2017 International Symposium on Cocoa Research (ISCR), Lima, Peru, 13-17 November 2017

Candidate SSR tags for fruit and seed traits of *Theobroma cacao* L. in the International Cocoa Genebank Trinidad.

L.A. Motilal[1], D. Zhang[2], S. Mischke[2], L.W. Meinhardt[2], and P. Umaharan[1].

[1]Cocoa Research Centre, Sir Frank Stockdale Bldg., The University of the West Indies, St. Augustine, 330912, Trinidad, Trinidad and Tobago.
[2]USDA/ARS, NEA, Beltsville Agricultural Research Center, SPCL, 10300 Baltimore Avenue, Bldg. 001, Room 223, BARC-W, Beltsville, MD 20705, USA.

Abstract

Increasing yield is a prominent feature of crop breeding programmes including the economically important cacao (*Theobroma cacao* L.). As a tropical tree crop, the time and acreage needed for selection of improved varieties are limiting factors. Selection at an early seedling stage in a marker-assisted selection programme is desirable. Candidate molecular microsatellite markers were identified under an association mapping approach for five fruit (fruit mass, husk mass, fruit length, fruit girth and fruit volume) and three seed (length, width and size of fresh peeled seeds) traits. Nine microsatellite markers (mTcCIR 19, 30, 40, 43, 57, 60, 126, 184 and 275) were consistently obtained under general and mixed linear models and explained between 4.68 – 12.87% of the observed variation. Markers mTcCIR60, mTcCIR126 and mTcCIR184 were most significantly associated with the reproductive traits. The adoption of these markers is recommended to the international cacao community.

Keywords: association mapping; microsatellite markers; marker-assisted selection; reproductive traits

Introduction

Cacao (*Theobroma cacao* L.), a diploid (2n = 20) tree in the family Malvaceae *sensu lato* (Alverson et al. 1999; Bayer et al. 1999), is an economically important plantation crop for many tropical countries worldwide (Eyre 2007). The centre of origin and diversity of this crop is in Amazonian South America (Cuatrecasas 1964; Motamayor et al. 2008). Genetic resources of cacao are established as field gene banks in national or universal collections (Butler and Umaharan 2004). These collections are good repositories to obtain breeding material for crop improvement. Breeding programmes have focussed on the economy of production by selecting for yield and disease resistance (Kennedy et al. 1987; Lockwood and Yin 1993; Lopes et al. 2011). Cacao breeding, though, is a long-term process due to the long reproductive cycle and the duration required for field trials. It took over 60 years in Trinidad to obtain the popular TSH cultivars (Gonsalves 1996; Maharaj et al. 2011) and 17 years in Brazil to obtain 41 farmer varieties (Lopes et al. 2011).

Acceleration of breeding goals has improved with the advent of molecular methods. With simulation studies, Crouzillat et al. (2000) demonstrated that, in cacao, the use of molecular markers alone or in combination with phenotypic selection was more effective than phenotypic evaluation only. This molecular breeding approach termed marker-assisted selection, marker-aided selection or marker-assisted breeding, uses a marker or set of markers associated with quantitative trait loci (QTL) to tag a trait of interest, thereby identifying improved individuals (Michelmore et al. 1991; Collard et al. 2005). Reviews on QTL analyses, and the application of markers in marker-assisted selection have been published (Paterson et al. 1991; Tanksley 1993; Hospital 2003; Peleman et al. 2005). Classical QTL analysis makes little or no use of ancestry information unlike admixture and association mapping. Admixture mapping (Rife 1954), is premised on population differentiation between ancestral populations, uses the local phenotype-ancestry correlation and is applied for recent (<20 generations) admixture (Shriner 2013). The approach is best used when different proportions of the allele affecting the trait are present in a recently admixed population derived from two known progenitors (Darvasi and Shifman 2005). Using admixture mapping, Marcano et al. (2007) identified 15 genomic regions that influenced seed and fruit mass variation using 101 microsatellites (SSRs) on 150 germplasm accessions and 92 SSRs on 291 plantation individuals. Similarly, Marcano et al. (2009), using 257 individuals and 92 SSRs identified several SSR markers linked to productivity, yield, bean dimensions, pigmentation, pubescence and fruit rugosity.

