Genome-wide association mapping and predictive modeling of wet bean mass in a diverse cacao collection

Jishnu Bhatt[1,†], Insuck Baek[2,†], Seunghyun Lim[1], Dongho Lee[3], Jae Hee Jang[1], Stephen P. Cohen[1], Amelia H. Lovelace[1], Moon S. Kim[2], Lyndel W. Meinhardt[1], Sunchung Park[1], Ezekiel Ahn[1,*]

[1] Sustainable Perennial Crops Laboratory, Agricultural Research Service, United States, Department of Agriculture, Beltsville, MD, 20705, USA

[2] Environmental Microbial and Food Safety Laboratory, Agricultural Research Service, United States, Department of Agriculture, Beltsville, MD, 20705, USA

[3] Soybean Genomics & Improvement Laboratory, Agricultural Research Service, United States, Department of Agriculture, Beltsville, MD, 20705, USA

[†] The authors contributed equally to this work.
* Author to whom correspondence should be addressed (email: Ezekiel.ahn@usda.gov).

Abstract

Yield improvement is a critical breeding target for the chocolate tree (*Theobroma cacao* L.), the source of cocoa beans. We investigated the genetic architecture of flower, fruit, and seed traits in 346 diverse cacao accessions using an integrated approach, combining phenotypic data (27 traits), genome-wide SNP genotyping (671 SNPs), and advanced machine learning. Our Bootstrap Forest-based GWAS identified novel SNP-trait associations, providing candidate genes linked to key agronomic traits like pod index, seed number, and cotyledon mass. Phenotypic variation exhibited a complex, continuous relationship with genetic relatedness, not fully explained by predefined accession groups. Our Neural Network model (R^2 = 0.715 in validation) found that cotyledon mass (portion = 0.520) and cotyledon length (portion = 0.213) were the major phenotypic contributors that predict cacao total wet bean mass, a direct measurement of yield. This finding suggests that simple measurement of these early-stage phenotypic attributes may be useful in identifying productive clones of cacao. Overall, this study enhances our understanding of the genetic basis of trait variation in cacao and provide a powerful framework for implementing genomic selection to develop improved cacao varieties with enhanced yield potential.

Author summary

Improving yield in the chocolate tree, Theobroma cacao, is vital but complex, as yield results from the interplay of many genes. Identifying these genes and predicting which trees will be most productive remains a major challenge for breeders. In this study, we combined genetic information (DNA markers) with detailed measurements of 27 traits from 346 diverse cacao

trees. Using advanced machine learning techniques, we pinpointed specific genetic regions associated with key yield components like pod characteristics and seed number. Furthermore, we developed a computational model that accurately predicts a tree's total wet bean mass – a direct measure of yield. Surprisingly, this model revealed that the mass and length of the initial seed leaves (cotyledons) are the strongest predictors of future productivity. Our findings provide new insights into the genetic control of cacao yield and offer a potentially simple, early-stage tool to help breeders select superior trees more efficiently.

Introduction

Theobroma cacao L., the source of cocoa beans, is a globally significant crop that underpins a multi-billion dollar chocolate industry and supports the livelihoods of millions of smallholder farmers, primarily in developing countries [1,2]. These smallholders are responsible for approximately 80% of global cocoa production [3]. Originating in the upper Amazon basin [4], cacao exhibits rich genetic diversity, contributing to a wide range of flavors, aromas, and other desirable traits [5]. However, cacao production faces increasing challenges from climate change, emerging diseases and pests, and a growing global demand for chocolate [6,7]. These challenges necessitate the development of improved cacao varieties with enhanced yield potential, disease resistance, and climate resilience, while maintaining and enhancing quality attributes.

Efforts to understand genetic diversity and identify genes controlling important traits from the cacao genome have been ongoing for decades. Studies have indicated that genetic differentiation within cacao lineages results from spatial variation in selection pressures, genetic drift, or a combination of both processes [8]. Yield stability over time is a critical trait in cacao breeding [3], yet studies focusing on temporal stability remain limited [9]. A study evaluating 34 cacao hybrid families over six years in Costa Rica revealed substantial variability in annual yield, with coefficient of variation values ranging from 113% to 642% for different years, particularly notable in years two and three, when 80% of trees produced no pods [9]. Cacao yield was contingent on interrelated factors such as pod health and disease incidence, shown through a positive correlation between the number of healthy pods and overall yield efficiency [3]. When low yield variability is optimized for in breeding programs, newly developed cultivars are susceptible to reduced yield efficiencies, necessitating broadly informed selection of genetic traits to sustain productivity over time [3]. Genetic factors significantly influence pod development and quality traits. For instance, F1 progenies from specific parental combinations, such as TSH858 x DR1, have demonstrated superior pod girth and weight, surpassing the average weights reported in other cacao-producing regions like Côte d'Ivoire [10]. The selection of progeny based on pod index, calculated as the number of pods required to produce one kilogram of dried cocoa [11], and weight can improve overall tree productivity and efficiency in production processes [10]. Monitoring genetic identity is essential as progenies from mislabeled or pollen-contaminated clones confound breeding efforts towards the development of superior varieties [12].

Traditional breeding methods, relying on phenotypic selection and controlled crosses, have contributed to significant improvements in cacao varieties [13,14]. However, these methods

are unable to take into account the complex interplay and inheritance of many traits [15]. The advent of genomic tools, such as genome-wide association studies (GWAS) and genomic selection, offers new opportunities to understand inherited trait complexity and accelerate cacao improvement [16,17]. GWAS allows for the identification of marker-trait associations, revealing genomic regions and potential candidate genes influencing traits of interest [18]. Previous studies in cacao have employed GWAS to identify loci associated with disease resistance, yield components, and bean quality attributes, demonstrating the potential of this approach [11,16,17].

A comprehensive understanding of the genetic architecture of complex traits in cacao such as those related to yield potential, has remained elusive. Only a few studies have integrated genomic data with advanced machine learning techniques to enhance the prediction and understanding of complex traits in cacao, such as with bean quality [19], fermentation level [20], canopy architecture [21], yield potential [22], aroma [23], or caffeine content [24]. This study addresses these limitations by leveraging a large and diverse collection of 346 cacao accessions, utilizing publicly available phenotypic and genotypic data from the ICGT Cacao Germplasm Database [11]. These accessions, conserved *ex-situ* at the International Cocoa Genebank, Trinidad (ICGT), represent a broad range of genetic and phenotypic variation [11]. We employed a multifaceted approach, combining detailed phenotypic evaluations of 27 traits, encompassing flower, fruit, and seed characteristics, with genome-wide SNP genotyping, phylogenetic analysis, and advanced statistical methods. Specifically, we performed a Bootstrap Forest (equivalent to Random Forest)-based GWAS to identify significant marker-trait associations and explored the relationships among traits using hierarchical clustering [25]. Furthermore, based on various phenotypic data, we trained a machine-learning model to predict wet bean mass, a direct indicator of yield potential.

The primary objectives of this study were to: (1) investigate the relationship between genetic relatedness and phenotypic variation in a diverse collection of cacao accessions; (2) identify genomic regions and potential candidate genes associated with key traits, including flower and fruit morphology, seed characteristics, and yield potential, using a Bootstrap Forest-based GWAS approach; (3) explore the relationships among traits and identify groups of traits with shared genetic associations through hierarchical clustering; and (4) develop and evaluate a machine learning model for predicting wet bean mass based on phenotypic data. We hypothesized that the diverse cacao collection would exhibit a complex pattern of phenotypic variation that is only partially explained by broad phylogenetic relationships. We further hypothesized that the GWAS would allow identification of novel genomic regions and candidate genes associated with the traits of interest, and that the machine learning model would accurately predict wet bean mass, demonstrating the potential for genomic selection in cacao.

Results

Correlation Between Patristic Distances and Phenotypic Traits

To investigate the relationship between genetic relatedness and phenotypic variation in the cacao accessions, we calculated pairwise Pearson's correlation coefficients between patristic

distances, derived from the phylogenetic analysis, and the log-transformed values of all phenotypic traits. The correlation matrix revealed a complex pattern of relationships between genetic relatedness (patristic distance) and phenotypic traits (Figure 1). Overall, many traits exhibited weak correlations with patristic distance, suggesting that genetic divergence, as measured by this metric, is not a primary driver of phenotypic variation for these traits (e.g.: sepal length (FL_SEP_LEN), anthocyanin intensity on the ligule (FL_LIG_COL), and anthocyanin intensity in the filament (FL_FIL_COL)), in this collection of cacao. However, several notable correlations were observed ($p < 0.05$). For example, pod index showed a moderate positive correlation with patristic distance ($r = 0.38$), suggesting genetic divergence as a driver of pod index variation.

Several traits exhibited weak to moderate negative correlations with patristic distance, including total wet bean mass (FR_WABW; $r = -0.33$), seed number (FR_BEN; $r = -0.30$), cotyledon mass (FR_BW; $r = -0.21$), cotyledon length (FR_BL; $r = -0.26$), cotyledon width (FR_BEW; $r = -0.22$), flower ovule number (FL_OV_NUM; $r = -0.23$), and fruit wall hardness (FR_HH; $r = -0.21$). These negative correlations indicate that for these traits, the phenotypic variation was greater among more genetically related accessions. This pattern could be attributed to convergent evolution, where distant lineages have independently evolved similar trait values due to similar environmental pressures or selection for yield-related traits. Alternatively, functional constraints related to seed development or plant physiology might limit the range of variation in these traits, even across diverse genetic backgrounds. Traits such as flower style length (FL_STY_LEN; $r = 0.16$) and fruit ridge disposition (FR_PFD; $r = 0.12$) showed weak positive correlations with patristic distance. These positive correlations suggest that more genetically divergent accessions exhibit greater differences in these traits. The variation in flower style length, for instance, might reflect adaptation to different pollinators in the geographic regions where these diverse lineages evolved.

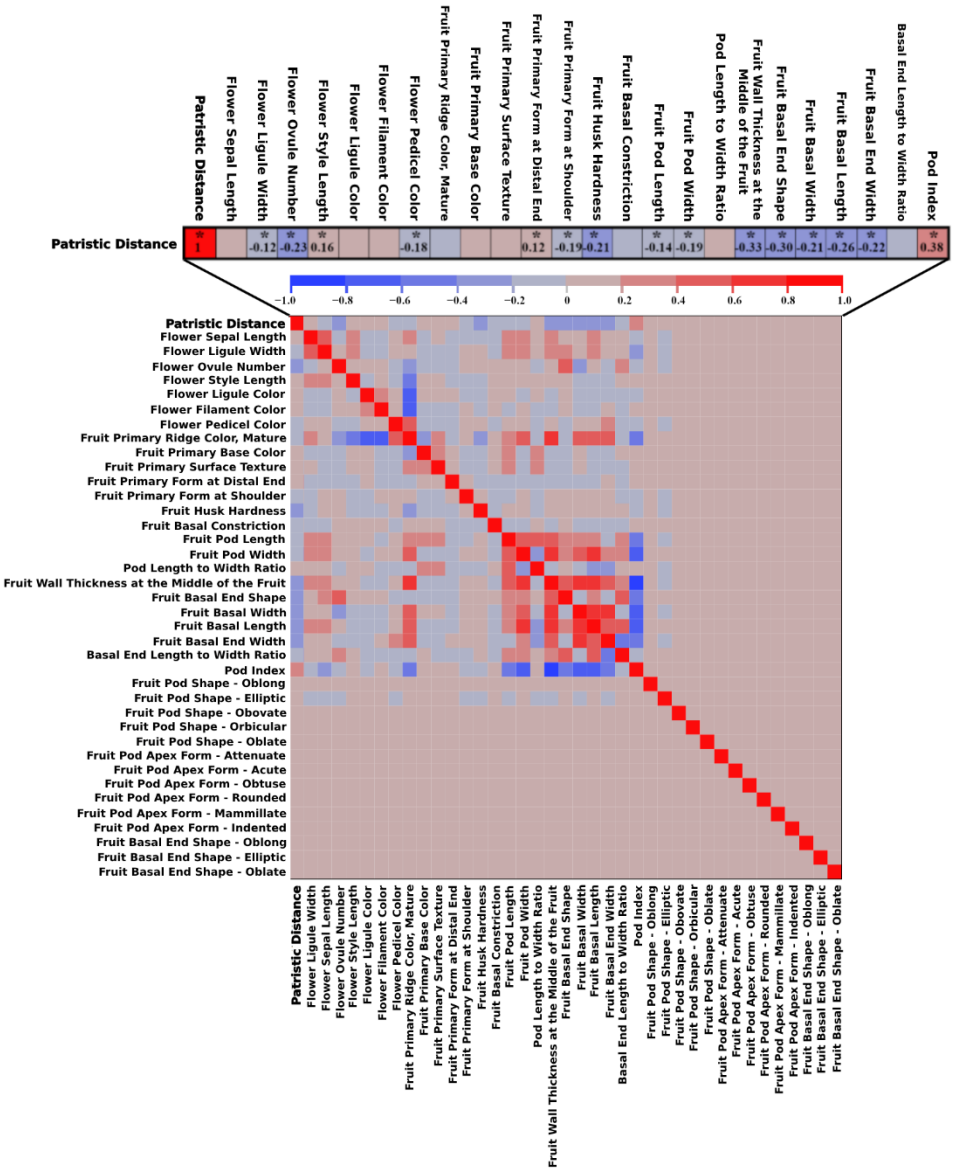


Fig 1. Heatmap visualization of the Pearson's correlation coefficients between patristic distances and log-transformed phenotypic traits in cacao accessions. The color intensity and direction represent the strength and nature of the correlation: red indicates a positive correlation, blue indicates a negative correlation, and darker shades represent stronger correlations. Asterisks in the top row denote statistically significant correlations ($p < 0.05$). Pod index, a key agronomic trait that reflects yield potential in cacao, is also included as a composite trait reflecting yield potential.

