South America and then spread to Central America and Mexico through Native American trade networks [51]. In addition to the interest in understanding the historical domestication of cacao, there is tremendous agronomist importance in assessing how development of land races and varietals as well as outcrossing to genetically diverse germplasm has shaped diversity in modern cacao crops [33,52]. The results from our models are consistent with the general idea that Cacao Criollo was domesticated in Mesoamerica, but more detailed information about the possible alternative routes will require additional genotyping of plants along the alternative spatial paths accompanied with appropriate analyses.

For the estimation of confidence intervals, we performed 1000 bootstraps of the observed dataset employed to perform estimations of the demographics with the site frequency spectrum. We re-estimated parameters under an isolation with migration model for each one of the bootstrapped datasets and we finally used the empirical cumulative distribution (ecd) for the parameters to estimate the 95%

confidence interval as those values that fell within the 0.025 and 0.975 quantiles of the empirical distribution. Figure S16 shows the ecd for the time since split estimation of the separation between Curaray and Criollo populations and Figure S17 shows the ecd for the fraction of the ancestral Curaray population that likely served as a seed for Criollo domestication. For our estimations we followed the same assumptions for the mutation rate (7.1×10^{-9} mutations per base pair per generation) [41,42] and generation time (5 years) [27]; although we relaxed the assumption of the generation time to accommodate the observation by breeders that in order to better understand how deviations from this assumption would affect our estimations of the time of divergence between the Curaray and Criollo populations. Assuming a generation time of 15 years, the expected time of Divergence between Curaray and Criollo would be set back to 10,861 years BP, and the confidence interval for the time of divergence between poulations would be 95% CI = 7444.1 – 32708.2 years BP. The estimated time under a longer generation time provides estimates of divergence between populations much older than our expectation for the domestication given the anthropological evidence that supports a timeline for the peopling of South America around 13,000 years BP and additional evidence that suggests that human settlements were able to develop major crops only 8,000 years BP. We further explored other possible generation times of 50 or 70 years per generation and maximum likelihood estimations of divergence are even more unlikely than those estimated under a 15 years per generation ($time_{50\ years\ gen}$ = 36,203 years BP, $time_{70\ years\ per\ gen}$ = 50,685 years BP).

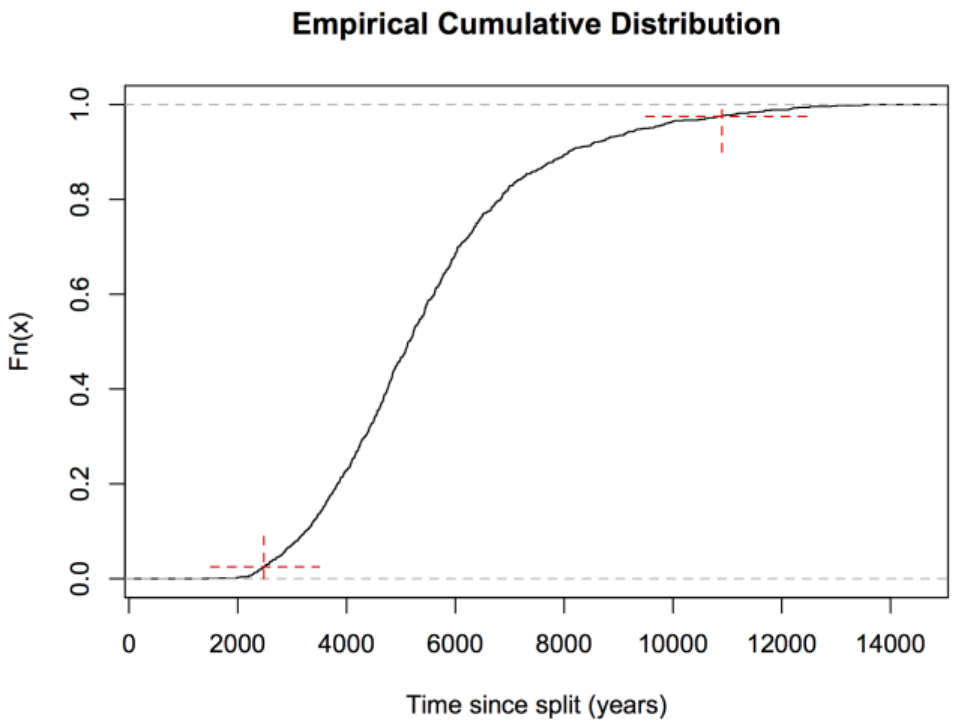

Figure S16 | Empirical cumulative distribution for the time since split of Curary and Criollo populations. Red crosses indicate the 95% confidence boundaries.

