G3 Genes | Genomes | Genetics

INVESTIGATION

Genome-Wide Association Study Reveals Novel Candidate Genes Associated with Productivity and Disease Resistance to *Moniliophthora* spp. in Cacao (*Theobroma cacao* L.)

Jaime A. Osorio-Guarín,*[,1] Jhon A. Berdugo-Cely,[†,1] Roberto A. Coronado-Silva,[‡] Eliana Baez,[‡] Yeirme Jaimes,[‡] and Roxana Yockteng*[,§,2]

*Centro de Investigación Tibaitatá, Corporación Colombiana de Investigación Agropecuaria, Agrosavia, Bogotá, Colombia, [†]Centro de Investigación Turipaná Corporación Colombiana de Investigación Agropecuaria, Agrosavia, Montería, Colombia, [‡]Centro de Investigación La Suiza Corporación Colombiana de Investigación Agropecuaria, Agrosavia, Rionegro, Colombia, and [§]Muséum National d'Histoire Naturelle, UMR-CNRS 7205, Paris, France

ORCID IDs: 0000-0002-5486-2680 (J.A.O.-G.); 0000-0003-2133-6139 (R.Y.)

ABSTRACT Cacao (*Theobroma cacao* L.), the source of chocolate, is one of the most important commodity products worldwide that helps improve the economic livelihood of farmers. Diseases like frosty pod rot caused by *Moniliophthora roreri* and witches' broom caused by *Moniliophthora perniciosa* limit the cacao productivity, this can be solved by using resistant varieties. In the current study, we sequenced 229 cacao accessions using genotyping-by-sequencing to examine the genetic diversity and population structure employing 9,003 and 8,131 single nucleotide polymorphisms recovered by mapping against two cacao genomes (Criollo B97-61/B2 v2 and Matina 1-6 v1.1). In the phenotypic evaluation, three promising accessions for productivity and 10 with good tolerance to the frosty pod rot and witches' broom diseases were found. A genome-wide association study was performed on 102 accessions, discovering two genes associated with productivity and seven to disease resistance. The results enriched the knowledge of the genetic regions associated with important cacao traits that can have significant implications for conservation and breeding strategies like marker-assisted selection.

KEYWORDS
Association mapping
Cacao
Genotyping-by-sequencing
Healthy pods
Monilia
Witches' broom

Theobroma cacao L (cacao) is an economically important perennial crop. It was originally domesticated from wild ancestors in Central America (Cuatrecasas 1964; Motamayor *et al.* 2002) and is currently produced commercially in more than 50 countries of tropical regions of Central and South America, Asia and Africa (Wickramasuriya and Dunwell 2018). Commercial cacao was initially classified into three groups based on morphological traits: Criollo, Forastero, and Trinitario (Cheesman 1944). The Criollo group shows the finest flavor but is highly susceptible to diseases (Wood and Lass 1985). The Forastero group is less susceptible to diseases, provides higher yield, and represents the most commonly grown cacao worldwide (Wood and Lass 1985). The Trinitario group is a hybrid resulting from the cross-pollination between Criollo and Forastero (Wood and Lass 1985). However, with the development of molecular markers, it is now recognized that the species *T. cacao* is composed by 10 major genetic clusters as follows: Marañon, Curaray, Criollo, Iquitos, Nanay, Contamana, Amelonado, Purús, Nacional and Guiana (Motamayor *et al.* 2008).

Cacao beans comprise the raw material of the multibillion-dollar industry that produces chocolate and is the primary income for about 6 million smallholder farmers globally (World Cocoa Foundation 2014). In Colombia, cacao cultivation occupies an area of 177 thousand hectares, and the country is classified as the tenth producer worldwide with a yield close to 400 kg/ha and a production of 54,000

doi: https://doi.org/10.1534/g3.120.401153
Manuscript received January 16, 2020; accepted for publication March 10, 2020; published Early Online March 11, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.10028999.
[1] These authors contributed equally to this work.
[2] Corresponding author: Km 14 vía Mosquera-Bogotá, Corporación Colombiana de Investigación Agropecuaria (Agrosavia). E-mail: ryockteng@agrosavia.co

metric tons of beans (Ríos *et al.* 2017). Predictions of the cocoa market estimate that the demand will continue to increase (Fountain and Hütz-Adams 2018); on the other hand, up to 40% of the cacao production is estimated to be lost annually because of diseases produced by pathogens, such as fungi and oomycetes (Bowers *et al.* 2001; Álvarez *et al.* 2014).

Among the fungal pathogens that attack cacao, some of the most important are: 1) *Moniliophthora roreri* that causes frosty pod rot disease (FPRD) commonly known as monilia that invades growing pods (Bailey and Meinhardt 2016), and 2) *Moniliophthora perniciosa* that causes witches' broom disease (WBD); this disease produces hypertrophy and hyperplasia of distal tissue forming abnormal stems, such as a flower cushion broom and deformed branches (Meinhardt *et al.* 2008). To control the incidence of FPRD and WBD, disease control methods that include cultural practices, application of fungicides, biological control, and development of varieties with disease resistance have been attempted (Bateman *et al.* 2005; Loguercio *et al.* 2009; Jaimes and Aranzazu 2010; Acebo-Guerrero *et al.* 2012; Tirado-gallego *et al.* 2016).

The development of resistant or high-yielding varieties could be accelerated by using marker-assisted selection (MAS), which refers to the selection of individuals based on molecular markers associated with quantitative trait loci (QTL) that control defense mechanism resistance (Collard and Mackill 2008). During recent decades, many QTL for productivity (Crouzillat *et al.* 2000; Clement *et al.* 2003) and resistance to FPRD or WBD (Brown *et al.* 2005; Faleiro *et al.* 2006; Queiroz *et al.* 2003; Motilal *et al.* 2016) have been reported. Also, based on data array of single nucleotide polymorphisms (SNPs), Royaert *et al.* (2016) described genomic regions involved in WBD resistance, and McElroy *et al.* (2018) and Romero-Navarro *et al.* (2017) identified several markers related to productivity and disease resistance to *Moniliophthora* spp.

However, these studies have two disadvantages. First, the use of bi-parental populations produce a low mapping resolution because only a limited number of recombination events can be evaluated (Korte and Farlow 2013). The second disadvantage is that the use of a SNP array does not allow finding new genetic variants (Ganal *et al.* 2012). To solve these disadvantages, a method known as genotyping-by-sequencing (GBS) allows identifying thousands of SNPs (Elshire *et al.* 2011; Ganal *et al.* 2012; Romay *et al.* 2013). This method is simple, reproducible, and a considerable number of samples can be multiplexed, being suitable for population studies, germplasm characterization, and trait mapping (Davey *et al.* 2011). The analysis of GBS data in a genome-wide association study (GWAS) takes advantage of the historical recombination events accumulated over thousands of generations, resulting in a high-resolution mapping (Brachi *et al.* 2011; Xu *et al.* 2017).

Our study is the first to explore GWAS using GBS data to identify new SNP variants distributed across the genome and associate them with important agronomic traits in *T. cacao*. The aims of this work were: 1) to assess the genetic diversity and population structure; 2) to characterize Agrosavia's diverse collection of cacao germplasm with regards to productivity and disease resistance to FPRD and WBD; 3) to identify marker-trait associations, and finally 4) to identify genomic regions that have undergone selection.

MATERIALS AND METHODS

Plant material

A total of 229 cacao accessions from the national germplasm bank located at the research center C.I. La Suiza (7°22'12" N, 73°11'39" W) of Corporación Colombiana de Investigación Agropecuaria (Agrosavia) (previously known as Corpoica) were used (Table S1). Agrosavia s collection is a representation of the diversity of the species and conserves not only native materials but also improved materials (Osorio-Guarín *et al.* 2017). Cacao trees were planted in 1998 under an agroforestry system in lanes with a distance of 2.5 m (Arguello *et al.* 1999). Trees from the species *Cordia gerascanthus L.* and *C. alliodora* L. were distributed in lanes every 12 m to provide shade. The accessions selected were those that represented the genetic diversity of the four clusters reported by Osorio-Guarín *et al.* (2017) which were created based on a population structure analysis of a collection of 565 samples. Besides, the accessions selected in the present study span different regions of Colombia as well as of the upper Amazon region of Brazil, Peru, and Ecuador; furthermore, 22 accessions from the breeding programs of Costa Rica and Trinidad were also used.

DNA isolation and sequencing analysis

Genomic DNA of the 229 accessions was extracted from young leaf tissue using the DNeasy Plant Mini Kit (QIAGEN, Germany). The DNA concentration and quality were estimated using a Qubit Fluorometer v2.0 (Life Technologies, Thermo Fisher Scientific Inc.) and by electrophoresis on 1% agarose. Genomic DNA was digested using the enzymes *Bsa*XI ($(N)_9AC(N)_5CTCC(N)_{10}$) (New England Biolabs) and *Csp*CI ($(N)_{10\text{-}11CAA}(N)_5GTGG(N)_{12\text{-}13}$) (New England Biolabs), with 2.0 units each according to the library preparation protocol proposed by Osorio-Guarín *et al.* (2018). Fragments between 200–300 bp were size-selected and the resulted libraries were checked on the Bioanalyzer using a High Sensitivity DNA chip (Agilent). The pooled barcoded samples were sequenced using a paired-end strategy with 100 bp in eight lanes on an Illumina HiScan SQ instrument carried out in the Molecular Genetics Laboratory at the research center C.I. Tibaitatá (4°41'45" N - 74°12'12" W) of Agrosavia.

In silico digestion

To identify all restriction cut-site positions for the combination of the *Bsa*XI and *Csp*CI enzymes in the two reference genomes (Argout *et al.* 2010; Motamayor *et al.* 2013), we used the restrict package from the Emboss v6.5.7.0 software (Rice *et al.* 2000). The total number of fragments along the genomes were ordered per chromosome and those with size range from 200 to 700 bp were counted and plotted with the software R (R development core team 2008). The number of fragments produced with the *in silico* predictions was compared to the resulting sequenced alignments using the package genomecov from BEDtools v2.27.0 (Quinlan and Hall 2010).