Visualization of phenotypic and genotypic diversity using t-SNE

t-SNE analysis was used to visualize the patterns of variation within the phenotypic, genotypic, and combined datasets (Figure 2). The t-SNE plot based on phenotypic data alone (Figure 2a) reveals a relatively dispersed distribution of accessions, with no strong clustering evident. However, there is a tendency for accessions from group 8 (purple) and group 14

(red) to be apart from one another, suggesting some degree of phenotypic dissimilarity between these two groups.

The t-SNE plot based on genotypic data (Figure 2b) shows two main clusters connected in a roughly linear pattern. The lower cluster is primarily composed of accessions from groups 8 (purple), 14 (red), and 15 (gray), along with accessions from other groups. The upper cluster mainly consists of genotypes that have not been phenotyped and do not have an assigned accession group. This pattern suggests a degree of genetic structure within the collection, with the lower cluster potentially representing a distinct genetic lineage.

The t-SNE plot based on the combined phenotypic and genotypic data (Figure 2c) provides the most comprehensive view of the variation. In this plot, accessions are colored according to their assigned accession groups. Some degree of separation between groups is apparent, particularly for groups 7 (dark purple), 8 (purple), 14 (red), and 15 (gray) (Figure 2c). However, there is also overlap between several groups, and the overall structure remains complex, indicating that while both phenotype and genotype contribute to the clustering patterns, neither data type alone is sufficient to fully delineate all accession groups.

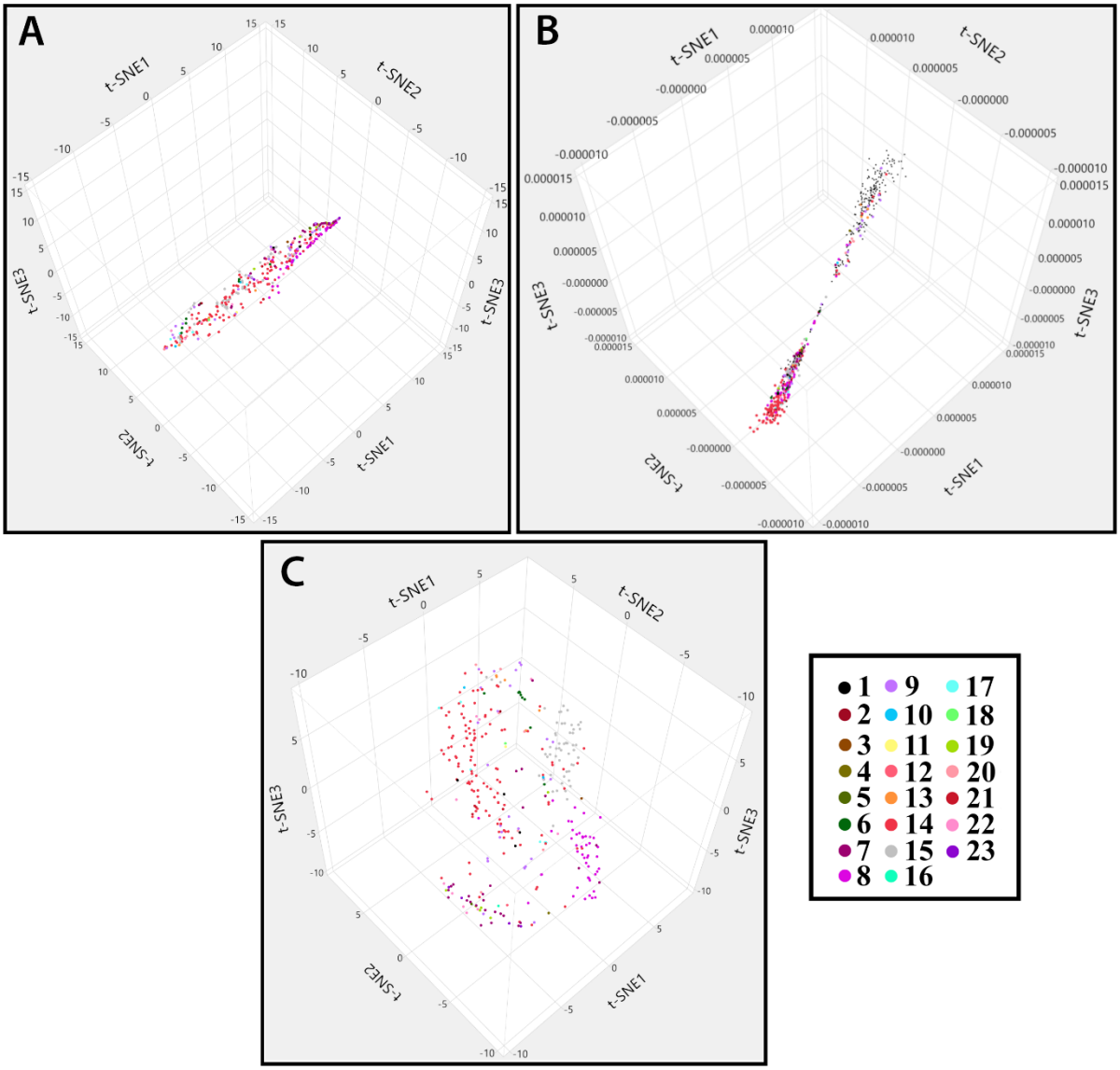


Fig 2. Visualization of cacao accessions using t-distributed Stochastic Neighbor Embedding (t-SNE) based on different data types. (a) all phenotypic traits, (b) SNP genotype data, and (c) a combined dataset of phenotypic traits and SNP genotypes. Each point represents a single cacao accession. In (c), colors correspond to the 23 distinct accession groups, and numbers on the legend are pertaining to each group.

Bootstrap Forest-based GWAS of flower traits

To identify genomic regions associated with variation in flower morphology in cacao, we performed a GWAS using a Bootstrap Forest model. Manhattan plots of the GWAS results for key flower traits (anthocyanin intensity on the ligule, anthocyanin in the column of the pedicel, anthocyanin in the filament, ovule number, sepal length, ligule width, and style length) revealed SNPs with importance scores that suggest significant associations with phenotypic variation (Figure 3). The top-ranked SNPs for each trait, along with their putative candidate genes and associated gene ontology (GO) terms, are presented in Table 1. For example, the analysis of anthocyanin intensity on the ligule identified TcSNP174 (chromosome 4) as a top-ranked SNP. The nearest gene to TcSNP174, Tc04v2_g009180,

encodes a 6-phosphogluconate dehydrogenase, an enzyme involved in the pentose phosphate pathway, which may play a role in pigment biosynthesis. As another example, for ovule number, TcSNP697 (chromosome 9) was a top-ranked SNP, located near Tc09v2_g017450, which encodes a WLIM1 protein potentially involved in floral development. These SNPs highlight potential candidate genes involved in the regulation of flower morphology and pigmentation in cacao.

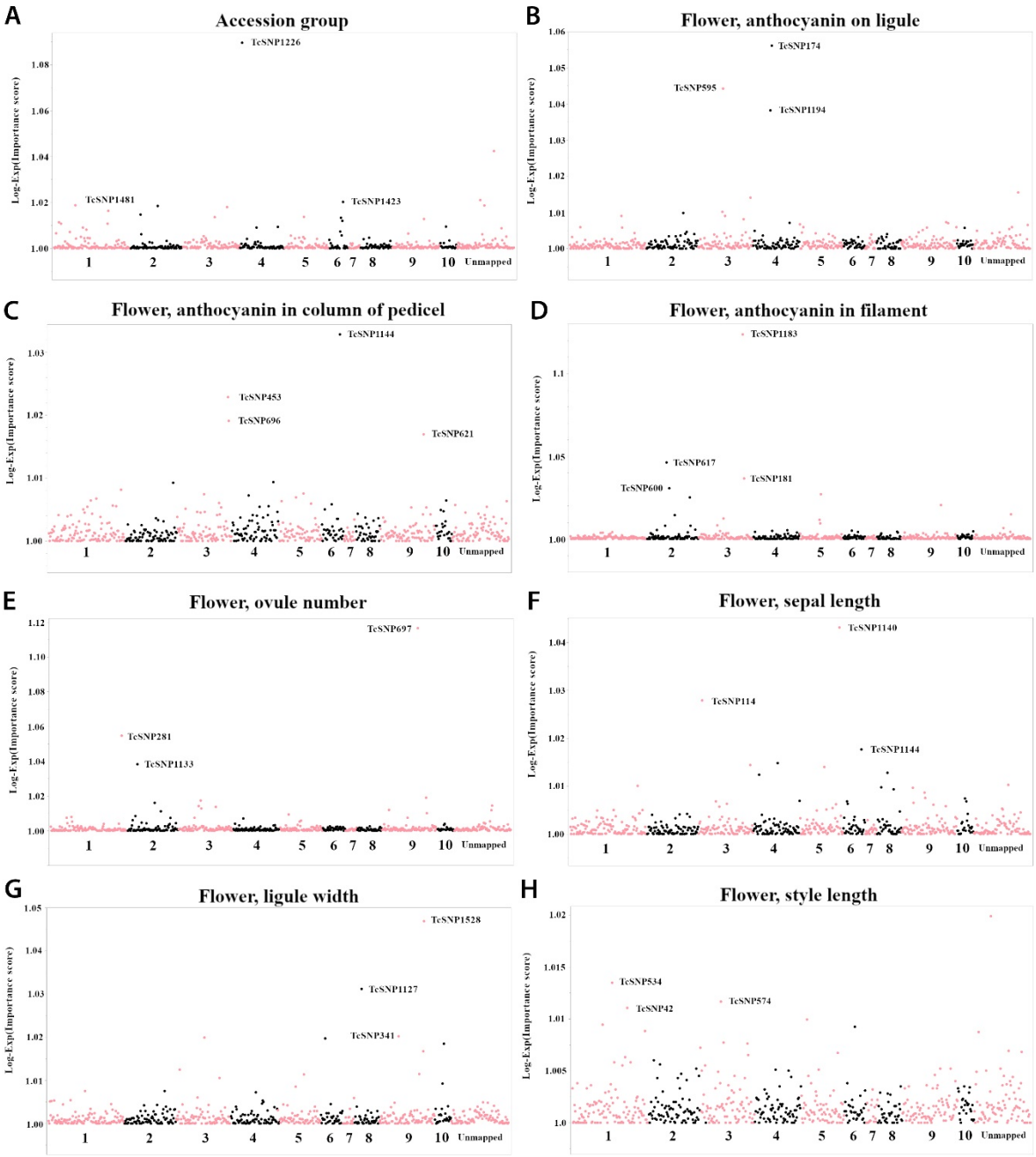


Fig 3. Manhattan plots of the Bootstrap Forest-based genome-wide association study for accession group and flower traits in cacao: (a) accession group, (b) anthocyanin on ligule, (c) anthocyanin in column of pedicel, (d) anthocyanin in

212 filament, (e) ovule number, (f) sepal length, (g) ligule width, and (h) style length. Each
213 point represents a SNP locus, plotted according to its chromosomal position (x-axis)
214 and its importance score (Log-Exp(Portion)) on the y-axis. The Portion score reflects
215 the contribution of each SNP to the prediction of the trait in the Bootstrap Forest
216 model. The top SNPs for each trait are labeled with their corresponding IDs.

217

Trait	**SNP ID**	**Chr**	**Nearest genes**	**GO annotation**	**Impartance score (portion)**
Accession group	TcSNP1226	4	*Tc04v2_g001730* S tudor-like domain (1253806..1259080)	Transcription	0.086
	TcSNP1423	6	-	-	0.042
	TcSNP1481	1	*Tc01v2_g009400* RING-H2 zinc finger protein RHA2a (5824131..5824989)	Metal-binding, zinc	0.021
Flower, anthocyanin on ligule	TcSNP174	4	*Tc04v2_g009180* 6-phosphogluconate dehydrogenase, decarboxylating 3 (19060678..19063761)	Carbohydrate degradation; pentose phosphate pathway; D-ribulose 5-phosphate from D-glucose 6-phosphate (oxidative stage)	0.055
	TcSNP595	3	*Tc03v2_g013080* Putative 21 kDa protein (27560742..27561854)	-	0.043
	TcSNP1194	4	*Tc04v2_g007120* Expansin-A5 (15625391..15627150)	-	0.038
Flower, anthocyanin in column of pedicel	TcSNP1144	6	*Tc06v2_g017980* Polyadenylate-binding protein RBP45B (25207100..25211042)	mRNA binding, mRNA processing	0.032
	TcSNP453	3	*Tc03v2_g025740* Binding partner of ACD11 1 (35482980..35485981)	-	0.023

	TcSNP696	3	-	-	0.019
	TcSNP621	9	*Tc09v2_g026960* Triosephosphate isomerase, cytosolic (36324529..36327221)	-	0.017
Flower, anthocyanin in filament	TcSNP1183	3	*Tc03v2_g021350* Acid beta-fructofuranosidase (32970469..32974667)	-	0.12
	TcSNP617	2	*Tc02v2_g008810* Probable cinnamyl alcohol dehydrogenase (5456852..5459045)	-	0.045
			Tc02v2_g008800 Putative Probable cinnamyl alcohol dehydrogenase (5455488..5456064)		
	TcSNP181	3	*Tc03v2_g022940* Osmotin-like protein OSM34 (33908266..33909227)	-	0.036
	TcSNP600	2	*Tc02v2_g009700* LIM domain-containing protein WLIM1 (5907737..5909583)	GATA type zinc finger transcription factor family protein. Actin cytoskeleton	0.03
Flower, ovule number	TcSNP697	9	-	-	0.11
	TcSNP281	1	*Tc01v2_g034810* 60S ribosomal protein L31 (37121671..37123110)	Cytosolic large ribosomal subunit, Structural constituent of ribosome	0.053
	TcSNP1133	2	-	-	0.038
Flower, sepal length	TcSNP1140	5	*Tc05v2_g026200* Uncharacterized protein (38465158..38468576)	-	0.042
	TcSNP114	3	Tc03v2_g001390 60S ribosomal protein L27a-3 (1332911..1333621)	Cytosolic large ribosomal subunit, mRNA binding	0.028

	TcSNP1144	6	*Tc06v2_g017980* Polyadenylate-binding protein RBP45B (25207100..25211042)	mRNA binding, mRNA processing	0.018
Flower, ligule width	TcSNP1528	9	*Tc09v2_g027450* Histone H2A (36574922..36575859) *Tc09v2_g027460* Glycine-rich RNA-binding protein 4, mitochondrial (36576252..36578701)	Nucleosome, heterochromatin formation, protein heterodimerization activity -	0.046
	TcSNP1127	8	*Tc08v2_g006380* E3 ubiquitin-protein ligase UPL2 isoform 1 (3501139..3516412) *Tc08v2_g006370* Di-glucose binding protein with Kinesin motor domain (3493845..3500575)	-	0.031
	TcSNP341	9	-	-	0.02
Flower, style length	TcSNP534	1	*Tc01v2_g018030* 40S ribosomal protein S8 (15417456..15420006)	Cytosolic small ribosomal subunit, maturation of SSU-rRNA from tricistronic rRNA transcript	0.02
	TcSNP574	3	-	-	0.013
	TcSNP42	1	*Tc01v2_g027510* Putative Thiol protease aleurain-like (32880730..32884255)	-	0.012

Table 1. Putative candidate genes near top SNPs associated with flower traits in cacao. Details of the most important SNPs associated with eight flower traits in cacao, as identified by the Bootstrap Forest-based classification. The table lists the SNP ID, chromosome (Chr), nearest gene (and its genomic location), GO annotation (if available), and the importance score (Portion) for each SNP. The importance score reflects the contribution of each SNP to the prediction of the trait in the Bootstrap Forest model. A dash (-) indicates that no information was available for that particular cell.