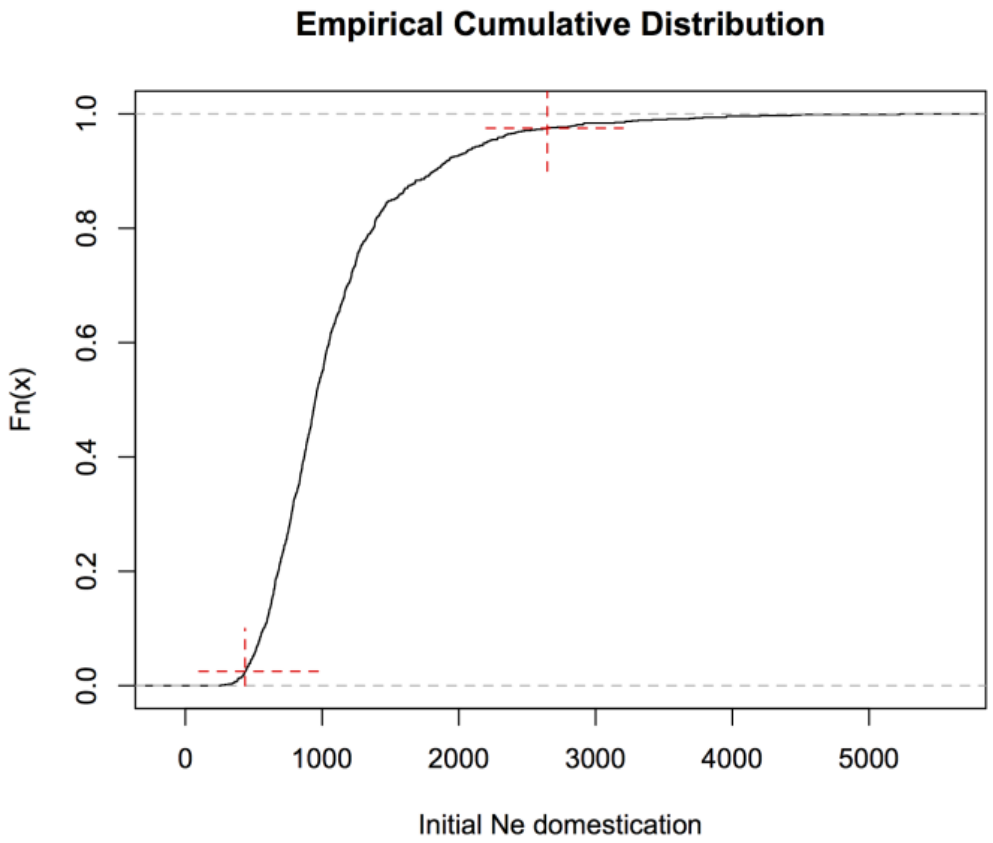

Figure S17 | Empirical cumulative distribution for the fraction of ancestral Curary population used for domestication of the Criollo variety. Red crosses indicate the 95% confidence boundaries.

Genes in genomic regions under selection

Analyses performed with XP-CLR[53] to detect local deviations from the genome-wide site frequency spectrum were performed assuming windows of 0.05 cM, for 200 SNPs and grid size of 2 Kb. For these analyses, we used the Curaray population as reference and took the top 1% windows with significant XPCLR score. We then intersected the windows in which significant signatures of selection were detected with the current annotation for the Matina reference genome to identify putative genes. Genes that overlapped with windows in which selection was detected are reported in Table S5.

Table S5 | Genes within regions of the genome identified under directional selection in Criollo, when compared to Curaray. Only genes found within windows that show significant signatures of selection at $p < 0.005$ were considered.

gene ID	*Predicted protein product*
Thecc1EG003330t1	Uncharacterized protein
Thecc1EG003331t1	Uncharacterized protein
Thecc1EG003333t1	Cysteine-rich RLK (RECEPTOR-like protein kinase) 8
Thecc1EG004046t1	Myb domain protein 13, putative
Thecc1EG020889t1	Myb domain protein 13, putative
Thecc1EG032927t2	Myb domain protein 58
Thecc1EG008253t1	Unknown
Thecc1EG001779t1	S-domain-2 5, putative
Thecc1EG014432t1	S-domain-2 5, putative
Thecc1EG014433t1	Signal peptide peptidase
Thecc1EG014433t2	Signal peptide peptidase
Thecc1EG014433t3	Signal peptide peptidase
Thecc1EG014912t1	Leucine-rich repeat receptor-like protein kinase family protein
Thecc1EG014913t1	Uncharacterized protein
Thecc1EG014914t1	Uncharacterized protein
Thecc1EG004030t2	Structural maintenance of chromosome 1 protein, putative (Genomic stability)
Thecc1EG014915t1	Structural maintenance of chromosomes (SMC) family protein

