Tree Genetics & Genomes
DOI 10.1007/s11295-013-0601-4

ORIGINAL PAPER

Complex origin of Trinitario-type *Theobroma cacao* (Malvaceae) from Trinidad and Tobago revealed using plastid genomics

Ji Yong Yang · Moira Scascitelli · Lambert A. Motilal · Saemundur Sveinsson · Johannes M. M. Engels · Nolan C. Kane · Hannes Dempewolf · Dapeng Zhang · Kamaldeo Maharaj · Quentin C. B. Cronk

Received: 5 July 2012 / Revised: 19 December 2012 / Accepted: 9 January 2013

Abstract Trinidad and Tobago has a long history of producing high-quality cacao (*Theobroma cacao* L.). Cacao genotypes in Trinidad and Tobago are of a highly distinctive kind, the so-called "Trinitario" cultivar group, widely considered to be of elite quality. The origin of Trinitario cacao is unclear, although it is generally considered to be of hybrid origin. We used massive parallel sequencing to identify polymorphic plastidic single nucleotide polymorphisms (cpSNPs) and polymorphic plastidic simple sequence repeats (cpSSRs) in order to determine the origin of the Trinitario cultivar group by comparing patterns of polymorphism to a reference set of ten completely sequenced chloroplast genomes (nine *T. cacao* and one outgroup, *T. grandiflorum* (Willd. ex Spreng.) Schum). Only three cpSNP haplotypes were present in the Trinitario cultivars sampled, each highly distinctive and corresponding to reference genotypes for the Criollo (CRI), Upper Amazon Forastero (UAF) and Lower Amazon Forastero (LAF) varietal groups. These three cpSNP haplotypes likely represent the founding lineages of cacao to Trinidad and Tobago. The cpSSRs were more variable with eight haplotypes, but these clustered into three groups corresponding to the three cpSNP haplotypes. The most common haplotype found in farms of Trinidad and Tobago was LAF, followed by UAF and then CRI. We conclude that the Trinitario cultivar group is of complex hybrid origin and has derived from at least three original introduction events.

Keywords *Theobroma cacao* · Chloroplast · Microsatellites · Single nucleotide polymorphisms · Trinitario

Communicated by R. Sederoff

J. Y. Yang (✉) · M. Scascitelli · S. Sveinsson · N. C. Kane · H. Dempewolf · Q. C. B. Cronk
Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
e-mail: albertoyang@yahoo.com

L. A. Motilal
Cocoa Research Unit, The University of the West Indies, St. Augustine, Trinidad, West Indies, Republic of Trinidad and Tobago

J. M. M. Engels
Biodiversity International, 00057 Maccarese Rome, Italy

D. Zhang
SPCL, USDA-ARS, Building 001 BARC-West, 10300 Baltimore Avenue, Beltsville, MD 20705, USA

K. Maharaj
Ministry of Agriculture, Food Production Land, and Marine Resources Affairs, Central Experiment Station, Centeno, Via Arima P.O., West Indies, Republic of Trinidad and Tobago

Introduction

Theobroma cacao L. (cacao) is an economically important crop as its seeds are the sole source of commercial chocolate. Cacao flowers are mainly pollinated by midges (*Forcipomyia* species) and exhibit a high degree of outcrossing (Silva et al. 2010). Commercial cacao has two main ancestral groups, Criollo and Forastero, and a derived group, Trinitario (Cheesman 1944). Criollos were likely domesticated by the Mayas in Central America (Toxopeus 1985; Motamayor et al. 2002) and were spread to the Caribbean and South America after the mid-16th century (Wood 1985). Criollos are now known to be highly susceptible to several diseases and of the three cacao groups (Criollo, Forastero, and Trinitario) they are the least vigorous (Toxopeus 1985; Wood 1985). Forastero is

Published online: 21 February 2013

an umbrella group referring to germplasm from the Upper and Lower Amazon basin. It is more disease resistant and vigorous than Criollo (Toxopeus 1985) and has a higher yield. It is now the most commonly grown cacao in the world (Wood 1985). However, Forasteros are less valuable because their fermented beans are not generally considered to produce a high quality flavor (Wood 1985), although there are exceptions such as the "Nacional" variety of Forastero from Ecuador (Loor et al. 2009).

History of cacao cultivation in Trinidad and Tobago

In Trinidad most of the trees were supposedly destroyed by disease in the 18th century (Shephard 1932; Cheesman 1944) especially by a disastrous outbreak in 1727. A new cultigen called Trinitario was later documented in Trinidad, seemingly having arisen through natural hybridization between the local surviving Criollo trees and the Forastero planted between the remaining Criollo (Pound 1931; Cheesman 1944; Engels 1986; Coe and Coe 1996; Bartley 2005). The hybrids, throughout the island of Trinidad, showed variation in their combination of parental characters and a selection program based primarily on yield led to the identification of 100 Trinitario trees now known as the Imperial College Selection (ICS) accessions (Pound 1933, 1934, 1935, 1936; Cheesman 1934). A more recent germplasm collection of accessions from relic estates in Trinidad (TRD) has also been made (Bekele et al. 2007). In the 1940s, a breeding program was established by the Ministry of Agriculture in Trinidad under which locally selected Trinitarios were crossed with genotypes from South America. This work generated the Trinidad Selected Hybrids (TSH) and has provided a basis for all modern cacao breeding programs (Freeman 1969). However, modern breeding programs have been affected by the lack of comprehensive characterization of the available cacao genetic resources, coupled with problems of accession mislabeling (Motilal et al. 2009, 2011; Irish et al. 2010).

Previous studies on the origin of the Trinitario cacao

With the accumulation of molecular analyses, primarily microsatellite data, new scenarios for the origin of the Trinitario varietal group have been suggested. Motamayor et al. (2003) proposed that Trinitarios originated by hybridization between a small number of Lower Amazon Amelonado-Forastero (LAF) genotypes and Criollo, while less than 10 % of Trinitario accessions bore alleles from the other main Forastero variety, the Upper Amazon Forastero (UAF). Johnson et al. (2009) found that Criollo played an important role in the origin of Trinitario accessions and they considered that the limited number of Trinitarios with UAF alleles resulted from intensive breeding efforts to integrate UAF disease resistance traits into Trinitario. Motilal et al. (2010) downplayed the contribution of Criollo and postulated a greater contribution of Forastero lineage to the genetic composition of Trinitarios, suggesting that the cacao cultivated in Trinidad prior to 1727 was "introgressed Criollo" (i.e., with a genetic contribution from Forastero) rather than what is now known as "pure Criollo."

Utility of plastid haplotype markers in determining the origin of Trinitario cacao

Genetic markers from uniparentally inherited and non-recombining organelles (e.g., chloroplasts or mitochondria) are generally of greater utility than nuclear markers in phylogeographical studies because there is no recombination that hinders the study of parental origin (Petit and Vendramin 2007). Chloroplast (plastid) markers have proven useful for distinguishing among different cultivars in several crop species including soya (Powell et al. 1996), rice (Provan et al. 1997), barley (Provan et al. 1999) and grapevine (Arroyo-Garcia et al. 2002). Thus, plastid markers should be useful to untangle the complex origin of the Trinitario trees currently found in Trinidad and Tobago. A recent sequencing effort characterized the full chloroplast genome of ten cacao genotypes (Kane et al. 2012), allowing the discovery of 78 single nucleotide polymorphisms (SNPs). We used ten of these SNPs to determine the chloroplast contributions of reference genotypes of UAF, LAF, and Criollo to the cacao trees farmed across Trinidad and Tobago. In addition we analyzed plastid DNA variation at nine polymorphic simple sequence repeats (SSRs) to assess the genetic composition and haplotype diversity of Trinitario cacao from Trinidad and Tobago.