In contrast, direct association mapping, tests the genotype-phenotype correlation, and is premised on similar allele frequencies across multiple ancestries allowing for fine-scale localization (Buckler and Thornsberry 2002; Flint-Garcia et al. 2003; Shriner 2013). In association mapping (association analysis or linkage disequilibrium mapping), the identified markers have the advantage of being broad-based in application instead of being restricted to a population or populations (Yu and Buckler 2006). Linkage disequilibrium (LD) is the higher-than-normal or lower-than-normal occurrence of natural non-random combinations of alleles at two or more loci. Association mapping is reliant on LD to examine the correlation between phenotypic variation and genetic polymorphisms (Flint-Garcia et al. 2003; Yu and Buckler 2006). Spurious or false associations may arise due to population structure and were minimised by accounting for population stratification and relatedness (Aranzana et al. 2005; Price et al. 2006; Yu et al. 2006). Association mapping studies on a wide range of plants have been reviewed (Flint-Garcia et al. 2003; Zhu et al. 2008; Soto-Cerda and Cloutier 2012; Gupta et al. 2014). Association mapping studies in cacao, although limited, have found markers for fruit colour (Motamayor et al. 2013; Stack et al. 2015), resistance to frosty pod disease (Romero Navarro et al. 2017), number of seeds and resistance to blackpod and witches' broom disease (Motilal et al. 2016). This study was therefore undertaken to search for SSR markers that may be linked fruit and seed traits of economic value in *T. cacao* L.

Materials and Methods

Phenotyping

Fruits were harvested, primarily from the main trunk, but also from primary and secondary branches from selected trees of 398 accessions in the International Cocoa Genebank Trinidad (ICGT). A minimum of three fruits of a unique accession was sampled. Fruits were sometimes harvested from multiple trees that were deemed equivalent from multilocus molecular profiles. Eight quantitative traits (Table 1) were evaluated in the laboratory. Fruit mass (FM) was determined on the same day of collection using an ACBplus-1500 top-loading balance with sensitivity of ± 0.05 g (Adam Equipment Co. Ltd., USA). The husk mass (HM) was obtained by subtracting the mass of the placental body from the fruit mass. Fruit length (FL) and fruit girth (FG) were measured with the aid of a tailor measuring tape. The tape was run along the maximum curvature of the fruit to obtain the length and at the equator or maximum girth of the fruit to obtain the fruit girth. The fruit volume (FV) was calculated using the FL and FG and treating the fruits as ellipsoidal forms. From each fruit, five seeds were selected from one of the five loculi. Seeds at the very apical and basal ends were avoided and when sufficient seeds were available, alternate seeds along the loculus were selected; otherwise contiguous seeds were sampled. The mucilaginous pulp (aril) of each seed was hand-peeled and the fresh bean length (FBL) and fresh bean width (FBW) of each peeled seed (unit of embryo with pair of cotyledons) was determined using a digital calliper (Mitutoyo Corporation, Japan). The sizes of the fresh peeled seeds (FBS) were determined from the corresponding lengths and widths.

Table 1 Fruit and seed quantitative traits of *Theobroma cacao* under study

Fruit Trait	Acronym	Unit	Formula[1]
Fresh bean length	FBL	mm	none
Fresh bean size	FBS	mm^2	FBL × FBW
Fresh bean width	FBW	mm	none
Fruit girth	FG	cm	none
Fruit length	FL	cm	none
Fruit mass	FM	g	none
Fruit volume	FV	cm^3	$(7 \times FL \times FG^2)/66$
Husk mass	HM	g	FM – mass of placental body

[1]Derived traits are those whose values are determined from formulae

Genotyping and population structure

Multilocus SSR profiles were obtained from 95 loci for each of the 398 samples on a Beckman Coulter 8000 or 8800 capillary sequencer. Population structure was determined independently from a set of 27 or 52 SSRs using a burn-in of 500,000 and 1×10^6 MCMC runs were performed for 20 iterations at K = 2-17 using Structure v2.3.4 (Pritchard et al. 2000).

Association mapping analysis

Trait ancestry and marker data were taken into TASSEL v4.2.1 (Bradbury et al. 2007). Association analysis can be configured as in Figure 1. In this YEAST model, the system information ($S_\alpha t$) can be taken from the ancestry information ($S_Q t$), the multivariate analysis based on molecular data ($S_M t$) or the kinship relationship based on molecular data ($S_K t$). Incorporation of a kinship matrix turns a general linear model (GLM) into a mixed linear model (MLM). The genotype file was filtered to remove alleles with frequency < 0.01 and retained for further manipulation. The filtered genotype file was used to create the kinship matrix in Tassel v4.2.1 (Bradbury et al. 2007). The filtered genotype file was collapsed and markers with >10% missing data were identified for exclusion from the un-collapsed filtered genotype file. After removal of these markers, the pruned file was collapsed and missing values were imputed from unweighted averages of three nearest neighbours, using a Manhattan distance. The principal components matrix was created from the repopulated collapsed file using a covariance method and eigenvectors were retained for three axes. The ancestry file was used as a covariate and one of the populations was removed from the analyses. Markers used for determination of ancestry were excluded from the allele file for association mapping analysis. Datasets were joined using the intersect function to minimise the incidence of missing phenotypic values across genotypes or allelic information for phenotypes. General linear models using a least squares solution (Searle 1987) on trait data were run independently using the default settings of 1000 permutations and the permutation test of Anderson and Ter Braak (2003). Mixed linear models were run independently using optimum level compression (Yu et al. 2006; Zhang et al. 2010) and P3D estimation of the variance component (Zhang et al. 2010). The strategies employed are presented in Table 2. Sample sizes within the SSR dataset/model combinations ranged from 195-277 for HM; 212-300 for FG, FL FM, and FV; and 140-200 for dimensions of fresh beans.