Genome-wide association study of fruit traits

We extended our analysis to fruit traits by conducting a GWAS using the Bootstrap Forest model. Six key fruit traits were shown: fruit length, fruit width, fruit length to width ratio, fruit surface anthocyanin in mature ridges, fruit surface texture, and fruit pod wall hardness (Figure 4 and Table 2). Manhattan plots of the GWAS results identified SNPs with high importance scores, suggesting strong linkage with phenotypic variation in these fruit traits. The top-ranked SNPs for each trait, along with their putative candidate genes and associated GO terms, are detailed in Table 2.

For fruit length, the SNP TcSNP642 (chromosome 8) exhibited a strong association and is located near *Tc08v2_g015120*, which encodes a peroxidase 12, potentially involved in cell wall modification during fruit development. Interestingly, for fruit surface anthocyanin in mature ridges, TcSNP174 (chromosome 4) was found again as a top-ranked SNP, located near *Tc04v2_g009180*, encoding a 6-phosphogluconate dehydrogenase, indicating that this enzyme plays a key role in both floral and fruit pigmentation. These findings, along with the other candidate genes identified (Table 2), provide a starting point for functional studies aimed at elucidating the genetic mechanisms underlying fruit development and quality in cacao.

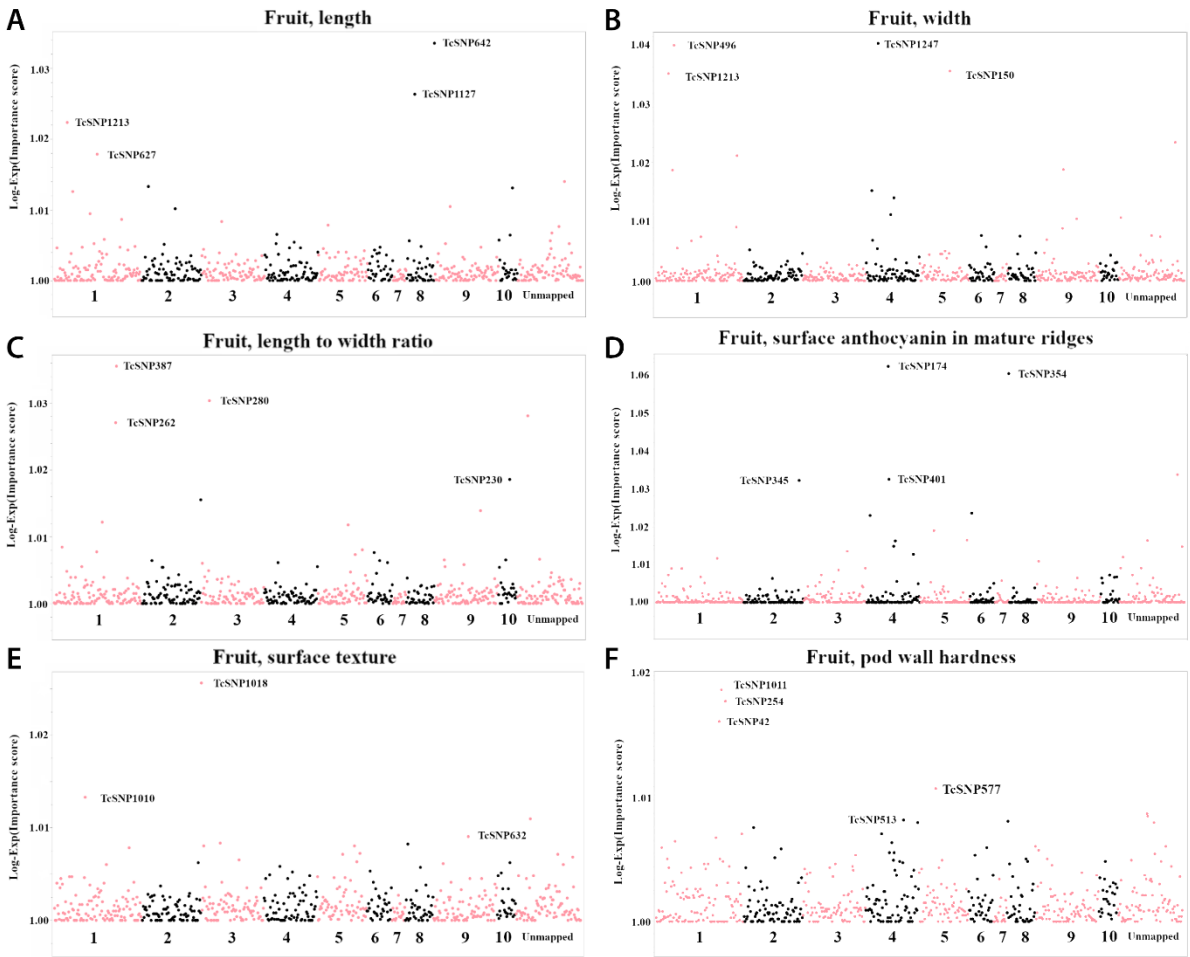


Fig 4. Manhattan plots of the Bootstrap Forest-based GWAS for fruit traits in cacao. Each point represents an SNP, plotted by its position on the chromosome (x-axis) and its importance score (Log-Exp(Portion)) on the y-axis. The top SNPs associated with each trait are labeled with their corresponding IDs.

Trait	SNP ID	Chr	Nearest genes and function	GO annotation	Impartance score (portion)
Fruit, length	TcSNP642	8	*Tc08v2_g015120* Peroxidase 12 (16398271..16401309)	-	0.033
	TcSNP1127	8	*Tc08v2_g006380* E3 ubiquitin-protein ligase UPL2 isoform 1 (3501139..3516412)	-	0.026
			Tc08v2_g006370 Di-glucose binding protein with Kinesin motor domain (3493845..3500575)		
	TcSNP1213	1	*Tc01v2_g005450* Probable cysteine protease RD21B (3027564..3029931)	-	0.022
	TcSNP627	1	*Tc01v2_g017220* Ribonuclease 2 (14489234..14495600)	-	0.018
Fruit, width	TcSNP1247	4	*Tc04v2_g003190* Protein transport protein Sec61 subunit gamma (2379921..2381593)	SecE/sec61-gamma protein transport protein, Ssh1 translocon complex	0.039
			Tc04v2_g003180 50S ribosomal protein L27, chloroplastic (2377934..2379191)	-	
	TcSNP496	1	*Tc01v2_g008340* 40S ribosomal protein S6 (5017545..5020015)	Ribonucleoprotein complex, structural constituent of ribosome, translation	0.039

	TcSNP150	5	*Tc05v2_g012150* Probable calcium-binding protein CML35 (24900882..24902223)	-	0.035
	TcSNP1213	1	*Tc01v2_g005450* Probable cysteine protease RD21B (3027564..3029931)	-	0.034
Fruit, length to width ratio	TcSNP387	1	-	-	0.035
	TcSNP280	3	*Tc03v2_g003930* Probable calcium-binding protein CML27 (5206898..5207881)	Calcium ion binding	0.03
	TcSNP262	1	*Tc01v2_g026930* S-adenosylmethionine synthase 1 (32474115..32476941)	ATP binding, S-adenosylmethionine biosynthetic process	0.028
			Tc01v2_g026940 O-Glycosyl hydrolases family 17 protein (32476724..32478711)	O-Glycosyl hydrolases family 17 protein, Transmembrane	
	TcSNP230	10	*Tc10v2_g009820* DnaJ protein homolog (15197459..15200209)	cytoplasm, response to heat, protein-folding chaperone binding, Hsp70 protein binding	0.027
Fruit, surface anthocyanin in mature ridges	TcSNP174	4	*Tc04v2_g009180* 6-phosphogluconate dehydrogenase, decarboxylating 3 (19060678..19063761)	Carbohydrate degradation; pentose phosphate pathway; D-ribulose 5-phosphate from D-glucose 6-phosphate (oxidative stage)	0.061
	TcSNP354	8	*Tc08v2_g000870* Patatin-like protein 2 (504500..507100)	-	0.059
	TcSNP401	4	*Tc04v2_g010060* ATP-dependent Clp protease proteolytic subunit 3, chloroplastic (20482061..20486150)	-	0.033

			Tc04v2_g010070 RNA polymerase II transcription elongation factor, putative isoform 1 (20487006..20489395)		
	TcSNP345	2	*Tc02v2_g029510* NAC transcription factor 29 (40054951..40056432)	Regulation of DNA-templated transcription	0.032
Fruit, surface texture	TcSNP1018	3	*TcUnv2_g002440* Signal peptidase I (9584804..9589980)	-	0.025
	TcSNP1010	1	-	-	0.013
	TcSNP632	9	*Tc09v2_g021090* 60S ribosomal protein L7-4 (31437584..31440037)	Ribonucleoprotein , Ribosomal protein	0.011
Fruit, pod wall hardness	TcSNP1011	1	*Tc01v2_g028610* Aspartyl protease family protein 2 (33641404..33643373)	-	0.018
	TcSNP254	1	*Tc01v2_g030070* Aquaporin PIP1-2 (34491268..34494325)	Plasma membrane intrinsic protein 1B, plasma membrane, response to water deprivation	0.018
	TcSNP42	1	*Tc01v2_g027510* Putative Thiol protease aleurain-like (32880730..32884255)	-	0.016
	TcSNP577	5	-	-	0.011
	TcSNP513	4	*Tc04v2_g019130* Putative Protein MARD1 (28025910..28029002)	NAD(P)H-quinone oxidoreductase subunit H, putative isoform 1, response to glucose, response to sucrose	0.0081

Table 2. Putative candidate genes near top SNPs associated with fruit traits in cacao. Summary of the top SNPs associated with six fruit-associated traits identified through the Bootstrap Forest-based GWAS, along with their nearest genes and associated GO annotations. A dash (-) indicates that no information was available for that particular cell.

Genome-Wide association study of seed traits, wet bean mass, and pod index. As a further step in dissecting the complex genetic architecture of cacao, we performed a Bootstrap Forest based GWAS focusing on seed characteristics, wet bean mass, and pod index as a measure of yield potential with nine important traits: seed number, seed cotyledon color, cotyledon mass, cotyledon length, cotyledon width, cotyledon length to width ratio, seed shape, wet bean mass, and pod index. The top-scored SNPs for each trait, along with their putative candidate genes and associated GO terms, are detailed in Figure 5 and Table 3.

Analysis of pod index revealed top-ranked SNPs on chromosomes 1, 2, and 8 (TcSNP1213, TcSNP597, and TcSNP642, respectively). These SNPs are located near genes encoding a cysteine protease RD21B (*Tc01v2_g005450*), 21 kDa proteinase inhibitor seed protein (*Tc02v2_g004390),* and peroxidase 12 (*Tc08v2_g015120*), respectively, all of which could potentially influence pod development or seed-related traits. For seed number, a top-ranked SNP (TcSNP1431) was found on chromosome 1, near a gene (*Tc01v2_g034670*) encoding an ADP, ATP carrier protein 1, potentially involved in energy metabolism during seed development.

A top-ranked SNP for cotyledon mass (TcSNP17) was identified on chromosome 2, in proximity to a gene (*Tc02v2_g029690*) encoding a RING-type E3 ubiquitin transferase, which may play a role in protein degradation during seed development. The top SNPs associated with additional traits are detailed in Supplementary Figure S1 and Table S1. Collectively, these findings provide a comprehensive set of candidate genes that may contribute to variation in seed characteristics, yield potential, and other complex traits in cacao.

