Unique haplotypes of cacao trees as revealed by *trnH-psbA* chloroplast DNA

Nidia Gutiérrez-López, Isidro Ovando-Medina, Miguel Salvador-Figueroa, Francisco Molina-Freaner, Carlos H. Avendaño-Arrazate, Alfredo Vázquez-Ovando

Cacao trees were domesticated in Mesoamerica approximately 4000 years ago and are still grown today. In this study, we analyzed sequence variation in the chloroplast DNA *trnH-psbA* intergenic spacer from 28 cacao trees from different farms in the Soconusco region in southern Mexico. Genetic relationships were established by two analysis approaches based on geographic origin (five populations) and genetic origin (based on a previous study). We identified six polymorphic sites, including five insertion/deletion (indels) types and one transversion. The overall nucleotide diversity was low for both approaches (geographic = 0.0032 and genetic = 0.0038). Conversely, we obtained moderate to high haplotype diversity (0.66 and 0.80) with 10 and 12 haplotypes, respectively. The common haplotype (H2) for both networks included cacao trees from all geographic locations (geographic approach) and four genetic groups (genetic approach). This common haplotype (ancient) derived a set of intermediate haplotypes and singletons interconnected by one or two mutational steps, which suggested directional selection and event purification from the expansion of narrow populations. No genetic differentiation (AMOVA, F_{ST} = 0) was found, and the SAMOVA F_{ST} value (0.04339) was not large enough to show moderate differentiation between the populations. Only one population showed a high haplotype frequency; thus, this population could be considered an important reservoir of genetic material. The indels located in the *trnH-psbA* intergenic spacer of cacao trees could be useful as markers for the development of DNA barcoding.

Unique haplotypes of cacao trees as revealed by *trnH-psbA* chloroplast DNA

Nidia Gutiérrez-López[1], Isidro Ovando-Medina[1], Miguel Salvador-Figueroa[1], Francisco Molina-Freaner[2], Carlos H. Avendaño-Arrazate[3], Alfredo Vázquez-Ovando[1]

[1] Instituto de Biociencias, Universidad Autónoma de Chiapas, Tapachula, Chiapas, Mexico

[2] Departamento de Ecología de la Biodiversidad, Instituto de Ecología, Universidad Nacional Autónoma de México, Hermosillo, Sonora, Mexico

[3] Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, C. E. Rosario Izapa, Tuxtla Chico, Chiapas, Mexico

Corresponding Author:

Alfredo Vázquez-Ovando

Boulevard Príncipe Akishino sin número. Colonia Solidaridad 2000. Tapachula, Chiapas, CP 30798, México

Email address: jose.vazquez@unach.mx

Abstract

Cacao trees were domesticated in Mesoamerica approximately 4000 years ago and are still grown today. In this study, we analyzed sequence variation in the chloroplast DNA *trnH-psbA* intergenic spacer from 28 cacao trees from different farms in the Soconusco region in southern Mexico. Genetic relationships were established by two analysis approaches based on geographic origin (five populations) and genetic origin (based on a previous study). We identified six polymorphic sites, including five insertion/deletion (indels) types and one transversion. The overall nucleotide diversity was low for both approaches (geographic = 0.0032 and genetic = 0.0038). Conversely, we obtained moderate to high haplotype diversity (0.66 and 0.80) with 10 and 12 haplotypes, respectively. The common haplotype (H2) for both networks included cacao trees from all geographic locations (geographic approach) and four genetic groups (genetic approach). This common haplotype (ancient) derived a set of intermediate haplotypes and singletons interconnected by one or two mutational steps, which suggested directional selection and event purification from the expansion of narrow populations. No genetic differentiation (AMOVA, $F_{ST} = 0$) was found, and the SAMOVA F_{ST} value (0.04339) was not large enough to show moderate differentiation between the populations. Only one population showed a high haplotype frequency; thus, this population could be considered an important reservoir of genetic material. The indels located in the *trnH-psbA* intergenic spacer of cacao trees could be useful as markers for the development of DNA barcoding.