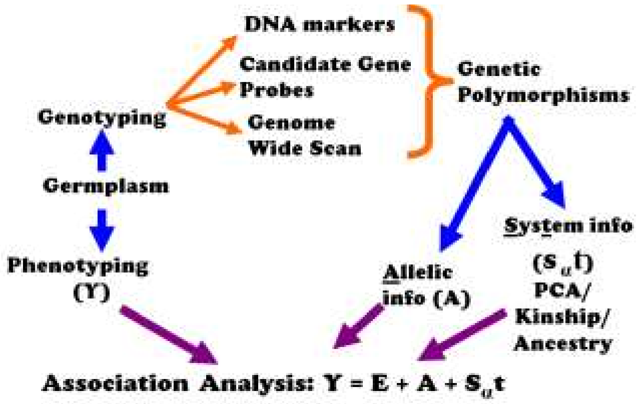

Figure 1 Schematic of association mapping
Modified from Zhu et al. (2008). The trait (Y) is modelled by allocating information to the molecular background ($S_\alpha t$) as ancestry ($S_Q t$), principal components ($S_M t$) or kinship relatedness ($S_K t$). E represents the error term and A, the allelic profiles. General Linear Models (GLM) do not contain kinship matrices unlike Mixed Linear Models (MLM). Differing combinations of $S_\alpha t$ terms may be present in the modelled equation.

Table 2 Synopsis of association mapping strategy in the present study

SSR Dataset[1]	# Ancestry markers	Ancestry filter[2]	# Markers for PCA & kinship[3]	# Tassel Markers	Models[4]
B_all	none	none	43 (3 PCA axes)	95	GLM2, MLM3
B1	52	Q10 to Q9	21 (3 PCA axes)	43	GLM2, 3; MLM3, 4
B2	27	Q10 to Q9	32 (3 PCA axes)	68	GLM2, 3; MLM3, 4

[1] SSR – microsatellite;
[2] Q – general population code
[3] PCA – principal coordinate analysis
[4] GLM – general linear model; MLM – mixed linear model;
GLM2: $Y = E + A + S_M t$; GLM3: $Y = E + A + S_Q t + S_M t$; MLM2: $Y = E + A + S_Q t + S_K t$;
MLM3: $Y = E + A + S_M t + S_K t$; MLM4: $Y = E + A + S_Q t + S_M t + S_K t$.
where Y = response, E = error, A = allelic information, $S_Q t$ = population ancestry matrix, $S_M t$ = principal components, $S_K t$ = kinship relatedness.

Selecting associated markers

Probability values from Tassel output were compared to Bonferroni (Bonferroni 1936; Dunn 1959, 1961) corrected p-values at the 5% level of significance. Final selection of associated markers employed the following criteria: (a) present in more than one dataset; (b) present in at least two models; (c) most constraining model or dataset chosen from (a) and (b) preceding; (d) if LD as r^2 is ≥ 0.1 or if they were within the LD decay distance (9.3 cM for chromosomes 1-9 and 2.5 cM for chromosome 10; Motilal et al. 2016), then only one marker was chosen; and (e) multiple markers from (d) were reduced by retaining the smallest set of markers to represent the total set, selecting markers with lowest p-values, selecting markers common to more than one trait and selecting markers with at least five observations in the effect output of the Tassel run.

Results

The majority of the associated markers were obtained under GLM rather than MLM models (Table 3). Over the dataset/model combinations, between 1-11 markers could be tagged to the studied traits, with FG having the most potentially associated markers. Consistently reported markers across models within a dataset and across datasets were present. For example, under both GLM2 and MLM3 models in the B_all dataset, common markers for FBL (mTcCIR60 and mTcCIR126), FBS (mTcCIR60), FG (mTcCIR184), FM (SHRSTc44) and FV (SHRSTc44) were found. Several significant markers had to be discarded because the effect size from the Tassel output was based on less than five observations. Further reduction was possible when LD was considered. For instance, six SSR loci (mTcCIR 40, 77, 126, 184, 275; SHRSTc44) could be retained for FV based on dataset/model considerations but since mTcCIR77, mTcCIR126 and SHRSTc44 were in LD, only one locus (mTcCIR126) was chosen to represent this set.