Thecc1EG014915t2	Structural maintenance of chromosomes 2
Thecc1EG041166t1	Structural maintenance of chromosomes (SMC) family protein
Thecc1EG004030t2	Structural maintenance of chromosome 1 protein, putative
Thecc1EG014915t1	Structural maintenance of chromosomes (SMC) family protein
Thecc1EG014915t2	Structural maintenance of chromosomes 2
Thecc1EG041166t1	Structural maintenance of chromosomes (SMC) family protein
Thecc1EG041166t2	Structural maintenance of chromosomes (SMC) family protein
Thecc1EG005256t1	WRKY DNA-binding protein 56
Thecc1EG008635t1	WRKY DNA-binding protein 75
Thecc1EG014673t1	WRKY-type DNA binding protein 1
Thecc1EG017248t1	WRKY DNA-binding protein 75
Thecc1EG019896t1	Uncharacterized protein
Thecc1EG019897t1	Ankyrin repeat family protein, putative
Thecc1EG019898t1	Uncharacterized protein
Thecc1EG019899t1	Uncharacterized protein
Thecc1EG019900t1	Uncharacterized protein
Thecc1EG019901t1	Transport, ribosome-binding, bacterial, putative
Thecc1EG019901t2	Transport, ribosome-binding, bacterial, putative
Thecc1EG019901t3	Transport, ribosome-binding, bacterial-like protein
Thecc1EG019902t1	Signal recognition particle 14 kDa protein
Thecc1EG024364t1	Uncharacterized protein
Thecc1EG024365t1	Uncharacterized protein
Thecc1EG024366t1	Uncharacterized protein
Thecc1EG024367t1	Uncharacterized protein
Thecc1EG030634t1	Uncharacterized protein
Thecc1EG030634t2	Uncharacterized protein
Thecc1EG030634t3	ARK-binding region
Thecc1EG030635t1	Nicotinate/nicotinamide mononucleotide adenyltransferase
Thecc1EG030635t2	Nicotinate/nicotinamide mononucleotide adenyltransferase

Thecc1EG030446t1	Nuclear transcription factor Y subunit B-6
Thecc1EG030636t1	Ccaat-binding transcription factor subunit A, putative
Thecc1EG030637t1	NAD-dependent deacetylase sirtuin-6 (Genomic stability)
Thecc1EG031781t1	Receptor like protein 53, putative
Thecc1EG031783t1	Receptor like protein 53, putative
Thecc1EG031801t1	Uncharacterized protein
Thecc1EG032107t1	Non-LTR retroelement reverse transcriptase, putative
Thecc1EG032108t1	Uncharacterized protein
Thecc1EG030240t1	Uncharacterized protein
Thecc1EG033489t1	Uncharacterized protein
Thecc1EG033490t1	Unknown
Thecc1EG035658t1	Uncharacterized protein
Thecc1EG035660t1	Uncharacterized protein
Thecc1EG036604t1	Secretory laccase, putative
Thecc1EG036604t2	Laccase/Diphenol oxidase family protein, putative
Thecc1EG036604t3	Laccase/Diphenol oxidase family protein, putative
Thecc1EG036608t1	Laccase 14, putative
Thecc1EG036608t1	Laccase 14, putative
Thecc1EG040227t1	Transferases, transferring hexosyl groups
Thecc1EG040227t2	Transferases, transferring hexosyl groups, putative
Thecc1EG040227t3	GPI mannosyltransferase 2
Thecc1EG040657t1	Uncharacterized protein
Thecc1EG040658t1	Xanthine dehydrogenase 1
Thecc1EG040657t1	Uncharacterized protein
Thecc1EG040660t1	Cysteine-rich RLK (RECEPTOR-like protein kinase) 8, putative
Thecc1EG040661t1	Uncharacterized protein
Thecc1EG040662t1	Tetratricopeptide repeat-like superfamily protein
Thecc1EG041218t1	Uncharacterized protein
Thecc1EG004502t1	BRI1 suppressor 1 (BSU1)-like 1

Thecc1EG041219t1	BRI1 suppressor 1 (BSU1)-like 2
Thecc1EG004502t1	BRI1 suppressor 1 (BSU1)-like 1