47

48 **Introduction**

49 The Neotropical cacao tree (*Theobroma cacao* L.) is widely cultivated in America, Africa and
50 Asia. It is considered an economically important crop because its seeds are used in the chocolate
51 industry (Wood, 2001). Trees are classified based on morphological traits and their geographic
52 origins as Criollo, Forastero and Trinitario (Cheesman, 1944). In Mesoamerica, the Criollo cacao
53 has been widely used as food for nearly 4000 years (De la Cruz *et al*., 1995; Whitkus *et al*.,
54 1998; Powis *et al*., 2011).

55 Motamayor *et al*. (2008) proposed 10 genetic cacao groups based on simple sequence repeat
56 (SSR) analysis. Currently, Criollo retains an identity as a separate group, whereas the other
57 proposed genetic groups comprise all trees from South America. This region has been reported to
58 contain the highest genetic diversity of cacao trees.

59 Conversely, the genetic diversity of cacaos in southern Mexico was reported to be moderate to
60 low in natural populations (Whitkus *et al*., 1998 using RAPD markers) and cultivated forms
61 (Vázquez-Ovando *et al*., 2014 using microsatellite markers), although a wide diversity in cacao
62 pod and seed morphologies was observed. In the Soconusco farms (Chiapas, Mexico), Vázquez-
63 Ovando *et al*. (2014) found moderate to high allelic richness and high levels of homozygosity.
64 The authors reported the presence of trees sharing genetic identity with those considered
65 "Ancient Criollo" but also reported the presence of private alleles. These alleles may be
66 associated with commercially relevant phenotypic traits that preserve their relationship with
67 other polymorphic regions of the DNA.

68 The chloroplast DNA (cpDNA) and its markers have been increasingly used for studies of the
69 genetic population structure, evolution, gene flow, haplotype frequency and phylogenetic

relationships. Due to its high conservation due to maternal uniparental inheritance, cpDNA is the main data source used for the construction of phylogenetic relationships in plants (Shaw & Small, 2005). However, the cpDNA contains variable DNA regions, which makes them useful for studies of population genetics and conservation issues (Shaw & Small, 2005; Shaw *et al*., 2007). These regions have been widely used to establish phylogeography patterns in alpine species (Wang *et al*., 2008), to gain insight into the center of origin of cultivated grape populations in Europe (Arroyo-Garcia *et al*., 2006) and to explain the diversity and population structure of cultivated Chinese cherries (Chen *et al*., 2013).

cpDNA is not commonly used for cocoa trees. Instead, this technique is primarily used to analyze population genetic variability and to elucidate the complex origins of the cocoa varieties. Yang *et al*. (2011) developed cpSSRs that were subsequently used together with cpSNP markers (developed by Kane *et al*., 2012) to untangle the genetic origins of the Trinitario cultivar in Trinidad and Tobago (Yang *et al*., 2013).

The most commonly used cpDNA intergenic spacer is *trnH-psbA,* which has shown high variability and can be used to elucidate genetic relationships at the intraspecific level (Azuma *et al*., 2001; Hamilton, Braverman & Soria-Hernanz, 2003). The *trnH-psbA* region sequences from 10 cacao accessions deposited in the NCBI database exposed only one haplotype (Kane *et al*., 2012), whereas Jansen *et al*. (2011) showed the presence of polymorphic sites located at a different haplotype. The main polymorphisms reported in the noncoding cpDNA region are inversions, transitions and transversions (Whitlock, Hale & Groff, 2010; Zeng *et al*., 2012). Few studies have reported the presence of insertions or deletions (indels), although indels are probably a common feature in the *trnH-psbA* spacer (Aldrich *et al*. 1988).

Nonetheless, the use of indels for diversity and phylogenetic analysis has been questioned (Bieniek, Mizianty & Szklarczyk, 2015; Whitlock, Hale & Groff, 2010) because the mechanism causing indels remains unclear. However, indels are informative characteristics because genetic variability detected using polymorphism due to indels or substitutions can be studied without distinction (Nei, 1987). Therefore, indels are useful markers. Moreover, the inclusion of indels in diversity and phylogenetic analyses enhances the discriminant power between species (Raymúndez *et al*., 2002; Hamilton, Braverman & Soria-Hernanz, 2003; Kress & Erickson, 2007; Sun *et al*., 2012) and even between conspecific individuals (Pérez-Jiménez *et al*., 2013). Therefore, the aim of this study was to evaluate the genetic diversity and describe the relationship between individuals of the *Theobroma cacao* L. Criollo type of the Soconusco region (Chiapas, Mexico) using the variations in chloroplast DNA revealed by the *trnH-psbA* spacer sequence.