SNP discovery

Raw sequence reads were checked for quality with FastQC (Andrews *et al.* 2012), and both ends were trimmed with Trim Galore v0.5.0 (Krueger 2018). Reads presenting a Phred quality score below 25 and a sequence length below 60 bp were removed. Samples were aligned with the software BWA v0.7.17 (Li and Durbin 2009) against two genomes, the Criollo B97-61/B2 v2 genome assembly which has a size of 314 Mpb in 10 chromosomes (Argout *et al.* 2010), and the Matina 1-6 v1.1 genome that has a size of 445 Mbp in 701 scaffolds; the last material represents the most commonly cultivated type of cacao worldwide (Motamayor *et al.* 2013). SNPs were discovered using Picard Tools v2.18.9 (Broad Institute 2019) and called with the GATK software v3.8.0 (McKenna *et al.* 2010). Finally, the SNPs were filtered to maintain bi-allelic SNPs with a minimum allele frequency of 0.05 and to remove missing data using VCFtools v4.2 (Danecek *et al.* 2011).

Genetic diversity, linkage disequilibrium, and population structure analyses

The observed (Ho) and expected heterozygosity (He) were obtained with Cervus v3.0.7 (Kalinowski *et al.* 2007). A linkage disequilibrium (LD) analysis was performed using the VCFtools v4.2 software (Danecek *et al.* 2011), applying the sliding window of 500 bp. The LD decay was established employing a loess regression generated from the plotting of pairwise LD (r^2) over an intermarker genetic distance with a threshold value of 0.1.

The population structure was analyzed using the intersected SNPs identified using a vcf file of 69 fully sequenced genomes representing the 10 recognized cacao genetic groups (Motamayor *et al.* 2008) (Table S1) generated mapping those genomes against the Criollo genome and then merged with the resulting vcf of our samples. Two analyses were carried out, a principal component analysis (PCA) and the maximum likelihood on the admixture v1.3 software (Alexander *et al.* 2009). The PCA was inferred with the vcfR (Knaus and Grünwald 2017) and adegenet (Jombart and Ahmed 2011) packages in the R Software. The admixture was run on a supervised mode in which the genetic group of each reference samples was indicated (Table S1). Visualization of the results was done with the package ggplot2 on the R software.

Phenotypic data and statistical analyses

Phenotypic data were collected for a subset of 102 accessions that presented promising data on productivity or disease resistance during previous assessments. The number of healthy and infected pods was registered weekly using two to six clones per accession during four harvest periods from September 2016 to May 2018 (Table S2).

With the resulting data, the area under the disease progress curve (AUDPC) for all harvest periods was calculated using the formula proposed by Shaner and Finney (1977):

$$AUDPC = \sum_{n-1}^{i=1}\left(\frac{y_{i+1}+y_i}{2}\right)\times(t_{i+1}+t_i)$$

Where *yi* refers to the counting of the disease in the nth observation, *ti* is the time in the nth observation, and *n* is the total number of observations.

The following four variables were evaluated:

1. Healthy pods (productivity).
2. AUDPC of pods infected by FPRD.
3. AUDPC of flower cushion broom caused by WBD.
4. AUDPC of deformed branches caused by WBD.

The results were compared among the accessions through an analysis of variance (ANOVA). Correlations among traits were calculated using Pearson's correlation coefficient (r) at $P \leq 0.05$. All statistical analyses were performed in the R software.

Furthermore, we conducted a principal component analysis (PCA) with the prcomp function of the R software using the AUDPC values for each genotype. Then, a cluster analysis using the Ward method and Euclidean distance was conducted in the R package factoextra using the two first components of the PCA.

Association analysis

A GWAS was performed for healthy pods (productivity) and AUDPC of the two diseases with the Genome Association and Prediction Integrated Tool (GAPIT) software package (Lipka *et al.* 2012). The mixed linear model (MLM) was used to minimize the risk of false association by incorporating population structure data (Q) and a kinship matrix (K) with the following equation:

$$y = Xa + Qb + Zu + e$$

Where y is the vector for phenotypes, a is the vector of marker fixed effects, b is a vector of fixed effects, u is the vector of random effects (the kinship matrix), and e is the vector of residuals. Moreover, X denotes the genotypes at the marker, Q is the Q-matrix, and Z is an identity matrix.

The Q-matrix was determined previously by the admixture v1.3 software, and the K-matrix was calculated using GAPIT with the Loiselle method. The quantile-quantile plots (Q-Q plots) were constructed to validate the appropriateness of the MLM. Further, Manhattan plots were generated using the $-\log_{10}(p)$ values for each trait.

Identification of genomic regions under selection

The site frequency spectrum and the LD pattern between polymorphic sites were calculated with the SweeD (Pavlidis *et al.* 2013) and OmegaPlus software (Kim and Nielsen 2004), respectively. The grid parameter was calculated for each chromosome to have a measure of the composite likelihood ratio (CLR) (SweeD) and the ω statistics (OmegaPlus) every 10,000 bp. The flanking region was fixed from 1,000 bp to 100,000 bp. The common outliers were found using an R script that includes the GFF file annotation to identify the genes in the region under selection.

Data availability

Table S1 contains the list of the accessions used. The phenotypic data are provided in Table S2. Table S3 contains a summary of the statistics for the sequenced data per individual. The candidate genes under positive selection per chromosome are provided in Table S4. Files S5 and S6 include the vcf files of the SNPs discovered for Criollo and Matina, respectively. Figure S1 exhibits an *in silico* analysis of the restriction enzymes. Figure S2 presents the number of sequenced reads of each accession. Figure S3 shows the correlation among traits of the entire population. Figure S4 contains the QQ-plots for each evaluated trait. The R script for the detection of common outliers in selective sweeps is available at http://pop-gen.eu/wordpress/wp-content/uploads/2013/12/combined_analysis.zip. The raw sequencing data of the reference samples were downloaded from the BioProject PRJNA486011 (Cornejo *et al.* 2018), via NCBI. Supplemental material available at figshare: https://doi.org/10.25387/g3.10028999.

RESULTS

SNP calling and in silico digestion

The sequencing of libraries generated by the GBS method resulted in 1,894 Gb raw reads. The observed read depth across samples ranged from 5.5 to 182.8 when mapped against the Criollo genome and from 7.1 to 179.6 when mapped against the Matina genome (Table S3). Genotyping the 229 accessions without missing data yielded a total of 9,003 SNPs using the Criollo genome as reference, whereas 8,131 SNPs were found with the Matina genome. The distance between SNPs markers ranged from 31.1 kb (chromosome 10) to 43.3 kb (chromosome 8), with an average of 39.8 kb.

The *in silico* digestion analysis showed a substantially higher number of fragments with optimal sizes for sequencing in an Illumina system (200 to 700 bp) in the Criollo genome (118,400) compared to the Matina genome (42,849). The number of base pairs was slightly lower in Criollo (8.5 million bp) compared to the value found in

Matina (9.1 million bp) (Figure S1). Finally, the number of sequenced fragments produced per accession was lower in Matina compared to those predicted *in silico* for Criollo (Figure S2).

Genetic diversity

Values for He and Ho of the 229 cacao accessions were slightly higher with the dataset generated using the Criollo genome as a reference (Ho = 0.350 and He = 0.317) compared with the dataset generated using the Matina genome (Ho = 0.316 and He = 0.314). Medium to high levels of genetic diversity and an excess of heterozygosity were found in this collection.

Linkage disequilibrium and population structure analyses

The LD decay was analyzed to characterize the mapping resolution for GWAS. In the studied collection, within a sliding window of 500 bp, the LD declined quickly at an average distance of ~320 bp (threshold of $r^2 = 0.1$) (Figure 1). The average mean r^2 value estimated between adjacent SNPs was 0.255.

The analyses were carried out with 3,712 SNPs that were successfully intersected between the reference panel of 69 fully sequenced genomes and the dataset generated in this study. PCA analysis showed that the first two components provided the most information and accounted 52% of the total variation, where each component explained 35%, and 17% of the overall variation, respectively. The most differentiated group was the Criollo genetic group and the rest were not fundamentally genetically distinct from each other, as the accessions were distributed evenly among the first two PCs (Figure 2).

Finally, the population structure on supervised mode enabled us to identify the ancestry of Agrosavia s samples and the results were consistent with the NJ analysis result. This allowed us to confirm the ancestry of well-known accessions (*e.g.*, IMC-67 (*Iquitos*), C58 (*Amelonado*), SCA-6 (*Contamana*)), and correctly assign the ancestry of previously uncharacterized accessions (Figure 3). The results showed that the principal genetic material found in the admixed samples came from the *Amelonado* and Criollo groups.

Phenotypic data

The cacao germplasm conserved in Agrosavia is located in a region with natural presence of inoculum of different pathogens allowing the evaluation of the disease resistance of the accessions. During the evaluation of the 102 accessions, we harvested a total of 38,178 pods with an average of 374.3 pods per plant. An imbalance between the production of healthy (6,141) and infected pods (30,477) was found. The percentage of pods that reached harvest was only 20.77%, whereas 79.23% of the pods were affected by FPRD. No pods infected by WBD were found during the harvest periods.

Correlations between phenotypic variables ranged from 0.12 to 0.54 and are shown in Figure S3. The WBD variables (flower cushion broom and deformed branches) were also positively correlated (0.54). In contrast, symptoms of FPRD and WBD diseases are weakly correlated. Finally, as expected, disease variables did not show a strong correlation with the productivity trait considered (healthy pods).

Mean, standard deviation, and coefficient of variation of the phenotypic data are shown in Table 1. The ANOVA showed significant variation between genotypes for all the analyzed traits (observed at a level of $P \leq 0.001$), suggesting that the accessions presented in the collection are highly diverse.

According to the analysis of the four harvest periods, the highest number of healthy pods was produced by the accessions GS-29 (279 pods), FCM-39 (208), and EET-8 (165) (Table S2). The genotypes CRICF-13, EBC-06, EBC-09, and SUI-72, were less affected by FPRD, as shown by the AUDPC values (Table S2). The floral cushions of genotypes EET-377, SCC-85, SCC-86, and UF-273 were less affected by WBD. The genotypes EET-377 and SCC-85 also showed the lowest number of branches affected by WBD, as well as the genotypes SUI-99 and FCM-19 (Table S2).