10. Accumulation of deleterious mutations

Sorting Intolerant from Tolerant (SIFT) 4G[54] was used for the prediction of the effect of nonsynonymous SNPs on protein functions. A custom database of predictions for all possible nonsynonymous SNPs was built using SIFT4G for *T. cacao*. SIFT outputs a SIFT score for each amino acid substitution, the score ranges from 0 to 1. The amino acid substitution is predicted deleterious if the score is ≤ 0.05, and tolerated if the score is > 0.05. Each prediction also provides SIFT median score which measures the diversity of the sequences used for prediction. SIFT median score ranges from 0 to 4.32, ideally the number would be between 2.75 and 3.5. A warning with low confidence occurs when the SIFT median score is greater than 3.25 because this indicates that the prediction was based on closely related sequences. The low confidence in SIFT score means that the protein alignment does not have enough sequence diversity because the position artificially appears to be conserved, an amino acid substitution may incorrectly predicted to be damaging. This score system was used to support the assignment of replacement substitutions as deleterious or tolerated for the rest of the analyses.

Prior to a Mantel-Hanzel test for specific effects, we fitted a generalized linear model to the count data for deleterious/tolerated mutations in Amelonado and Criollo, assuming a log-linear model. This model allowed us to test for general trends in the data and show that there is a significant difference in the number of deleterious mutations among Criollo and Amelonado along binned classes of minor allele frequency. Because we have differences in sample size between Criollo and Amelonado, we could not compare directly all the minor allele frequency classes and decided to bin them, making the direct comparison feasible. For each allele frequency class: rare (0-0.25], intermediate (0.25-0.375] and frequent (0.375,0.5] the number of predicted deleterious and tolerated mutations were identified using SIFT4G. Our model of the form:

$$Y_{ij} = I_i|freq + Pop_j|freq + \varepsilon_{ij},$$

was set to explain the counts of mutations as a function of their impact (deleterious/tolerated) and the population of origin, taking into account that different minor allele frequency classes will have different absolute counts in them. In our model Y_{ij} are the counts of the number of SNPs, I_{ij} corresponds to the impact (deleterious vs tolerated mutations) and Pop_i corresponds to the population (Amelonado and Criollo) and the comparisons are done conditional on frequency class bin of minor alleles. The link function is assumed to be

Poisson. The values of fitted coefficients are shown in Table S6 and Figure S18 presents a graphical representation of the mosaic plot that better describes the results.

Table S6 | Coefficients GLM model explaining differences in the rate of accumulation of deleterious mutations between Amelonado and Criollo.

Coefficients	*Estimate*	*Std._Error*	*z_value*	*Pr(>|z|)*
(Intercept)	8.66615	0.01236	701.006	<2e-16***
ImpactTOLERATED	1.24366	0.01377	90.306	<2e-16***
PopulationCriollo	-1.76858	0.01627	-108.714	<2e-16***
freq_class(0.25,0.375]	-2.28433	0.03594	-63.567	<2e-16***
freq_class(0.375,0.5]	-2.59115	0.03772	-68.702	<2e-16***
ImpactTOLERATED:freq_class(0.25,0.375]	0.01512	0.03808	0.397	0.691
ImpactTOLERATED:freq_class(0.375,0.5]	-0.03374	0.03794	-0.889	0.374
PopulationCriollo:freq_class(0.25,0.375]	1.44586	0.034	42.531	<2e-16***
PopulationCriollo:freq_class(0.375,0.5]	2.0974	0.03425	61.236	<2e-16***

