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

Genetic Diversity and Genetic Structure of Wild Cacao Collected in the Oriente using Single Nucleotide Polymorphisms

A. A. Sankar[1], L.A. Motilal[1], D. Zhang[2] and P. Umaharan[1]
[1]Cocoa Research Centre, The UWI, St. Augustine, Trinidad and Tobago
[2]Beltsville Agricultural Research Centre, ARS, USDA, Beltville, MD, USA

ABSTRACT

Cacao, the precious resource upon which a global multi-billion dollar industry depends is under threat from diseases such as Black Pod (a global scourge), Witches' Broom, and Frosty Pod. The latter two are confined to parts of the Americas but pose a threat in the entire cocoa world. Other major threats to cocoa production include adverse climate change predictions, aging trees, animal pests, and pollinator habitat loss. Due to these threats demand for the 'brown gold' that is cocoa has been predicted to outstrip supply in the not-too-distant future. What the cocoa industry needs may be locked in the genome of untapped germplasm such as the wild-type accessions collected during the London Cocoa Trade (LCT) Amazon project. The LCT accessions were held at the EEN station (thus the accession prefix: LCTEEN), but are now held at the Pichilingue station of INIAP in Ecuador, though some have been successfully transferred to the International Cocoa Genebank, Trinidad (ICGT) as part of the original project agreement. The original trees from which this collection arose may no longer exist due to the destruction of virgin forest in the Amazon. A timely intervention into the conservation of this collection is long overdue. Single Nucleotide Polymorphism (SNP) markers because of their high resolution have the capacity to provide valuable insight into the genetic constitution of these wild cacao accessions. The objectives of this research were to characterise available LCTEEN accessions with SNPs to create a genetic profile of the population and to use available geo-referencing information combined with the genetic information to assess spatial-genetic relationships. Results show LCTEENs are related to Purus, Nacional and Contamana populations and are good candidates for conservation.

INTRODUCTION

The genetic diversity of populations of a species impacts their future capacity to adapt to environmental threats (Frankham et al. 2004). Through breeding the genetic diversity existing in cacao germplasm collections can be harnessed to satisfy the needs of the cacao industry with regards to productivity increases, resistance to pests and diseases, environmental adaptation. Witches' Broom disease (WBD) in the Caribbean and Latin America, which wiped out entire industries along its path of dispersal (Silberner 2008, Smallman 2012, Farquharson 2014), was managed in Trinidad through the creation of the *Trinidad Selected Hybrids* (TSH) (Rudgard et al. 1993), the product of a multi-decade breeding effort (Bekele 2003) undertaken with material brought to Trinidad and Tobago (T&T) through a number of collaborative collection expeditions. Although the quest to improve resistance to WBD and broaden its genetic base is continuing, the TSH varieties represent a valuable source of tolerance to the disease combined with high productivity and award-winning flavour attributes.

Collection expeditions to the centre of diversity to obtain cacao to improve the level of genetic diversity available to breeders for development of new varieties is vital to address other ememerging threats to the cocoa industry. Industry collaboration has been the fuel to sustain such efforts. In 1979 one such industry stakeholder, the London Cocoa Terminal Market Association recognised the importance of collecting wild cacao and in celebrating its 50th anniversary made a generous donation that enabled the project to acquire wild cacao from Ecuador in the region known as The Oriente. The wild cacao that existed there was under threat from development activities (Allen 1983), therefore the project was a timely intervention.