Materials and methods

Cacao material studied

Ten complete chloroplast genome reference sequences were used for polymorphic plastidic single nucleotide polymorphism (cpSNP) development. These consisted of nine *T. cacao* accessions and one outgroup, *T. grandiflorum*. The *T. cacao* accessions included a Criollo genotype (accession: Criollo-22), a UAF genotype (Scavina-6, accession: MIA29885), and a LAF genotype (Amelonado, accession: TARS 16542). These three reference accessions were selected to represent standard exemplars of the three varietal groups. The remaining *T. cacao* accessions were of Trinitario genotypes (see Kane et al. 2012 for further details).

In total, 95 Trinitario cacao trees were genotyped in this study (see Tables 1 and 2). Sixty-four accessions were collected from 33 different farms in Trinidad (Fig. 1); eight accessions were collected from six different farms in Tobago (Fig. 2). The remaining 23 samples were reference Trinitario accessions from the International Cocoa Genebank,

Table 1 Haplotype diversity of Trinitarios collected from farms in Trinidad and Tobago, based on cpSSR and cpSNP analysis

Sample ID	Provenance	Farm	Location	cpSSR Haplotype	cpSNP Haplotype
K087	Trinidad	AF	Aripo	2	LAF
K088	Trinidad	AF	Aripo	2	LAF
K089	Trinidad	E	Aripo	7	UAF
K107	Trinidad	V	Aripo	8	LAF
K018	Trinidad	AC	Biche	7	UAF
K017	Trinidad	AD	Biche	2	LAF
K092	Trinidad	AD	Biche	7	UAF
K034	Trinidad	D	Biche	1	UAF
K033	Trinidad	D	Biche	2	LAF
K052	Trinidad	K	Biche	1	UAF
K152	Trinidad	Z	Biche	4	LAF
K051	Trinidad	Z	Biche	7	UAF
K024	Trinidad	Z	Biche	8	LAF
K153	Trinidad	Z	Biche	8	LAF
K013	Trinidad	A	Brasso Seco	2	LAF
K094	Trinidad	A	Brasso Seco	3	CRI
K117	Trinidad	A	Brasso Seco	8	LAF
K012	Trinidad	W	Brasso Seco	1	UAF
K114	Trinidad	W	Brasso Seco	4	LAF
K113	Trinidad	W	Brasso Seco	7	UAF
K011	Trinidad	W	Brasso Seco	8	LAF
K110	Trinidad	W	Brasso Seco	8	LAF
K111	Trinidad	W	Brasso Seco	8	LAF
K112	Trinidad	W	Brasso Seco	8	LAF
K131	Trinidad	W	Brasso Seco	8	LAF
K025	Trinidad	C	Brasso Venado	1	UAF
K108	Trinidad	C	Brasso Venado	1	UAF
K050	Trinidad	U	Carapal	1	UAF
K054	Trinidad	AA	Coromandel	8	LAF
K049	Trinidad	I	Coromandel	2	LAF
K035	Trinidad	S	Coromandel	1	UAF
K129	Trinidad	R	Cumana	5	LAF
K128	Trinidad	R	Cumana	8	LAF
K028	Trinidad	P	Gran Couva	2	LAF
K079	Trinidad	P	Gran Couva	4	LAF
K026	Trinidad	T	Gran Couva	7	UAF
K073	Trinidad	X	Gran Couva	8	LAF
K120	Trinidad	B	Lopinot	8	LAF
K071	Trinidad	H	Lopinot	1	UAF
K097	Trinidad	J	Lopinot	1	UAF
K109	Trinidad	N	Lopinot	3	CRI
K056	Trinidad	O	Lopinot	1	UAF
K106	Trinidad	AG	Moruga	8	LAF
K121	Trinidad	AG	Moruga	8	LAF
K059	Trinidad	G	Moruga	8	LAF
K141	Trinidad	G	Moruga	8	LAF
K142	Trinidad	G	Moruga	8	LAF
K143	Trinidad	G	Moruga	8	LAF
K146	Trinidad	G	Moruga	8	LAF
K147	Trinidad	G	Moruga	8	LAF
K148	Trinidad	G	Moruga	8	LAF
K149	Trinidad	G	Moruga	8	LAF
K151	Trinidad	G	Moruga	8	LAF
K070	Trinidad	Y	Moruga	2	LAF
K133	Trinidad	F	Tabaquite	7	UAF
K062	Trinidad	AB	Tableland	8	LAF
K135	Trinidad	L	Tableland	7	UAF
K136	Trinidad	L	Tableland	7	UAF
K137	Trinidad	L	Tableland	7	UAF
K134	Trinidad	L	Tableland	8	LAF
K055	Trinidad	AE	Tamana	8	LAF
K015	Trinidad	M	Vega de Oropouche	7	UAF
K016	Trinidad	M	Vega de Oropouche	7	UAF
K029	Trinidad	Q	Vega de Oropouche	8	LAF
K064	Tobago	AJ	Betsy's Hope	2	LAF
K130	Tobago	AJ	Betsy's Hope	2	LAF
K053	Tobago	AH	Moriah	7	UAF
K083	Tobago	AH	Moriah	8	LAF
K075	Tobago	AK	Moriah	3	CRI
K060	Tobago	AL	Moriah	1	UAF
K080	Tobago	AM	Roxborough	8	LAF
K067	Tobago	AN	Runnemede	1	UAF

UAF Upper Amazon Forastero, *CRI* Criollo, *LAF* Lower Amazon Forastero

Trinidad (ICG, T) collection. Of these, one sample was from Grenada Selection (GS), 14 samples from ICS, seven samples of relic clones from Trinidad (TRD) and one sample from Dominica (DOM). One of the TRD accessions (TRD 66) was a duplicate and one of the ICS accessions was mislabeled in the field (ICS 40). These two samples were excluded from further data analysis.

Chloroplast single nucleotide polymorphisms

The cpSNPs were characterized in a previous study by Kane et al. (2012). Briefly, whole plastid genomes of nine *T. cacao* accessions and *T. grandiflorum* were assembled from Illumina short read whole genome sequencing runs. Trimmed, cleaned paired-end sequences were then mapped to a previously generated reference cacao genome (SCAVINA 6) using MOSAIK (http://bioinformatics.bc.edu/marthlab/Mosaik) and SNP polymorphisms called using SAMTOOLS (Li et al. 2009;

Table 2 Haplotype diversity of the 21 reference Trinitario cultivars from the Trinidad cacao germplasm collection (ICG, T), based on cpSSR and cpSNP analysis