Material & methods

Plant material and sample collection

A total of 45 cacao samples were included in this study. Thirty-eight trees were sequenced and analyzed, and seven sequence accessions were downloaded from GenBank as references. A total of 28 of the 38 sequenced trees were selected from plantations in Soconusco (Chiapas, Mexico) based on a previous characterization (Vázquez-Ovando *et al*., 2014) using 10 SSR molecular markers. The individuals were selected based on fruit (pod) and seed traits that resembled those of the Criollo variety. The pods were elongated, deeply grooved and pointed at the pod end and

had a lumpy surface with a warty exterior appearance, white or slightly pigmented seeds and sweet mucilage. In agreement with the report by Vázquez-Ovando *et al*. (2014), the individuals were classified as 12 trees with high Criollo ancestry, 11 Non-Criollo group trees and five admixtures (Table 1). Additionally, 10 accessions were sequenced and included as references: two Forastero variety (Catongo and EET 399), one Trinitario variety (RIM 24) and seven wild Criollo [one collected in the Lacandon rainforest (SL01) and six obtained from the germplasm of the Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, México (Yaxcabá, Xocen, Lacandón 06, Lacandón 28, Lagarto and Carmelo); Table 1]. *Theobroma bicolor* was used as the outgroup. The average age of the trees was 30 years. Leaves were sampled and placed in plastic bags, taken to the laboratory (4 °C) and stored at -20 °C prior to processing.

DNA extraction, amplification and sequencing

The total DNA extraction was performed by modifying the method described by Doyle & Doyle (1990). The leaves were washed with sterile water and 70% ethyl alcohol. Approximately 200 mg of cacao leaves were ground with liquid nitrogen with 60 mg polyvinyl pyrrolidone and 1 mL of CTAB buffer [2% CTAB (w/v), 20 mM ethylenediaminetetraacetic acid (EDTA), 1.4 M NaCl, 100 mM Trizma® base, pH adjusted to 8 with HCl and 1% 2-mercaptoethanol (v/v)]. DNA extractions were performed with chloroform-isoamyl alcohol and precipitated with isopropanol. The extracted DNA was purified with a mixture of phenol:chloroform:isoamyl alcohol (25:24:1). The DNA integrity (dissolved in 60 µL of Milli-Q water) was verified by 0.8% agarose gel electrophoresis and quantified by spectrophotometry at 260 nm (Jenway, Genova). The purity was inferred by the 260/280 and 260/230 absorbance ratios.

The cpDNA amplification of the *trnH-psbA* intergenic spacer was conducted using the forward primer 5'-CGCGCATGGTGGATTCACAATCC-3' and reverse primer 5'-GTTATGCATGAACGTAATGCTC-3' (Shaw & Small, 2005). The PCR conditions were performed as described by Shaw & Small (2005) with modifications in the concentration of $MgCl_2$ (2 mM) and the use of the average value of the reported melting temperatures. The PCR was performed in a 25 µL reaction mixture containing 100 ng of genomic DNA, 4 µL of 10x PCR ViBuffer A (Vivantis™ Oceanside CA, USA), 1 µL of $MgCl_2$ (50 mM), 0.5 µL of dNTP Mix (10 mM, Promega), 0.05 mM of each primer and 2.5 U of Taq DNA polymerase (Vivantis™). Following one cycle of 5 min at 94 °C, 35 PCR cycles of 30 s at 94 °C, 30 s at 53 °C and 1 min at 72 °C and a 10 min 72 °C final extension were performed in a TC3000 thermal cycler (Techne, Cambridge, UK). The PCR products were separated on 6% polyacrylamide gels using 0.5X TBE buffer at 110 V for 210 min, stained with ethidium bromide (0.6 ng/µL) for 15 min, visualized under UV light and photographed with a Gel Doc™ EZ Imager gel documentation system (Bio-Rad, USA). Fragment sizes were estimated using Image Lab (v. 4.0.1, Bio-Rad Laboratories) and integrating the GeneRuler™ 100 bp DNA Ladder Plus (Fermentas®) as a molecular weight marker.