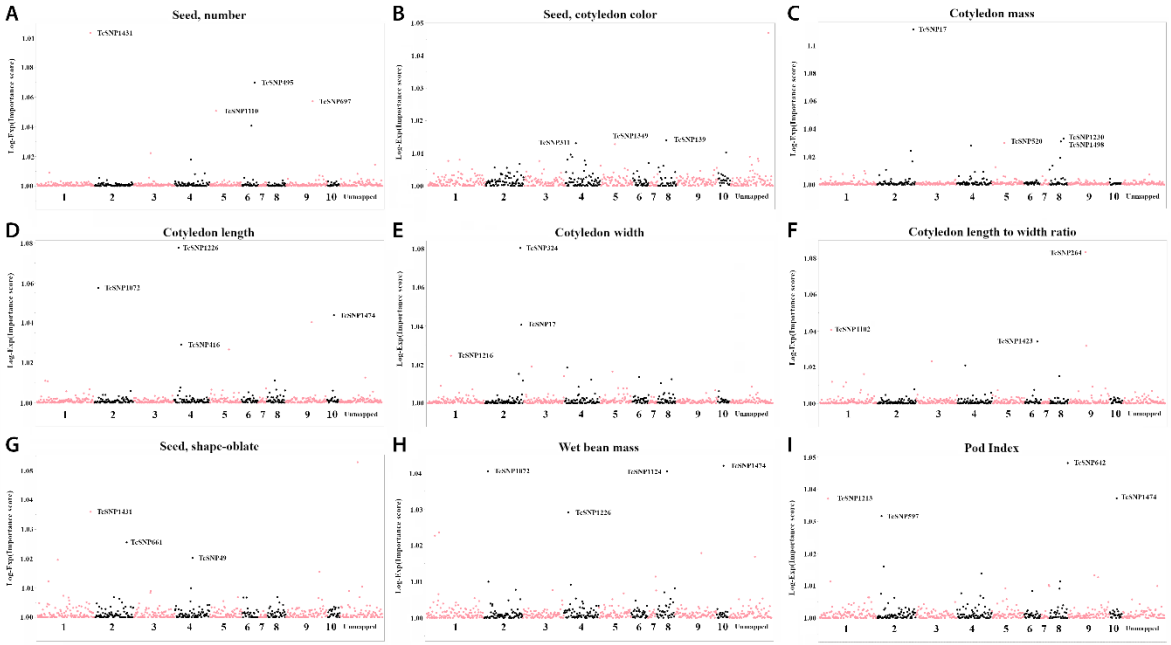


Fig 5. GWAS results for nine seed and pod traits in cacao, based on Bootstrap Forest analysis. Each point represents a SNP, plotted by its position on the

281 chromosome (x-axis) and its importance score (Log-Exp(Portion)) on the y-axis. The
282 top SNPs associated with each trait are labeled with their corresponding IDs.

283

Trait	SNP ID	Chr	Nearest genes and function	GO annotation	Impartance score (portion)
Seed, number	TcSNP1431	1	*Tc01v2_g034670* ADP,ATP carrier protein 1, mitochondrial (37067714..37070950)	-	0.098
	TcSNP495	6	*Tc06v2_g017610* Postsynaptic protein-related isoform 1 (25060968..25062477)	Cysteine-rich PDZ-binding protein, cytoplasmic microtubule organization	0.068
	TcSNP697	9	-	-	0.056
	TcSNP1110	5	*Tc05v2_g003770* THO complex subunit 4A (1968735..1971420)	-	0.05
Seed, cotyledon color	TcSNP139	8	*Tc08v2_g007790* Putative Cytochrome B5 isoform D (4514410..4515413)	-	0.014
	TcSNP311	4	*Tc04v2_g005940* AAA-ATPase ASD, mitochondrial (11953136..11955085)	-	0.013
	TcSNP1349	5	*Tc05v2_g007090* Fructose-bisphosphate aldolase cytoplasmic isozyme (4739626..4741948)	Fructose-bisphosphate aldolase activity, glycolytic process	0.013
Cotyledon mass	TcSNP17	2	*Tc02v2_g029690* RING/U-box superfamily protein, putative (40147535..40148635)	RING-type E3 ubiquitin transferase, regulation of protein complex stability, metal ion binding, protein polyubiquitinatio n	0.11
	TcSNP1230	8	*Tc08v2_g010820* Protein disulfide-isomerase (6639537..6644666)	-	0.032

	TcSNP1498	8	*Tc08v2_g008730* 14-3-3-like protein (5058053..5059576)	General regulatory factor 2, OMEGA, signal transduction	0.03
	TcSNP520	5	*Tc05v2_g006570* Threonine dehydratase biosynthetic, chloroplastic (4133868..4138898)	-	0.03
Cotyledon length	TcSNP1226	4	*Tc04v2_g001730* S tudor-like domain (1253806..1259080)	Transcription	0.075
	TcSNP1072	2	*Tc02v2_g004580* Probable aquaporin PIP2-8 (2902588..2904258)	Plasma membrane intrinsic protein 2,8 isoform 1	0.056
	TcSNP1474	10	-	-	0.043
	TcSNP416	4	*Tc09v2_g022340* Protein STRICTOSIDINE SYNTHASE-LIKE 3 (32687382..32691491)	-	0.04
Cotyledon width	TcSNP324	2	*Tc02v2_g026800* Inositol oxygenase 1 (38247661..38250652)	-	0.077
	TcSNP17	2	*Tc02v2_g029690* RING/U-box superfamily protein, putative (40147535..40148635)	RING-type E3 ubiquitin transferase, regulation of protein complex stability, metal ion binding, protein polyubiquitinatio n	0.04
	TcSNP1216	1	*Tc01v2_g014800* Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein isoform 1 (11219924..11221556)	-	0.024
Cotyledon length to width ratio	TcSNP264	9	*Tc09v2_g013370* Tetraspanin-8 (9508867..9510826)	-	0.08
	TcSNP1102	1	*Tc01v2_g006960* Cytochrome c oxidase subunit 6a, mitochondrial (3956725..3959234)	-	0.04

	TcSNP1423	6	-	-	0.034
Seed, shape-oblate	TcSNP1431	1	*Tc01v2_g034670* ADP, ATP carrier protein 1, mitochondrial (37067714..37070950)	-	0.052
	TcSNP661	2	*Tc02v2_g022140* Ubiquitin-40S ribosomal protein S27a (32406042..32406768)	Ribosomal protein S27a / Ubiquitin family protein, modification-dependent protein catabolic process	0.035
	TcSNP49	4	*Tc04v2_g014230* Probable fructokinase-4 (24603163..24606710)	-	0.025
Wet bean mass	TcSNP1474	10	-	-	0.041
	TcSNP1072	2	*Tc02v2_g004580* Probable aquaporin PIP2-8 (2902588..2904258)	Plasma membrane intrinsic protein 2,8 isoform 1	0.04
	TcSNP1124	8	*Tc08v2_g008690* Cytochrome c oxidase subunit 6b-1 (5033275..5037237)	-	0.04
	TcSNP1226	4	*Tc04v2_g001730* S tudor-like domain (1253806..1259080)	Transcription	0.029
Pod Index	TcSNP642	8	*Tc08v2_g015120* Peroxidase 12 (16398271..16401309)	-	0.047
	TcSNP1474	10	-	-	0.037
	TcSNP1213	1	*Tc01v2_g005450* Probable cysteine protease RD21B (3027564..3029931)	-	0.036
	TcSNP597	2	*Tc02v2_g004390* 21 kDa kunitz proteinase inhibitor seed protein (2756812..2757680)	-	0.031

284 **Table 3. Putative candidate genes near top SNPs associated with seed traits, wet bean**
285 **mass, and pod index in cacao.** Summary of the top SNPs associated with six fruit traits

identified through the Bootstrap Forest-based GWAS, along with their nearest genes and associated GO annotations. A dash (-) indicates that no information was available for that particular cell.

Correlation of SNP importance scores across traits

To understand whether similar SNPs were associated with multiple traits, we performed a correlation analysis of the SNP importance scores (Portion) derived from the Bootstrap Forest-based GWAS. Pairwise Pearson's correlation coefficients were calculated between the Portion scores of all SNPs for each pair of traits. The resulting correlation matrix is visualized as a heatmap in Figure 6, and the detailed correlation coefficients and associated *p*-values are presented in Supplementary Data 1. The heatmap reveals a complex pattern of correlations between the importance scores for different traits. Notably, the analysis revealed significant positive correlations between the importance scores for accession group and several other traits. For example, accession group showed a strong positive correlation with cotyledon length (FR_BL; $r = 0.46$, $p < 0.0001$) and sepal length (FL_SEP_LEN; $r = 0.25$, $p < 0.0001$). A strong positive correlation was also observed between seed number (FR_BEN) and ovule number (FL_OV_NUM) ($r = 0.60$, $p < 0.0001$). Fruit basal constriction (FR_PBC) and fruit length to width ratio (PODL_W) exhibited a positive correlation ($r = 0.50$, $p < 0.0001$). In addition to these strong correlations, numerous other pairs of traits exhibited moderate positive correlations, as evident in the heatmap. These findings suggest that some SNPs may have pleiotropic effects, influencing multiple traits, or that shared genetic pathways may underlie the observed associations (Supplementary data 1).

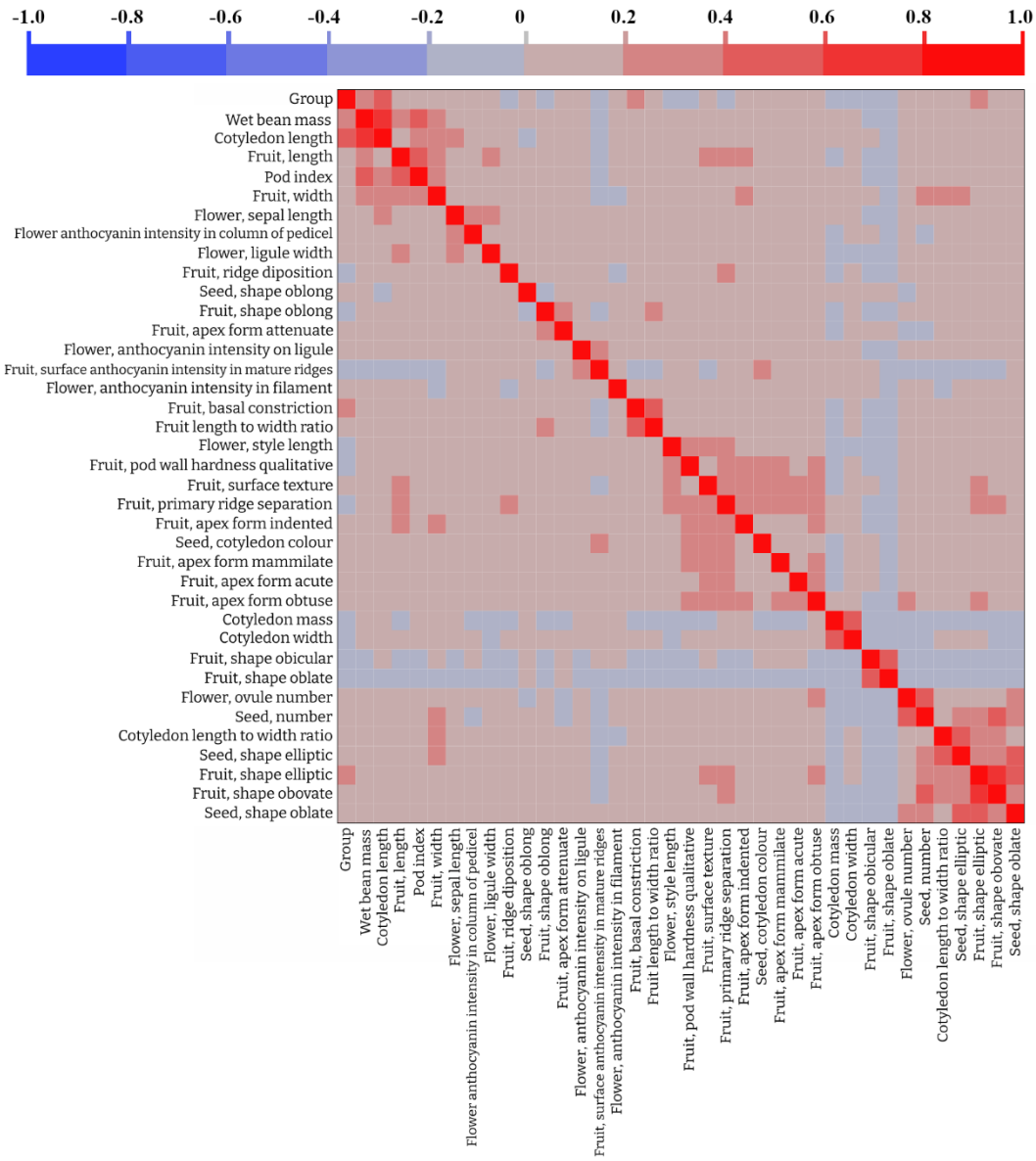


Fig 6. Correlation analysis of SNP importance scores across all traits, derived from the Bootstrap Forest-based GWAS. The heatmap displays the pairwise Pearson's correlation coefficients between the scores of all SNPs for each pair of traits. The color gradient ranges from red for positive correlations to blue for negative correlations, with darker shades indicating stronger correlations. This analysis reveals the extent to which similar SNPs are associated with multiple traits, suggesting potential pleiotropic effects or shared genetic pathways.

Hierarchical clustering of SNP loci and traits based on importance scores

To elucidate the relationships between traits and their underlying genetic architecture, we conducted hierarchical clustering analysis using SNP importance scores (Portion) derived from the Bootstrap Forest GWAS. Figure 7 presents the resulting clusters for both SNP loci and traits. The heatmap (Figure 7) visualizes the trait values for each SNP across all traits, with a color gradient from red (high values) to blue (low values). The right-side dendrogram

displays the hierarchical clustering of SNPs based on their importance scores across all traits, using Ward's method. This analysis resolved 20 distinct SNP clusters. The bottom dendrogram depicts the clustering of traits based on their importance score profiles across the SNP set. Several noteworthy trait clusters emerged from the analysis (bottom dendrogram, Figure 7). One cluster includes cotyledon length-to-width ratio (BENL_W), seed shape-elliptic (FR_BES_elliptic), fruit shape-obovate (FR_POSH_obovate), fruit shape-elliptic (FR_POSH_elliptic), and seed shape-oblate (FR_BES_oblate). This clustering suggests that shape-related morphological traits may share common genetic influences. Pod index was found to cluster closely with fruit length (FR_POL) followed by fruit width (FR_POW), which may be expected if fruit dimensions influence pod index within this dataset. Interestingly, cotyledon mass (FR_BW) and seed number (FR_BEN) are located further from pod index in the dendrogram, despite being primary variables in pod index calculations, implying different genetic influences affecting these traits. A final prominent cluster comprises cotyledon length (FR_BL) and total wet bean mass (FR_WABW), which are also positioned near the accession group, indicating a potential genetic link between these traits and the genetic background of the accessions. The clusters of SNP loci and traits revealed in this analysis provide valuable insights into the complex genotype-phenotype relationships within this extensive cacao collection. The distinct SNP clusters likely represent genomic regions influencing specific trait combinations, while the trait clusters suggest potential pleiotropic effects or shared underlying genetic pathways.