Sample ID	Location	cpSSR Haplotype	cpSNP Haplotype
DOM 13	Field 5A, Nursery Plot 2, Plot 28, Tree 1	1	UAF
GS 29	Field 6B, Section A Plot 36, Tree 5	3	CRI
ICS 111	Field 6B, Section B, Plot 99, Tree 15	3	CRI
ICS 15	Field 4A, Section C, Plot 302, Tree 3	3	CRI
ICS 16	Field 6B, Section E, Plot 345, Tree 2	1	UAF
ICS 17	Field 6B, Section E, Plot 301, Tree 5	3	CRI
ICS 43	Field 6B, Section E, Plot 299, Tree 1	3	CRI
ICS 45	Field 6B, Section B, Plot 113, Tree 4	3	CRI
ICS 46	Field 6B, Section E, Plot 289, Tree 1	6	UAF
ICS 60	Field 6B, Section E Plot 332, Tree 7	3	CRI
ICS 84	Field 6B, Section E, Plot 329, Tree 10	3	CRI
ICS 85	Field 6B, Section E, Plot 343, Tree 11	1	UAF
ICS 92	Field 6B, Section E, Plot 347, Tree 1	3	CRI
ICS 95	Field 6B, Section B, Plot 98, Tree 7	1	UAF
ICS 97	Field 6B, Section B, Plot 110, Tree 4	8	LAF
TRD 115	Field 4A, Section A, Plot 167, Tree 2	3	CRI
TRD 23	Field 4A, Section A, Plot 118, T1	1	UAF
TRD 35	Field 4A, Section A, Plot 83, Tree 1	3	CRI
TRD 52	Field 4A, Section A, Plot 123, Tree 1	3	CRI
TRD 66	Field 4A, Section A, Plot 50, Tree 1	3	CRI
TRD 86	Field 4A, Section A, Plot 38, Tree 1	1	UAF

ICG, *T* Trinidad cacao germplasm collections (*DOM* Dominica, *GS* Grenada Selection, *ICS* Imperial College, *TRD* Trinidad); *UAF* Upper Amazon Forastero, *LAF* Lower Amazon Forastero, *CRI* Criollo

http://samtools.sourceforge.net/). Two criteria were used to select SNPs for PCR-based genotyping: they were selected (1) in plastid regions with high read depth to ensure that the SNPs detected were real, and (2) in regions which maximized the density of SNPs per amplified fragment. Primers were then developed using PRIMER3 (Rozen and Skaletsky 2000;

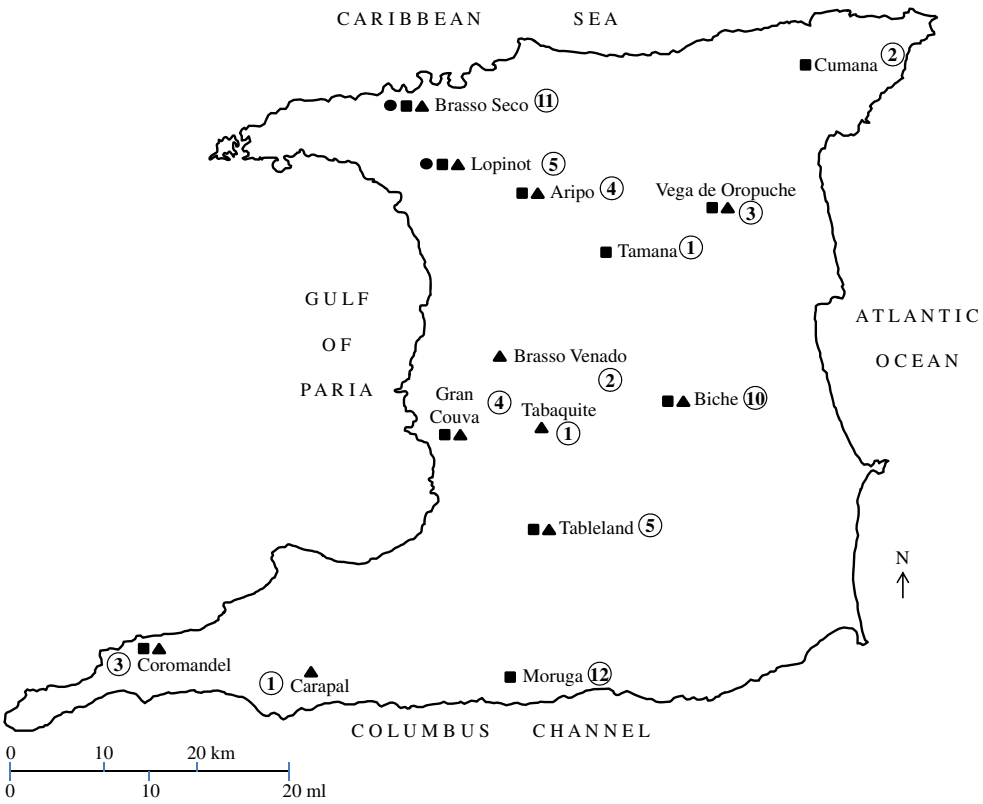

Fig. 1 Collection sites of Trinidad farms. The number of samples for each location is reported *within circles*. The *filled circles*, *squares* and *triangles* correspond to the three cpSNP haplotypes defined in this study: Criollo (*CRI*), Lower Amazon Forastero (*LAF*) and Upper Amazon Forastero (*UAF*), respectively

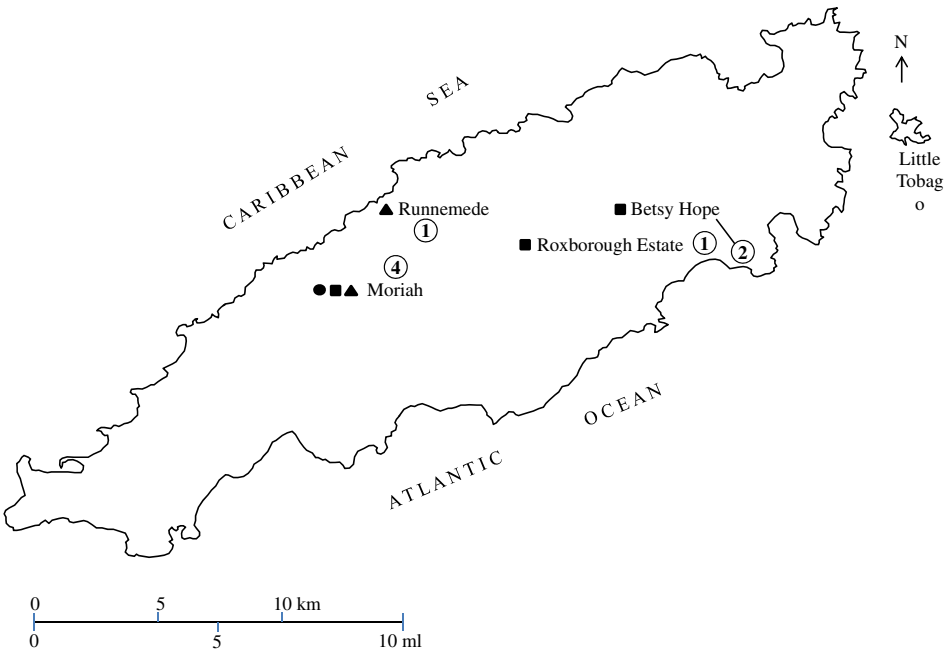

Fig. 2 Collection sites of Tobago farms. The number of samples for each location is reported *within circles*. The *filled circles*, *squares* and *triangles* correspond to the three cpSNP haplotypes defined in this study: Criollo (*CRI*), Lower Amazon Forastero (*LAF*) and Upper Amazon Forastero (*UAF*), respectively

http://frodo.wi.mit.edu/). Using two primer pairs and thus two chloroplast DNA regions, we were able to examine ten SNPs (see Table 3). The first three SNPs (5236, 5246, and 5410) were located in the spacer region between the genes *trnK* (*matK*) and *rps16*. The next seven SNPs (63472, 63473, 63479, 62486, 63803, 63901 and 63932) were located in the *accD-psaI* spacer. The SNP ID numbers represent their positions on the chloroplast. Although we chose these regions independently on the basis of whole plastid genomes, both regions have been noted as variable in previous studies: *trnK-rps16* in *Raphanus* (Kim et al. 2009) and *accD-psaI* in *Gossypium* (Small et al. 1998).