The PCR products were directly sequenced using the Dye Terminator Cycle Sequencing with Quick Start Kit (GenomeLab™) on a CEQ™ 8000 automatic DNA sequencer (Beckman Coulter™). To validate the results, the DNA was extracted twice and amplified independently. The sequences were verified by comparison with their forward and reverse sequences when applicable.

Sequence alignment and data analysis

160 The sequence quality was checked and the electropherograms were edited using BioEdit© (Hall,
161 1999). Sequences were limited at the ends to avoid the presence of variable sites due to the
162 introduction of sequencing artifacts by the polymerase (approx. 40 bp) and aligned with
163 ClustalW 1.81 (Thompson, Higgins & Gibson, 1994). Visual inspection and manual editing of
164 the sequences was performed to confirm the variable sites. We used two different analytical
165 approaches based on the geographic origin and the genetic origin of the samples (Table 1). In
166 both approaches, molecular diversity indices including the number of segregating sites (S), the
167 number of haplotypes, the haplotype diversity (Hd) and the nucleotide diversity (πd) were
168 estimated following the methods of Nei (1987) in DnaSP© 5.1 (Rozas *et al*., 2010).
169 To infer evolutionary relationships at the intraspecific level, we evaluated network building. The
170 method used was median-joining (MD) based on parsimony criteria (Bandelt, Forster & Röhl,
171 1999; Polzin & Daneshmand, 2012) and was performed with the software Network© 4.6.1.3.
172 Analysis of molecular variance (AMOVA), pairwise *Fst* values and statistical analyses of
173 molecular variance [F_{CT} (test performed by permuting individuals within populations), F_{ST} (test
174 performed by permuting genotypes among populations but within groups) and F_{SC} (test
175 performed by permuting genotypes among groups)] were estimated using Arlequin© version 3.0
176 (Excoffier, Laval & Schneider, 2005). Significance was evaluated by 99,999 random sequence
177 permutations. To determine whether sample sites clustered on a population level, a spatial
178 analysis of variance (SAMOVA) was conducted (Dupanloup, Schneider & Excoffier, 2002)
179 using haplotype data and the geographic coordinates of each of the 5 sample sites. The
180 SAMOVA was run for K = 2–5 putative populations to determine the maximum F_{ST} value and
181 the highest proportion of differences between populations due to genetic variation.

The neutral evolution of chloroplast DNA was evaluated to examine whether any population had experienced historic demographic changes using Tajima´s D test (Tajima, 1989) with Arlequin© version 3.0 (Excoffier, Laval & Schneider, 2005). We evaluated changes using the overall geographic approach as well as the populations.

Seven accessions from the NCBI database were included as references in the genetic origin approach analysis: Matina 1/6 (HQ336404.2), Criollo-22 (JQ228379.1) Amelonado (JQ228380.1) Scavina 6 (JQ228382.1), ICS 1 (JQ228381.1), ICS 6 (JQ228383.1) and ICS 39 (JQ228387.1).

Results

Sequence characterization and genetic diversity

The *trnH-psbA* intergenic spacer sequences from 45 *Theobroma cacao* samples (Table 1) were aligned with a consensus length of 526 bp. Six segregating polymorphic sites (Table 2) were present as five indels (Figure 1) and one transversion (T↔A event at position 134). These polymorphisms resulted in 12 haplotypes, of which four were singletons represented by a unique sequence in the sample (Table 2). The nucleotide composition of the fragment revealed that is was AT-rich (A+T, 75.52%).

The results based on the geographic approach revealed that the overall average haplotype diversity (Hd) and nucleotide diversity (πd) values were 0.66 and 0.0032, respectively (Table 3). We identified 10 haplotypes. The most frequent haplotype (H2) was shared by 19 trees from seven geographic populations formed *a priori* (Table 2). Four trees belonging to Population 1

(one tree), Population 3 (one tree) and Population 5 (two trees) formed the second most common haplotype (H1). Overall, 50% of the haplotypes were singletons (Table 2). The analysis showed that most of the genetic diversity was found in Population 4 (Mazatán), which contained the highest values for all indices; Population 4 included 41.7% of the identified haplotypes. The other populations maintained moderate Hd and low πd values that were similar for each population (Table 3). The Yucatán and Selva Lacandona populations (wild) exhibited Hd values of 1 and 0, respectively, although these data were influenced by the low numbers of reference individuals.