The final set of retained SSR loci that were tagged to traits were found on chromosomes 1, 2, 3, 4 and 9, explained between 4.68-12.87% of trait variation and had an overall mean of 8.08 ± 0.42% (Table 4).Traits were tagged with one (FBW, FBS, FL), three (FM), four (FBL, FV, HM) or six (FG) loci. Five loci were most informative, tagging three (mTcCIR40, mTcCIR275), four (mTcCIR60, mTcCIR184) and six (mTcCIR126) traits. Applying a significance threshold of 5×10^{-5} identified three SSRs (mTcCIR60, 126, 184) that were strongly associated within the set of markers identified in the association analysis (Table 4). The locus mTcCIR60 was associated with all seed traits and mTcCIR126 was associated with all fruit traits. The locus mTcCIR184 was associated with all fruit traits except for fruit length.

Discussion

Eight SSRs were found under an association mapping approach to tag eight traits based on fruit and seed phenotypes in *T. cacao*. The markers identified represent possible sets as correlated markers (LD as multiallelic $r^2 > 0.1$, markers within decay distance) and imprecise markers (less than 5 observations in effect size) were discarded. It was therefore possible those potentially useful markers were not selected and that the identified sets represented the best minimum number of associated markers in the current study. The accumulation of more phenotypic data across all traits is therefore recommended so as to reduce the incidence of missing data and to increase the number of phenotyped individuals. This should improve the chances of getting more than five observations per genotypic state in the effects file. It would also substantially improve the power of association mapping studies as increasing the number of phenotyped individuals is more effective than increasing the number of SNPs (Long and Langley 1999; Myles et al. 2009). The number of markers was variably increased depending on the trait involved and the model employed, with GLM having more associated markers. The use of MLM models together with PCA have been reported in the literature (Price et al. 2006; Yu et al. 2006; Zhao et al. 2007; Raman et al. 2010). MLM models have been shown to be more effective in controlling for spurious association than GLM models (Yu et al. 2006; Zhao et al. 2007; Raman et al. 2010; Soto-Cerda and Cloutier 2012) with false positives being primarily due to population structure and relatedness (Thornsberry et al. 2001; Aranzana et al. 2005; Price et al. 2006; Yu et al. 2006). Stich et al. (2005) suggested that MLM models and using the ancestry file for genetic structure did not correct for LD caused by selection and genetic drift. The markers presented in Table 4 were considered to be likely trait tags because they were common across datasets; they were common across models; population structure was variably accounted for by ancestry, kinship and/or principal coordinate analyses; and some markers were at a highly significant *p*-threshold, tenfold lower than that indicated by Bonferroni adjustment.

Marcano et al. (2007) found that mTcCIR184 and mTcCIR275 were linked to QTL for fruit mass. Like Marcano et al. (2009) associations were found for mTcCIR30 with FBL and mTcCIR157 with FBW supporting the reliability of these markers. However, the mTcCIR60 marker identified in the present study for FBL, FBMF and FBW was 3 cM distant from mTcCIR253, a locus absent from the present study but found by Marcano et al. (2009) for FBMF, bean length and bean width. Since mTcCIR60 and mTcCIR253 were within the LD decay distance, the reliability of mTcCIR60 is supported. The marker mTcCIR60 was also within 4 cM of a flanking marker for a QTL for bean length (Clement et al. 2003a, 2003b). The SSR locus mTcCIR60 which was associated with quantitative fruit and seed traits in this study was also found to be associated with productivity (Schnell et al. 2005). The markers found associated to the traits may be used as candidate markers for trait expression. These can be employed in a MAS programme to help identify promising progeny at the seedling stage and reduce the number of plants required for phenotypic evaluation. The efficiency of marker-assisted vs. phenotype-assisted selection is higher for traits of low heritability (Collard et al. 2005). Narrow and broad sense heritabilities for a variety of fruit and seed traits (fruit length, fruit diameter, fruit mass, fruit wall width, number of seeds, wet mass of seeds, husk mass, pod index and seed index) ranged from 0.36 – 0.79 and 0.54 – 0.93, respectively, with fruit wall thickness having the lowest heritabilities (Mora et al. 1987).