Based on the conglomerate analysis, the cacao accessions were divided into three main groups with different levels of productivity and susceptibility/resistance responses to WBD and FPRD (Figure 4). The first group (I) comprised 14 accessions highly susceptible to

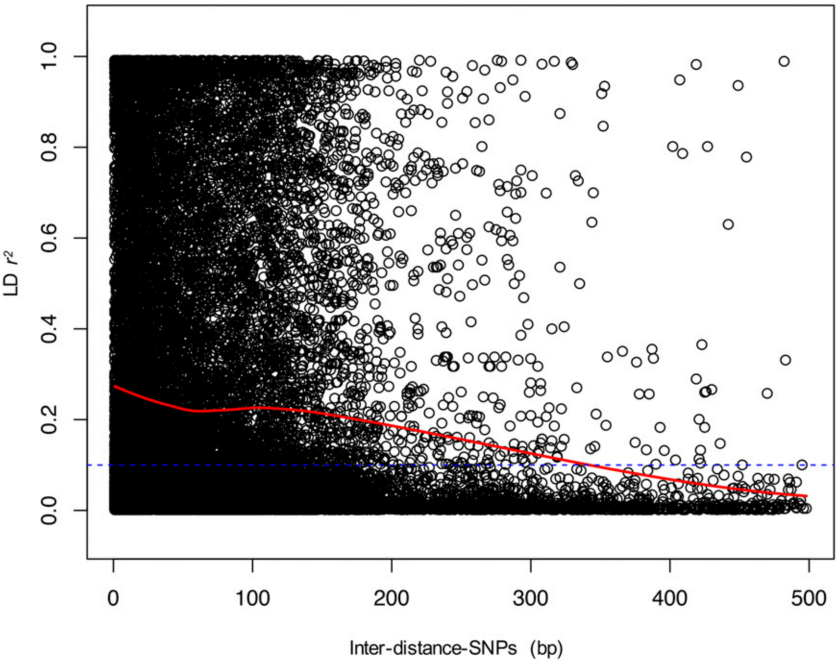

Figure 1 Linkage disequilibrium (LD) analysis. LD decay (r^2) as a function of physical distance on all chromosomes. Only r^2 values with $P \leq 0.05$ are shown. The LD threshold of 0.1 is indicated with a blue dashed line.

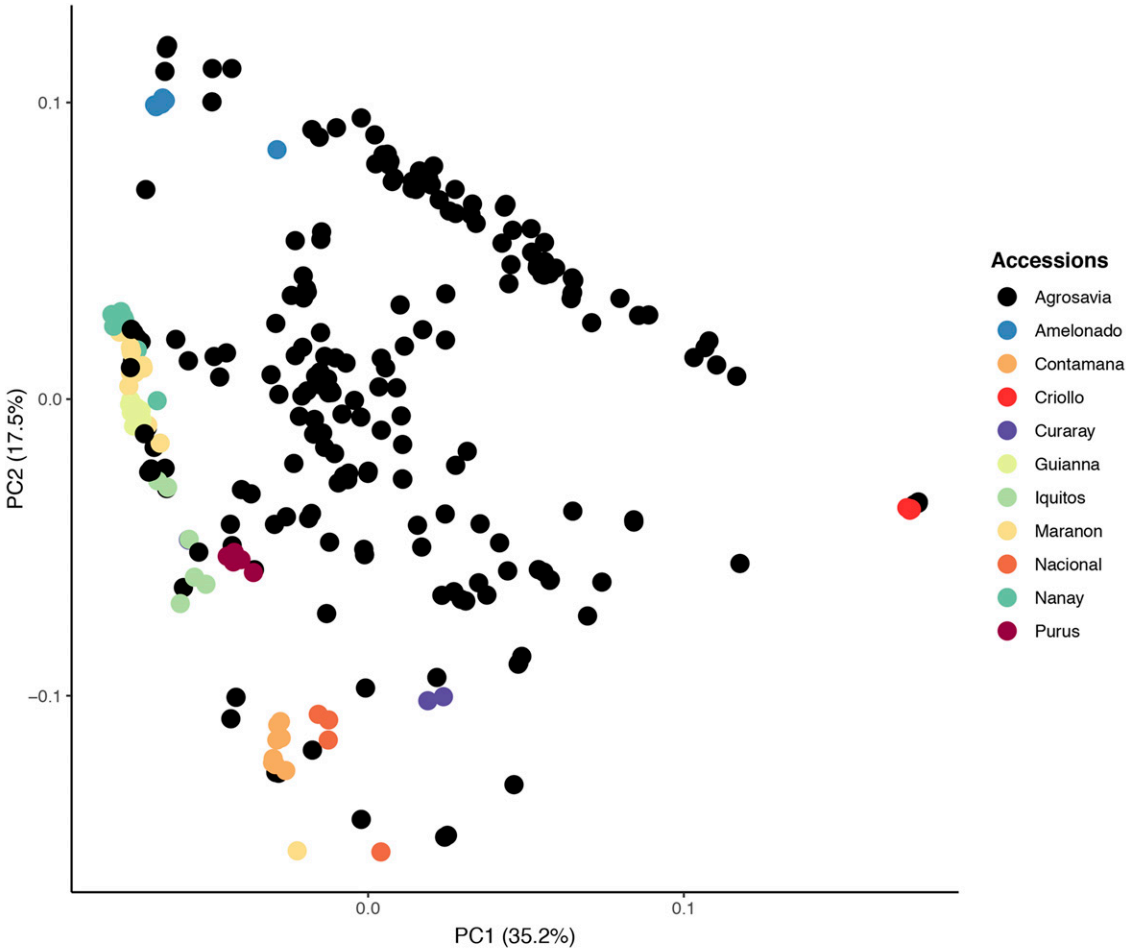

Figure 2 Principal component analysis (PCA) of genetic relatedness of 229 cacao accessions of the Agrosavia's collection and 69 of the reference genetic groups, using a panel of 3,712 SNPs. Percentage of the variation captured by each component is given on the axis labels.

pathogens with the highest values of infected pods by FPDR and organs infected by WBD. The second group (II) included 35 accessions with higher production of healthy pods, and (III) contained 53 accessions that showed the lowest values for the traits related to diseases (Table 2).

Association analysis

The associated SNPs were distributed in five of the 10 cacao chromosomes (Table 3). The Q-Q plots supported the association of the SNPs with the traits ($P \leq 0.005$) and suggested that the population structure was adequately controlled in the GWAS model (Figure S4).

Manhattan plots showing the $\log_{10}(p)$-values for the SNP markers according to their positions in each chromosome are shown in Figure 5. The GWAS identified four loci on chromosome 2 and one on chromosome 3 associated with WBD. The annotation of these loci was exactly the same using both reference genomes (Table 3).

Loci related to FPRD were located on chromosome 1 using both genomes as reference. The SNP S1_18354976 was upstream 11.4 kb to the transcript Tc00cons_t021470.1 of the gene consensus model of the Criollo

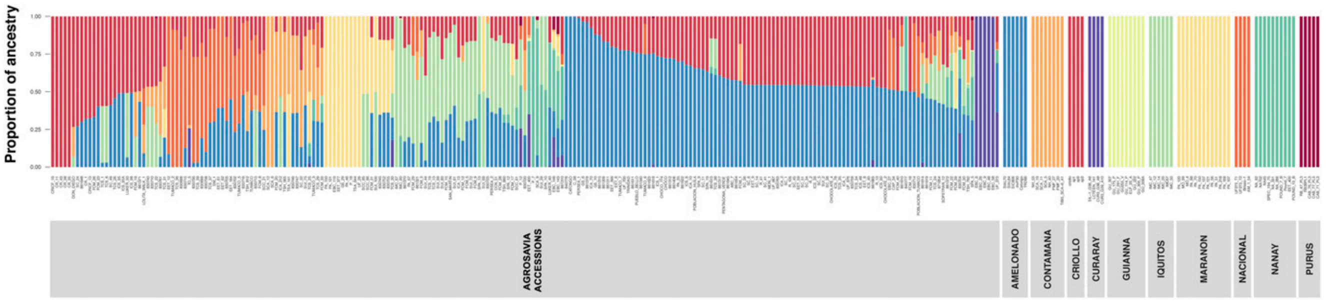

Figure 3 Population structure of *T. cacao* of the germplasm maintained in Agrosavia using single nucleotide polymorphisms (SNPs) mapped against the Criollo genome. The color in each bar corresponds to the probability of a genotype belonging to an assigned group. The pure clusters on right side correspond to the reference samples.

Table 1 Means, standard deviation, coefficients of variation, and ANOVA's test for the four tested variables

Trait	Min	Max	Mean	SD [a]	R^2	CV [b]	ANOVA p-value
Healthy pods	1	279	60.2	45.1	0.52	75.13	$\leq 1.0^{E-4}$
Pods FPRD	14	1304	298.8	243.2	0.74	51.19	$\leq 1.0^{E-4}$
Flower cushion broom WBD	0	310.5	49.3	70.5	0.73	92.28	$\leq 1.0^{E-4}$
Deformed branches WBD	0	59	14.8	12.1	0.55	80.0	$\leq 1.0^{E-4}$

[a] Standard deviation; [b] Coefficient of variation.

genome and probably corresponded to the same SNP, *i.e.*, S1_27025499, identified with the Matina genome. A second SNP located in the gene Thecc1EG003830 was identified using the Matina genome (Table 3).

Regarding productivity, three genomic regions were identified on chromosomes 2, 4, and 8. Two candidate genes found on the Criollo genome were not located in the coding region, but they were located downstream of the gene Tc08v2_g012850 and upstream of the Tc00cons_t021570.1 transcript. Based on the Matina genome, two associated SNPs were located in chromosome 8 in genes without a functional annotation, and one SNP was located on chromosome 4 (Table 3).

Identification of regions under selection

The presence of positive selection was tested by scanning the genome for i) reduced variability regions, and ii) local patterns of high linkage disequilibrium. The highest number of genes under positive selection was identified in chromosome 3 (39), and the lowest number was found in chromosome 1 (5) (Table S4). Figure 6 showed the identified genes under positive selection throughout the genome detected by the softwares SweeD and OmegaPlus independently and the consensus of the two approaches.

The genes under selection in chromosome 1 allowed identifying a selective sweep on a region comprised between 11.1 and 22.6 million base pairs (Mbp), in which the previous associated gene Tc00cons_t021470.1 was located (Table 3). The region of chromosome 2 under selection had a size of 29.1 Mbp and is located close to the genes Tc02v2_g014130 and Tc00cons_t021570.1 associated to the response to WBD and number of healthy pods, respectively. The selective sweep on chromosome 3 covered a genomic region of 27.1 Mbp from the position 8.8 to 36 Mbp in which the gene Tc03v2_g015970 associated to the response to WBD is located. The region in chromosome 4 under selection was not related to any of the candidate genes found by the GWAS. Finally, the selective sweep for chromosome 8 included from 9.3 to 18 Mbp in which the gene Tc08v2_g012850 was associated with healthy pods.