John B. Allen, the scientist tasked with the execution of the LCT Amazon Project (LCTAP) under the technical supervision of R. Anthony Lass, ensured the collection was made at random and recorded global positioning system (GPS) co-ordinates. In collecting systematically, he achieved for the project, something hitherto un-attempted, the accurate sampling of a true wild cacao population with geographical positioning information. Allen would later oversee the establishment of the clones, (prefix LCTEEN in honour of the project sponsors and the research station), at the National Institute of Agricultural Research (INIAP) in Ecuador and record morphological characterisation data on the established material. The LCTEENs were described as comprising greater than 50% Criollo ancestry, producing large pods with white beans (Allen 1983). In accordance with the project terms, transfer of the LCTEENs was initiated from the INIAP station to Cocoa Research Unit (now Cocoa Research Centre), T&T, through the Barbados Cocoa Quarantine Station (Allen 1987). Successful transfer to T&T was not achieved for the entire collection, but a significant portion of the material made it to Trinidad and survived the stressful process to be established in the ICGT after repeated efforts. Though the LCTEEN accessions were extensively evaluated by Allen in terms of physical traits, molecular work has not been published to date for this collection although a small subset was evaluated with SSR markers (Loor Solorzano et al. 2013). The preliminary conclusion of genetic diversity based on the quantitative traits measured is likely to be an incomplete estimate of the gene diversity present in the population. The main observation which will still hold relate to the presence of white beans similar to pods of South and Central American Criollo ancestry. Molecular markers such as SNPs will be useful to characterise the genetic diversity of these accessions and their interrelationships to other cacao genetic groups. SNP markers have been used for genetic mapping (Argout et al. 2010) genetic identification, ancestry and parentage studies in cacao (Ji et al. 2013, Takrama et al. 2014, Lukman et al. 2014).

SNP markers are ideal for many reasons (Motilal et al. 2017), in particular their capacity for high throughput and abundance in plants including cacao. A thorough genetic characterisation of wild samples such as the LCTEEN is particularly warranted. Such an undertaking will be strategically important as the industry seeks to prepare for the potentially disastrous effects of climate change on cocoa production as well as other current and imminent threats including the spread of such devastating diseases such as Frosty Pod Rot (caused by *Moniliophthora roreri*) which require a broadening of the narrow genetic base through the use of wild cacao such as the LCTEENs. For this study, a collection of samples, including both seedlings and mother trees will be evaluated using SNP markers as a preliminary exploratory survey of the molecular genetic diversity and structure of the LCTEENs.

MATERIALS AND METHODS

DNA was extracted from leaf samples (location data of collections as described in Allen, 1983) and diluted (as needed) for genotyping on a nanofluidic Fluidigm system. SNP markers, selected for their coverage over all the chromosomes of cacao (Ji et al. 2012), were used in the Fluidigm system (Fluidigm Inc.) to genotype approximately 500 samples of LCTEEN cacao clones (at the Beltsville Agricultural Research Centre in Maryland, USA). Duplicates and reference samples were included in each plate for cross checking along with negative controls. The initial set of 96 SNPs was reduced to 82 after initial data curation. Data was analysed in GenAlEx Version 6.5 (Peakall and Smouse 2006, 2012) and STRUCTURE Version 2.3.4 (Pritchard et al. 2000). STRUCTURE runs were performed at a *burnin* of 100000 and *MCMC* 100000 using the admixture model for both *Correlated* and *Independent* allele frequency models and runs were performed using a range of K=1-5. The best K was selected using mainly the Evanno method (Evanno et al. 2005) and another manual method (Motilal 2016) involving ΔlnPr vs ΔK, but also with Structure Harvester (Earl and vonHoldt 2012).

Computation of descriptive data statistics was done in GenAlEx and using the multi-locus matching option, pairwise comparisons were made of the genotypes and probabilities of identities were calculated.

Table 1. List of six reference cacao germplasm groups (120 individual accessions) and their origin

Population/group	**Origin**	**Sample size**	**Source**
Nacional	Ecuador	20	INIAP (Ecuador)
IMC (Iquitos)	Peru	20	ICGT (T&T)
Parinari	Peru	20	ICGT (T&T)

Ucayali (Contamana)	Peru	20	ICGT (T&T) and ICT (Peru)
Purus	Brazil	20	SPCL, TARS, USDA; CATIE (Costa Rica)
French Guiana	French Guiana	20	CIRAD (France)
Total		120	

A genetic distance matrix was computed in GenAlEx and used to make PCoA plots with and without seedlings and reference samples. In the plot with reference samples, 20 samples from each of the reference groups listed in the PCoA diagram (Table 1) enabled an interpretation of the diversity of the LCTEENs. After initial analyses, the sample set was culled to 509 by removing duplicates and samples with more than 10% missing values. This culled data set was used for a Mantel test in GenAlEx to determine the presence of spatial effects. PowerCore (v1.0) (Biotechnology 2006) was used to select a minimal set of accessions (core set) that can capture the maximum molecular genetic diversity to represent the complete set of LCTEENs. AMOVA was performed to compare the core vs the remainder.