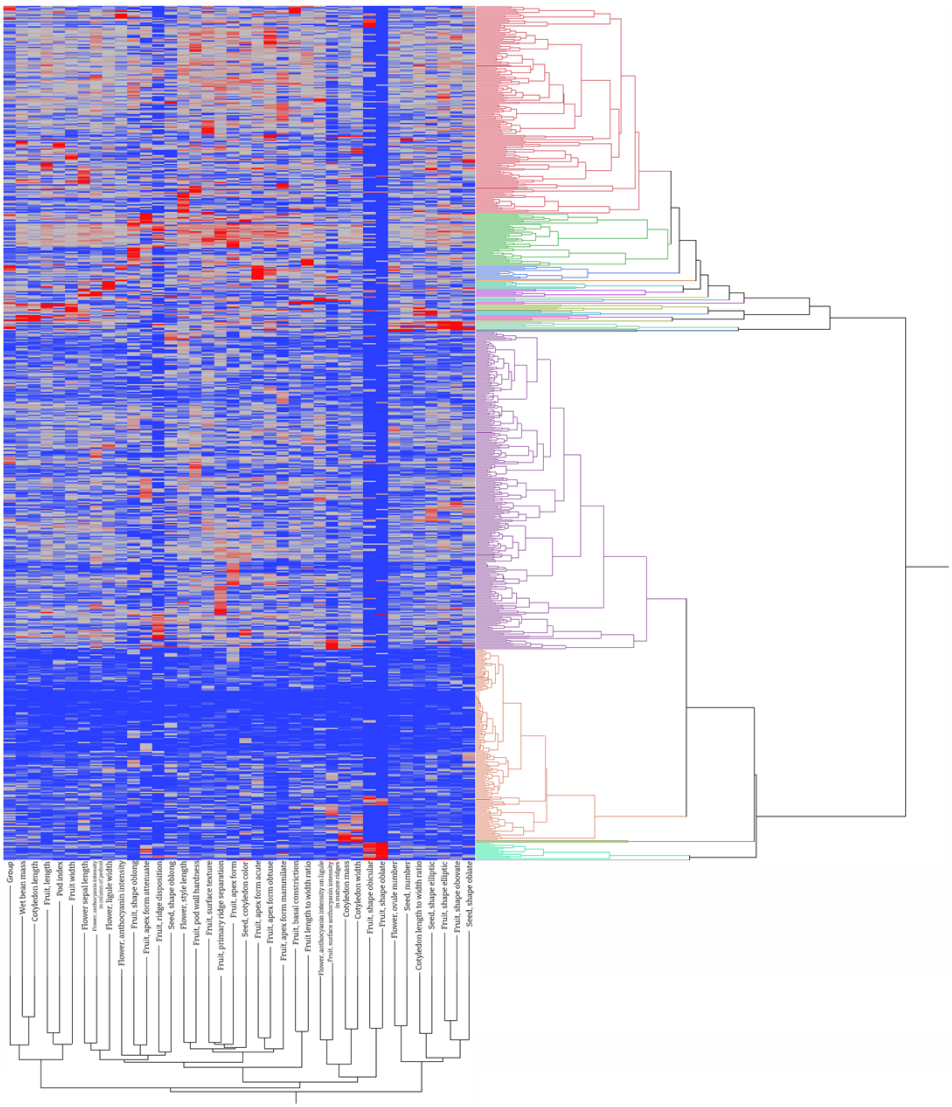


Fig 7. Hierarchical clustering of SNP loci and traits based on their importance, derived from a Bootstrap Forest-based GWAS. The heatmap visualizes trait values using a color gradient, ranging from red (high) to blue (low). The right-side dendrogram illustrates the clustering of SNP loci based on their importance scores (portion values) using Ward's method, forming 20 distinct groups where each group is represented by a different color branch. The bottom dendrogram depicts the clustering of traits according to their importance score profiles across all SNP loci.

Predictive modeling of wet bean mass using a Neural Network model

Pod index is a key trait reflecting yield potential in cacao, however it is an indirect score calculated using dried cotyledon mass and average seed number which can confound phenotypic associations. Wet bean mass provides a more direct assessment of yield, so we employed a machine learning approach using neural network model to identify predictive phenotypic traits while excluding obvious correlating traits such as seed number, fruit length, fruit width, and fruit length to width ratio. The performance of the trained Neural Boosted

model is shown in Figure 8. The model exhibited strong predictive accuracy on both the training set (R-squared = 0.790 +/- .08) with and the validation set (R-squared = 0.715 +/- .12). The scatter plots of actual vs. predicted wet bean mass values show a close alignment of the points with the diagonal line representing perfect prediction, indicating that the model accurately captures the relationship between the predictor variables and wet bean mass. Further analysis using the Neural Boosted model identified cotyledon mass as the primary driver of wet bean mass, with a remarkably high importance score of 6 (Table 4). Cotyledon length emerged as the second most influential predictor, receiving an importance score of 2. Cotyledon length-to-width ratio showed slight predictive power with an importance score of 1, whereas cotyledon width and other phenotypic traits showed little to no predictive influence (Table 4). Cotyledon mass displayed a substantial positive influence on wet bean mass, evidenced by its main effect of 0.484 and total effect of 0.5198. This suggests that cotyledon mass impacts wet bean mass both directly and indirectly through interactions with other variables. Cotyledon mass (portion = 0.520) and cotyledon length (portion = 0.213) accounted for the vast majority (73.3%) of the predictive power in the model, highlighting their dominant roles in determining wet bean mass.

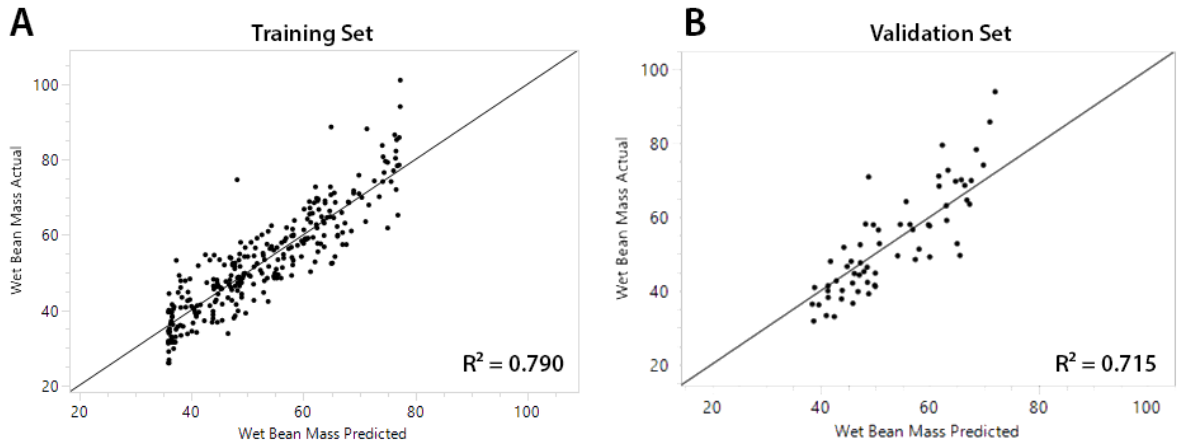


Fig 8. Evaluation of the Neural Boosted model for predicting wet bean mass in cacao, based on phenotypic and genotypic data. (a) Actual vs. predicted values for the training set. (b) Actual vs. predicted wet bean mass values for the validation set. The black line in each plot represents a perfect prediction (actual = predicted). The proximity of the points to the line indicates the accuracy of the model's predictions. The model's performance was assessed using the coefficient of determination (R-squared) (n= 283 for the training set, and n= 63 for the validation set). Plots are the averages of 20 validation columns obtained from random seed 20 and 13, respectively.

Trait	Main effect	Total effect	Importance score
Cotyledon mass	0.4843	0.5198	6

Cotyledon length	0.1864	0.2133	2
Cotyledon length to width ratio	0.0867	0.1101	1
Cotyledon Width	0.04815	0.06895	0

Table 4. Most important traits for wet bean mass prediction based on Neural Boosted model analysis. Relative importance of the top five traits in predicting wet bean mass, as determined by the Neural Network model trained on 346 cacao accessions and validated on 116 accessions. The table lists the trait, its main effect, total effect, and importance score. The importance score reflects the contribution of each trait to the model's predictive accuracy. Higher importance scores indicate a stronger influence on wet bean mass prediction. Cotyledon mass had the highest importance score as the most important variable.

Discussion

This study employed a comprehensive approach, integrating phenotypic data, phylogenetic analysis, dimensionality reduction, genome-wide association studies, cluster analysis, and machine learning, to dissect the complex relationships between genotype and phenotype in a diverse collection of 346 cacao accessions. The weak overall correlation observed between patristic distances and most phenotypic traits suggests that broad phylogenetic relationships, based on genome-wide SNP diversity, are not the primary determinants of phenotypic variation in this diverse collection of cacao. This finding aligns with the work of [26], who demonstrated that phylogenetic signal, a measure of the statistical dependence among species' trait values due to shared ancestry, is not directly related to the rate of evolution under simple models of genetic drift, but rather is more strongly influenced by the underlying evolutionary process. Indeed, different evolutionary processes can produce similar levels of phylogenetic signal [26]. Furthermore, our results highlight the limitations of likelihood-based approaches for comparing evolutionary models in highly multivariate datasets, particularly as the dimensionality of the data increases. It was found that the estimation of the likelihood becomes increasingly unstable as the number of trait dimensions grows, due to the evolutionary covariance matrix (V) becoming more ill-conditioned [27]. This "curse of dimensionality" renders likelihood-based model comparisons unreliable when the number of traits approaches or exceeds the number of species, a common scenario in many morphometric studies, and a situation not uncommon in genomic analyses as well [27,28]. These computational challenges are consistent with the complex genetic architecture of many quantitative traits, where multiple loci with small effects, rather than a few major genes, often underlie phenotypic differences [18,29–31]. Despite the generally weak correlation between patristic distance and most phenotypic traits, we observed significant associations for specific traits. Notably, pod index exhibited a positive correlation with patristic distance ($r = 0.38$, $p < 0.0001$), while total wet bean mass ($r = -0.33$, $p < 0.0001$), seed number ($r = -0.30$, $p = 0.0005$), and cotyledon characteristics (mass: $r = -0.21$, $p = 0.0001$; length: $r = -0.26$, $p < 0.0001$; and width: $r = -0.22$, $p < 0.0001$) showed negative correlations. These associations

426 suggest that genetic relatedness does play a role in shaping variation in these ecologically
427 relevant traits, potentially reflecting the influence of specific genomic regions with larger
428 effects that are, to some extent, captured by the phylogenetic analysis [32]. These regions
429 may be under different selective pressures, potentially driving local adaptation to specific
430 environments, or subject to genetic constraints. Furthermore, the t-SNE analysis provided a
431 complementary perspective, revealing a complex pattern of variation within the cacao
432 collection. When combining phenotypic and genotypic data, the t-SNE plot (Fig. 2c) suggests
433 the presence of roughly four clusters, although with some overlap and within-group variation.
434 While predefined accession groups exhibited some degree of clustering, particularly for
435 groups 7, 8, 14, and 15, the substantial within-group variation and the overlap between
436 several other groups underscore the limitations of relying solely on broad genetic
437 classifications for understanding phenotypic diversity. This continuous variation is likely a
438 consequence of the complex interplay of multiple genetic factors, environmental influences,
439 and the history of hybridization and selection in cacao [18,29,31,33–35].

440 Building upon the insights gained from the initial analyses, the GWAS provided a more
441 detailed view of the genetic architecture of key traits in cacao. By employing a machine
442 learning approach based on Bootstrap Forest models, we were able to identify numerous
443 SNPs that exhibited significant associations with flower, fruit, and seed traits, as well as pod
444 index (Figures 3, 4, 5, and Supplementary Figure S1). These associations pinpoint specific
445 genomic regions that may harbor genes influencing these traits. For example, the analysis of
446 flower traits revealed significant SNPs near genes encoding proteins involved in various
447 biological processes, including carbohydrate metabolism, stress response, and signal
448 transduction (Table 2).

449 The SNP most strongly associated with anthocyanin content in the ligule and mature ridges,
450 TcSNP174, is located near a gene (Tc04v2_g009180) encoding a 6-phosphogluconate
451 dehydrogenase. This enzyme participates in the oxidative pentose phosphate (OPP) pathway,
452 a fundamental metabolic route that generates reducing power (NADPH) and precursors for
453 various biosynthetic pathways. Among these is the shikimate pathway, which is responsible
454 for the synthesis of aromatic amino acids, including phenylalanine . The shikimate pathway is
455 a major branch point in plant metabolism, channeling a significant portion of fixed carbon
456 (20-50%) into the production of essential amino acids and a vast array of secondary
457 metabolites [36]. Phenylalanine serves as a crucial precursor for the biosynthesis of
458 flavonoids, including anthocyanins, the pigments responsible for red to blue coloration in
459 many plants [37]. Recent work by Tang et al. demonstrated that 6-phosphogluconate
460 dehydrogenase 2 (PGD2) can serve as a link between the OPP and shikimate pathways in
461 Arabidopsis, influencing the production of aromatic amino acids [38]. The analysis of fruit
462 traits identified SNPs near genes with diverse putative functions, including those involved in
463 protein turnover, calcium signaling, heat response, and cell wall modification (Table 2).
464 Notably, the top SNP for fruit length and pod index, TcSNP642, is located near a gene
465 (*Tc08v2_g015120*) encoding a peroxidase 12 (Table 2 & Table 3). These plant-specific Class
466 III peroxidases (CIII Prxs) are predominantly localized to the cell wall and play diverse roles
467 in cell wall loosening and stiffening, influencing processes like cell elongation and overall

468 plant development [39]. *Tc08v2_g015120* could be involved in cell wall modifications during
469 cacao fruit expansion, potentially contributing to the observed variation in fruit length and,
470 consequently, affect fruit yield.