DNA regions containing SNPs were amplified by polymerase chain reactions (PCRs) on 50 ng of DNA in a 50-μl PCR mix containing 1× buffer, 200 μM of each dNTP, 1.5 mM $MgCl_2$, 2 U of Taq DNA polymerase and 0.5 μM of forward and reverse primers. Reactions were performed using a Bio-Rad thermal cycler with a touchdown program: 95 °C for 3 min, followed by 9 cycles of 94 °C for 30 s, 65 °C for 30 s (temperature decreased by 1 °C for every cycle) and 72 °C for 45 s, followed by 29 cycles of 94 °C for 30 s, 55 °C for 30 s, 72 °C for 45 s, and finally followed by a final extension at 72 °C for 20 min. PCR products were diluted at 1:10 and submitted for purification

Table 3 Genotypes of the three cpSNP haplotypes (Ha 1, 2 and 3) of *T. cacao* in Trinidad and Tobago and *T. grandiflorum* (*T. grand.*)

Polymorphic position	Primer sequences 5′ to 3′	Ha 1 (UAF) allele	Ha 2 (CRI) allele	Ha 3 (LAF) allele	*T. grand.* allele
Primer pair 1	F: GGTTTTGTCGTTATAGAACACGGTA				
	R: GCCGTACGAGGAGAAAACTTC				
SNP 5236		G	G	T	T
SNP 5246		G	T	G	T
SNP 5410		T	T	T	G
Primer pair 2	F: CTGCTCTTGGATCGGATTCT				
	R: TCCGTGGCATCTAAGTCTTG				
SNP 63472		A	T	T	T
SNP 63473		A	C	C	A
SNP 63479		A	A	A	C
SNP 63486		C	A	C	A
SNP 63803		C	C	C	T
SNP 63901		T	G	G	G
SNP 63932		T	T	C	T

Primer sequences for the two primer pairs used to obtain the ten SNPs are shown. The SNP ID numbers represent their positions on the chloroplast reference genome

UAF Upper Amazon Forastero, *CRI* Criollo, *LAF* Lower Amazon Forastero

and sequencing in both forward and reverse directions per DNA regions to Macrogen USA (www.macrogenusa.net). Four out of 95 Trinitario cacao trees sequenced produced unreadable sequences due to poor DNA quality. These were excluded from further SNP analysis. Sequences for the remaining 91 samples at each of the ten SNPs were aligned using Bioedit (http://www.mbio.ncsu.edu/bioedit/bioedit.html). SNPs obtained by Sanger sequencing were validated using the Illumina assemblies. No discrepancies were found between these two methods.

Chloroplast microsatellites

The nine chloroplast microsatellite markers (polymorphic plastidic SSRs [cpSSRs]) were developed previously by Yang et al. (2011). These were obtained by directly comparing the two complete cacao chloroplast genomes available at the time (Sveinsson et al. 2010; Jansen et al. 2011). They were specifically selected for their polymorphism in the same panel of individuals as used here (Yang et al. 2011). Details of cpSSR amplification and microsatellite scoring are reported by Yang et al. (2011).

Neighbor-joining tree construction

For cpSSRs, we generated a matrix of Nei's genetic distance (D_A; Nei et al. 1983; Takezaki and Nei 1996), with 1,000 bootstrap replicates, using MSA v4.0.5 (Dieringer and Schlötterer 2003). The consensus Neighbor Joining tree was then obtained using an extended majority-rule method, as implemented in the programs NEIGHBOR and CONSENSE, available in the package PHYLIP v3.69 (Felsenstein 1989).

Plastid SNPs were analyzed under maximum likelihood (Felsenstein 1973) using GARLI 2.0 (Zwickl 2006). TIM1ef was selected as a base substitution model, based on a model search using jModelTest 2.0.1 (Guindon and Gascuel 2003; Posada 2008). A majority rule consensus tree was constructed from a 100 bootstrap replicates using scripts in the DendroPy package (Sukumaran and Holder 2010) and drawn as an unrooted tree using FigTree v.1.3.1 (Rambaut 2006–2009; http://tree.bio.ed.ac.uk/software/figtree/).

Results

Chloroplast SNP analysis

Two primer pairs were used to amplify the two plastid regions containing the ten SNPs (see Table 3). Of the ten SNPs, eight were polymorphic within *T. cacao* and the remaining two were polymorphic between *T. cacao* and the outgroup *T. grandiflorum*. Only three *T. cacao* SNP haplotypes were distinguished in all samples examined from farms in Trinidad and Tobago. Each of these haplotypes had an identical cpSNP profile to one of our three reference accessions (i.e., Criollo-22, Scavina-6 and Amelonado). A set of Trinitario accessions therefore had the identical SNP profile to the reference Criollo genotype (accession: Criollo-22), a second set to the reference LAF genotype (Amelonado, accession: TARS 16542) and a third set to the reference UAF genotype (Scavina-6, accession: MIA 29885) (Fig. 3). These three haplotypes are denoted Criollo (CRI), LAF, and UAF, respectively, in this study (see Tables 1 and 2). *T. grandiflorum* had a divergent haplotype, and its relationship to the three *T. cacao* haplotypes is shown in Fig. 3.

Chloroplast microsatellite analysis

The nine chloroplast microsatellite (cpSSR) markers allowed discrimination of eight different haplotypes. All nine loci had between two and seven alleles, with an average of 2.9 alleles per locus. The eight different haplotypes clustered into three groups (Fig. 4). These groups corresponded to the three haplotypes recovered from the cpSNP analysis. Tables 1 and 2 and Fig. 5 show the correspondence of the cpSSR haplotypes to the cpSNP haplotypes.

The cpSSRs revealed a moderate level of genetic variation among individual trees within farms and within localities in Trinidad and Tobago. Considering only farms where two or more accessions were sampled, an average of 1.71 haplotypes per farm was found. Few farms lacked haplotype diversity. An exception was farm G in Moruga, Trinidad (Table 1), which had a single haplotype of LAF origin for all of the nine plants sampled. In general, samples within farms were fairly heterogeneous. As an example, eight individuals sampled from farm W in Brasso Seco (Table 1) showed four different haplotypes consisting of two cacao varietal groups, LAF and UAF. If the other farm (A) from Brasso Seco was included, six cpSSRs haplotypes and all three cacao varietal types were present in this locale. Because of the observed level of haplotype variability within farms, it has not been possible to find farm- or location-specific genetic identifiers. The LAF and UAF haplotypes are geographically well dispersed as illustrated in Figs. 1 and 2.

Based on these uniparentally inherited chloroplast markers, of the 72 farms sampled, only three farms contained the Criollo varietal type, 44 farms had the LAF type and 25 farms had the UAF type (Table 1). The 21 reference Trinitario accessions comprised Criollo (13), UAF (7) and LAF (1) ancestries (Table 2).