RESULTS AND DISCUSSION

Missing data for three of the SNP loci was between 10 and 15%. Summary statistics from GenAlEx computation are presented in Table 2. Mean and standard error values for $\mathbf{H_e}$ were similar to $\boldsymbol{u}\mathbf{H_e}$ with the observed heterozygosity being slightly lower. Heterozygosity estimates were lower than that reported by Lukman et al. (2014) for varieties in Indonesia or reference clones, or by Takrama et al. (2014) for cacao in Ghana but comparable to that observed by Ji et al. (2012).

Table 2. GenAlEx descriptive statistics for the LCTEENs analysed with SNP markers (N_e = Number of effective alleles = 1 / (Sum p_i^2), I = Shannon's Information Index = -1* Sum (p_i * *Ln* (p_i)), H_o = Observed Heterozygosity = No. of Hets / N, H_e = Expected Heterozygosity = 1 - Sum p_i^2, uH_e = Unbiased Expected Heterozygosity = (2N / (2N-1)) * H_e, F = Fixation Index = (H_e - H_o) / H_e = 1 - (H_o / H_e) Where p_i is the frequency of the *i*th allele for the population & Sum p_i^2 is the sum of the squared population allele frequencies.)

LCT		**N**	**N_e**	**I**	**H_o**	**H_e**	**uH_e**	**F**
	Mean	504.256	1.409	0.422	0.235	0.265	0.266	0.147
	SE	1.448	0.030	0.019	0.017	0.015	0.015	0.028

The positive Fixation Index (**F**) of 0.147 indicates greater homozygosity. Homozygotes are useful for breeding; therefore the LCTEEN accessions will be a valuable addition to pre-breeding germplasm pools to exploit this feature. Homozygous individuals should be typed using more loci and the most homozygous should be conserved and fully characterised for future breeding use. A wide range of heterozygosity levels was observed among the individual samples (Figure 1) and there were over 100 samples (~20 %) with heterozygosity values of 0.1 or lower.

GenAlEx multi-locus matching revealed the possibility the 82 loci either could not separate all of the LCTEEN clones or there were some synonymous clones in the set carrying a different clone name. The latter is more probable since instances of duplication and mislabelling have been reported for the ICGT (Motilal et al. 2011). Some matches were as expected among the seedling variants e.g. LCTEEN 54/S3 and 54/S4.

There are likely to be genetically distinct groups among the LCTEEN samples indicated by the STRUCTURE output. Results obtained using both the *Correlated* and the *Independent* model parameter settings demonstrate the presence of at least two groups. The *Independent* model results were used to select a run for interpretation of the STRUCTURE analysis since the most data was available for that job vs the job done with the *Correlated* model, that is, more iterations were requested and a larger range of *K* was set.

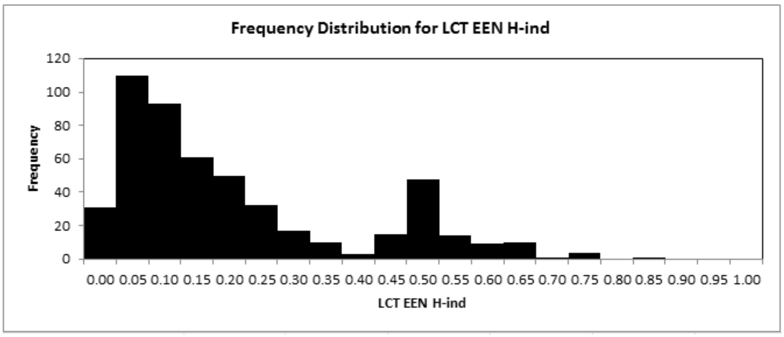

Figure 1. Frequency distribution of individual heterozygosity (GENALEX output)