471 The GWAS of seed traits, wet bean mass, and pod index revealed associations with SNPs
472 near genes involved in protein synthesis, lipid metabolism, and stress response (Table 3). For
473 example, a significant SNP for seed number and seed shape (oblate) (TcSNP1431) is located
474 near a gene (*Tc01v2_g034670*) encoding an ADP, ATP carrier protein, highlighting the
475 importance of energy metabolism for seed development. These proteins facilitate the
476 transport of ATP and ADP across the mitochondrial membrane, playing a crucial role in
477 maintaining cellular energy levels. It was demonstrated that ATP levels in germinating seeds
478 are closely related to seed vigor, with higher ATP levels correlating with better germination
479 and seedling establishment [40]. Perl suggested that ATP accumulation in seeds is a complex
480 interplay between synthesis and utilization, and that high ATP levels can result from either
481 high synthesis (in high-quality seeds) or low utilization (in low-quality seeds) [41]. Thus, the
482 association between TcSNP1431 and seed number suggests that variation in the efficiency of
483 ADP/ATP transport, potentially influenced by this SNP, could impact ATP availability
484 during cacao seed development, ultimately affecting seed viability and number.

485 Several top SNPs were located near genes with potential roles in pod development and
486 morphology. These include Tc08v2_g015120, encoding a peroxidase 12, *Tc01v2_g005450*,
487 encoding a probable cysteine protease (RD21B), and *Tc02v2_g004390*, encoding a 21 kDa
488 proteinase inhibitor seed protein. Again, peroxidases are known to participate in cell wall
489 modifications, including both loosening and stiffening, depending on the specific isoform and
490 context [39]. These modifications are not only crucial for normal cell growth and expansion
491 but also for plant responses to abiotic stresses like drought or salinity [42]. Under stress
492 conditions, reactive oxygen species (ROS) production is often increased, and peroxidases, in
493 concert with ROS, can mediate cell wall remodeling, either by cross-linking cell wall
494 components or by inducing cell wall loosening through the action of hydroxyl radicals [42].
495 Given the importance of cell wall remodeling during fruit growth and expansion, it is
496 plausible that peroxidase 12 contributes to pod size and shape, ultimately influencing yield
497 traits such as pod index which is directly affected by pod size and seed number. Cysteine
498 proteases, such as RD21B, play diverse roles in plant development and responses to biotic
499 and abiotic stresses [43–45]. In the context, RD21B might be involved in tissue remodeling
500 or defense responses that indirectly affect pod morphology and, consequently, pod index.

501 The 21 kDa protein encoded by *Tc02v2_g004390* is identified as a proteinase inhibitor and
502 seed storage protein. While its precise function in cacao pod development is unknown, it
503 likely plays a role in protection during animal consumption as well as a role in providing
504 nutrients for seedling development. Other seed storage proteins have also been shown to
505 possess antimicrobial activities and may contribute to plant defense mechanisms [46]. For
506 instance, some storage proteins exhibit insecticidal properties or can inhibit the growth of
507 phytopathogenic fungi [46]. Furthermore, our Bootstrap Forest analysis identified numerous
508 other candidate genes associated with the various traits examined in this study. Interestingly,

509 we observed only one overlapping gene, *Tc04v2_g019280*, between our study and a previous
510 GWAS [11]. This gene, which encodes a vicilin-like seed storage protein, was associated
511 with fruit apex form in our analysis and with pod index in the previous study, which
512 employed TASSEL MLM and FarmCPU approaches. The limited overlap between the two
513 studies, along with the different traits associated with *Tc04v2_g019280*, could be attributed to
514 differences in the GWAS methodologies used (Bootstrap Forest vs. TASSEL MLM and
515 FarmCPU) or might reflect the complex genetic architecture of these traits, potentially
516 involving numerous loci with small individual effects and/or epistatic interactions. The
517 numerous other genes identified through our Bootstrap Forest analysis, although not
518 discussed in detail here, represent a valuable resource for future investigations.

519 The hierarchical clustering analysis based on SNP importance scores provided further
520 insights into the relationships among traits and the potential genetic architecture underlying
521 trait variation (Figure 7). The 20 distinct clusters of SNP loci identified likely represent
522 groups of genetically linked variants that influence specific combinations of traits. The notion
523 of pleiotropy and shared genetic pathways is further supported by the correlation analysis of
524 SNP importance scores across obvious correlating traits (Figure 6). For example, significant
525 positive correlations were observed between the importance scores for seed number and
526 ovule number ($r = 0.60$, $p < 0.0001$) which are expectedly interrelated. Additionally, pod
527 index correlates with total wet bean mass ($r = 0.6$, $p < 0.0001$) which provide credence
528 towards their non-obvious phenotypic association with fruit length ($r = 0.5$, $p < 0.0001$)
529 (Figure 6, Supplementary Data 1). Another example of obviously correlating phenotypic
530 traits that share a genetic underpinning is the close relationship between pod index, fruit
531 length, and fruit width in the trait dendrogram (Figure 7) which corroborates the importance
532 of these fruit characteristics as key determinants of yield potential. The clustering of traits
533 such as cotyledon length and total wet bean mass, along with their proximity to the accession
534 group in the dendrogram (Figure 7), supports the notion that these traits may be potentially
535 influenced by common genetic factors related to the genetic background of the accessions.
536 Overall, hierarchical clustering of phenotypic traits indicate how underlying SNPs influence
537 multiple traits simultaneously in both obvious and non-obvious ways that can be used to
538 inform marker-assisted selection.

539 The successful prediction of wet bean mass using a Neural Network model (Figure 8), well-
540 suited for capturing complex, non-linear relationships often found in biological systems [47],
541 demonstrates the potential of leveraging phenotypic data for developing predictive tools for
542 complex traits in cacao. The model's high accuracy (R-squared = 0.790 for training, 0.715 for
543 validation) highlights the strong relationships between the predictor variables and wet bean
544 mass. Notably, cotyledon mass and cotyledon length were identified as the two most
545 important predictors (Table 4) of wet bean mass, a direct measure of yield. Larger cotyledons
546 on young cacao seedlings may be indicators of vigor and higher photosynthetic capacity
547 which over the long-term result in more productive cacao trees. This suggests that simple
548 measurement of these two traits could provide a surprisingly accurate, yet cost-effective,
549 method for predicting productivity in cacao, which has significant implications for breeding
550 programs which suffer from an inability to screen for productive clones. Targeting specific

551 combinations of traits, guided by the identified clusters (Figure 7) and the machine learning
552 model (Figure 8), could lead to even more efficient selection strategies for improving yield
553 potential. Furthermore, these results, combined with our GWAS findings, suggest that
554 genomic selection, incorporating information from multiple loci across the genome, could be
555 a powerful tool for accelerating cacao improvement.

556 The findings of this study, particularly the successful application of a Neural Network model
557 for predicting wet bean mass, provide a springboard for several promising avenues of future
558 research in cacao. Given the potential impact on breeding strategies, prioritizing the
559 validation of candidate genes identified through the GWAS would be beneficial. This could
560 involve gene expression studies, functional analyses using transgenic approaches, or
561 CRISPR-Cas9-mediated gene editing to assess the effects of altering gene expression or
562 protein function on phenotypic variation. Second, expanding the association mapping and
563 genomic selection studies to larger and more diverse populations of cacao, including different
564 genetic groups and geographic origins, would help to confirm the identified associations and
565 identify additional loci contributing to trait variation across a wider range of germplasm.
566 Applying advanced machine learning methods to these larger datasets could further enhance
567 the accuracy and efficiency of genomic selection in cacao. This could also involve exploring
568 genotype-by-environment interactions by evaluating the performance of diverse accessions in
569 multiple locations in collaboration with international partners. Third, investigating the role of
570 other types of genetic variants, such as copy number variations and structural variants, could
571 be integrated into machine learning models to improve their predictive power. Fourth,
572 integrating the genomic data with other omics data, such as transcriptomic, proteomic, and
573 metabolomic profiles, leveraging advanced mass spectrometry and sequencing platforms,
574 could further elucidate the molecular mechanisms underlying trait variation and identify key
575 regulatory networks. These multi-omics datasets could then be used to train more complex
576 and comprehensive machine learning models capable of capturing the intricate interplay of
577 genetic, environmental, and molecular factors that influence complex traits. Finally, the
578 development of more sophisticated machine learning models, potentially incorporating gene-
579 gene and gene-environment interactions, along with feature selection algorithms to identify
580 the most informative predictors, could further enhance the accuracy of predicting complex
581 traits and facilitate the identification of superior genotypes for breeding programs. Finally,
582 exploring different neural network architectures, such as deep learning models, could provide
583 even greater accuracy in predicting complex traits from high-dimensional genomic and
584 phenotypic data.
585

586 # Conclusion

587

588 This study comprehensively analyzes the phenotypic and genotypic variation in a diverse
589 collection of cacao accessions, revealing complex relationships between genetic relatedness,
590 population structure, and key traits related to flower and fruit morphology, seed
591 characteristics, and yield potential. The t-SNE analysis highlighted the continuous nature of
592 variation in the collection, while the GWAS identified numerous SNP-trait associations and
593 potential candidate genes involved in determining these traits. The hierarchical clustering

analysis further illuminated the relationships among traits and suggested the existence of shared genetic factors influencing specific trait combinations. The development of a highly accurate machine-learning model for predicting wet bean mass demonstrates the potential of integrating phenotypic and genotypic data for enhancing cacao breeding strategies. These findings contribute significantly to our understanding of the genetic architecture of important traits in cacao and provide valuable resources for future research and improvement efforts aimed at developing superior cacao varieties with enhanced yield potential and other desirable characteristics.

Materials and methods

Phenotype and genotype data

Phenotypic and genotypic data used in this study were obtained from a publicly available, large-scale assessment of cacao germplasm detailed in [11]. The study utilized 421 cacao accessions from the ICGT, a diverse collection that includes 263 wild genotypes, 23 accession groups (e.g.: wild cacao types such as AMAZ, GU, IMC, MO, NA, PA, POUND, RB, SCA and SPEC accession groups), and most of the genetic groups described by [11,48].

Phenotypic evaluations for 27 traits, encompassing flower, fruit, and seed characteristics, were derived from the dataset published by [11]. Flower traits included anthocyanin intensity in the column of the pedicel (FL_PCOL_C), sepal length (FL_SEP_LEN), anthocyanin intensity on the ligule (FL_LIG_COL), ligule width (FL_LIG_WID), anthocyanin intensity in the filament (FL_FIL_COL), style length (FL_STY_LEN), and ovule number (FL_OV_NUM). Fruit traits included shape (FR_POSH), basal constriction (FR_PBC), apex form (FR_PAF), surface texture (FR_PST), surface anthocyanin intensity in mature ridges (FR_PRCM), ridge disposition (FR_PFD), primary ridge separation (FR_PFS), pod wall hardness (FR_HH), fruit length (FR_POL), fruit width (FR_POW), and fruit length to width ratio (FR_PODL_W). Seed traits included number (FR_BEN), shape (FR_BES), cotyledon color (FR_BEC), total wet bean mass (FR_WABW), cotyledon length (FR_BL), cotyledon width (FR_BEW), cotyledon mass (FR_BW), and cotyledon length to width ratio (FR_BENL_W). Pod index, a key agronomic trait that reflects yield potential in cacao, was also included as a composite trait [11].

Complete phenotypic data were available for 346 of the 421 accessions [11]. For the genotypic data, Bekele et al. targeted a total of 836 SNPs located in coding sequences, selected based on their similarity to known protein sequences. After excluding SNPs with low call rates, 671 high-quality SNPs were retained. Consequently, the final dataset used for analysis in this study consisted of 671 SNPs and phenotypic data for 346 accessions.

Analysis of genetic relatedness and phenotypic variation

To investigate the relationship between genetic relatedness and phenotypic variation, we calculated pairwise patristic distances between all cacao accessions. The SNP data were imported into Geneious Prime (version 2023.0.3), and a distance matrix was generated using the Tamura-Nei genetic distance model [49]. A phylogenetic tree was constructed using the Neighbor-Joining method to visualize these relationships, with alignment type set to global

alignment with free end gaps and the cost matrix set to 65% similarity [50]. Patristic distances (the sum of branch lengths separating two taxa on the tree) were then calculated in Geneious Prime. These patristic distances, representing genetic divergence, were correlated with log-transformed phenotypic data for all traits using JMP Pro 17 software (SAS Institute Inc., Cary, NC) [51]. Pearson's correlation coefficients were computed for each pair of patristic distance and a phenotypic trait.

t-distributed stochastic neighbor embedding analysis

To visualize and explore the underlying structure of the phenotypic and genotypic variation among the cacao accessions, t-distributed Stochastic Neighbor Embedding (t-SNE) was employed through JMP Pro 17. t-SNE is a non-linear dimensionality reduction technique that visualizes high-dimensional data in fewer dimensions while preserving local relationships [52]. Separate t-SNE analyses were performed on the phenotypic data (the all phenotypic traits measured for each accession), the genotypic data (the 671 SNP genotype data for each accession), and the combined phenotypic and genotypic data. For all t-SNE analyses, the following parameters were used: output dimensions were set to 3, perplexity was 200, maximum iterations were 10,000, initial principal component dimensions were 100, convergence criterion was 0.0001, initial scale was 0.0001, Eta (learning rate) was 200, inflate iterations were 250, and the random seed was 123.