DISCUSSION

The main objective of the cacao breeding program is to accumulate favorable alleles for productivity and disease resistance (Rodriguez-Medina *et al.* 2019). However, the development of an improved cacao variety could take several years because it is a perennial species with juvenile stages, which can vary from 1.5 to 3 years, depending on the genotype (de Almeida and Valle 2007). For this reason, to accelerate the selection of promising materials, it is necessary to identify genes or molecular markers associated with genomic regions involved in disease resistance or productivity.

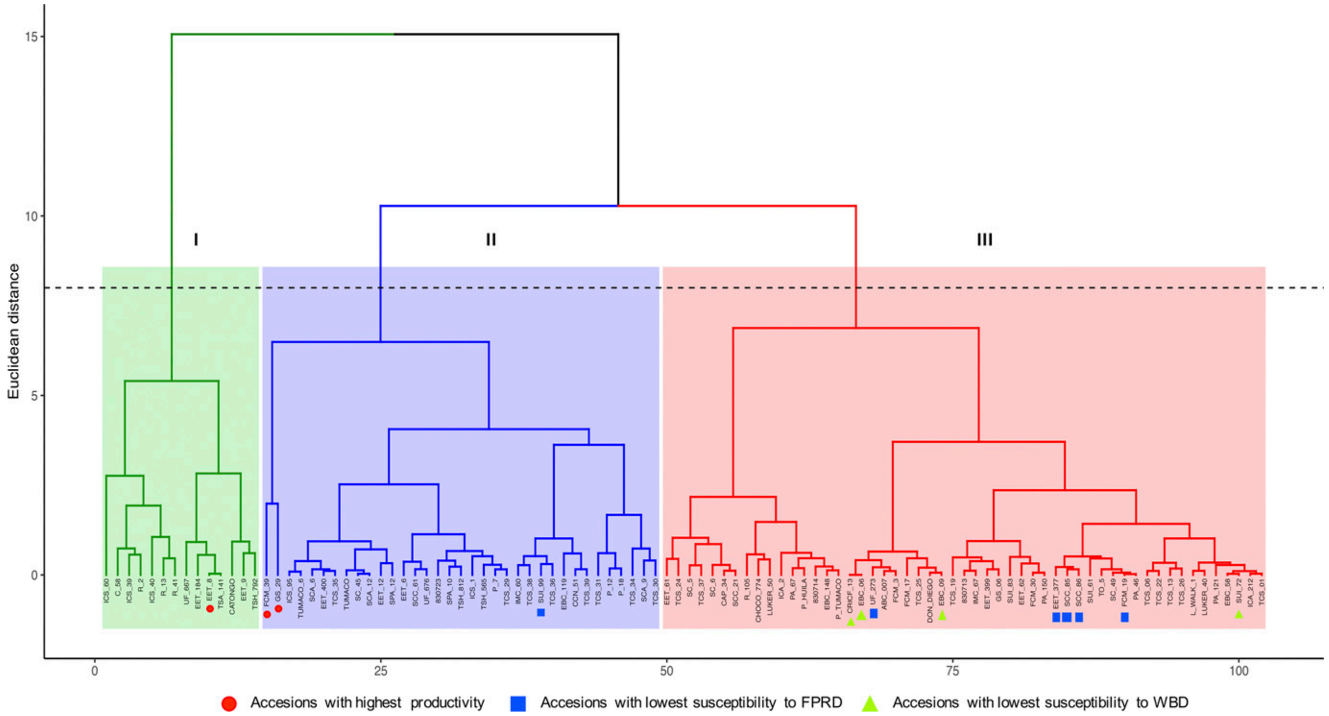

Figure 4 Conglomerate analysis of phenotypic data. Principal component analysis conducted using as variables the total counts of healthy pods and area under the disease progress curve (AUDPC). The best accessions for the variables evaluated are highlighted with indicators.

Table 2 Statistics of evaluated traits for each group

Trait	Cluster	Mean	SD	CV	Min	Max
Healthy pods	Cluster 1	79.86	38.46	48.16	33	165
	Cluster 2	87.97	53.48	60.79	12	279
	Cluster 3	36.68	22.73	61.96	1	95
Pods with FPRD	Cluster 1	523.21	263.84	50.43	264.5	1304
	Cluster 2	422	253.75	60.13	116	1225.5
	Cluster 3	158.15	107.08	67.71	14	414.5
Flower with cushion WBD	Cluster 1	187.61	91.77	48.92	65	310.5
	Cluster 2	35.29	27.73	78.59	1	97
	Cluster 3	22.09	31.84	144.11	0	142
Branches with WBD	Cluster 1	32.82	14.39	43.85	8.5	59
	Cluster 2	13.33	7.05	52.93	1	37.5
	Cluster 3	11.09	9.96	89.78	0	41.5

GBS protocol

Originally, GBS used single-enzyme digestion of frequent cuts like *Pst*I and *Apek*I (Elshire *et al.* 2011). However, this method has the disadvantage of producing a high number of short fragments and low variability in the number of reads obtained per individual (Poland *et al.* 2012). To address this issue, a study conducted by Cooke *et al.* (2016) demonstrated that the use of enzymes that cut far from the recognition site called CutSmart enzymes, like *Bsa*XI, increases the diversity of the GBS library. Osorio-Guarín *et al.* (2018) proposed a modified protocol in cacao using single-enzyme digestion with *Bsa*XI to generate a significant number of informative SNPs. In our study, we used a combination of the two enzymes (*Bsa*XI and *Csp*CI) and to confirm that this was suitable to study the cacao genome, we performed an *in silico* digestion using the two reference genomes and recovered 3 million bp more than the study of Osorio-Guarín *et al.* (2018). Moreover, we found that the combination of the two enzymes is more frequently found in the Criollo genome than in the Matina genome. The raw SNPs number in the presented study was 16,773, which is a higher value compared to the datasets obtained by Osorio-Guarín *et al.* (2018) who discovered a raw number of 12,457 SNPs in 32 samples and by Lachenaud *et al.* (2018) that reported a raw number of 9,187 SNPs using 264 samples with double-digestion using *Pst*I and *Mse*I. The results of the current study indicated the suitability of using a double-digestion for GBS libraries for cacao employing CutSmart enzymes (*Bsa*XI and *Csp*CI).

Genotypic analysis

The genetic analysis showed that the cacao population conserved in the germplasm bank of Agrosavia has a medium to high level of genetic diversity (Ho = 0.350 and He= 0.317), that is equivalent to other studies assessing cacao diversity (Ji *et al.* 2013; Cosme *et al.* 2016; Gopaulchan *et al.* 2019; De Wever *et al.* 2019). Osorio-Guarín *et al.* (2017) found a higher heterozygosity value analyzing 565 accessions; although we used the same collection, we analyzed only 229 accessions generating lower heterozygosity results.

Table 3 Significant marker–trait associations for evaluated traits

Genome	Trait	SNP Position	Chromosome	p-value	FDR Adjusted p-values	Gene	SNP position relative to the candidate gene[a]	Candidate gene annotation
Criollo	Deformed branches WBD	9,316,991	2	2 $^{E-08}$	0.138	Tc02v2_g014130	0	G-type lectin S-receptor-like serine/threonine protein
		29,572,898	3	5 $^{E-09}$	0.182	Tc03v2_g015970	0	Protein IWS1
	Flower cushion broom WBD	1,064,432	2	4 $^{E-07}$	2 $^{E-3}$	Tc02v2_g001650	0	Ion channel DMI1
		770,436	2	4 $^{E-07}$	2 $^{E-3}$	Tc02v2_g001090	0	Xyloglucan galactosyltransferase
		4,523,324	2	2 $^{E-07}$	2 $^{E-3}$	Tc02v2_g007330	0	MLO-like protein
	Pods FPRD	18,354,976	1	2 $^{E-04}$	0.280	Tc00cons_t021470.1	−11.4 kb	S-locus lectin protein kinase family
	Harvested healthy pods	10,111,356	8	4 $^{E-07}$	0.330	Tc08v2_g012850	+5.2 kb	RNA-directed DNA polymerase
		21,845,071	2	7 $^{E-09}$	0.330	Tc00cons_t021570.1	−22,0 kb	Ty3-gypsy retrotransposon
Matina	Deformed branches WBD	9,415,422	2	6 $^{E-09}$	0.300		Unkown	
		27,458,385	3	1 $^{E-04}$	0.300	Thecc1EG015342	0	Protein IWS1
	Flower cushion broom WBD	1,107,779	2	5 $^{E-07}$	2 $^{E-3}$	Thecc1EG006180	0	Ion channel DMI1
		794,296	2	5 $^{E-07}$	2 $^{E-3}$	Thecc1EG006099	0	Xyloglucan galactosyl transferase
		4,567,399	2	2 $^{E-08}$	4 $^{E-3}$	Thecc1EG006939	0	MLO-like protein
	Pods FPRD	18,890,740	1	3 $^{E-04}$	0.655	Thecc1EG003068	0	Unknown
		27,025,499	1	4 $^{E-04}$	0.655	Thecc1EG003830	0	Retrotransposon-like protein
	Harvested healthy pods	9,466,348	8	1 $^{E-04}$	0.309		Unkown	
		8,295,774	8	3 $^{E-04}$	0.309			
		9,491,291	4	3 $^{E-04}$	0.309	Thecc1EG017982	0	Ty3-gypsy retrotransposon

[a] SNP position relative to the closest candidate gene: upstream and downstream SNPs of candidate genes are specified with "–" and "+," respectively. 0 indicates that SNPs are located within the candidate gene.