Determination of best '*K*' number of groups done with the Evanno method which was compared with another manual determination (Motilal 2016). Both Evanno's method and Structure Harvester concur for two major groupings. Allen (1983) also suggested this possibility from phenotypic observations however sub-clustering was detected from using the manual method pointing to more than two groups (perhaps four) being present. These subgroups may confirm divergence due to geographic boundaries or other factors. Higher numbers of iterations will be used in future comparing all models to confirm these preliminary findings.

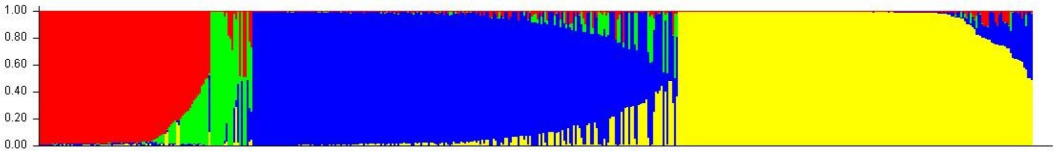

Figure 2. Structure bar plot showing population divergence among the LCTEENs (sorted by *q;* best *K*)

PCoA plots are presented in Figures 3 & 4. Figure 3 represents the PCoA without reference clones and shows evidence of at least three groups in agreement with the results obtained using the method of *K* determination of Motilal (2016). The first three axes account for 57.8 % of the variation. In the PCoA plot with reference samples (Figure 4) 55.89% of the variation was explained by the first three axes. The LCTEEN mother trees are fairly distinct as a group and are most closely related to Nacional, Ucayali and Purus (View 1 0vs 2, not shown) populations.

A core set of 36 accessions was obtained with PowerCore and complete congruency was achieved in selection of the core set. AMOVA results confirmed the core successfully represented a high proportion of the total molecular variability (80%) in comparison to the remainder accessions. The core contained none of the absolute homozygotes or the most heterozygous accessions but these can be used to supplement the core list. No private alleles were detected. Further evaluation of the core, in particular for self and cross compatibility is recommended.