When the data analysis was based on the genetic origins, the highest Hd (1.0) was found in the Admixture group (Table 3). In contrast, the Trinitario-reference group had the lowest Hd value (0.5). The πd was low (0.0025 to 0.006) for all groups, which was similar to the results obtained with another approach. The Forastero-reference and Trinitario-reference groups did not present singletons (Table 3). Sequences from the NCBI database were grouped into one haplotype (H12) with the exception of MATI 1/6, which grouped in H11 with EET399 corresponding to the Forastero-reference group.

Intraspecific relationship haplotype

Figures 2 and 3 show the haplotype networks built with data from the geographic (Figure 2) and genetic approaches (Figure 3). The individuals belonging to each haplotype are also included. Both networks show a star arrangement. The general base has a common haplotype for the two networks (H2) that includes cacao trees from all geographic populations (Figure 2) and four of six groups based on the genetic approach (Figure 3). A unique set of intermediate haplotypes are derived from this common haplotype (H2) and are interconnected by one or two mutational steps

in both networks. The H4-H6 haplotypes were farthest from the central clade (i.e., newly created haplotypes; Figures 2 and 3). Haplotypes H3 and H8-H10 were singletons.

Population genetic structure

The analysis of molecular variance (AMOVA) was not significant and had a value of F_{ST}=0. In the spatial analysis of molecular variance (SAMOVA), the value K=2 extended the F_{ST} to 0.04339 (Table 4) and generated two clusters: the first contained only Population 4 (Mazatán) and the second cluster grouped the other geographic populations (Table 4).

The neutrality tests showed non-significant values in the Tajima´s D with the exception of Population 4, in which the Tajima's D value was negative (D = -0.93302). All other populations showed values of D = 0; however, the overall value for this test was D = -0.13329 (P> 0.1).

Discussion

In this study, high haplotype variation was found in the chloroplast DNA from cacao trees grown in the Soconusco region. No inversions or transitions were found, although they were reported to be common in other plants (Whitlock, Hale & Groff, 2010; Zeng *et al*., 2012). However, we found insertions or deletions (indels) in three poly-A regions and one A↔T transversion (Figure 1). This result agreed with the findings reported by Jansen *et al*. (2011) in the 1/6 MATI accession and supported the affirmation by Aldrich *et al*. (1988) that indels were a presumably common feature in the *trnH-psbA* region. In the data analysis, we included the indels as informative character states, and the high interspecific divergence of the spacer region allowed their use as a marker for DNA barcoding (Kress & Erickson, 2007). The molecular diversity

indices determined in the present study were similar to the results of Zeng *et al*. (2012) using the same intergenic spacer, which revealed 11 haplotypes for 35 *Thinopyrum intermedium* samples, low nucleotide diversity ($\pi d = 0.00473$) and moderately high haplotype diversity (Hd = 0.7331) (our results for the geographic populations were $\pi d = 0.0032$ and Hd = 0.66). The results of these authors supported the use of one intergenic spacer to reveal nucleotide polymorphisms.

Our haplotype diversity results are contrary to those reported by Vázquez-Ovando *et al*. (2014). These authors reported low genetic diversity in individuals from the same region (in particular Population 4 in Mazatán) using microsatellite markers. One reason for the discrepancy may be that a larger number of individuals with Criollo ancestry was included in that study, resulting in a higher degree of homozygotes and lower population genetic diversity. Our study also included individuals from other cacao varieties that possessed greater genetic diversity, at least at the nuclear DNA level. However, the low nucleotide diversity found in this study was supported by the low genetic variability found using nuclear microsatellites. Individuals included in both studies showed great morphological pod variability that resembled the Criollo type (e.g., different degrees of roughness, color and deep grooves). This finding could reveal a greater association between the morphological variability of the cacao pod with the reported allelic richness (Vázquez-Ovando *et al*., 2014) and the variability of the haplotypes found in our study.