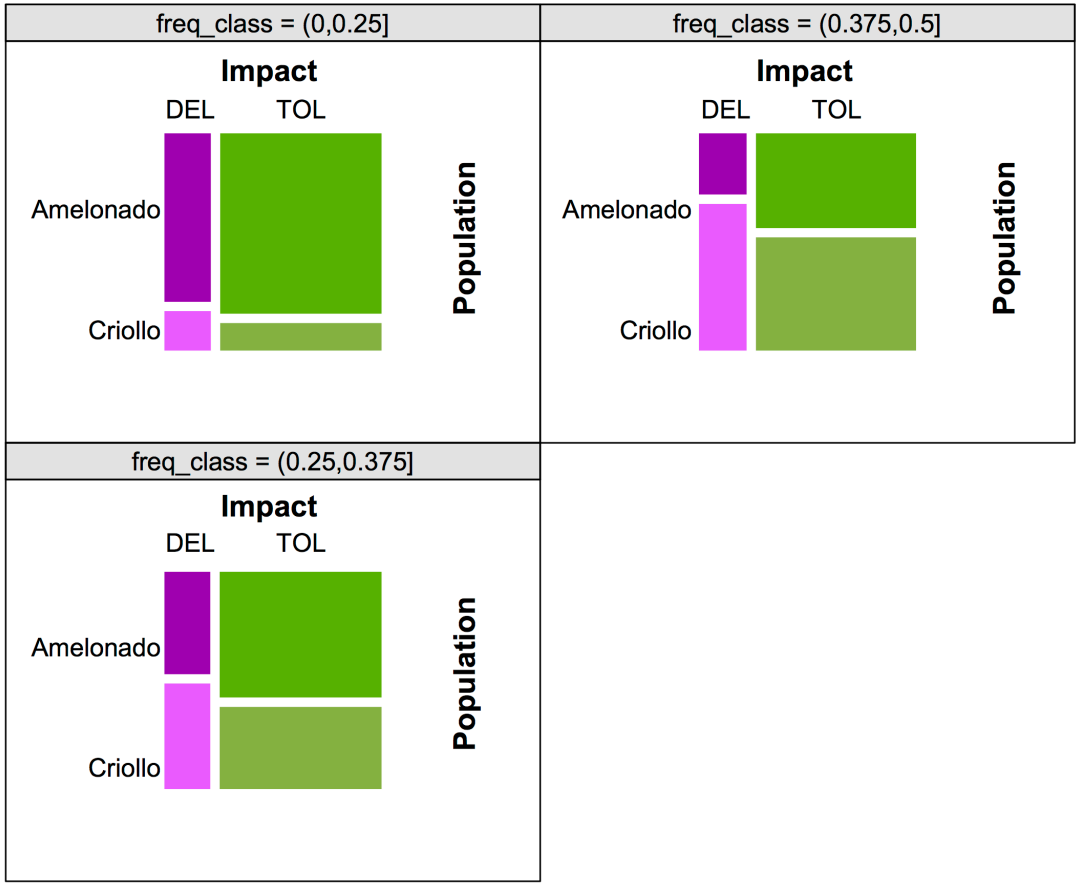

Figure S18 | Mosaic plot showing the proportional distribution of deleterious (DEL; shades of magenta) and tolerated (TOL; shades of green) among Criollo (light colors) and Amelonado (darker colors) for each minor allele frequency class.

Everything else being equal, with similar selfing rates across populations, it is expected that Amelonados and Criollos will present remarkable differences in the accumulation of deleterious mutations as the result of the differences in the magnitude of the population size reductions and the impact of domestication in Criollo. We decided to use Amelonado as a contrasting group because we could control for the similar frequency of selfing in the populations. Everything else being equal, with similar selfing rates across populations, it is expected that Amelonados and Criollos will present remarkable differences in the accumulation of deleterious mutations as the result of the differences in the magnitude of the population size reductions and the impact of domestication in Criollo. Our analyses aimed at understanding the pattern of accumulation of deleterious mutations in populations with similar levels of inbreeding driven by selfing and yet very different domestication pressures produces a pattern that has been revealed in other systems like maize, rice and composite flowers [55-57]. The pattern we show in *T. cacao* is consistent with other observations, we extend the work to test how this reflects in the fitness (measured as productivity of seeds) in *T. cacao*, something that has not been tested in long lived arboreal crops like cacao. In the next section we show how the increase in proportion of criollo ancestry and thus the relative frequency of deleterious mutations impact the productivity, a proxy for measuring the reproductive component of fitness.

Association between Criollo ancestry and productivity.

We show that Criollo populations sustain deleterious mutations at a higher frequency than Amelonado, even though both populations present a high frequency of self-compatible individuals. It remained to be tested what was the phenotypic effect of the proportional increased accumulation of deleterious mutations in the Criollo populations. For this, we used an additional dataset of plants for which productivity (measured as yield of beans per hectare per year) had been measured. We genotyped these plants with a Fluidigm array developed based on SNPs that were generated from some of the samples from the 200 genomes.

After genotyping, we merged the SNPs from newly genotyped individuals with SNPs from the individuals clearly assigned to each one of the 10 populations. The intersected dataset resulted in 7,621 SNPs. We then used *ADMIXTURE*[29] using a supervised assignment mode to estimate the proportion of ancestries to each one of the 10 populations. For each individual in the newly genotyped dataset, we also used vcftools[19] to estimate inbreeding coefficients. The proportion of ancestry assigned to each one of the newly genotyped individuals can be seen in Figure S19.