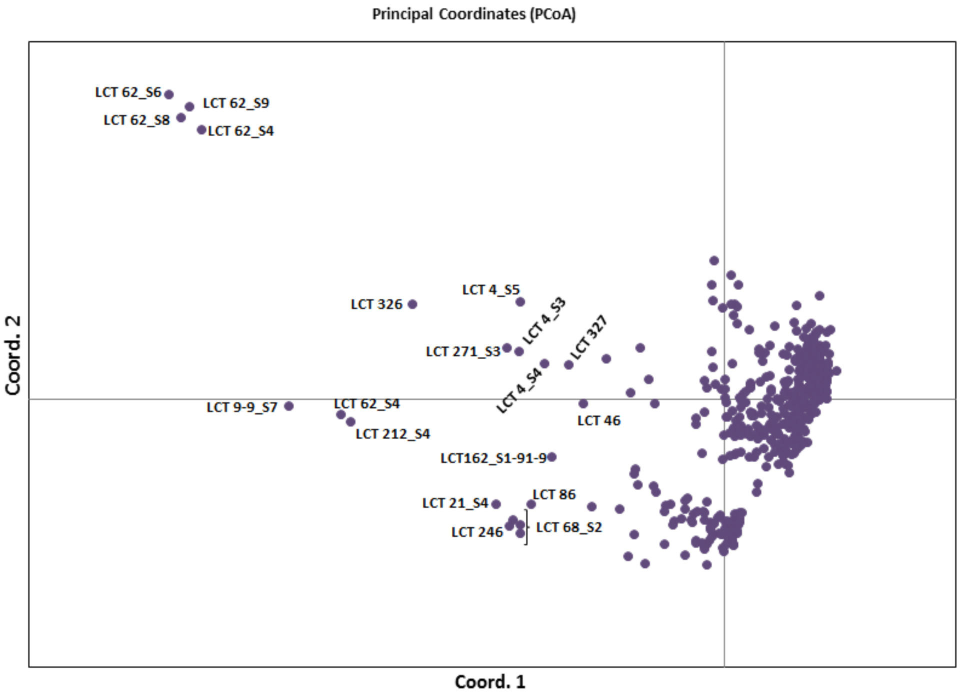

Figure 3. PCoA plot: Clustering of LCTEEN seedlings and mother trees from Ecuador.

D_{est} values for estimating differentiation obtained with G_{st} analysis were found to be similar to the Nei's unbiased measure. The Mantel test (Figures 5 & 6) result showed spatial structure was significant and a weak but significant isolation by distance effect.

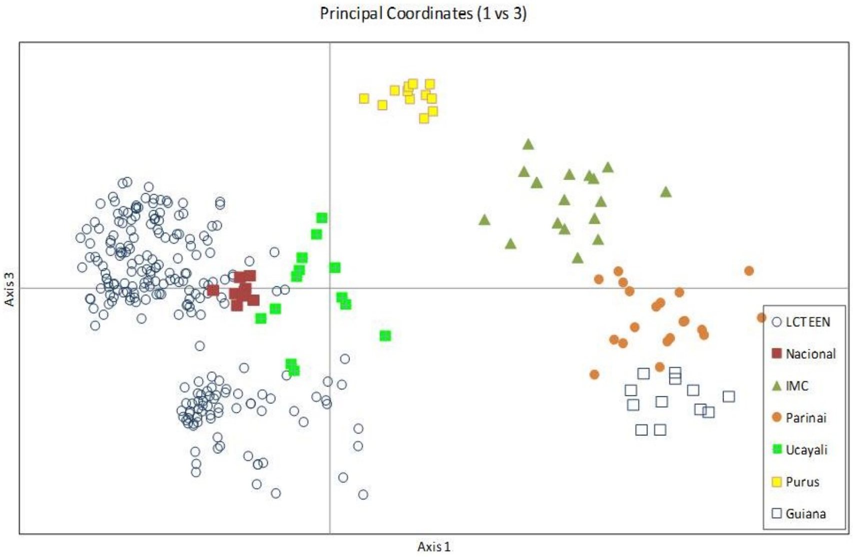

Figure 4. PCoA plot: LCTEEN mother trees and reference clones: Nacional, IMC, Parinari, Ucayali, Purus, and Guiana.