CaCrSSR1, a useful marker

The most variable marker was the one with the pentameric repeat (CaCrSSR1; Yang et al. 2011), and this is likely to be the most useful for cacao genetic identification. This marker alone had seven different alleles, which could be used to distinguish all of the eight haplotypes except for haplotypes

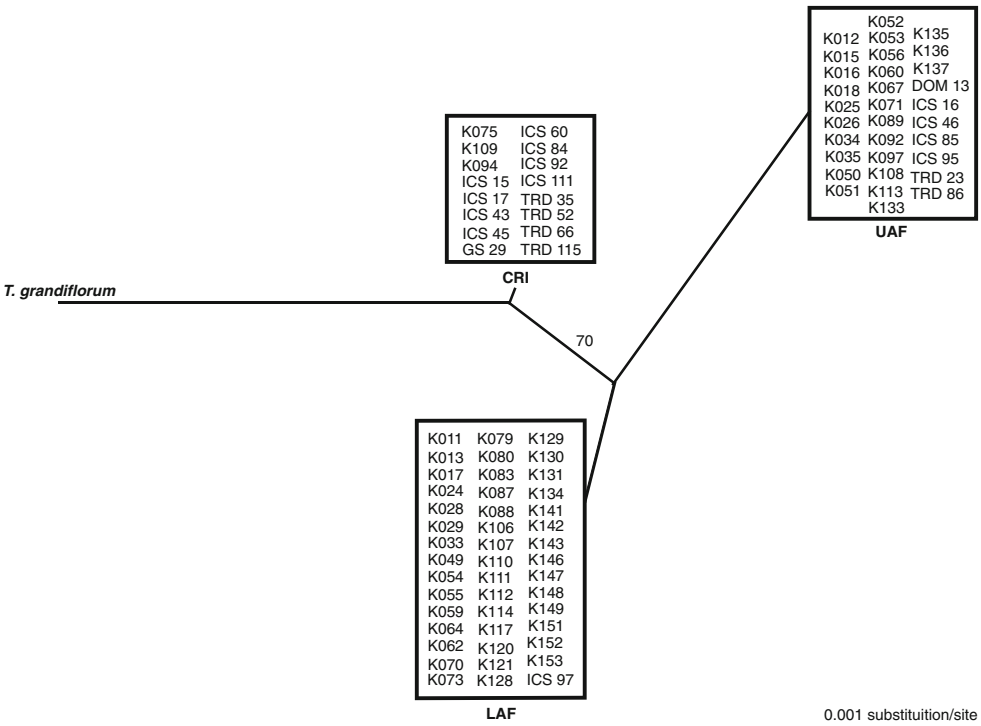

Fig. 3 Phylogenetic tree representing relationships of the three cpSNP haplotypes (Criollo [*CRI*], Lower Amazon Forastero [*LAF*] and Upper Amazon Forastero [*UAF*]) and *T. grandiflorum* based on the sequence of two variable intergenic spacers in the plastid genome of *Theobroma cacao*

6 and 7 (Fig. 6). Another advantage is that the 5-bp allelic size differences allow complete resolution of all alleles using 2 % high performance agarose gel electrophoresis (Fig. 7). An extra ("ghost") band of weaker intensity at 362 bp always accompanied the main band (Figs. 6 and 7), but did not create ambiguities in scoring since the extra band was separate and monomorphic in all the 95 samples analyzed (see Yang et al. 2011). Attempts to identify its origin by cutting out the band from the gel, and subsequent PCRs to improve its concentration before sequencing, did not produce clean readable sequences (data not shown). This was probably due to its proximity to the pentameric repeat in the gel.

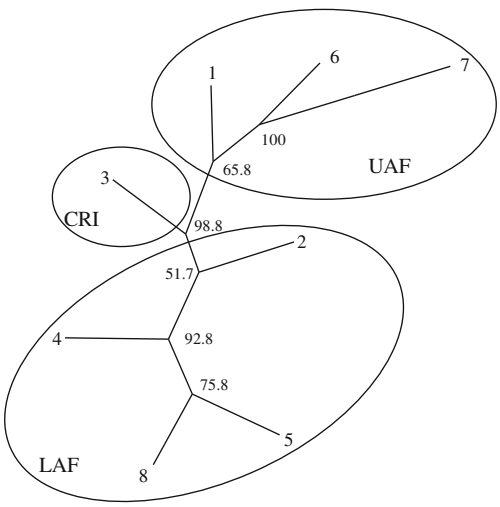

Fig. 4 Neighbor joining tree of the eight haplotypes obtained with the nine cpSSRs using Nei's genetic distance (D_A) after 1,000 bootstrap replicates. Bootstrap values are reported as percentages. *Numbers 1–8* represent the SSRs haplotype types, while the three cpSNP varietal types are *circled*

Discussion

The Trinitario cultivar group has a complex origin

The three cpSNP haplotype groups corresponding to the reference Criollo, UAF and LAF genotypes were found to be scattered across the two islands and among farms. We expect our chloroplast DNA haplotypes to be maternally inherited, as is usual in Angiosperms (Mogensen 1996). All members of the Malvaceae (the family that includes cacao) examined to date have maternally inherited plastids including the well-studied *Gossypium* (Corriveau and Coleman 1988; Wendel 1989; Takayama et al. 2006). There is no reason to suppose that chloroplast DNA is not also maternally inherited in *T. cacao*. Given that, chloroplast markers allow us to track the maternal lineages of cacao in Trinidad and Tobago from multiple seed introductions. It is known that cacao cultivation in Trinidad and Tobago started around 1678, following introductions from Venezuela (Shephard 1932). After the disaster of 1727, more disease-resistant Forastero types were introduced in 1756 (Shephard 1932; Coe and Coe 1996). It is thought that the introduction of germplasm from multiple sources, including the lower and upper Amazon region (Bartley 2005), allowed natural hybridization to occur, forming Trinitario lineages with

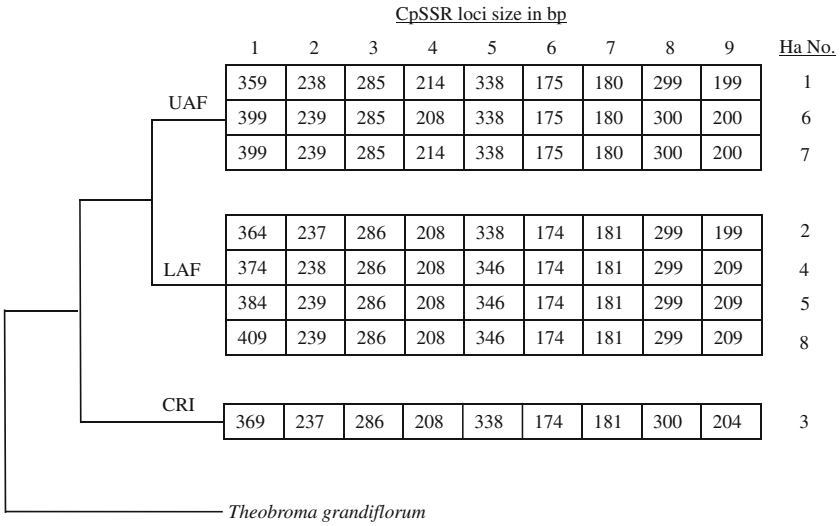

Fig. 5 Correspondence of the three cpSNP haplotypes (UAF, LAF and CRI), compared against the outgroup *T. grandiflorum*, to the eight cpSSR haplotypes (Ha No.). Numbers within squares represent the fragment size (bp) scored for each of the nine SSR loci. The eight cpSSR haplotypes showed the same fragment size pattern as in the work of Yang et al. (2011), except for CpSSR 2 for Ha 1 which should be 238, not 239 as shown in Table 1 of Yang et al. (2011)

combinations of desirable characteristics primarily yield, pest resistance and superior flavor. This scenario is consistent with the haplotype data presented here.