Table 3 Microsatellite markers associated with cacao phenotypes

Trait	Dataset	Model[1]	Microsatellite marker[2]
Fresh bean length (mm)	B_all	GLM2	m30, m40, m60, m126, m140; S51
	B_all	MLM3	m60, m126
	B1	GLM2	m30
	B2	GLM2	m40, m60; S51
	B2	MLM3	m60; S51
Fresh bean size (mm^2)	B_all	GLM2	m60; S51
	B_all	MLM3	m60
	B1	GLM2	m43
	B2	GLM2	m60; S51
	B2	GLM3	m157
	B2	MLM3	m60
Fresh bean width (mm)	B_all	GLM2	m60
	B2	GLM2; MLM3	m60
	B2	GLM3	m157
Fruit girth (cm)	B_all	GLM2	m19, m40, m43, m60, m77, m90, m126, m184, m225, m275; S44
	B_all	MLM3	m184
	B1	GLM2	m43, m77, m184; S44
	B2	GLM2	m19, m40, m60, m90, m126, m184, m275
	B2	MLM3	m37
Fruit length (cm)	B_all	GLM2	m126; S44
	B1	GLM2	S44
	B1	MLM4	m225
	B2	GLM2	m126, m275
Fruit mass (g)	B_all	GLM2	m43, m77, m90, m126, m184, m275; S44
	B_all	MLM3	S44
	B1	GLM2	m77, m184; S44
	B1	MLM3	m184
	B2	GLM2	m90, m126, m184, m275
Fruit volume (cm^3)	B_all	GLM2	m40, m43, m77, m126, m184, m225, m275; S44
	B_all	MLM3	m126; S44
	B1	GLM2	m77, m184; S44
	B1	MLM3	m184
	B2	GLM2	m19, m40, m126, m184, m275
	B2	GLM3	m275
Husk mass (g)	B_all	GLM2	m43, m57, m77, m126, m184, m275; S44
	B1	GLM2	m57, m77, m184; S44
	B1	MLM3	m184
	B2	GLM2	m90, m126, m184, m275
	B2	MLM3	m184

[1] GLM – general linear model; MLM – mixed linear model; Y = trait value, E = error, A = allele information, S_Qt = population ancestry matrix, S_Kt = kinship matrix, S_Mt = principal component matrix
GLM2: $Y = E + A + S_Mt$; MLM3: $Y = E + A + S_Mt + S_Kt$; MLM4: $Y = E + A + S_Qt + S_Mt + S_Kt$

[2] m = mTcCIR; S = SHRSTc

Details of dataset can be found in Table 2

Table 4 Selected microsatellite markers significantly associated with cacao (Theobroma cacao L.) phenotypic traits

Trait	Marker[1]	High effect	Chrom[2]	Position (cM)	Dataset/ Model[3]	%Var(P)[4]
Fresh bean length (mm)	CIR30	176/184	9	22.1	B_all/GLM2	8.60
	CIR40	288/288	3	17.1	B_all/GLM2	10.00
	CIR60	195/215	2	54.6	B_all/MLM3	12.87
	CIR126	214/214	9	9.70	B_all/MLM3	9.63
Fresh bean width (mm)	CIR60	195/215	2	54.6	B2/MLM3	9.34
Fresh bean size (mm^2)	CIR60	195/215	2	54.6	B_all/MLM3	12.39
Fruit girth (cm)	CIR19	376/376	2	14.6	B_all/GLM2	6.61
	CIR40	288/288	3	17.1	B_all/GLM2	7.27
	CIR43	208/208	4	33.4	B_all/GLM2	6.23
	CIR60	195/195	2	54.6	B_all/GLM2	6.72
	CIR126	208/208	9	9.7	B_all/GLM2	10.90
	CIR184	117/117	1	2.0	B_all/MLM3	6.26
Fruit length (cm)	CIR126	208/208	9	9.7	B_all/GLM2	7.11
Fruit mass (g)	CIR126	208/208	9	9.7	B_all/GLM2	10.50
	CIR184	117/117	1	2.0	B_all/GLM2	8.13
	CIR275	146/146	1	81.4	B_all/GLM2	7.10
Fruit volume (cm^3)	CIR40	288/288	3	17.1	B_all/GLM2	6.67
	CIR126	208/208	9	9.7	B_all/MLM3	6.45
	CIR184	117/117	1	2.0	B1/MLM3	6.85
	CIR275	146/146	1	81.4	B2/GLM3	4.68
Husk mass (g)	CIR57	251/255	4	53.6	B_all/GLM2	7.65
	CIR126	208/208	9	9.7	B_all/GLM2	7.83
	CIR184	117/117	1	2.0	B1/MLM3	7.64
	CIR275	146/146	1	81.4	B_all/GLM2	6.66

[1] CIR = mTcCIR; S = SHRSTc; entries with $p \leq 5 \times 10^{-5}$ bolded;

[2] Chromosome and map position from SSR/SNP consensus map of CocoaGenDb (http://cocoagendb.cirad.fr/) except for S44 obtained from Kuhn et al. (2006)

[3] Datasets as in Table 9.1; Models are general linear models (GLM) or mixed linear models (MLM); GLM2: $Y = E + A + S_{M}t$; GLM3: $Y = E + A + S_{Q}t + S_{M}t$; MLM3: $Y = E + A + S_{M}t + S_{K}t$; MLM4: $Y = E + A + S_{Q}t + S_{M}t + S_{K}t$.