The number of haplotypes was higher than the number of polymorphic sites (Table 2). This feature is associated with ancestral species that have sufficiently diverged to accumulate mutations among different haplotypes (Roger, 1995). The haplotype number detected in the present study is unusually striking compared with other works. For example, Yang *et al*. (2013) found only three haplotypes based on three cpSNP markers. However, that study exclusively

analyzed nucleotide substitutions, whereas in this study five indel regions were included; this difference may explain the high haplotype diversity found here. Indels have been reported to have a high mutation rate compared with other regions of the cpDNA (Igvarsson, Ribstein & Taylor, 2003), especially when they are repeated locally (Yamane, Yano & Kawahara, 2006) such as in region 309-310 of our sequences (Figure 1).

Several explanations are possible for the presence of more than one Criollo haplotype. First, only the maternal line gave rise to the eight Criollo haplotypes by mutation. Second, the "Criollo" phenotype had multiple provenances, indicating that the ancient haplotypes persisted over time in the Soconusco cacao farms. Third, some samples were misclassified as "Criollo" (especially MAJH02 and Carmelo, which were the most divergent "Criollo" individuals; haplotypes 4 and 6, respectively, Figure 3). These samples possibly belong to the Admixture group rather than the Criollo. However, they are also contenders for the Modern Criollo group (i.e., individuals classified as Criollo that might have been introgressed with Forastero genes) (Motamayor *et al*., 2002) and preserve phenotypic traits of the ancient Criollo. Finally, heteroplasmy and haplotype polymorphisms of plastid genomes within and among individuals were documented in Malvaceae (Wolfe & Randle, 2004). These phenomena could be present in the *Theobroma cacao*. To test those hypotheses, additional studies are needed using high-throughput sequencing of chloroplast genomes.

Population 7 (Selva Lacandona) exhibited no haplotype diversity (Hd = 0). However, haplotype H2 located in this population is considered the common ancestor because it is shared by all populations (Figure 2). In contrast, the two individuals belonging to Population 6 (Yucatán), which exhibited different haplotypes (H2 and H9) from one another, were interrelated by only a

297 mutational step (Figure 2). This result shows that an individual tree belonging to a Yucatán
298 population eventually descended from other individuals in this region where the Maya people
299 grew cacao.
300
301 The low nucleotide polymorphism levels could be explained by rapid population expansion
302 events in the distribution range, whereas high haplotype diversity might be due to the continuous
303 introduction of individuals from different locations. Populations recently introduced or expanded
304 from a small number of founders would have a common haplotype shared by most individuals
305 and many rare haplotypes connected to the main population by a few independent mutations
306 (Slatkin & Hudson, 1991; Avise, 2000) such as observed in the present study (Figure 2). A
307 similar argument was proposed based on the use of microsatellite markers (Vázquez-Ovando *et*
308 *al*., 2014).
309 The relatively low variability in the cultivated cacao populations was supported by the lack of
310 neutrality revealed by the global Tajima test. Specifically, the negative Tajima´s D value (-
311 0.93302) in Population 4 (Mazatán) could be related to a "bottleneck" event, which would
312 indicate population expansion and not natural expansion because it was a cultivated population.
313 The occurrence of unclear events in the past (disease, volcanic eruptions or other natural events)
314 may have caused the almost complete disappearance of populations established by the people in
315 the Mesoamerican region (De Sahagún, 2009, *Codex Florentino*). Rapid expansion due to
316 recolonization of the populations and the probable introduction of other varieties of cacao trees
317 not native to the region would have subjected the populations to a bottleneck events in very
318 recent periods. However, these are presumptive weak inferences of the population history based
319 on a single locus. The bottleneck event could also be related to the loss of alleles (haplotypes;

especially rare alleles), which is much greater than the loss of genetic variance *per se*. Although these rare alleles contribute little to the total genetic variability, they can provide unique responses against evolutionary challenges similar to the high number of unique haplotypes found in this study (3 haplotypes in Population 4). The presence of both common and rare haplotypes can be the result of a directional-purifying selection process or expansion events from small populations (Hedrick, 2005). The H3 and H6-H10 haplotypes (cultivated populations) are singletons. This finding agreed with Crandall & Templeton (1993), who reported that the singletons identified in this study were connected to haplotypes from the same population. Population 4 (Mazatán) shows the highest haplotype diversity, which makes this population an important reservoir of genetic material at the chloroplast and possibly phenotypic levels based on the abundance of pod morphologies observed in this population.