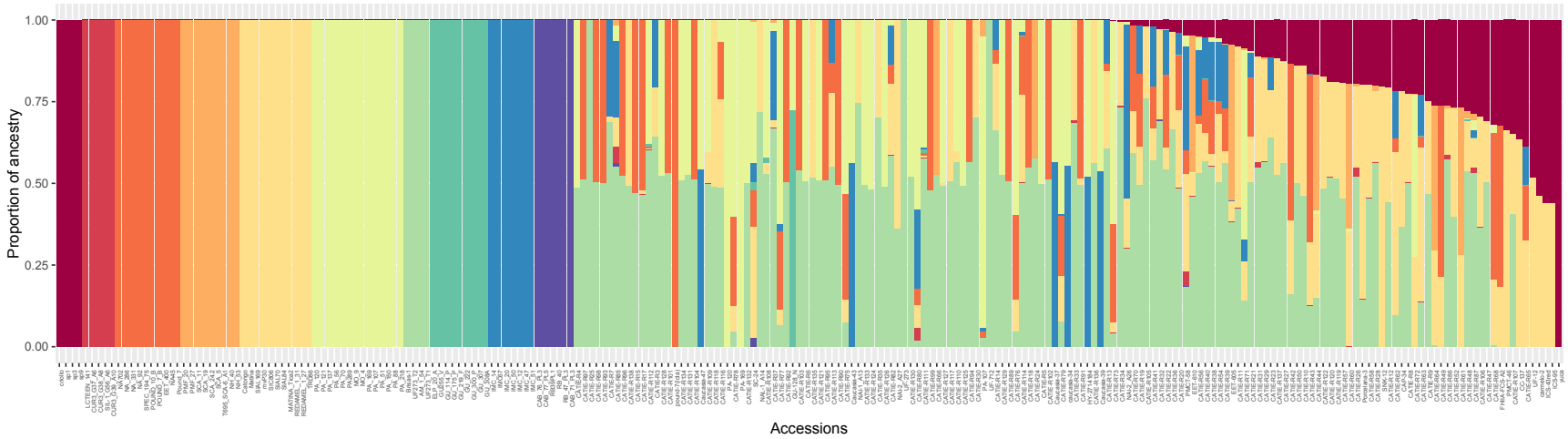

Figure S19 | Ancestry assignment (K=10) for original ten clusters and newly genotyped admixed individuals. The order of the groups corresponds to the assignment in Figure 5 of the main manuscript. From left to right, the colors correspond to the following groups: Criollo (dark red), Curaray (red), Nanay (dark orange), Contamana (orange), Amelonado (light orange), Marañon (light green), Nacional (green), Guianna (dark green), Iquitos (blue) and Purus (purple). The assignment of newly genotyped individuals to each ancestry allowed us to study the relationship of Criollo ancestry and the accumulation of deleterious mutations.

Following the estimation of ancestry, we estimated if the proportion of Criollo ancestry is associated with a reduction in the productivity using a simple linear model, while controlling for inbreeding. We built a generalized linear model assuming a Gaussian family of the form:

$Y = \beta_0 + \beta_1 + \beta_2 + \varepsilon$, where Y corresponds to the yield, β_0 corresponds to the intersect, $\beta_1$corresponds to the proportion of Criollo ancestry and $\beta_2$is the coefficient of inbreeding F, estimated for each individual.

We compared the model that considers inbreeding and a reduced version $Y = \beta_0 + \beta_1 + \varepsilon$ that considers only the proportion of Criollo ancestry.

Fitting of the full model suggests that as the proportion of Criollo increases, the reduction of yield is highly significant

Coefficients:

Intercept	Criollo_ancestry	F (inbreeding)
452.8	-555.2	-124.6

Degrees of Freedom: 145 Total (i.e. Null); 143 Residual
(4 observations deleted due to missingness)
Null Deviance: 8249000
Residual Deviance: 7420000
AIC: 2004

A likelihood ratio test suggests that the full model marginally explains the data better than the reduced model

Analysis of Deviance Table

Model 1: Yield.kg.ha.year. ~ Criollo_ancestry
Model 2: Yield.kg.ha.year. ~ Criollo_ancestry + F

Model	Resid	Df Resid.	Dev Df	Deviance	Pr(>Chi)
1	144	7565221			
2	143	7419722	1	145499	0.09402

The difference between the two models is not statistically significant. We also noticed that the size of the effect for the Criollo ancestry on productivity is at least 4.4 times larger than the size of the effect for inbreeding to explain the differences in yield. Taken together, we can say confidently that the proportion of Criollo ancestry and thus the increase in higher frequency deleterious mutations have a strong impact on productivity in cacao. Diagnostic plots for the fitting of the model are provided in Diagnostic1.zip. These results have a special appeal given that there is no appreciable association between Criollo ancestry and Pop Index (number of pods required 1 kg of dried cocoa without testa). The lack of association between Criollo ancestry and Pod Index is consistent with our interpretation that the accumulation of deleterious mutations decreased the fitness (Kilograms of beans per Hectare), but not the overall quality and ability to prepare chocolate from the cacao trees.