As cpSSRs are known to be more variable than cpSNPs (Hale et al. 2004), it is not surprising that more haplotypes were detected by the former method. The presence of eight distinct haplotypes suggests three hypotheses: (1) that the eight maternal lineages (corresponding to the eight cpSSR haplotypes) were introduced in multiple (at least eight) introductions; (2) that each haplotype group, Criollo, UAF and LAF corresponding to three different geographical origins, putatively Central America, Peru, and Venezuela, respectively, was the result of a single introduction event, but that multiple maternal lineages (seeds) were introduced in each event; or (3) that three maternal lineages survived after 1727 from early germplasm introductions to Trinidad and Tobago, corresponding to the three cpSNP haplotypes, and that the diversity of cpSSR haplotypes arose subsequently by mutations occurring within Trinidad and Tobago. Although chloroplast DNA is generally more conserved than nuclear DNA, cpSSRs have been shown to possess high levels of intraspecific variability (reviewed by Provan et al. 2001). Marshall et al. (2002) found cpSSR mutation rate in lodgepole pine on the order of 10^{-3} mutations per site per generation. If *T. cacao* has a similarly high cpSSR mutation rate, then it is not impossible that at least some of the cpSSR haplotypes could have arisen in situ within the last 300 years. Further haplotype sampling in the potential source areas will be needed to elucidate which of these three hypotheses best explains the origin and history of the Trinitario cacao.

On farm haplotype variation in cacao from Trinidad and Tobago

The plastid haplotype composition of different accessions within farms and within localities in Trinidad and Tobago revealed relatively high on-farm diversity making it impossible to find farm-specific or locality-specific genetic identifiers with these markers. One surprising result was the relative paucity of cpSNP Criollo haplotypes in the farms of Trinidad and Tobago. Of the 72 Trinitario farm accessions sampled, only three had Criollo haplotypes (Table 1). Interestingly, Motamayor et al. (2003), using nuclear RFLP and SSR markers, found Criollo allelic admixture in 49 % of the Trinitarios sampled. This difference is not too surprising since the two studies used two different types of markers: plastid versus nuclear. A hybrid plant will carry the plastid haplotype of the maternal tree. It may be that the remnant Criollo trees acted predominantly as pollen donors to LAF and UAF trees from which fruits were collected for propagation.

Though the Criollo haplotypes were found to be rare on the farms, they were more common in the Trinitario accessions of the ICG, T collection. Eight of the 13 ICS accessions and four out of six TRD samples had Criollo haplotypes, which suggests that the TRD series may be a particularly valuable collection in capturing a greater percentage material carrying a Criollo maternal lineage. One explanation as to the low number of Criollo haplotypes on the farms could be that trees with Forastero cytoplasm have preferentially been used as seed sources over multiple generations. Alternatively Criollo genotypes generally with their higher susceptibility to disease may have suffered preferential mortality on the farms of Trinidad and Tobago. Interestingly, the three Criollo haplotypes were all collected from farms in regions of high elevation, ranging from 226 to 363 m. Moreover, the TRD Criollo accessions were also collected in the high relief region of north east Trinidad (Bekele et al. 2007). It is likely that these mountainous regions, being remote and isolated, may have served as

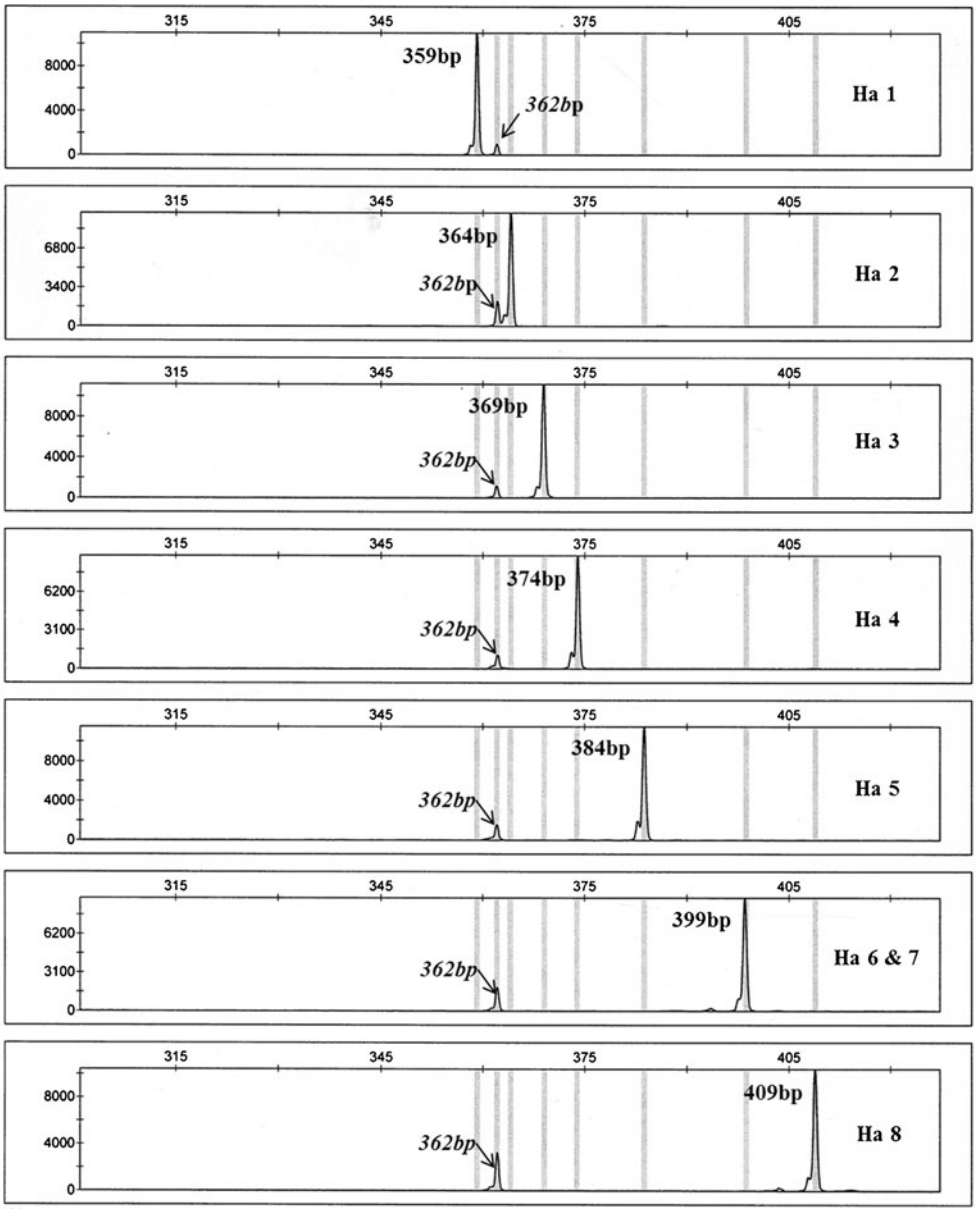

Fig. 6 Chromatograms of the seven alleles from the pentameric SSR locus (CaCrSSR1), ranging from 359 to 409 bp. A low-intensity peak of 362 bp fragment size was found in all samples and might represent the product of cross-amplification from the nuclear DNA. This peak never presented a problem in the scoring of the main alleles. Haplotype (*Ha*) 1, Ha6 and Ha7=LAF; Ha2, Ha4 and Ha5=UAF; Ha3=CRI

an area of refuge for the Criollo types. An important point to note is that it is entirely possible that accessions with Forastero cytoplasmic haplotypes could have abundant Criollo ancestry in their nuclear genome.