[4] pecentage of phenotypic variation explained

These results suggested that MAS may not have significant advantage over phenotypic selection for traits with high heritability in cacao. However, phenotypic evaluations is often time-consuming, difficult or costly (Dreher et al. 2003; Young 1999; Yu et al. 2000). Current trends indicate cost reduction for SNP genotyping which should make MAS more cost-effective and therefore more favourable than phenotypic selection. The limited availability of land resources in terms of quantity and issues of tenure may also weigh against phenotypic selection due to the long vegetative phase and number of years needed to obtain productivity values. It would be more cost-effective to screen progenies at the greenhouse stage under an MAS scenario and select the most promising ones for phenotypic validation. Moreover, the approach of genomic selection may be more promising as the cost of next generation sequencing continues to decrease. In contrast to MAS which utilizes markers to track small numbers of loci with large effects, genomic selection uses large set of marker information distributed across the whole genome to predict breeding values of individuals. Once the prediction model is established based on training populations, the selection can be based on markers only without known phenotype (Isik, 2014).

Acknowledgements

This study was partially funded by a grant from the Government of Trinidad and Tobago Research Development Fund and from The UWI-Trinidad and Tobago Research and Development Impact Fund from The University of the West Indies. Field evaluations were made possible by several persons including Mr. Leon Ali, Ms. Zainab Ali, Ms. Sarah Bharath, Ms. Rena K. Kalloo, Ms. Carelene Lakhan, Mr. Gangadeen Ramdhanie, Ms. Cassandra Shaw, Mr. Valmiki Singh and Mr. Eusebius Solozano.

References

Alverson, W.S., B.A. Whitlock, R. Nyffler, C. Bayer, and D.A. Baum. 1999. Phylogeny of the core Malvales: evidence from ndhF sequence data. *American Journal of Botany* 86: 1474-1486.

Anderson, M.J., and C.J.F. Ter Braak. 2003. Permutations tests for multifactorial analysis of variance. *Journal of Statistical Computation and Simulation* 73: 85-113.

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics* 1:e60. doi: 10.1371/journal.pgen.0010060.

Bayer, C., M.F. Fay, P.Y. De Bruijn, V. Savolainen, C.M. Morton, K. Kubitzki, W.S. Alverson, and M.W. Chase. 1999. Support for an expanded family concept of Malvaceae within a recircumscibed order Malvales: a combined analysis of plastid atpB and rbcL DNA sequences. *Botanical Journal of the Linnaean Society* 129: 267-303.

Bonferroni, C.E. 1936. *Teoria statistica delle classi e calcolo delle probabilità*, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19): 2633-2635. doi: 10.1093/bioinformatics/btm308.

Buckler, E.S., and J.M. Thornsberry. 2002. Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology* 5: 107-111.

Butler, D.R., and P. Umaharan. 2004. Working with cocoa germplasm. In *Cocoa Futures. A source book of some important issues facing the cocoa industry*, edited by J. Flood and R. Murphy, 54-63. Chinchiná, Colombia: CABI-FEDERACAFÉ, USDA.

Clément, D., A.M. Risterucci, J.C. Motamayor, J. N'Goran, and C. Lanaud. 2003a. Mapping QTL for yield components, vigor and resistance to *Phytophthora palmivora* in *Theobroma cacao* L. *Genome* 46: 204-212.

Clément, D., A.M. Risterucci, J.C. Motamayor, J. N'Goran, and C. Lanaud. 2003b. Mapping quantitative trait loci for bean traits and ovule number in *Theobroma cacao* L. *Genome* 46: 103-111.

Collard, B.C.Y., M.Z.Z. Jahufer, J.B. Brouwer, and E.C.K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142: 169-196. doi:10.1007/s10681-005-1681-5.

Crouzillat, D., B. Menard, A. Mora, W. Phillips, and V. Petiard. 2000a. Quantitative trait analysis in *Theobroma cacao* using molecular markers. Yield QTL detection and stability over 15 years. *Euphytica* 114: 13-23.