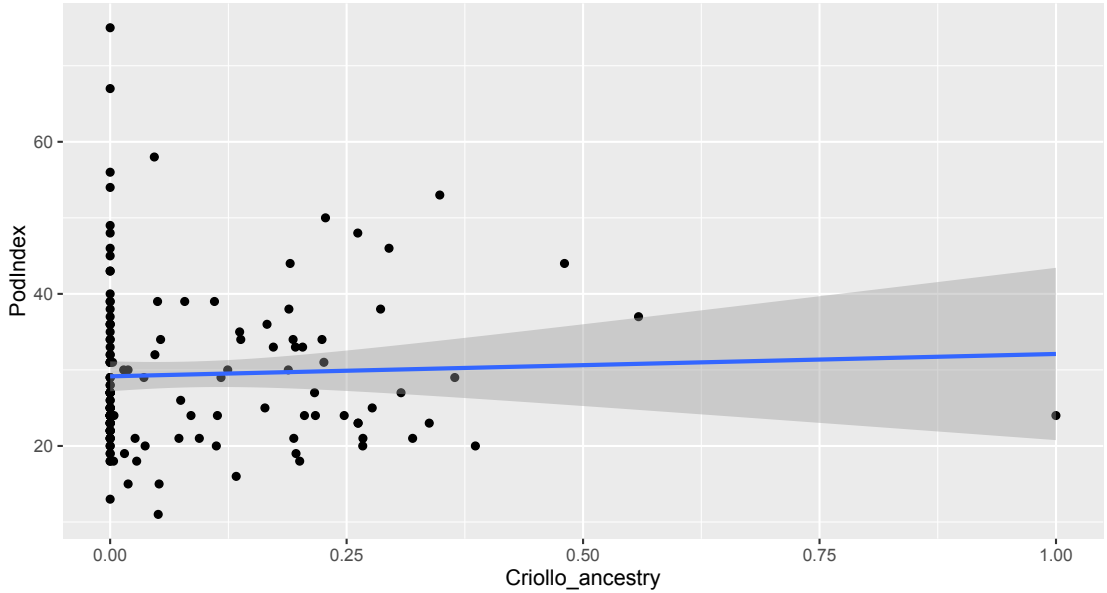

Figure S20 | We found no association between the proportion of Criollo ancestry and the Pod Index

References

1 S., A. *FastQC: a quality control tool for high throughput sequence data.*, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).

2 Krueger, F. *Trim Galore: a wrapper script to automate quality and adapter trimming as well as quality control*, <http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/> (2017).

3 Martin, M. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet.journal* **17**, 10-12 (2011).

4 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 Biol* **14**, r53, doi:10.1186/gb-2013-14-6-r53 (2013).

5 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

6 Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691-1692, doi:10.1093/bioinformatics/btr174 (2011).

7 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

8 Institute, B. *Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF*, <https://broadinstitute.github.io/picard/> (2016).

9 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

10 Cleary, J. G. *et al.* Joint Variant and De Novo Mutation Identification on Pedigrees from High-Throughput Sequencing Data. *Journal of Computational Biology* **21**, 405-419, doi:doi:10.1089/cmb.2014.0029. (2014).

11 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-181, doi:10.1038/nmeth.1785 (2011).

12 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).

13 Cingolani, P. *et al.* Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics* **3**, 35 (2012).

14 Jo, B. S. & Choi, S. S. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform* **13**, 112-118, doi:10.5808/GI.2015.13.4.112 (2015).

15 Rodrigues, J. P. *et al.* REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. *P Natl Acad Sci USA* **98**, 1030-1035 (2001).

16 Ryu, W. S. & Mertz, J. E. Simian virus 40 late transcripts lacking excisable intervening sequences are defective in both stability in the nucleus and transport to the cytoplasm. *Journal of Virology* **63**, 4386-4394 (1989).

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

18 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).

19 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158, doi:10.1093/bioinformatics/btr330 (2011).

20 Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585-595 (1989).

21 Charlesworth, D. & Charlesworth, B. Inbreeding Depression and Its Evolutionary Consequences. *Annu Rev Ecol Syst* **18**, 237-268, doi:DOI 10.1146/annurev.ecolsys.18.1.237 (1987).