Our study confirms that the Trinitario germplasm originated from UAF, LAF and Criollo stock with a strong contribution from the Forastero groups as suggested by Motilal et al. (2010). However, we find a more prominent role of LAF in Trinitario development. The greater preponderance of cpSNP LAF over UAF haplotypes (61 % vs. 35 %) would support the idea of a prevalence of LAF haplotypes existing in Trinidad as relic germplasm. The proximity of Venezuela to Trinidad and the island's early political affiliation with the mainland would have increased the likelihood of LAF germplasm being brought to the island from the lower Amazon basin and Venezuela (Bartley 2005).

Pentameric cpSSR and its possible applications

CaCrSSR1, a microsatellite with a pentameric repeat motif, was so variable that seven of the eight plastid haplotypes found in our samples could be distinguished by using this marker alone. The main chloroplast fragment was always accompanied by a single monomorphic weak band; however, this did not produce a source of genotyping bias as the two bands clearly differed in intensity. The weaker intensity of the monomorphic band suggests that it might have nuclear DNA origin, reflecting the differential representation of organelle and

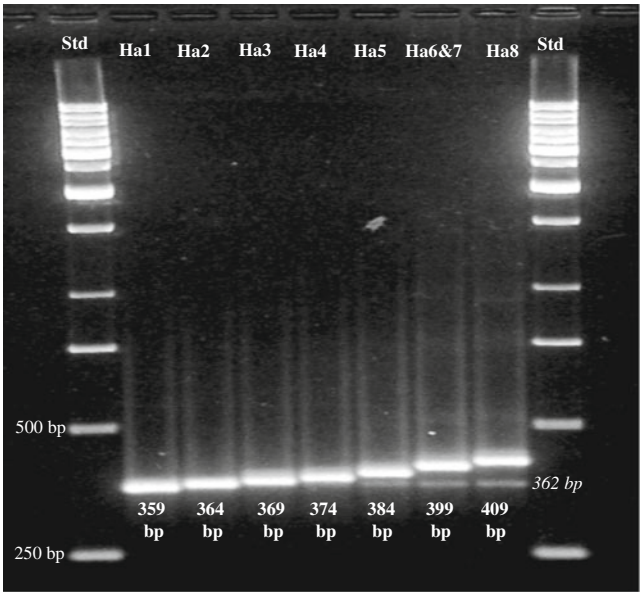

Fig. 7 Gel image of the pentameric SSR amplified fragments. Seven haplotypes could be distinguished on a 2 % high performance agarose gel. The allele size ranged from 359 to 409 bp in multiples of 5 bp. A 1-kb DNA ladder (Std) was used to compare the allele sizes

nuclear DNA. Multiple copies of the plastid genome would be amplified rather than two copies of nuclear genes (or at most a few copies depending on the occurrence of endopolyploidy). The presence of the extra monomorphic band should not overshadow the usefulness of this pentameric microsatellite as a quick and efficient genotyping tool. The CaCrSSR1 marker would provide a time saving and cost-effective resource for genetic identification of cacao accessions (see Yang et al. (2011) for more details).

Transfer of chloroplast genes and large fragments of the chloroplast into the nuclear genome is a common phenomenon in angiosperms (Huang et al. 2005; Matsuo et al. 2005). This raises the question of whether nuclear copies of plastid regions could be a confounding factor in our study. Our method of chloroplast genomics (Kane et al. 2012) allows us to distinguish true chloroplast sequences by a depth of coverage criterion. Our low-pass genome sequences of leaf tissue gives very low coverage for nuclear genes whereas true plastid genomes, represented in each mesophyll cell by hundreds of organelles each with multiple genomes, are present with up to×1,000 coverage (Kane et al. 2012). By setting coverage thresholds very high for plastid genome assembly, we can be sure we are using true plastid sequence as a reference. Plastidic transfers to the nucleus undergo mutational decay fairly rapidly (Huang et al. 2005) and therefore our highly specific organelle primers will not necessarily amplify nuclear copies. Furthermore, nuclear PCR template is likely to be swamped by the copy number differences in favor of chloroplast sequences during amplification. If nuclear copies are amplified we would expect to see heterozygotes, which, despite careful analysis, we do not (with the exception of the invariant ghost band of the pentameric repeat). There is therefore no evidence for confounding nuclear markers in our system.

Future directions

These chloroplast markers could help with fine-scale phylogeographical elucidation of the precise source of Trinitario maternal lineages now found in Trinidad and Tobago. For example, LAF cacao types occur widely in the lower Amazonian region, i.e., the Lower Amazon basin, the Orinoco basin, and the Guyanas (Motamayor et al. 2003). Broader surveys could possibly narrow the region of origin of the cpSSR haplotypes that correspond to LAF in Trinidad and Tobago (Hap 2, 4, 5, and 8). In addition, these chloroplast markers, given their variability and their ability to discriminate samples within populations, could also be used to address questions at an agronomic scale, allowing, for instance, a deeper analysis of the composition of material in West Africa and other regions to which cacao has been introduced. Furthermore, the combined use of nuclear and chloroplast markers would enhance the ability to identify the parental pair of a given cacao tree. This has considerable utility in cacao breeding and biology, including determining seed provenance, identifying superior parental combinations in elite farmer selections and studying gene flow in natural populations.

Acknowledgements We thank the World Bank for a Development Marketplace grant awarded to QC, HD and Johannes Engels and coordinated by Bioversity International. The Staff of Cocoa Research, Centeno, Trinidad is highly acknowledged for identifying and collecting leaves from the farmers' cacao trees, sometimes under very challenging conditions. We thank Chris Grassa for analytical assistance, and Brian Irish (USDA, ARS) and Kyle Wallick (USBG) for providing leaf samples. Stephen Pinney (USDA, ARS) and Kasey Gordon (CRU) are thanked for assistance with DNA extractions. Jon Armstrong and Jarret Glasscock (Cofactor Genomics, St. Louis) provided valuable sequencing assistance. We also gratefully acknowledge the laboratory support from an NSERC (Canada) discovery grant to Quentin Cronk.

Data Archiving Statement Short-read sequences generated by the Illumina GA-II platform to assemble the *T. cacao* chloroplast genome have been deposited in the NIH Sequence Read Archive (SRA). The SRA accession number is SRA048198. The whole chloroplast genome sequences of the nine *T. cacao* genotypes and one *T. grandiflorum* genotype were deposited in GenBank. The accession numbers for these sequences are JQ228379-JQ228389. The DNA sequences used to obtain the ten SNPs were deposited in GenBank as well. The GenBank accession numbers for the first SNP primer pair are JX413598–JX413691; and the GenBank accession number for the second primer pair are JX413692–JX413781. The GenBank accession numbers for the SSR primers are JF979116–JF979124.