Cuatrecasas, J. 1964. Cacao and its allies. A taxonomic revision of the genus *Theobroma*. In *Contributions to the U.S. National Herbarium* 35(6): 375-614. Washington, DC: Smithsonian Institution.

Darvasi A., and S. Shifman. 2005. The beauty of admixture. *Nature Genetics* 37: 118-119.

Dreher, K., M. Khairallah, J. Ribaut, and M. Morris. 2003. Money matters (I): Costs of field and laboratory procedures associated with conventional and marker-assisted maize breeding at CIMMYT. *Molecular Breeding* 11: 221-234.

Dunn, O.J. 1959. Estimation of the medians for dependent variables. *The Annals of Mathematical Statistics* 30: 192-197. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177706374.

Dunn, O.J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association* 56: 52-64. http://sci2s.ugr.es/keel/pdf/algorithm/articulo/1961-Bonferroni_Dunn-JASA.pdf.

Eyre, C. 2007. Cocoa demand on the up, July 27, 2007. https://www.confectionerynews.com/Article/2007/07/27/Cocoa-demand-on-the-up (accessed August 02, 2007).

Flint-Garcia, S.A., J.M. Thornberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54: 357-374.

Gonsalves, C. 1996. History of cocoa breeding in the Ministry of Agriculture, Trinidad. *Cocoa Research Unit Newsletter* 3: 4-6.

Gupta, P.K., P.L. Kulwal, and V. Jaiswal. 2014. Association mapping in crop plants: opportunities and challenges. *Advances in Genetics* 85: 109-147.

Hospital, F. 2003. Marker-assisted breeding. In *Plant Molecular Breeding*, edited by H. J. Newbury, 30-59. Oxford, UK: CRC Press, Blackwell.

Isik, F. 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forest* 45:379–401

Kennedy, A.J., G. Lockwood, G. Mossu, N.W. Simmonds, and G.Y. Tan. 1987. Cocoa breeding: past, present and future. *Cocoa Growers' Bulletin* 38:5-22.

Kuhn, D.N., G. Narasimhan, K. Nakamura, J.S. Brown, R.J. Schnell, and A.W. Meerow. 2006. Identification of cacao TIR-NBS-LRR resistance gene homologues and their use as genetic markers. *Journal of the American Society of Horticultural Science* 131: 806-813.

Lockwood, G., and J.P.T. Yin. 1993. Utilization of cocoa germplasm in breeding for yield. In *Proceedings of the International Workshop on Conservation, Characterisation and Utilisation of Cocoa Genetic Resources in the 21st Century*, 13-17 September, 1992, 198-214. Port of Spain, Trinidad: The Cocoa Research Unit.

Long, A.D., and C.H. Langley. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research* 9: 720-731.

Lopes, U.V., W.R. Monteiro, J.L. Pires, D. Clement, M.M. Yamada, and K.P. Gramacho. 2011. Cacao breeding in Bahia, Brazil – strategies and results. *Crop Breeding and Applied Biotechnology* S1: 73-81.

Maharaj, K., P. Maharaj, F.L. Bekele, D. Ramnath, G.G. Bidaisee, I. Bekele, C. Persad, K. Jennings, and R. Sankar. 2011. Trinidad selected hybrids: An investigation of the phenotypic and agro-economic traits of 20 selected cacao cultivars. *Tropical Agriculture (Trinidad)* 88: 175-185.

Marcano M., S. Morales, M.T. Hoyer, B. Courtois, A.M. Risterucci, O. Fouet, and T. Pugh. 2009. A genomewide admixture mapping study for yield factors and morphological traits in a cultivated cocoa (*Theobroma cacao* L.) population. *Tree Genetics & Genomes* 5: 329–337. doi: 10.1007/s11295-008-0185-6.

Marcano, M., T. Pugh, E. Cros, S. Morales, E.A. Portillo Paez, B. Courtois, J.C. Glaszmann, et al. 2007. Adding value to cocoa (*Theobroma cacao* L.) germplasm information with domestication history and admixture mapping. *Theoretical and Applied Genetics* 114: 877-884.

Michelmore, R.W., I. Paran, R.V. Kesseli. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proceedings of the National Academy of Sciences USA* 88: 9828-9832.

Mora, L.G.R. 1987. Herencia de ciertos caracteres la mazorca y del arbol de cacao (*Theobroma cacao* L). M.Sc. thesis, CATIE, Turrialba, Costa Rica.

Motamayor, J.C., P. Lachneaud, J.W. da Silva e Mota, R. Loor, D.N. Kuhn, J.S. Brown, and R.J. Schnell. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS ONE* 3(10): e3311. doi: 10.1371/journal.pone.0003311.