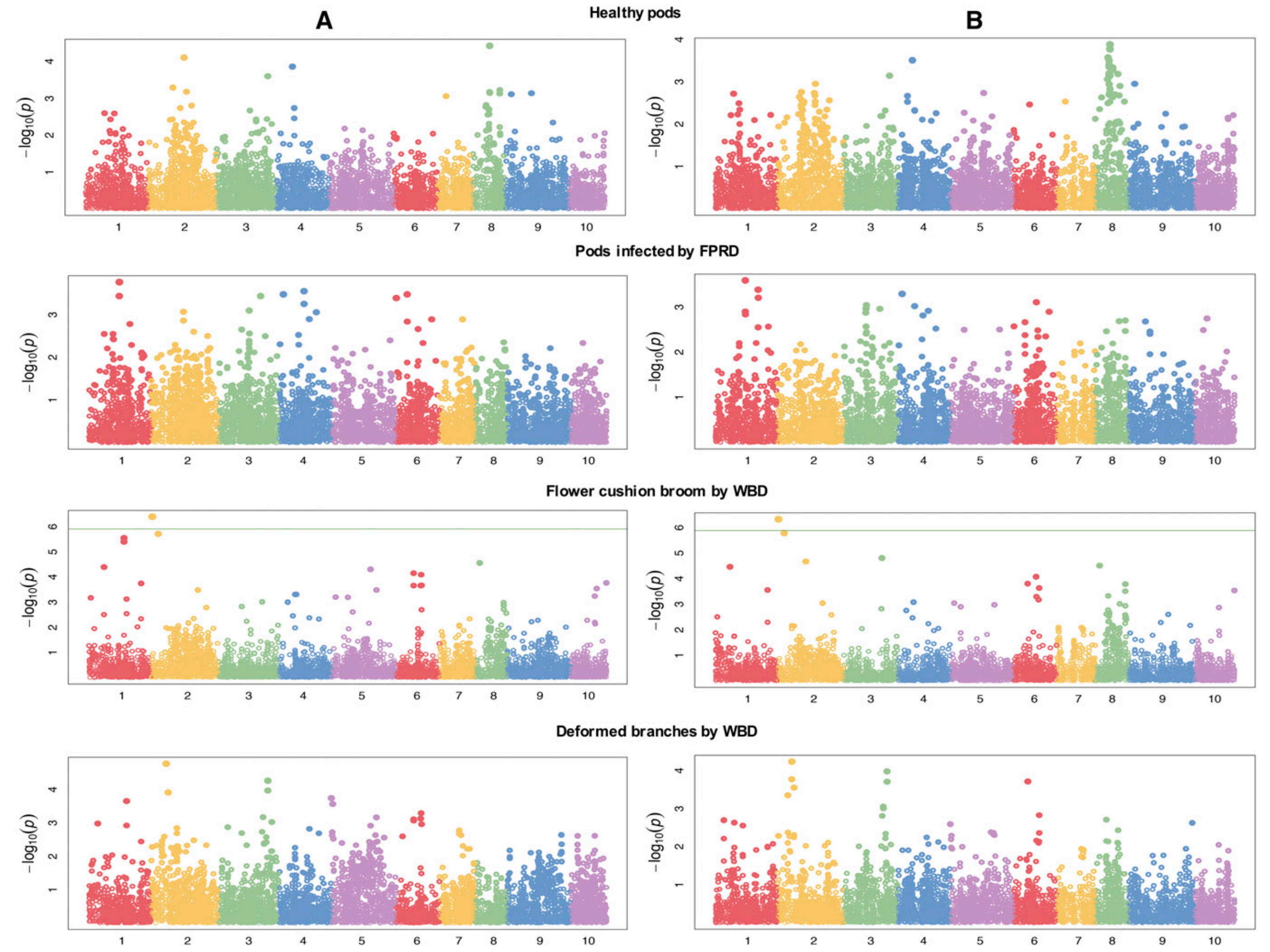

Figure 5 Manhattan plots of marker-trait associations for productivity, witches' broom disease (WBD), and frosty pod rot disease (FPRD) for both reference genomes. A. Criollo. B. Matina. The green-dotted horizontal line represents the genome-wide significance threshold of $P < 5.0 \times 10^{-3}$.

High-density SNP mapping facilitates understanding the genetic determinants of complex traits in GWAS. In the current study, the genome-wide LD pattern was explored using 9,003 SNPs, which provided a valuable resource for association mapping. In our association panel, the LD decay at ~320 bp using a sliding window of 0.5 kb (threshold of $r^2 = 0.1$) indicated that the LD blocks were small, which was expected considering the out-crossing nature of the species. In contrast, Stack *et al.* (2015) identified large LD blocks with a LD decay within 5.0 - 10 Mb (threshold of $r^2 = 0.1$), possible due to the use of SSRs. The LD r^2 average of our study was 0.255, higher than values reported by McElroy *et al.* (2018) in the populations Ganaderia and Las Tecas (average r^2 values of 0.147 and 0.188, respectively), but lower compared than the population Malvinas (average r^2 value of 0.418). Distinct methodologies, number of markers, population sizes, genetic origins, and standard errors among the studies may account for the different results.

The population structure analyses demonstrated that the collection is diverse and has a good representation of different cacao genotypes, consistent with previous analyses (Osorio-Guarín *et al.* 2017). The spread of Agrosavia accessions in the PCA (Figure 2) is consistent with our admixture analysis (Figure 3) in which individuals are principally hybrids between Criollo and other Amazonian groups (Amelonado, Iquitos, Contamana, etc). The genetic group Criollo is the only one that is separated by the PCA analysis. This result is consistent with the study reported by (Cornejo *et al.* 2018) in which they conclude that this differentiation could be to the early diversification or recent domestication of the Criollo population from the rest of the genetic clusters. The other cacao samples do not form distinguishable groups consistent with the study of McElroy *et al.* (2018) but the distribution probably corresponds to a geographical gradient as explained by Cornejo *et al.* (2018).

The highest representation of the Amelonado and Criollo ancestries in the Agrosavia s samples (Figure 3) could be explained by the extensive use of these materials in the breeding programs to obtain high-quality flavor from Criollo genotypes combined with high yields and disease resistance from Amelonado genotypes (Wood and Lass 1985). In contrast, the Purus and Curaray genetic groups were less represented in the collection. This could be an opportunity to explore the genetic diversity of these two groups for generating new varieties in the breeding program.

Phenotypic analysis

All the phenotypic traits evaluated in our study were positively correlated. The highest correlation (0.54) was found between flower cushion brooms and deformed branches infected both by WBD. It has been observed that *M. perniciosa* infects indiscriminately these

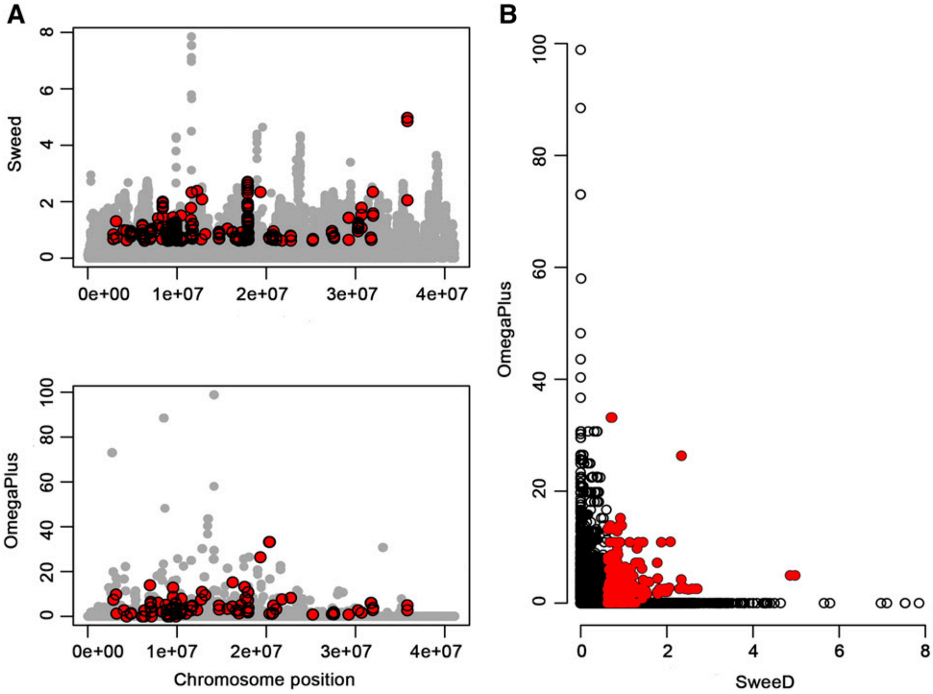

Figure 6 Selective sweep analysis for each chromosome. A. The x-axis denotes the position on the chromosomes, and the y-axis shows the composite likelihood ratio (CLR) evaluated with SweeD (upper panel) and the ω statistic (bottom panel) assessed with OmegaPlus. B. The joint plot for SweeD and OmegaPlus. Red dots denote outliers at a significance level of 5%.

two plant structures. In addition, low correlation was found between FPRD and WBD. Similar results were reported by McElroy *et al.* (2018) on different cacao populations evaluated. A hypothetical explanation could be that the two pathogens are competing for infecting healthy tissue. As expected, there is a low correlation between healthy pods and infected pods by the two diseases.

One important goal of this study was to identify accessions in the germplasm collection with excellent productivity or disease resistance response to FPRD and WBD. The evaluation during the four harvest periods allowed identifying three accessions with the highest productivity, GS-29 (279 pods), FCM-39 (208), and EET-8 (165) (Table S2). The accession GS-29 from Grenade was previously identified as a high productivity clone (Hunter 1990). The EET-8 accession from Ecuador was highlighted in a study conducted by Aranzazu *et al.* (2009) due to its high yield of 1,500 kg/ha, and the FCM-39 accession is part of the germplasm bank and has not be used commercially.

As expected, the incidence of FPRD was elevated because the germplasm collection is conserved at the research center La Suiza, located in the Magdalena Valley region of Colombia, considered the diversity center for *Moniliopthora roreri* (Jaimes *et al.* 2016). Four accessions, SUI-72, CRICF-13, EBC-06, and EBC-09, showed promising results of resistance against *M. roreri* (Table S2). Accession SUI-72 has additional qualities reported by Parra and Duarte (2007), such as a grain index (total grain weight per pod/number of grains per pod) of 1.73 and a pod index (number of cacao pods required to yield one pound of dry beans) of 22 that is not far from the commercial genotype CCN-51 that has the highest productivity in Colombia with a pod index of 15.2 (Jaimez *et al.* 2018). In comparison to cultivated Accessions CRICF-13, EBC-06, and EBC-09 are native Colombian genotypes not used commercially, and further studies are necessary to confirm their resistance response.