22 Charlesworth, D., Morgan, M. T. & Charlesworth, B. Mutation Accumulation in Finite Outbreeding and Inbreeding Populations. *Genet Res* **61**, 39-56 (1993).
23 Piganeau, G. & Eyre-Walker, A. Evidence for variation in the effective population size of animal mitochondrial DNA. *PLoS One* **4**, e4396, doi:10.1371/journal.pone.0004396 (2009).
24 Cornejo, O. E., Fisher, D. & Escalante, A. A. Genome-wide patterns of genetic polymorphism and signatures of selection in Plasmodium vivax. *Genome Biol Evol* **7**, 106-119, doi:10.1093/gbe/evu267 (2014).
25 Gossmann, T. I., Keightley, P. D. & Eyre-Walker, A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol* **4**, 658-667, doi:10.1093/gbe/evs027 (2012).
26 Gossmann, T. I., Woolfit, M. & Eyre-Walker, A. Quantifying the variation in the effective population size within a genome. *Genetics* **189**, 1389-1402, doi:10.1534/genetics.111.132654 (2011).
27 Bartley, B. G. D. *The genetic diversity of cacao and its utilization*. (CABI Publishing, 2005).
28 Cope, F. W. The mechanism of pollen incompatibility in *Theobroma cacao* L. *Heredity* **17**, 157-182 (1962).
29 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664, doi:10.1101/gr.094052.109 (2009).
30 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).
31 Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238-242, doi:10.1038/nature09103 (2010).
32 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, doi:10.1093/bioinformatics/btm233 (2007).
33 Motamayor, J. C. *et al.* Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PLoS One* **3**, e3311, doi:10.1371/journal.pone.0003311 (2008).
34 Zhang, D. *et al.* Dissecting Genetic Structure in Farmer Selections of Theobroma Cacao in the Peruvian Amazon: Implications for on Farm Conservation and Rehabilitation. *Tropical Plant Biology* **4**, 106-116, doi:10.1007/s12042-010-9064-z (2011).
35 Zhang, D. *et al.* Genetic diversity and spatial structure in a new distinct Theobroma cacao L. population in Bolivia. *Genetic Resources and Crop Evolution* **59**, 239-252, doi:10.1007/s10722-011-9680-y (2011).
36 Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190, doi:10.1371/journal.pgen.0020190 (2006).
37 Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, doi:doi:10.2307/2408641 (1984).
38 Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967, doi:10.1371/journal.pgen.1002967 (2012).
39 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).

40 Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-496, doi:10.1038/nature10231 (2011).
41 Lynch, M. Evolution of the mutation rate. *Trends Genet* **26**, 345-352, doi:10.1016/j.tig.2010.05.003 (2010).
42 Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* **327**, 92-94, doi:10.1126/science.1180677 (2010).
43 Kidd, J. M. *et al.* Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *Am J Hum Genet* **91**, 660-671, doi:10.1016/j.ajhg.2012.08.025 (2012).
44 Wright, S. Coefficients of inbreeding and relationship. *American Naturalist* **56**, 330–338 (1922).
45 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695, doi:10.1371/journal.pgen.1000695 (2009).
46 Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial selffertilization. *Genetics* **154**, 923–929 (2000).
47 Nordborg, M. & Donnelly, P. The coalescent process with selfing. *Genetics* **146**, 1185–1195 (1997).
48 Motamayor, J. C. *et al.* Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity (Edinb)* **89**, 380-386, doi:10.1038/sj.hdy.6800156 (2002).
49 Henderson, J. S., Joyce, R. A., Hall, G. R., Hurst, W. J. & McGovern, P. E. Chemical and archaeological evidence for the earliest cacao beverages. *P Natl Acad Sci USA* **104**, 18937-18940, doi:10.1073/pnas.0708815104 (2007).
50 Powis, T. G., Cyphers, A., Gaikwad, N. W., Grivetti, L. & Cheong, K. Cacao use and the San Lorenzo Olmec. *P Natl Acad Sci USA* **108**, 8595-8600, doi:10.1073/pnas.1100620108 (2011).
51 Schultes, R. E. in *Pre-Columbian plant migration, Papers of the Peabody Museum of Archaeology and Ethnology* Vol. 76 (ed D. Stone) 69–83 (Harvard University Press, 1984).
52 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, doi:10.1371/journal.pone.0048438 (2012).
53 Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res* **20**, 393-402, doi:10.1101/gr.100545.109 (2010).
54 Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat Protoc* **11**, 1-9, doi:10.1038/nprot.2015.123 (2016).
55 Lu, J. *et al.* The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics* **22**, 126-131 (2006).
56 Renaut, S. & Rieseberg, L. H. The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops. *Molecular Biology and Evolution* **32**, 2273-2283, doi:10.1093/molbev/msv106 (2015).

57 Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)* **4**, 163-171, doi:10.1534/g3.113.008870 (2014).

Additional Figures not inserted in the document are included in zip file