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

Motilal, L.A., D. Zhang, S. Mischke, L.W. Meinhardt, M. Boccara, O. Fouet, C. Lanaud, and P. Umaharan. 2016. Association mapping of seed and disease resistance traits in *Theobroma cacao* L. *Planta* 244(6): 1265-1276. doi 10.1007/s00425-016-7.

Myles S, J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, and E.S. Buckler. 2009. Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21: 2194-2202.

Paterson, A., S. Tanksley, and M.E. Sorrels. 1991. DNA markers in plant improvement. *Advances in Agronomy* 44: 39-90.

Peleman, J.D., A.P. Sørensen, and J. R. van der Voort. 2005. Breeding by design: exploiting genetic maps and molecular markers through marker-assisted selection. In: *The handbook of plant genome mapping. Genetic and physical mapping*, edited by K. Meksem and G. Kahl, 109-129. Weinheim: WILEY-VCH Verlag GmbH & Co. KGaA.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and R. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38: 904-909. doi: 10.1038/ng1847.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure from multilocus genotype data. *Genetics* 155: 945-959.

Raman, H., B. Stodart, P. Ryan, E. Delhaize, L. Emebiri, R. Raman, N. Coombes, and A. Milgate. 2010. Genome-wide association analysis of common wheat (*Triticum aestivum* L.) germplasm identifies multiple loci for aluminium resistance. *Genome* 53: 957-966.

Rife, D.C. 1954. Populations of hybrid origin as source material for the detection of linkage. American *Journal of Human Genetics* 6: 26-33.

Romero Navarro, J.A., W. Phillips-Mora, A. Arciniegas-Leal, A. Mata-Quirós, N. Haiminen, G. Mustiga, D. Livingstone III, H. van Bakel, D.N. Kuhn, L. Parida, A. Kasarskis, and J.C. Motamayor. 2017. Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. *Frontiers in Plant Science* 8: 1905. doi: 10.3389/fpls.2017.01905

Schnell, R.J., C.T. Olano, J.S. Brown, A.W. Meerow, C. Cervantes-Martinez, C. Nagai, and J.C. Motamayor. 2005. Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. *Journal of the American Society of Horticultural Science* 130: 181-190.

Searle, S.R. 1987. *Linear models for unbalanced data*. New York: Wiley.

Shriner, D. 2013. Overview of admixture mapping. *Current Protocols in Human Genetics* 76:1.23:1.23.1–1.23.8. doi: 10.1002/0471142905.hg0123s76.

Soto-Cerda, B. J., and S. Cloutier. 2012. *Association mapping in plant genomes, genetic diversity in plants*, edited by Mahmut Caliskan, InTech. http://www.intechopen.com/books/genetic-diversity-in-plants/association-mapping-in-plant-genomes (accessed February 13, 2013).

Stack, J.C., S. Royaert, O. Gutiérrez, C. Nagai, I.S.A. Holanda, R. Schnell, and J.C. Motamayor. 2015. Assessing microsatellite linkage disequilibrium in wild, cultivated, and mapping populations of *Theobroma cacao* L. and its impact on association mapping. *Tree Genetics Genomes* 11:19. doi:10.1007/s11295-015-0839-0

Stich, B., A.E. Melchinger, M. Frisch, H.P. Maurer, M. Heckenberger, and J.C. Reif. 2005. Linkage disequilibrium in European elite maize germplasm investigated with SSRs. *Theoretical and Applied Genetics* 111: 723-730.

Tanksley, S.D. 1993. Mapping polygenes. *Annual Review of Genetics* 27: 205-233.

Thornsberry, J., M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. Buckler. 2001. Dwarf polymorphisms associate with variation in flowering time. *Nature Genetics* 28:286-289.

Young, N.D. 1999. A cautiously optimistic vision for marker-assisted breeding. *Molecular Breeding* 5: 505-510.

Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology* 17: 155-160. doi: 10.1016/j.copbio.2006.02.003.

Yu, K., S. Park, and V. Poysa. 2000. Marker-assisted selection of common beans for resistance to common bacterial blight: Efficacy and economics. *Plant Breeding* 119: 411-415.

Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38: 203-208. doi: 10.1038/ng1702.

Zhang, Z., E. Ersoz, C.-Q. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, P.J. Bradbury, et al. 2010. Mixed linear model approach adapted for genome wide association studies. *Nature Genetics* 42: 355-60. doi: 10.1038/ng.546.

Zhao, K., M. Aranzana, and S. Kim. 2007. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genetics* 3(1): e4.

Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and prospects of association mapping in plants. *The Plant Genome* 1: 5-20. doi: 10.3835/plantgenome2008.02.0089.