Bootstrap Forest-based GWAS to associate SNPs with single traits

To identify genomic regions associated with phenotypic variation in the cacao accessions, particularly those influencing complex traits that might be difficult to detect using traditional GWAS methods, we employed a machine learning approach based on Bootstrap Forest models implemented in JMP Pro 17. The GWAS was conducted for all phenotypic traits, including those related to accession group, flower morphology, fruit morphology, seed characteristics, cotyledon and bean mass, and pod index. Pod index is a key agronomic trait that reflects yield potential, calculated as 1000/(average dried cotyledon mass x average seed/bean number) [11]. Importantly, it is a composite trait that takes into account both pod size and seed number, with lower pod index values indicating higher yield potential [11]. Separate Bootstrap Forest models were trained to predict each trait using the full set of SNP markers as predictors. These default settings included the following parameters for the Bootstrap Forest models: number of trees in the forest was set to 100, the number of terms sampled per split was 1, the bootstrap sample rate was 1, the minimum number of splits per tree was 10, the maximum number of splits per tree was 2,000, and the minimum size split was 5. The top SNPs associated with each trait were identified based on their importance scores, calculated as the "Portion" in JMP Pro 18. To explore potential candidate genes underlying the observed associations, the nearest genes to the top-ranked SNPs for each trait were identified, when available, using the Criollo v2 reference genome (GCF_000208745.1). The CocoaGen DB genome browser (http://cocoa-genome-hub.southgreen.fr/jbrowse) and UniProtKB (https://www.uniprot.org/uniprot/) were used as resources to gather information on gene annotations (GO) and putative functions.

Correlation analysis of GWAS results to identify associations among traits

Following the GWAS using Bootstrap Forest models, a correlation analysis was conducted to investigate the relationships between the importance scores of SNP markers across the different traits analyzed. As described in the previous section, the GWAS yielded an importance score (Portion) for each SNP for each trait, reflecting the contribution of that SNP to the prediction of the trait in the Bootstrap Forest model. To assess whether SNPs were associated with multiple traits, pairwise Pearson's correlation coefficients were calculated between the Portion scores of all SNPs for each pair of traits. The resulting correlation matrix provides a measure of the linear relationship between the SNP importance profiles of different traits. For example, a strong positive correlation between two traits would indicate that SNPs with high importance for one trait tend to also have high importance for the other, suggesting potential shared genetic influences or pleiotropic effects. To further investigate these relationships, we performed hierarchical clustering analysis using the Ward method based on the SNP importance scores (Portion) derived from the Bootstrap Forest models for all traits.

Development of machine learning model for predicting wet bean mass

To further investigate the complex relationships between the various measured traits and wet bean mass, a machine learning approach employing neural network models was utilized [53]. The primary goal was to train a predictive model for wet bean mass, a key agronomic trait in cacao, based on the suite of phenotypic characteristics measured in this study. Model selection and training were performed using JMP Pro 18. An initial model screening step was conducted using the built-in model screening function to evaluate the performance of various algorithms and identify the most promising model type for predicting pod index. The Neural Boosted model with settings NTanH(3) and NBoost(20) emerged as the best-performing model and was therefore selected for further analysis (data not shown). For model training, a validation column with a random seed associated with the column number was set and dataset was randomly partitioned into two subsets: 80.0% for training and 20.0% for validation. This partitioning was performed randomly with a random seed designation respective to the validation column. Training was done using 20 validation columns. The Neural Boosted model, with the specified NTanH(3) and NBoost(20) settings was then trained on the training set to predict wet bean mass using the other phenotypic traits as predictors. Obvious correlating phenotypic traits that would lead to spurious associations were excluded such as seed number, fruit length, fruit width, and fruit length to width ratio, as to identify non-obvious phenotypic predictors. The performance of the trained model was assessed using the coefficient of determination (R-squared) on both the training and validation sets. In addition to evaluating overall model performance, the importance of individual features (traits) in the Neural Boosted model was assessed to gain insights into the factors most strongly influencing pod index prediction.

Data Availability: All phenotypic and genotypic data analyzed in this study are publicly available. These data were originally generated and published by Bekele et al. (2022) and can be found within the Supporting Information files associated with that publication (PLoS One 17(1): e0260907), accessible via: https://doi.org/10.1371/journal.pone.0260907.

Funding: This work is supported by the U.S. Department of Agriculture, Agricultural Research Service, In-House Projects No. 8042-21220-258-000-D and 8042-21000-303-000-D.

Competing interests: The authors declare no competing interests.

Acknowledgements

We are also grateful to the reviewers for their constructive feedback. Mention of any trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U. S. Department of Agriculture. USDA is an equal opportunity provider and employer, and all agency services are available without discrimination.

References

1. Díaz-Montenegro J. Livelihood strategies and risk behavior of cacao producers in Ecuador: effects of national policies to support cacao farmers and specialty cacao landraces. 2019.

2. Hall JN. Applying a One Health approach to study the livelihoods of cocoa farming communities in Bougainville. 2022.

3. Bekele F, Phillips-Mora W. Cacao (Theobroma cacao L.) breeding. Adv Plant Breed Strateg Ind Food Crops Vol 6. 2019; 409–487.

4. Alden D. The significance of cacao production in the Amazon region during the late colonial period: an essay in comparative economic history. Proc Am Philos Soc. 1976;120: 103–135.

5. Gardea AA, García-Bañuelos ML, Orozco-Avitia JA, Sánchez-Chávez E, Sastré-Flores B, Ávila-Quezada G. Cacao (Theobroma cacao L.). Fruit Veg Phytochem Chem Hum Health 2nd Ed. 2017; 921–940.

6. Kongor JE, Owusu M, Oduro-Yeboah C. Cocoa production in the 2020s: Challenges and solutions. CABI Agric Biosci. 2024;5: 102.

7. Walters D. Chocolate crisis: Climate change and other threats to the future of cacao. University Press of Florida; 2020.

8. Boza EJ, Motamayor JC, Amores FM, Cedeno-Amador S, Tondo CL, Livingstone DS, et al. Genetic characterization of the cacao cultivar CCN 51: its impact and significance on global cacao improvement and production. J Am Soc Hortic Sci. 2014;139: 219–229.

9. Mustiga GM, Gezan SA, Phillips-Mora W, Arciniegas-Leal A, Mata-Quirós A, Motamayor JC. Phenotypic description of Theobroma cacao L. for yield and vigor traits from 34 hybrid families in Costa Rica based on the genetic basis of the parental population. Front Plant Sci. 2018;9: 808.

10. Izzah NK, Tresniawati C, Rubiyo R, Pranowo D, Randriani E, Firdaus NK, et al. Improvement of Cacao Pod Characteristics and its Molecular Characterization in 4 F1 Cacao Populations. 2024.

11. Bekele FL, Bidaisee GG, Allegre M, Argout X, Fouet O, Boccara M, et al. Genome-wide association studies and genomic selection assays made in a large sample of cacao (Theobroma cacao L.) germplasm reveal significant marker-trait associations and good predictive value for improving yield potential. Plos One. 2022;17: e0260907.

12. Bediako KA, Padi FK, Obeng-Bio E, Ofori A. Genetic diversity and parentage of cacao (Theobroma cacao L.) populations from Ghana using single nucleotide polymorphism (SNP) markers. Plant Genet Resour. 2025;23: 40–47.

13. Gutiérrez OA, Campbell AS, Phillips-Mora W. Breeding for disease resistance in cacao. Cacao Dis Hist Old Enemies New Encount. 2016; 567–609.

14. Rodriguez-Medina C, Arana AC, Sounigo O, Argout X, Alvarado GA, Yockteng R. Cacao breeding in Colombia, past, present and future. Breed Sci. 2019;69: 373–382.

15. Wickramasuriya AM, Dunwell JM. Cacao biotechnology: current status and future prospects. Plant Biotechnol J. 2018;16: 4–17.

16. McElroy MS, Navarro AJ, Mustiga G, Stack C, Gezan S, Peña G, et al. Prediction of cacao (Theobroma cacao) resistance to Moniliophthora spp. diseases via genome-wide association analysis and genomic selection. Front Plant Sci. 2018;9: 343.

17. Romero Navarro JA, Phillips-Mora W, Arciniegas-Leal A, Mata-Quirós A, Haiminen N, Mustiga G, et al. Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. Front Plant Sci. 2017;8: 1905.

18. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90: 7–24.

19. Omas-as AM, ARBOLEDA ER, DAANG JAM. Machine Learning as a Strategic Tool: A Comprehensive Literature Review for Advancing Agricultural Analysis, with Emphasis on the Cocoa Bean Quality Assessment. Int J Sci Res Eng Dev.

20. Tan J, Balasubramanian B, Sukha D, Ramkissoon S, Umaharan P. Sensing fermentation degree of cocoa (Theobroma cacao L.) beans by machine learning classification models based electronic nose system. J Food Process Eng. 2019;42: e13175.

21. Duarte-Carvajalino JM, Paramo-Alvarez M, Ramos-Calderón PF, González-Orozco CE. Estimation of canopy attributes of wild cacao trees using digital cover photography and machine learning algorithms. IForest-Biogeosciences For. 2021;14: 517.

22. Lamos-Díaz H, Puentes-Garzón D-E, Zarate-Caicedo D-A. Comparison between machine learning models for yield forecast in cocoa crops in santander, Colombia. Rev Fac Ing. 2020;29.

23. Fuentes S, Chacon G, Torrico DD, Zarate A, Gonzalez Viejo C. Spatial variability of aroma profiles of cocoa trees obtained through computer vision and machine learning modelling: A cover photography and high spatial remote sensing application. Sensors. 2019;19: 3054.

24. Collazos-Escobar GA, Bahamón-Monje AF, Gutiérrez-Guzmán N. Mid-infrared spectra of dried and roasted cocoa (Theobroma cacao L.): A dataset for machine learning-based classification of cocoa varieties and prediction of theobromine and caffeine content. Data Brief. 2025;58: 111243.

25. Breiman L. Random forests. Mach Learn. 2001;45: 5–32.

26. Revell LJ, Harmon LJ, Collar DC. Phylogenetic signal, evolutionary process, and rate. Syst Biol. 2008;57: 591–601.

27. Adams DC, Collyer ML. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. Syst Biol. 2018;67: 14–31.

826 28. Bellman R. Dynamic programming. science. 1966;153: 34–37.

827 29. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits:
828 from polygenic to omnigenic. Cell. 2017;169: 1177–1186.

829 30. Lindhout P. The perspectives of polygenic resistance in breeding for
830 durable disease resistance. Euphytica. 2002;124: 217–226.

831 31. Rockman MV. The QTN program and the alleles that matter for
832 evolution: all that's gold does not glitter. Evolution. 2012;66: 1–17.

833 32. Felsenstein J. Phylogenies and the comparative method. Am Nat.
834 1985;125: 1–15.

835 33. Mallet J. Hybridization as an invasion of the genome. Trends Ecol
836 Evol. 2005;20: 229–237.

837 34. Rieseberg LH. Hybrid origins of plant species. Annu Rev Ecol Syst.
838 1997;28: 359–389.

839 35. Via S, Gomulkiewicz R, De Jong G, Scheiner SM, Schlichting CD,
840 Van Tienderen PH. Adaptive phenotypic plasticity: consensus and
841 controversy. Trends Ecol Evol. 1995;10: 212–217.

842 36. Tohge T, Watanabe M, Hoefgen R, Fernie AR. Shikimate and
843 phenylalanine biosynthesis in the green lineage. Front Plant Sci.
844 2013;4: 62.

845 37. Tanaka Y, Sasaki N, Ohmiya A. Biosynthesis of plant pigments:
846 anthocyanins, betalains and carotenoids. Plant J. 2008;54: 733–749.

847 38. Tang Q, Huang Y, Shen Z, Sun L, Gu Y, He H, et al. 6-
848 Phosphogluconate dehydrogenase 2 bridges the OPP and shikimate
849 pathways to enhance aromatic amino acid production in plants. Sci
850 China Life Sci. 2024; 1–11.

851 39. Francoz E, Ranocha P, Nguyen-Kim H, Jamet E, Burlat V, Dunand C.
852 Roles of cell wall peroxidases in plant development. Phytochemistry.
853 2015;112: 15–21.

854 40. Lunn G, Madsen E. ATP-levels of germinating seeds in relation to
855 vigor. Physiol Plant. 1981;53: 164–169.

856 41. Perl M. ATP synthesis and utilization in the early stage of seed
857 germination in relation to seed dormancy and quality. Physiol Plant.
858 1986;66: 177–182.

859 42. Tenhaken R. Cell wall remodeling under abiotic stress. Front Plant
860 Sci. 2015;5: 771.

861 43. Li C, Luo S, Feng L, Wang Q, Cheng J, Xie J, et al. Protist ubiquitin
862 ligase effector PbE3-2 targets cysteine protease RD21A to impede
863 plant immunity. Plant Physiol. 2024;194: 1764–1778.

864 44. Rustgi S, Boex-Fontvieille E, Reinbothe C, von Wettstein D,
865 Reinbothe S. Serpin1 and WSCP differentially regulate the activity of
866 the cysteine protease RD21 during plant development in Arabidopsis
867 thaliana. Proc Natl Acad Sci. 2017;114: 2212–2217.

868 45. Shindo T, Misas-Villamil JC, Hörger AC, Song J, van der Hoorn RA.
869 A role in immunity for Arabidopsis cysteine protease RD21, the
870 ortholog of the tomato immune protease C14. PloS One. 2012;7:
871 e29317.

872 46. de Souza Cândido E, Pinto MFS, Pelegrini PB, Lima TB, Silva ON,
873 Pogue R, et al. Plant storage proteins with antimicrobial activity: novel
874 insights into plant defense mechanisms. FASEB J. 2011;25: 3290–
875 3305.

876 47. Hesami M, Naderi R, Tohidfar M, Yoosefzadeh-Najafabadi M.
877 Development of support vector machine-based model and comparative
878 analysis with artificial neural network for modeling the plant tissue
879 culture procedures: effect of plant growth regulators on somatic
880 embryogenesis of chrysanthemum, as a case study. Plant Methods.
881 2020;16: 1–15.

882 48. Motamayor JC, Lachenaud P, Da Silva e Mota JW, Loor R, Kuhn DN,
883 Brown JS, et al. Geographic and genetic population differentiation of
884 the Amazonian chocolate tree (Theobroma cacao L). PloS One.
885 2008;3: e3311.