During the four harvests carried out in this research, differences were observed in the total number of healthy and infected pods. Besides, the incidence of WBD was significantly lower compared to FPRD. These results are probably because in Colombia, during 2016 and 2017, the "El Niño" phenomenon caused adverse conditions for the production of pods and the development of these diseases that require highly humid conditions and rain all year round for its fast spread and survival (Purdy and Schmidt 1996). The evaluation of the variables related to WBD allowed identifying six tolerant accessions (SCC-85, SCC-86, SUI-99, EET-377, UF-273, and FCM-19) (Table S2). The SCC genotypes (Arguello *et al.* 1999) are elite cultivars selected from regional hybrids from the department of Santander (Colombia), while accession SUI was sampled in the department of Antioquia (Colombia) and is a cacao genotype probably introduced from Central America. The EET-377 genotype has an Ecuadorian origin and is derived from the cross between EET-156 and Scavina 6 (End *et al.* 1992). The latter is a member of the Contamana ancestral group (Motamayor *et al.* 2008) and was previously identified as a resistant genotype against WBD in Brazil (Royaert *et al.* 2016). The UF-273 has beforehand been reported as a resistant genotype to FPRD (Romero-Navarro *et al.* 2017). Finally, the FCM-19 accession is part of the germplasm bank and is not used commercially.

The accessions that presented the highest values for the evaluated variables were highlighted in the conglomerate analysis (Figure 4), in which the group II present the accessions with good productivity and the group III the ones with good resistance to diseases. However, the EET-8 accession reported in our study as highly productive genotype was presented in the group I that regrouped the accessions most susceptible to diseases. This could be due to the fact that during the domestication process when selecting a characteristic of interest such as high productivity, the diversity of resistance-related genes was possibly lost (Smýkal *et al.* 2018). In fact, Yamada *et al.* (2013) found a positive correlation between number of pods and signs of infection of WBD.

Besides, the comparison of the admixture and the conglomerate analyses (Figure 3) allowed us to determine that the group I of the conglomerate analysis presented the highest percentage (31%) of Criollo ancestry, which has been reported as the group most susceptible to diseases (Albores-Flores *et al.* 2018). In contrast, the group

II presented the lowest ancestry of criollo (15%) with high percentage of Amelonado type (previously classified as the Forastero genetic group) which is known for high productivity (Wood and Lass 1985). Finally, the group III has less representation of Criollo ancestry (22%) than group I and has better representation of the other groups that can explain the presence of less susceptible materials in this group III.

Association analysis

The associations found were significant ($P \leq 0.05$) before the FDR correction; however, only the associations for flower cushions with WBD were maintained significant after the correction. The non-significant p-values using the FDR correction is possibly due to the reduced sample size used in this study. QTL and association studies are often limited by the relatively small population sizes mapped, resulting in low statistical power and thus, rendering small- or even medium-effect QTL that are statistically non-significant and difficult to detect. Such statistically underpowered populations may also suffer from severe inflation of effect size estimates (the so-called Beavis effect) (Beavis 1998). Hence, increasing the population size and marker density is required to enable estimations that are unbiased by the Beavis effect and achieve higher statistical power (Beavis 1998; Klein 2007; Hong and Park 2012). Although in our work the marker density is high, the studied species is a perennial plant (long generation time) with limited offspring numbers, therefore to study large populations would require a considerable investment.

Bi-parental mapping studies from several F_1 and F_2 populations reported QTL for disease resistance to FPRD on chromosomes 1, 2, 5, 7, 8, 9 and 10 (Faleiro *et al.* 2006; Brown *et al.* 2007; Queiroz *et al.* 2003; Lanaud *et al.* 2009; McElroy *et al.* 2018) and WBD on chromosomes 1, 2, 4, 6, 7, 8, and 9 (Brown *et al.* 2005; Lanaud *et al.* 2009; McElroy *et al.* 2018). Furthermore, Royaert *et al.* (2016) identified seven candidates genes related to WBD, and Romero-Navarro *et al.* (2017) reported six genes related to FPRD. In this study, loci associated with disease resistance traits were found on the same previously reported chromosomes (*i.e.*, 1, 2, 4, 5, 6, 7, 8, and 9); however, the SNPs were located in different genomic regions (Figure 5 and Table 3).

For WBD resistance response traits, three statistically significant genes were found. The first SNP was related to a G-type lectin S-receptor-like serine/threonine protein (GsSRK) that is associated with the receptor-like protein kinases (*RLKs*) gene family. In plants, *RLKs* show important roles in pathogen resistance induced after the activation of the recognition receptors of microbe-associated molecular patterns (Singh and Zimmerli 2013). Previous studies showed that *GsSRK* exhibited plant defense function in *Nicotiana glutinosa* (Kim *et al.* 2009) and *Capsicum annuum* (Kim *et al.* 2015), and they are also co-expressed with BIR2 kinase that confers resistance to *Arabidopsis thaliana* (Blaum *et al.* 2014). The second gene associated was the *DMI1* (Doesn't Make Infections 1) related to ion channels (Zimmermann *et al.* 1999). Studies with ion channels in species like tomato have demonstrated their possible relationship with plant defense (Zhang *et al.* 2018). The last identified gene was related to the mildew resistance Locus O (*MLO)* protein. The *MLO* is a gene family specific to plants and plays significant roles in the resistance to powdery mildew and response to a variety of abiotic stresses, plant growth, and development (Liu *et al.* 2017). Studies on *MLO* genes have demonstrated that these confer resistance to species such as *Arabidopsis* and tomato (Acevedo-Garcia *et al.* 2014). Other candidate genes were also found for WDB resistance (*IWS1* and Xyloglucan galactosyltransferase), but further studies are necessary to determine their role in plant defense.

Two candidate genes were related to resistance response against FPRD. The first one was associated with the S-locus lectin protein, a member of the *RLKs* gene family, which is related to plant immunity (Singh and Zimmerli 2013), as explained before. The second gene was related to retrotransposons, and according to a study conducted by Zervudacki *et al.* (2018), certain retrotransposons behave as an immune-responsive gene during pathogen defense in *Arabidopsis*.

QTL for productivity have been previously reported on chromosomes 1, 2, 4, 5, 9, and 10 (Clement *et al.* 2003). Recently, McElroy *et al.* (2018) reported QTL related to the fresh weight of cacao pods on chromosomes 1, 3, 7, 8, 9, and 10. Previous studies on healthy pods did not find candidate genes related to the trait (Romero-Navarro *et al.* 2017). In this study, we found two candidate genes associated to the healthy pod trait (Tc08v2_g012850 and Tc00cons_t021570.1); however, further studies are necessary to confirm its role in cacao productivity.

Selective sweep analysis

A selective sweep analysis was conducted to identify nucleotide variation that can be associated with beneficial traits for plant adaptation (Stephan 2010). Recent studies demonstrated that genomic regions that exhibit selection signatures are also enriched in genes associated with biologically important traits (Wen *et al.* 2015; Xie *et al.* 2015). The discover and intepretation of regions under selection depends on the biological background of a population that is affected by different factors such as the effective population size, population structure, migration, introgression, etc. (Crisci *et al.* 2013). A comprehensive study of the evolutionary history of a population should be addressed using different approaches.

Therefore, two approaches were used to investigate regions under selection. First, the SweeD software calculates the site frequency spectrum (SFS) which describes the frequency of a beneficial mutation that shifts toward a high or low frequency derived alleles (Fay and Wu 2000). The second approach uses the OmegaPlus software that recognizes a pattern of linkage disequilibrium which is expected to increase in regions flanking a selected site, but not across it (Nielsen 2005). In order to avoid false positives errors, a consensus between the results of the LD-based method (OmegaPlus) and the results of the SFS-based method (SweeD) was done.

The chromosomes 1 and 3 showed the largest regions under selection comprising 5 and 39 genes, correspondingly. These genes were involved in different biological process; in particular genes Tc01v2_g016260, Tc03v2_g026440, and Tc03v2_g006130 could have an importance for pod productivity because they have a role in the ovule development (Byzova *et al.* 1999), the control of root growth (Pinosa *et al.* 2013) and fruit dehiscence, respectively (Fulton *et al.* 2009). The detection of selective sweeps could help unravel the genetic structure of complex traits (Chen *et al.* 2016). In our study, five out of 13 genes that were associated genes with productivity and disease resistance were located in chromosome regions under positive selection, indicating that natural selection is probably occurring in this cacao collection due to the natural incidence of FPRD and WBD in the Magdalena River valley region (Álvarez *et al.* 2014). However, further studies on evolution and domestication processes are necessary to validate these hypotheses.

Our findings expand previous genetic mapping efforts and allow increasing the mapping resolution of the regions responsible for productivity and disease resistance traits. Leveraging this information could contribute to select promising accessions in juvenile stages in greenhouse conditions, and consequently, reduce breeding cycles.

CONCLUSIONS

This research reported the first GWAS study of disease resistance trait for WBD and FPRD in cacao, based on SNPs produced by the GBS method. In total, two candidate genes related to productivity and seven related to disease resistance response to FPRD and WBD were detected. We also identified a total of 10 promising accessions that presented a degree of resistance response, and three accessions with promising productivity values. The current work provides new knowledge on genomic regions involved in the productivity and disease resistance response, which, after functional validation, could be useful in MAS for cacao breeding programs.

LITERATURE CITED

Acebo-Guerrero, Y., A. Hernández-Rodríguez, M. Heydrich-Pérez, M. El Jaziri, and A. N. Hernández-Lauzardo, 2012 Management of black pod rot in cacao (*Theobroma cacao* L.): a review. Fruits 67: 41–48. https://doi.org/10.1051/fruits/2011065

Acevedo-Garcia, J., S. Kusch, and R. Panstruga, 2014 Magical mystery tour: MLO proteins in plant immunity and beyond. New Phytol. 204: 273–281. https://doi.org/10.1111/nph.12889

Albores-Flores, V. J., G. García-Guzmán, F. J. Espinosa-García, and M. Salvador-Figueroa, 2018 Degree of domestication influences susceptibility of *Theobroma cacao* to frosty pod rot: a severe disease devastating Mexican cacao. Bot. Sci. 96: 84–94. https://doi.org/10.17129/botsci.1793

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19: 1655–1664. https://doi.org/10.1101/gr.094052.109

Álvarez, J. C., S. C. Martínez, and J. Coy, 2014 Estado de la moniliasis del cacao causada por *Moniliophthora roreri* en Colombia. Acta Agron. 63: 388–399. https://doi.org/10.15446/acag.v63n4.42747

Andrews, S., F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger *et al.*, 2012 FastQC: a quality control tool for high throughput sequence data.

Aranzazu, F., N. Martínez, G. Palencia, R. Coronado, and D. Rincón, 2009 Mejoramiento genético para incrementar la producción y productividad del sistema de cacao en Colombia FEDECACAO–CORPOICA. 128.