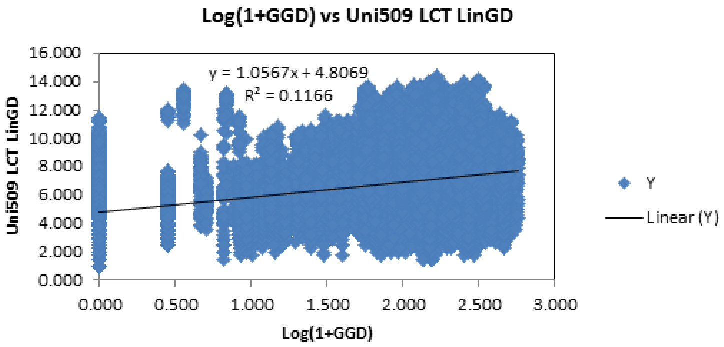

Figure 5. Mantel test scatter plot of LCT EEN from Ecuador

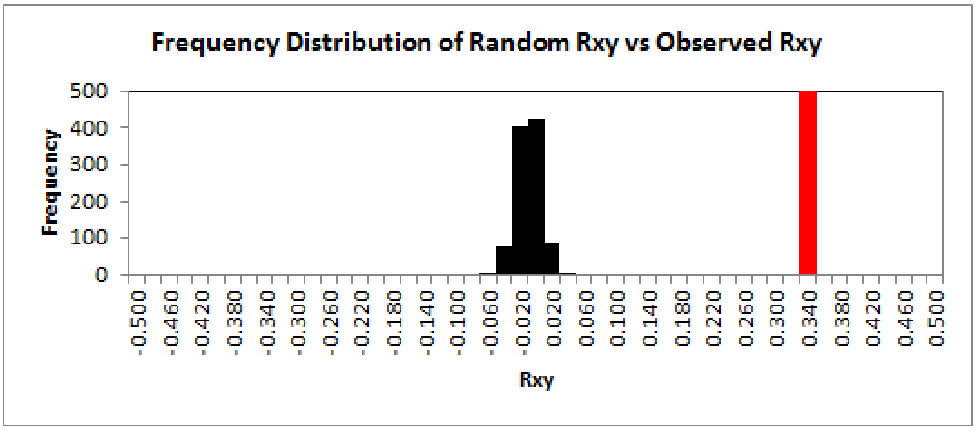

Figure 6. Mantel test (r = 0.341, p = 0.001, 999 runs) frequency distribution for LCTEENs

CONCLUSION AND FUTURE WORK

DNA samples of LCTEEN cacao clones were genotyped successfully with SNPs. GenAlEx and STRUCTURE analyses revealed the presence of genetically distinct clusters among the samples. A core set of the accessions was established that represents most of the diversity of the group. Future work is planned with a larger complement of SNPs to get an even better picture of the genetics of this group of accessions. Data analysis will also be completed with more stringent parameters for STRUCTURE to verify initial findings and select the most appropriate number of clusters. STRUCTURE and other software runs are planned to include the use of reference samples as done with GenAlEx and at a larger range of *K* for confirmation. DIVA-GIS analysis will be done to make further use of the GPS co-ordinates recorded to analyse the impact of the geographic location and geographic barriers on gene flow and investigate the spatial arrangement of the genetic diversity using this software as a means to confirm and expand on initial findings.

The evidence of population structure and genetic diversity suggests this group of samples consisting of two main populations and potentially other subpopulations is worthy of further exploration to obtain the genetic information required to properly exploit this wild cacao germplasm for the future fitness of the crop. A complete genetic profile of this population will be attempted along with the selection of a core sample as a priority group for conservation within the ICGT and this list will be recommended to the curators at INIAP in Ecuador. Further work will be done to evaluate adaptability of some accessions to their environmental pressures.

REFERENCES CITED

Allen, J. B. “London Cocoa Trade Amazon Project: Final Report Phase Two.” In Special Issue of Cocoa Grower's Bulletin edited by R. A. Lass and Sue Sanderson, Vol. 39. Birmingham: Haines Clark & Co. Ltd., 1987.

Allen, J.B. “London Cocoa Trade Amazon Project: Phase 1” In Special Issue of Cocoa Grower's Bulletin edited by R. A. Lass and Sue Sanderson, Vol. 33. Birmingham: Haines Clark & Co. Ltd., 1983.

Bekele, Frances. “The History of Cocoa Production in Trinidad and Tobago” In Proceedings of the APASTT Seminar - Exhibition entitled Revitalisation of the Trinidad and Tobago Cocoa Industry, 4-12. St. Augustine: APASTT, 2003.

Earl, Dent A., and Bridgett M. vonHoldt. “STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method.” Conservation Genetics Resources 4, no. 2 (2012): 359-361.

Evanno, G., S. Regnaut, and J. Goudet. “Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study.” Molecular Ecology 14 (2005): 2611-2620.

Farquharson, Kathleen L. “Scientists seek cure for devastating witches' broom disease of the chocolate tree” Science Daily, October 31, 2014. https://www.sciencedaily.com/releases/2014/10/141031150006.htm.

Frankham, R., J. D. Ballou and D. A. Briscoe. A Primer of Conservation Genetics. London: Cambridge University Press, 2004.