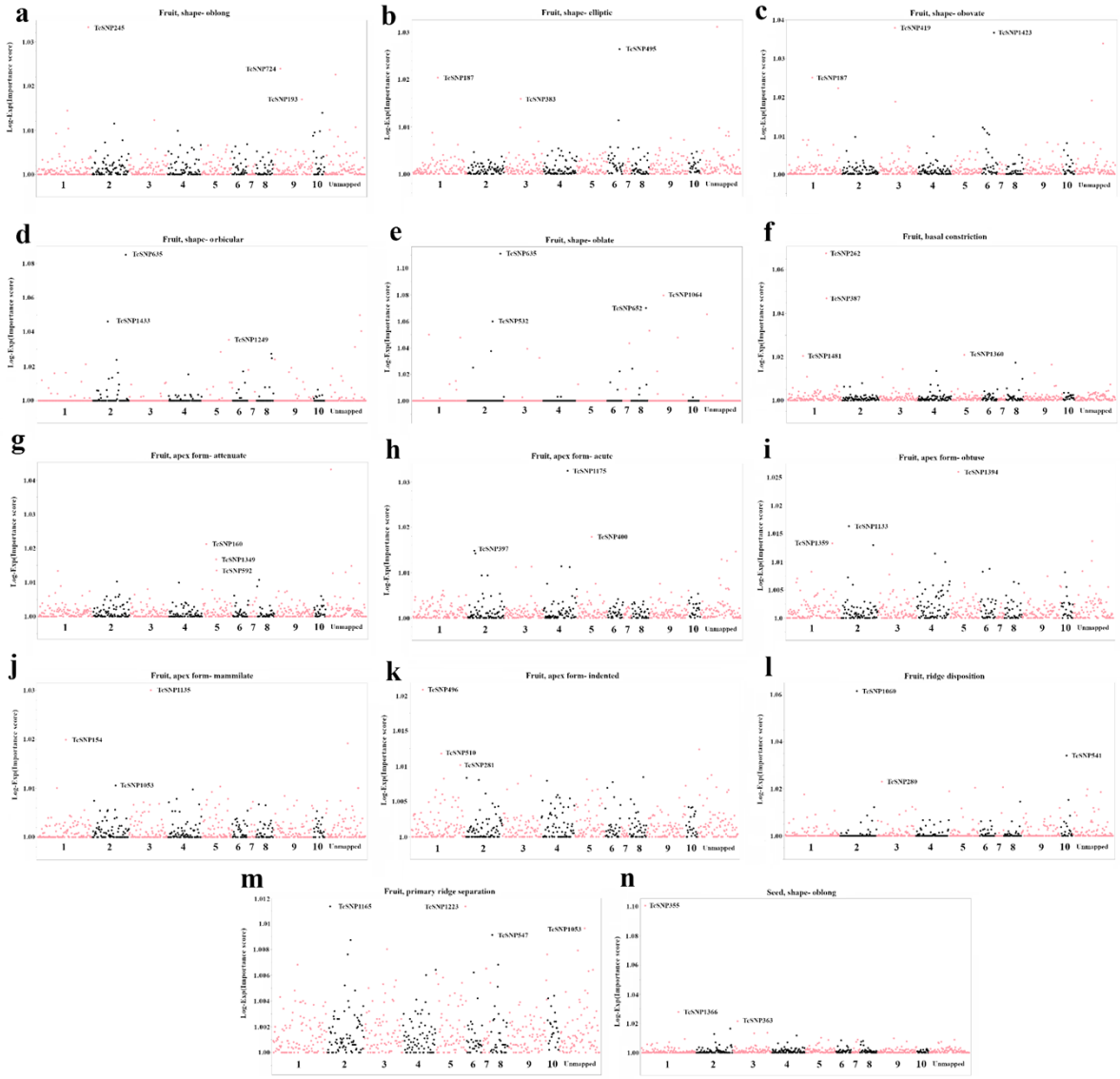
886 49. Tamura K, Nei M, Kumar S. Prospects for inferring very large
887 phylogenies by using the neighbor-joining method. Proc Natl Acad
888 Sci. 2004;101: 11030–11035.

50. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4: 406–425.

51. Klimberg R. Fundamentals of predictive analytics with JMP. Sas institute; 2023.

52. Hamid Y, Sugumaran M. A t-SNE based non linear dimension reduction for network intrusion detection. Int J Inf Technol. 2020;12: 125–134.

53. Schwenk H, Bengio Y. Boosting neural networks. Neural Comput. 2000;12: 1869–1887.



901 **Supplementary Figure S1. Genome-wide association study results for the remaining**
902 **traits in cacao, based on Bootstrap Forest analysis.** Panels (a) through (n) display the
903 importance scores (Log-Exp(Portion)) for all SNPs, plotted along the 10 chromosomes and
904 unmapped SNPs. The traits analyzed are: (a) Fruit shape-oblong, (b) Fruit shape-elliptic, (c)
905 Fruit shape-ovate, (d) Fruit shape-obovate, (e) Fruit basal constriction, (f) Fruit apex form,
906 (g) Fruit shape-obovate, (h) Fruit surface texture, (i) Fruit, ridge disposition, (j) Fruit, apex
907 form- acute, (k) Fruit, apex form- indented, (l) Fruit, ridge disposition, (m) Fruit, primary
908 ridge separation, (n) Fruit shape-oblong.
909

Trait	SNP ID	Chr	Nearest genes and function	Go annotation	Impartance score (portion)
Fruit, shape-oblong	TcSNP245	1	*Tc01v2_g033900* 40S ribosomal protein S11 (36635164..36637259)	-	0.033
	TcSNP724	9	*Tc09v2_g004320* Zinc finger CCCH domain-containing protein 20 (2513095..2514632)	-	0.024
	TcSNP193	9	*Tc09v2_g025980* 17.5 kDa class I heat shock protein (35825700..35826590)	HSP20-like chaperones superfamily protein, Stress response	0.022
Fruit, shape-elliptic	TcSNP495	6	*Tc06v2_g017610* Postsynaptic protein-related isoform 1 (25060968..25062477)	Cysteine-rich PDZ-binding protein, cytoplasmic microtubule organization	0.026
	TcSNP187	1	*Tc01v2_g015660* Fasciclin-like arabinogalactan protein 7 (12390486..12392512)	Plasma membrane, plant-type secondary cell wall biogenesis	0.02
	TcSNP383	3	*Tc03v2_g012710* 60S ribosomal protein L15 (27247828..27250100)	Ribonucleo protein	0.016
Fruit, shape-obovate	TcSNP419	3	*Tc03v2_g012710* 60S ribosomal protein L15 (27247828..27250100)	Ribonucleo protein	0.037
			Tc03v2_g012720 Lipid transporters isoform 1(27250512..27258189)	-	

	TcSNP1423	6	-	-	0.036
	TcSNP187	1	*Tc01v2_g015660* Fasciclin-like arabinogalactan protein 7 (12390486..12392512)	Plasma membrane, plant-type secondary cell wall biogenesis	0.025
Fruit, shape-orbicular	TcSNP635	2	*Tc02v2_g025000* Putative 21 kDa protein (36067420..36068518)	-	0.082
	TcSNP1433	2	*Tc02v2_g008930* Plant basic secretory protein (BSP) family protein (5513630..5514505)	-	0.045
	TcSNP1249	5	*Tc05v2_g025910* 40S ribosomal protein S21-2 (38343866..38345771)	-	0.035
Fruit, shape-oblate	TcSNP635	2	*Tc02v2_g025000* Putative 21 kDa protein (36067420..36068518)	-	0.11
	TcSNP1064	9	*Tc09v2_g001920* ER membrane protein complex subunit 7 homolog (1155585..1158234)	-	0.077
	TcSNP652	8	-	-	0.068
	TcSNP532	2	*Tc02v2_g017890* 60S ribosomal protein L17-2 (17471194..17473288)	Large ribosomal subunit Ribonucleo protein	0.058
Fruit, basal constriction	TcSNP262	1	*Tc01v2_g026930* S-adenosylmethionine synthase 1 (32474115..32476941)	ATP binding, S-adenosylmethionine biosynthetic process	0.065
			Tc01v2_g026940 O-Glycosyl hydrolases family 17 protein (32476724..32478711)	O-Glycosyl hydrolases family 17 protein, Transmembrane	

	TcSNP387	1	-	-	0.046
	TcSNP1360	5	*Tc05v2_g006960* Putative Histone deacetylase HDT1 (4516215..4519307)	-	0.021
	TcSNP1481	1	*Tc01v2_g009400* RING-H2 zinc finger protein RHA2a (5824131..5824989)	Metal-binding, zinc	0.02
Fruit, apex form- attenuate	TcSNP160	5	*Tc05v2_g002820* Aldo-keto reductase family 4 member C9 (1438324..1440675)	-	0.021
	TcSNP1349	5	*Tc05v2_g007090* Fructose-bisphosphate aldolase cytoplasmic isozyme (4739626..4741948)	Fructose-bisphosphate aldolase activity, glycolytic process	0.017
	TcSNP592	5	*Tc05v2_g007090* Fructose-bisphosphate aldolase cytoplasmic isozyme (4739626..4741948)	Fructose-bisphosphate aldolase activity, glycolytic process	0.013
Fruit, apex form- acute	TcSNP1175	4	*Tc04v2_g019280* Vicilin (28070761..28073194)	Storage protein	0.032
	TcSNP400	5	*Tc05v2_g007090* Fructose-bisphosphate aldolase cytoplasmic isozyme (4739626..4741948)	Fructose-bisphosphate aldolase activity, glycolytic process	0.018
	TcSNP397	2	*Tc02v2_g005940* V-type proton ATPase 16 kDa proteolipid subunit (3611563..3613553)	Hydrogen ion transport, vacuolar membrane	0.015
Fruit, apex form- obtuse	TcSNP1394	5	*Tc05v2_g005000* Transposable gene (2688034..2689543)	-	0.026
			Tc05v2_g004990 MLO-like protein 6 (2684247..2688006)	-	

	TcSNP1133	2	-	-	0.016
	TcSNP1359	1	*Tc01v2_g033930* Putative 20 kDa chaperonin, chloroplastic (36649828..36652181)	-	0.013
Fruit, apex form-mammilate	TcSNP1135	3	*Tc03v2_g015350* Heat shock protein 83 (29196217..29199339)	-	0.03
			Tc03v2_g015360 40S ribosomal protein S10-1 (29199844..29202045)	RNA binding Plectin/S10 domain-containing protein	
	TcSNP154	1	*Tc01v2_g018540* Xyloglucan endotransglucosylase/hydrolase protein 2 (16609918..16611431)	-	0.02
	TcSNP1053	2	*Tc02v2_g014590* Uncharacterized protein isoform 1 (9962662..9964553)	-	0.011
Fruit, apex form-indented	TcSNP496	1	*Tc01v2_g008340* 40S ribosomal protein S6 (5017545..5020015)	Ribonucleo protein complex, structural constituent of ribosome, translation	0.021
	TcSNP510	1	*Tc01v2_g020970* 60S ribosomal protein L23 (26581701..26583333)	Ribonucleo protein	0.012
	TcSNP281	1	*Tc01v2_g034810* 60S ribosomal protein L31 (37121671..37123110)	Cytosolic large ribosomal subunit, Structural constituent of ribosome	0.01

			Tc01v2_g020960 Glutathione S-transferase family protein (26579481..26581574)	-	
Fruit, ridge disposition	TcSNP1060	2	*Tc02v2_g010820* Peptidyl-prolyl cis-trans isomerase FKBP12 (6669481..6672179)	Isomerase, Rotamase	0.06
	TcSNP541	10	*Tc10_t005520* Zn-dependent exopeptidases superfamily protein isoform 1 (3986112..3997148)	-	0.034
	TcSNP280	3	*Tc03v2_g003930* Probable calcium-binding protein CML27 (5206898..5207881)	Calcium ion binding	0.023
Fruit, primary ridge separation	TcSNP1165	2	*Tc02v2_g002450* Pathogenesis-related protein 1 (1715993..1716709)	-	0.011
	TcSNP1223	5	*Tc05v2_g026730* GTP-binding protein SAR1A (38787425..38789642)	Secretion-associated RAS super family 2, ER-Golgi transport, membrane organization	0.011
	TcSNP547	8	*Tc08v2_g001820* Arabinosyltransferase RRA3 (1007297..1010003)	Glycosyltransferase, Golgi membrane, cell wall organization	0.0091
	TcSNP1053	2	*Tc02v2_g014590* Uncharacterized protein isoform 1 (9962662..9964553)	-	0.0087
Seed, shape-oblong	TcSNP355	1	-	-	0.1
	TcSNP1366	1	*Tc01v2_g025790* Uncharacterized protein isoform 1 (31723669..31730804)	Membrane, cellular component	0.028
	TcSNP363	3	*Tc03v2_g003930* Probable calcium-binding protein CML27 (5206898..5207881)	Calcium ion binding	0.021

910 **Table S1. Putative candidate genes near top SNPs associated with additional traits in**
911 **cacao.** Details of the most important SNPs associated with additional traits in cacao, as
912 identified by the Bootstrap Forest-based GWAS. The table lists the SNP ID, chromosome
913 (Chr), nearest gene (and its genomic location), Gene Ontology (GO) annotation (if available),
914 and the importance score (Portion) for each SNP. The importance score reflects the
915 contribution of each SNP to the prediction of the trait in the Bootstrap Forest model. A dash
916 (-) indicates that no information was available for that particular cell. The additional traits are
917 Fruit shape-oblong, Fruit shape-elliptic, Fruit shape-ovate, Fruit shape-orbicular, Fruit shape-
918 oblate, Fruit basal constriction, Fruit apex form-attenuate, Fruit apex form- acute, Fruit apex
919 form- obtuse, Fruit apex form- mammilate, Fruit apex form- indented, Fruit, ridge
920 disposition, Fruit, primary ridge separation, and Seed shape-oblong.