Argout, X., J. Salse, J.-M. Aury, M. J. Guiltinan, G. Droc *et al.*, 2010 The genome of *Theobroma cacao*. Nat. Genet. 43: 101–108. https://doi.org/10.1038/ng.736

Arguello, O., L. Mejia, N. Contreras, and J. Toloza, 1999 Manual de caracterización morfoagronómica de clones élite de cacao (*Theobroma cacao* L.) en el Nororiente Colombiano. 60.

Bailey, B. A., and L. W. Meinhardt, 2016 *Cacao diseases: a history of old enemies and new encounters*, Springer International Publishing, New York. https://doi.org/10.1007/978-3-319-24789-2

Bateman, R. P., E. Hidalgo, J. García, C. Arroyo, G. M. Ten Hoopen *et al.*, 2005 Application of chemical and biological agents for the management of frosty pod rot (*Moniliophthora roreri*) in Costa Rican cocoa (*Theobroma cacao*). Ann. Appl. Biol. 147: 129–138. https://doi.org/10.1111/j.1744-7348.2005.00012.x

Beavis, W., 1998 QTL analyses: power, precision, and accuracy, pp. 250–266 in *Molecular dissection of complex traits*, American Seed Trade Association, Chicago, IL.

Blaum, B. S., S. Mazzotta, E. R. Nöldeke, T. Halter, J. Madlung *et al.*, 2014 Structure of the pseudokinase domain of BIR2, a regulator of BAK1-mediated immune signaling in Arabidopsis. J. Struct. Biol. 186: 112–121. https://doi.org/10.1016/j.jsb.2014.02.005

Bowers, J., B. Bailey, P. Hebbar, S. Sanogo, and R. Lumsden, 2001 The impact of plant diseases on world chocolate production, Plant Heal. Prog. https://doi.org/10.1094/PHP-2001-0709-01-RV

Brachi, B., G. P. Morris, and J. O. Borevitz, 2011 Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol. 12: 232. https://doi.org/10.1186/gb-2011-12-10-232

Broad Institute, 2019 Picard toolkit.

Brown, J. S., W. Phillips-Mora, E. J. Power, C. Krol, C. Cervantes-Martinez *et al.*, 2007 Mapping QTLs for resistance to frosty pod and black pod diseases and horticultural traits in *Theobroma cacao* L. Crop Sci. 47: 1851–1858. https://doi.org/10.2135/cropsci2006.11.0753

Brown, J. S., R. J. Schnell, J. C. Motamayor, D. N. Kuhn, and J. W. Borrone, 2005 Resistance gene mapping for witches' broom disease in *Theobroma cacao* L. in an F2 population using SSR markers and candidate genes. J. Am. Soc. Hortic. Sci. 130: 366–373. https://doi.org/10.21273/JASHS.130.3.366

Byzova, M. V., J. Franken, M. G. Aarts, J. de Almeida-Engler, G. Engler *et al.*, 1999 Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. Genes Dev. 13: 1002–1014. https://doi.org/10.1101/gad.13.8.1002

Cheesman, E. E., 1944 *Notes on the Nomenclature, Classification and Possible Relationships of Cacao Populations*, IPC Science and Technology Press. 21: 144–159.

Chen, M., D. Pan, H. Ren, J. Fu, J. Li *et al.*, 2016 Identification of selective sweeps reveals divergent selection between Chinese Holstein and Simmental cattle populations. Genet. Sel. Evol. 48: 76. https://doi.org/10.1186/s12711-016-0254-5

Clement, D., A. M. Risterucci, J. C. Motamayor, J. N'Goran, and C. Lanaud, 2003 Mapping QTL for yield components, vigor, and resistance to Phytophthora palmivora in *Theobroma cacao* L. Genome 46: 204–212. https://doi.org/10.1139/g02-125

Collard, B. C. Y., and D. J. Mackill, 2008 Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. 363: 557–572.

Cooke, T. F., M.-C. Yee, M. Muzzio, A. Sockell, R. Bell *et al.*, 2016 GBStools: A statistical method for estimating allelic dropout in reduced representation requencing data. PLoS Genet. 12: e1005631. https://doi.org/10.1371/journal.pgen.1005631

Cornejo, O. E., M.-C. Yee, V. Dominguez, M. Andrews, A. Sockell *et al.*, 2018 Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. Commun. Biol. 1: 167. https://doi.org/10.1038/s42003-018-0168-6

Cosme, S., H. E. Cuevas, D. Zhang, T. K. Oleksyk, and B. M. Irish, 2016 Genetic diversity of naturalized cacao (*Theobroma cacao* L.) in Puerto Rico. Tree Genet. Genomes 12: 88. https://doi.org/10.1007/s11295-016-1045-4

Crisci, J. L., Y.-P. Poh, S. Mahajan, and J. D. Jensen, 2013 The impact of equilibrium assumptions on tests of selection. Front. Genet. 4: 235. https://doi.org/10.3389/fgene.2013.00235

Crouzillat, D., B. Ménard, A. Mora, W. Phillips, and V. Pétiard, 2000 Quantitative trait analysis in *Theobroma cacao* using molecular markers. Euphytica 114: 13–23. https://doi.org/10.1023/A:1003892217582

Cuatrecasas, J., 1964 Cacao and its allies: a taxonomic revision of the genus Theobroma. Contrib US Herb. 35: 379–614.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510. https://doi.org/10.1038/nrg3012

de Almeida, A.-A. F., and R. R. Valle, 2007 Ecophysiology of the cacao tree. Braz. J. Plant Physiol. 19: 425–448. https://doi.org/10.1590/S1677-04202007000400011

De Wever, J., H. Everaert, F. Coppieters, H. Rottiers, K. Dewettinck *et al.*, 2019 The development of a novel SNP genotyping assay to differentiate cacao clones. Sci. Rep. 9: 9512. https://doi.org/10.1038/s41598-019-45884-8

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379. https://doi.org/10.1371/journal.pone.0019379

End, M. J., R. M. Wadsworth, and P. Hadley, 1992 *International cocoa germplasm database*, University of Reading, London.

Faleiro, F. G., V. T. Queiroz, U. V. Lopes, C. T. Guimarães, J. L. Pires *et al.*, 2006 Mapping QTLs for witches' broom (*Crinipellis perniciosa*) resistance in cacao (*Theobroma Cacao* L.). Euphytica 149: 227–235. https://doi.org/10.1007/s10681-005-9070-7

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking Under Positive Darwinian Selection. Genetics 155: 1405–1413.

Fountain, A. C., and F. Hütz-Adams, 2018 Cocoa barometer. https://www.voicenetwork.eu/wp-content/uploads/2019/07/2018-Cocoa-Barometer.pdf

Fulton, L., M. Batoux, P. Vaddepalli, R. K. Yadav, W. Busch *et al.*, 2009 DETORQUEO, QUIRKY, and ZERZAUST represent novel components involved in organ development mediated by the receptor-like kinase STRUBBELIG in Arabidopsis thaliana. PLoS Genet. 5: e1000355. https://doi.org/10.1371/journal.pgen.1000355

Ganal, M. W., A. Polley, E.-M. Graner, J. Plieske, R. Wieseke *et al.*, 2012 Large SNP arrays for genotyping in crop plants. J. Biosci. 37: 821–828. https://doi.org/10.1007/s12038-012-9225-3

Gopaulchan, D., L. A. Motilal, F. L. Bekele, S. Clause, J. O. Ariko *et al.*, 2019 Morphological and genetic diversity of cacao (*Theobroma cacao* L.) in Uganda. Physiol. Mol. Biol. Plants 25: 361–375. https://doi.org/10.1007/s12298-018-0632-2

Hong, E. P., and J. W. Park, 2012 Sample Size and Statistical Power Calculation in Genetic Association Studies. Genomics Inform. 10: 117–122. https://doi.org/10.5808/GI.2012.10.2.117

Hunter, J. R., 1990 The status of cacao (*Theobroma cacao*, sterculiaceae) in the western hemisphere. Econ. Bot. 44: 425–439. https://doi.org/10.1007/BF02859775

Jaimes, Y., and F. Aranzazu, 2010 *Manejo de las enfermedades del cacao (Theobroma cacao L.) en Colombia*, *con énfasis en monilia (Moniliophthora roreri)*, Corpoica, Bogotá. https://doi.org/10.21930/978-958-740-034-2

Jaimes, Y. Y., C. Gonzalez, J. Rojas, O. E. Cornejo, M. F. Mideros *et al.*, 2016 Geographic differentiation and population genetic structure of *Moniliophthora roreri* in the principal cocoa production areas in Colombia. Plant Dis. 100: 1548–1558. https://doi.org/10.1094/PDIS-12-15-1498-RE

Jaimez, R. E., F. Amores Puyutaxi, A. Vasco, R. G. Loor, O. Tarqui *et al.*, 2018 Photosynthetic response to low and high light of cacao growing without shade in an area of low evaporative demand. Acta Biol. Colomb. 23: 95–103. https://doi.org/10.15446/abc.v23n1.64962

Ji, K., D. Zhang, L. A. Motilal, M. Boccara, P. Lachenaud *et al.*, 2013 Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. Genet. Resour. Crop Evol. 60: 441–453. https://doi.org/10.1007/s10722-012-9847-1

Jombart, T., and I. Ahmed, 2011 adegenet 1.3–1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27: 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Kalinowski, S. T., M. L. Taper, and T. C. Marshall, 2007 Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16: 1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kim, N. H., D. H. Lee, D. S. Choi, and B. K. Hwang, 2015 The pepper GNA-related lectin and PAN domain protein gene, CaGLP1, is required for plant cell death and defense signaling during bacterial infection. Plant Sci. 241: 307–315. https://doi.org/10.1016/j.plantsci.2015.07.003

Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. Genetics 167: 1513–1524. https://doi.org/10.1534/genetics.103.025387

Kim, Y.-T., J. Oh, K.-H. Kim, J.-Y. Uhm, and B.-M. Lee, 2009 Isolation and characterization of NgRLK1, a receptor-like kinase of *Nicotiana glutinosa* that interacts with the elicitin of *Phytophthora capsici*. Mol. Biol. Rep. 37: 717–727. https://doi.org/10.1007/s11033-009-9570-y

Klein, R. J., 2007 Power analysis for genome-wide association studies. BMC Genet. 8: 58. https://doi.org/10.1186/1471-2156-8-58

Knaus, B. J., and N. J. Grünwald, 2017 VcfR: a package to manipulate and visualize VCF format data in R. Mol Ecol Resour. 17: 44–53. https://doi.org/10.1111/1755-0998.12549