Ji, Kun, Dapeng Zhang, Lambert A. Motilal, Michel Boccara, Philippe Lachenaud, and Lyndel W. Meinhardt. “Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers.” Genetic Resources and Crop Evolution 60, no. 2 (February 2013): 441–453.

Loor Solorzano, R. G., O. Fouet, A. Lemainque, S. Pavek, M. Boccara, X. Argout, F. Amores, B. Courtois, A.M. Risterucci, and C. Lanaud. “Insight into the Wild Origin, Migration and Domestication History of the Fine Flavour Nacional *Theobroma cacao* L. Variety from Ecuador.” Plos One 7 no. 11 (November 2012): e48438. Doi: https://doi.org/10.1371/journal.pone.0048438.

Lukman, Dapeng Zhang, Agung W. Susilo, Diny Dinarti, Bryan Bailey, Sue Mischke, and Lyndel W. Meinhardt. “Genetic Identity, Ancestry and Parentage in Farmer Selections of Cacao from Aceh, Indonesia Revealed by Single Nucleotide Polymorphism (SNP) Markers.” Tropical Plant Biol. 7, no. 3-4 (2014): 133-143.

Motilal, L. A. “A molecular genetic study of the International Cocoa Genebank, Trinidad towards efficient conservation and utilisation.” PhD diss., The University of the West Indies, St. Augustine, 2016.

Motilal, Lambert A.; Antoinette Sankar, David Gopaulchan, and Pathmanathan Umaharan. "Cocoa" In Biotechnology of Plantation Crops edited by Pallem Chowdappa; Anitha Karun; M. K. Rajesh; S. V. Ramesh, 313-354. New Delhi: Daya Publishing House, 2017.

Motilal, Lambert A., Dapeng Zhang, Pathmanathan Umaharan, Sue Mischke, Stephen Pinney, and Lyndel W. Meinhardt. “Microsatellite fingerprinting in the International Cocoa Genebank, Trinidad:accession and plot homogeneity information for germplasm management.” Plant Genetic Resources: Characterization and Utilization, 2011: 1–9.

National Institute of Agricultural Biotechnology PowerCore (v. 1.0). “A program applying the advanced M strategy using heuristic search for establishing core or allele mining sets.” R. Korea: Rural Development Administration (RDA), 2006.

Peakall, R. and P. E. Smouse. “GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research.” Molecular Ecology Notes, 6 (2006): 288–295. doi:10.1111/j.1471-8286.2005.01155.x.

Peakall, R., and P. E. Smouse. “GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update.” Bioinformatics 28 (2012): 2537-2539.

Pritchard, JK, M. Stephens, and P. Donnelly. “Inference of population structure using multilocus genotype data.” Genetics, (June 2000): 945-59.

Rudgard, S. A., T. Andebrhan, A.C. Maddison, and R.A. Schmidt. "Future Prospects for Improvements in Disease Management" In Disease Management edited by S. A. Rudgard, A.C. Maddison, and T. Andebrhan, 213-216. Dordrecht: Springer Netherlands, 1993.

Silberner, Joanne. “A Not-So-Sweet Lesson from Brazil's Cocoa Farms” National Public Radio, Inc. [US], June 14, 2008. https://www.npr.org/templates/story/story.php?storyId=91479835.

Smallman, Shawn. “Witches’ Broom: The Mystery of Chocolate and Bioterrorism in Brazil” Introduction to International and Global Studies, March 17, 2012. https://www.introtoglobalstudies.com/2012/03/witches-broom-the-mystery-of-chocolate-and-bioterrorism-in-brazil/.

Takrama, Jemmy, Ji Kun, Lyndel Meinhardt, Sue Mischke, Stephen Y. Opoku, Francis K. Padi, and Dapeng Zhang. “Verification of genetic identity of introduced cacao germplasm in Ghana using single nucleotide polymorphism (SNP) markers.” African Journal of Biotechnology 13, no. 21 (2014): 2127-2136.