References

Arroyo-García R, Lefort F, de Andrés MT, Ibáñaez J, Borrego J, Jouve N, Cabello F, Martínez-Zapater JM (2002) Chloroplast microsatellite polymorphisms in *Vitis* species. Genome 45:1142–1149

Bartley BGD (2005) The genetic diversity of cacao and its utilization. CABI, Wallingford, UK

Bekele FL, Bidaisee GG, Bhola J (2007) A comparative morphological study of two Trinitario groups from the International Cocoa Genebank, Trinidad. Annual Report 2006, Cocoa Research Unit, University of the West Indies, Trinidad and Tobago, pp 34–42

Cheesman EE (1934) The botanical programme of 1933. In Third annual report on cacao research,1933. Government Printing Office, Port-of-Spain, Trinidad, pp 1–2

Cheesman EE (1944) Notes on the nomenclature, classification and possible relationships of cocoa populations. Trop Agric 21:144–159

Coe SD, Coe MD (1996) The true history of chocolate. Thames and Hudson, New York, USA

Corriveau JL, Coleman AW (1988) Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperms. Am J Bot 75:1443–1458

Dieringer D, Schlötterer C (2003) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Mol Ecol Notes 3:167–169

Engels JMM (1986) The systematic description of cacao clones and its significance for taxonomy and plant breeding. Dissertation, Agricultural University, Wageningen, Netherlands, 125p

Felsenstein J (1973) Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 22:240–249

Felsenstein J (1989) PHYLIP — phylogeny inference package (Version 3.2). Cladistics 5:164–166

Freeman WE (1969) Some aspects of the cacao breeding programme. Proceedings of Agricultural Society of Trinidad and Tobago, Dec 1968, pp 1–15

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Hale ML, Borland AM, Gustafsson MH, Wolff K (2004) Causes of size homoplasy among chloroplast microsatellites in closely related *Clusia* species. J Mol Evol 58:182–190

Huang CY, Grünheit N, Ahmadinejad N, Timmis JN, Martin W (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol 138:1723–33

Irish BI, Goenaga R, Zhang D, Schnell R, Brown S, Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS tropical agriculture research station cacao (*Theobroma cacao* L.) germplasm collection. Crop Sci 50:656–667

Jansen RK, Saski C, Lee SB, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol 28:835–847

Johnson ES, Bekele FB, Brown SJ, Song Q, Zhang D, Meinhardt LW, Schnell RJ (2009) Population structure and genetic diversity of the Trinitario cacao (*Theobroma cacao* L.) from Trinidad and Tobago. Crop Sci 49:564–572

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, Engels JMM, Cronk Q (2012) Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am J Bot 99:320–329

Kim S, Lee Y-P, Lim H, Ahn Y, Sung SK (2009) Identification of highly variable chloroplast sequences and development of cpDNA-based molecular markers that distinguish four cytoplasm types in radish (*Raphanus sativus* L.). Theor Appl Genet 119:189–198

Loor RG, Risterucci AM, Courtois B, Fouet O, Jeanneau M, Rosenquist E, Amores F, Vasco A, Medina M, Lanaud C (2009) Tracing the native ancestors of the modern *Theobroma cacao* L. population in Ecuador. Tree Genetics & Genomes 5:421–433

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079

Marshall HD, Newton C, Ritland K (2002) Chloroplast phylogeography and evolution of highly polymorphic microsatellites in lodgepole pine (*Pinus contorta*). Theor Appl Genet 104:367–378

Matsuo M, Ito Y, Yamauchi R, Obokata J (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast–nuclear DNA flux. The Plant Cell 17:665–675

Mogensen HL (1996) The hows and whys of cytoplasmic inheritance in seed plants. Am J Bot 83:383–404

Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication: I The origin of the cacao cultivated by the Mayas. Heredity 89:380–386

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication: II Progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322–330

Motilal LA, Zhang D, Umaharan P, Mischke S, Boccara M, Pinney S (2009) Increasing accuracy and throughput in large scale microsatellite fingerprinting of cacao field germplasm collections. Trop Plant Biol 2:23–37

Motilal LA, Zhang D, Umaharan P, Mischke S, Mooleedhar V, Meinhardt LW (2010) The relic Criollo cacao in Belize – genetic diversity and relationship with Trinitario and other cacao clones held in the International Cocoa Genebank Trinidad. Plant Genet Resource 8:106–115

Motilal LA, Zhang D, Umaharan P, Mischke S, Pinney S, Meinhardt LW (2011) Microsatellite fingerprinting in the International Cocoa Genebank, Trinidad: accession and plot homogeneity information for germplasm management. Plant Genet Resource 9:430–438

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170

Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction. In: Weiss S, Ferrand N (eds) Phylogeography of Southern European refugia. Springer, Dordrecht, Netherlands, pp 23–97

Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256

Pound FJ (1931) The genetic constitution of the cacao crop. First annual report on cacao research 1931. Government Printing Office, Port of Spain, Trinidad, pp 10–24

Pound FJ (1933) Criteria and methods of selection in cacao. In Second annual report on cacao research 1932. Government Printing Office, Port of Spain, Trinidad, pp 27–29

Pound FJ (1934) The progress of selection, 1933. In Third annual report on cacao research 1933. Government Printing Office, Port of Spain, Trinidad, pp 25–28

Pound FJ (1935) The progress of selection, 1934. In Fourth annual report on cacao research 1934. Government Printing Office, Port of Spain, Trinidad, pp 7–11

Pound FJ (1936) The progress of selection, 1935. In Fifth annual report on cacao research 1934. Government Printing Office, Port of Spain, Trinidad, pp 7–16

Powell W, Morgante M, Doyle JJ, McNicol JW, Tingey SV, Rafalski AJ (1996) Genepool variation in genus *Glycine* subgenus *Soja* revealed by polymorphic nuclear and chloroplast microsatellites. Genetics 144:793–803

Provan J, Corbett G, McNicol JW, Powell W (1997) Chloroplast DNA variability in wild and cultivated rice (*Oryza* spp.) revealed by polymorphic chloroplast simple sequence repeats. Genome 40:104–110

Provan J, Russell JR, Booth A, Powell W (1999) Polymorphic chloroplast simple sequence repeat primers for systematic and population studies in the genus *Hordeum*. Mol Ecol 8:505–511

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147

Rambaut A (2006–2009) FigTree. Tree figure drawing tool v.1.3.1, Institute of Evolutionary Biology, University of Edinburgh, UK

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, New Jersey, USA, pp 365–386

Shephard CY (1932) The cacao industry of Trinidad: some economic aspects, Part I. Government Printing office, Port of Spain, Trinidad, pp 95–100

Silva CRS, Albuquerque PSB, Ervedosa FR, Mota JWS, Figueira A, Sebbenn AM (2010) Understanding the genetic diversity, spatial genetic structure and mating system at the hierarchical levels of fruits and individuals of a continuous *Theobroma cacao* population from the Brazilian Amazon. Heredity 106:973–985

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction of a recently diverged plant group. Am J Bot 85:1301–1315

Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571

Sveinsson S, Kane NC, Dempewolf H, Zhang D, Cronk QC (2010) *Theobroma cacao* chloroplast, complete genome. Website http://www.ncbi.nlm.nih.gov/nuccore/HQ244500 Accessed 15 January 2012

Takayama K, Kajita T, Murata J, Yoichi T (2006) Phylogeography and genetic structure of *Hibiscus tiliaceus* – speciation of a pantropical plant with sea-drifted seeds. Mol Ecol 15:2871–2881

Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399

Toxopeus H (1985) Botany, types and populations. In: Wood GAR, Lass RA (eds) Cocoa, 4th edn. Longman, London, pp 11–37

Wendel JF (1989) New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci 86:4132–4136

Wood GAR (1985) History and development. In: Wood GAR, Lass RA (eds) Cocoa, 4th edn. Longman, London, pp 1–10

Yang JY, Motilal LA, Dempewolf H, Maharaj K, Cronk QC (2011) Chloroplast microsatellite primers for cacao (*Theobroma cacao*) and other Malvaceae. Am J Bot 98:e372–374

Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Dissertation, University of Texas at Austin, Texas, USA