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29. https://doi.org/10.1186/1746-4811-9-29

Krueger, F., 2018 Trim Galore. Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Lachenaud, P., D. Clément, X. Argout, S. Scalabrin, and F. Doaré, 2018 The Guiana cacao genetic group (*Theobroma cacao* L.): a new core collection in French Guiana. Bot. Lett. 165: 248–254. https://doi.org/10.1080/23818107.2018.1465466

Lanaud, C., O. Fouet, D. Clément, M. Boccara, A. M. Risterucci *et al.*, 2009 A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. Mol. Breed. 24: 361–374. https://doi.org/10.1007/s11032-009-9297-4

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lipka, A. E., F. Tian, Q. Wang, J. Peiffer, M. Li *et al.*, 2012 GAPIT: genome association and prediction integrated tool. Bioinformatics 28: 2397–2399. https://doi.org/10.1093/bioinformatics/bts444

Liu, L.-P., J.-W. Qu, X.-Q. Yi, and H.-H. Huang, 2017 Genome-wide Identification, classification and expression analysis of the mildew resistance locus O (MLO) gene family in sweet orange (*Citrus sinensis*). Braz. Arch. Biol. Technol. 60: e17160474. https://doi.org/10.1590/1678-4324-2017160474

Loguercio, L. L., A. C. de Carvalho, G. R. Niella, J. T. De Souza, and A. W. V. Pomella, 2009 Selection of *Trichoderma stromaticum* isolates for efficient biological control of witches' broom disease in cacao. Biol. Control 51: 130–139. https://doi.org/10.1016/j.biocontrol.2009.06.005

McElroy, M. S., A. J. R. Navarro, G. Mustiga, C. Stack, S. Gezan *et al.*, 2018 Prediction of cacao (*Theobroma cacao*) resistance to *Moniliophthora spp*. diseases via genome-wide association analysis and genomic selection. Front. Plant Sci. 9: 343. https://doi.org/10.3389/fpls.2018.00343

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303. https://doi.org/10.1101/gr.107524.110

Meinhardt, L. W., J. Rincones, B. A. Bailey, M. C. Aime, G. W. Griffith *et al.*, 2008 Moniliophthora perniciosa, the causal agent of witches' broom disease of cacao: what's new from this old foe? Mol. Plant Pathol. 9: 577–588. https://doi.org/10.1111/j.1364-3703.2008.00496.x

Motamayor, J. C., P. Lachenaud, J. Wallace, R. Loor, D. N. Kuhn *et al.*, 2008 Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). PLoS One 3: e3311. https://doi.org/10.1371/journal.pone.0003311

Motamayor, J. C., K. Mockaitis, J. Schmutz, N. Haiminen, D. Livingstone *et al.*, 2013 The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 14: r53. https://doi.org/10.1186/gb-2013-14-6-r53

Motamayor, J., A. Risterucci, P. Lopez, C. Ortiz, A. Moreno *et al.*, 2002 Cocoa domestication I: the origin of the cocoa cultivated by the Mayas. Heredity 89: 380–386. https://doi.org/10.1038/sj.hdy.6800156

Motilal, L. A., D. Zhang, S. Mischke, L. W. Meinhardt, M. Boccara *et al.*, 2016 Association mapping of seed and disease resistance traits in *Theobroma cacao* L. Planta 244: 1265–1276. https://doi.org/10.1007/s00425-016-2582-7

Nielsen, R., 2005 Molecular Signatures of Natural Selection. Annu. Rev. Genet. 39: 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420

Osorio-Guarín, J. A., J. Berdugo-Cely, R. A. Coronado, Y. P. Zapata, C. Quintero *et al.*, 2017 Colombia a source of cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. Front. Plant Sci. 8: 1994. https://doi.org/10.3389/fpls.2017.01994

Osorio-Guarín, J. A., C. R. Quackenbush, and O. E. Cornejo, 2018 Ancestry informative alleles captured with reduced representation library sequencing in *Theobroma cacao*. PLoS One 13: e0203973. https://doi.org/10.1371/journal.pone.0203973

Parra, D., and Y. Duarte, 2007 *Caracterización morfológica de 100 accesiones en la colección colombiana de cacao (Theobroma Cacao L.)*, Universidad Industrial de Santander, Colombia.

Pavlidis, P., D. Živkovic, A. Stamatakis, and N. Alachiotis, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30: 2224–2234. https://doi.org/10.1093/molbev/mst112

Pinosa, F., M. Begheldo, T. Pasternak, M. Zermiani, I. A. Paponov *et al.*, 2013 The Arabidopsis thaliana Mob1A gene is required for organ growth and correct tissue patterning of the root tip. Ann. Bot. 112: 1803–1814. https://doi.org/10.1093/aob/mct235

Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink, 2012 Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253. https://doi.org/10.1371/journal.pone.0032253

Purdy, L. H., and R. A. Schmidt, 1996 Status of cacao witches' broom: biology, epidemiology, and management. Annu. Rev. Phytopathol. 34: 573–594. https://doi.org/10.1146/annurev.phyto.34.1.573

Queiroz, V. T., C. T. Guimarães, D. Anhert, I. Schuster, R. T. Daher *et al.*, 2003 Identification of a major QTL in cocoa (*Theobroma cacao* L.) associated with resistance to witches' broom disease. Plant Breed. 122: 268–272. https://doi.org/10.1046/j.1439-0523.2003.00809.x

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

R development core team, 2008 R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna Austria. http://www.R-project.org.

Rice, P., I. Longden, and A. Bleasby, 2000 EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16: 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Ríos, F., A. Ruiz, J. Lecaro, and C. Rehpani, 2017 *Estrategias país para la oferta de cacaos especiales políticas e iniciativas privadas exitosas en el Perú*, Colombia y República Dominicana, Ecuador.

Rodriguez-Medina, C., A. Caicedo Arana, O. Sounigo, X. Argout, G. A. Alvarado *et al.*, 2019 Cacao breeding in Colombia, past, present and future. Breed. Sci. 69: 373–382. https://doi.org/10.1270/jsbbs.19011

Romay, M. C., M. J. Millard, J. C. Glaubitz, J. A. Peiffer, K. L. Swarts *et al.*, 2013 Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14: R55. https://doi.org/10.1186/gb-2013-14-6-r55

Romero Navarro, J. A., W. Phillips-Mora, A. Arciniegas-Leal, A. Mata-Quirós, N. Haiminen *et al.*, 2017 Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. Front. Plant Sci. 8: 1905. https://doi.org/10.3389/fpls.2017.01905

Royaert, S., J. Jansen, D. V. da Silva, S. M. de Jesus Branco, D. S. Livingstone *et al.*, 2016 Identification of candidate genes involved in Witches' broom disease resistance in a segregating mapping population of *Theobroma cacao* L. in Brazil. BMC Genomics 17: 107. https://doi.org/10.1186/s12864-016-2415-x

Shaner, G., and R. E. Finney, 1977 The effect of nitrogen fertilization on the expression of slow-mildewing resistance in knox wheat. Phytopathology 67: 1051–1056. https://doi.org/10.1094/Phyto-67-1051

Singh, P., and L. Zimmerli, 2013 Lectin receptor kinases in plant innate immunity. Front. Plant Sci. 4: 124. https://doi.org/10.3389/fpls.2013.00124

Smýkal, P., M. N. Nelson, J. D. Berger, and E. J. B. Von Wettberg, 2018 The impact of genetic changes during crop domestication. Agronomy (Basel) 8: 1–22.

Stack, J. C., S. Royaert, O. Gutiérrez, C. Nagai, I. S. A. Holanda *et al.*, 2015 Assessing microsatellite linkage disequilibrium in wild, cultivated, and mapping populations of *Theobroma cacao* L. and its impact on association mapping. Tree Genet. Genomes 11: 19.

Stephan, W., 2010 Genetic hitchhiking *vs.* background selection: the controversy and its implications. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1245–1253. https://doi.org/10.1098/rstb.2009.0278

Tirado-gallego, P. A., A. Lopera-álvarez, and L. A. Ríos-osorio, 2016 Moniliophthora perniciosa en *Theobroma cacao* L. Corpoica Cienc Tecnol Agropecu. Mosquera 17: 2500–5308.

Wen, Z., J. F. Boyse, Q. Song, P. B. Cregan, and D. Wang, 2015 Genomic consequences of selection and genome-wide association mapping in soybean. BMC Genomics 16: 671. https://doi.org/10.1186/s12864-015-1872-y

Wickramasuriya, A. M., and J. M. Dunwell, 2018 Cacao biotechnology: current status and future prospects. Plant Biotechnol. J. 16: 4–17. https://doi.org/10.1111/pbi.12848

Wood, G. A. R., and R. Lass, 1985 *Cocoa*, Longman, London.

World Cocoa Foundation, 2014 Cocoa market update.

Xie, W., G. Wang, M. Yuan, W. Yao, K. Lyu *et al.*, 2015 Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc. Natl. Acad. Sci. USA 112: E5411–E5419. https://doi.org/10.1073/pnas.1515919112

Xu, Y., P. Li, Z. Yang, and C. Xu, 2017 Genetic mapping of quantitative trait loci in crops. Crop J. 5: 175–184. https://doi.org/10.1016/j.cj.2016.06.003

Yamada, M. M., J. L. Pires, F. G. Faleiro, U. V. Lopes, and M. M. Macedo, 2013 Agronomic performance of 27 cocoa progenies and plant selection based on productivity, self-compatibility and disease resistance. Rev. Ceres 60: 514–518. https://doi.org/10.1590/S0034-737X2013000400010

Zervudacki, J., A. Yu, D. Amesefe, J. Wang, J. Drouaud *et al.*, 2018 Transcriptional control and exploitation of an immune-responsive family of plant retrotransposons. EMBO J. 37: e98482. https://doi.org/10.15252/embj.201798482

Zhang, X.-R., Y.-P. Xu, and X.-Z. Cai, 2018 SlCNGC1 and SlCNGC14 Suppress *Xanthomonas oryzae pv. oryzicola*-Induced Hypersensitive Response and Non-host Resistance in Tomato. Front. Plant Sci. 9: 285. https://doi.org/10.3389/fpls.2018.00285

Zimmermann, S., T. Ehrhardt, G. Plesch, and B. Mueller-Roeber, 1999 Ion channels in plant signaling. Cell. Mol. Life Sci. 55: 183–203. https://doi.org/10.1007/s000180050284

Communicating editor: I. Parkin