Overall, cacao trees with high ancestry were located in the center of the haplotype network. This result was supported by the coalescence theory that predicted that the ancient haplotype should be the most common and most distributed among the populations. In concordance, derived haplotypes would be less frequent and in many cases would be private; these haplotypes would be located in regions containing the latest cultivated cacao populations. The H4 and H5 haplotypes may have been recently created because they are located at the ends of the network, possibly due to germplasm exchange with traits of interest to cacao farmers. These anthropogenic activities may have had a strong impact on the levels of variation observed in the cpDNA sequences, which explains the observed lack of differentiation. Additionally, migration over long distances and the exchange by farmers contributed to the colonization of new regions

342 founded by a few individuals, thereby establishing different alleles via mutation and genetic
343 drift.
344 Furthermore, the F_{ST} = 0 value determined by AMOVA revealed that all of the molecular
345 variance occurred within populations. Indeed, the SAMOVA F_{ST} value (Table 4) was not
346 sufficient to show at least moderate differentiation between populations ($F_{ST} \geq 0.05$). This
347 finding provides some explanations regarding the demographic history of *T. cacao* trees,
348 indicating that the populations formed *a priori* and experienced gene flow, resulting in
349 population homogenization. The spatial analysis revealed the highest differentiation between
350 groups when K=2 was tested; K=3 (F_{ST} = 0.00088) grouped trees from the Yucatán, Selva and
351 Cacahoatán in the same genetic population. This grouping is unusual because the geographic
352 distance is longer among the three localities and may be associated with the distribution of trees
353 in the past [i.e., the ancestral haplotype (H2) grouped individuals from Selva; one mutational
354 step resulted in the origination of the individuals from the Yucatán, which in turn originated the
355 individuals at Cacahoatán by the same event (Figure 3)]. Following this criterion, H3 has a
356 greater correspondence with the Criollo genotype, although it was previously reported to be an
357 Admixture (Vázquez-Ovando *et al*., 2014).
358
359
360 **Conclusions**
361 Indels and one transversion located in the chloroplast DNA *trnH-psbA* spacer region of cacao
362 trees could allow the development of genetic marker barcodes. The molecular analysis showed
363 low nucleotide diversity but high haplotype diversity possibly due to population bottleneck
364 events. These results were confirmed by the negative Tajima's D and the arrangement of the

haplotype network into a star arrangement. We identified 10 different haplotypes (trees grown) of which H3 and H6-H10 resulted in singletons because they were not associated with other cacaos or with those reported in the molecular databases. The presence of these haplotypes accompanied by the low number of mutational steps might suggest a very short evolutionary history or events that led to disappearing-expanding populations in southern Mexico. One geographic population (Pop 4, Mazatán) consisted of high frequency haplotypes, which makes this zone an important reservoir of genetic material at the chloroplast and possibly phenotypic levels because an abundance of pod morphology was also observed in this population. The genetic differentiation between populations was zero, suggesting that gene flow homogenized the populations.

Acknowledgments

To Nancy Gálvez- Reyes for his advice on data analysis and comments on the manuscript. The authors thank the exhaustive revision of the three referees which helped to improve the manuscript.

References

Aldrich J, Cherney BW, Merlin E, Christopherson L. 1988. The role of insertion/deletions in the evolution of the intergenic region between *psbA* and *trnH* in the chloroplast genome. *Current Genetics* 14:137-146.

387 Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López MA, Arnold C, Ergul A,
388 Söylemezoğlu G, Uzun HI, Cabello F, Ibáñez J, Aradhya MK, Atanassov A, Atanassov I,
389 Balint S, Cenis JL, Costantini L, Goris-Lavets S, Grando MS, Klein BY, McGovern PE,
390 Merdinoglu D, Pejic I, Pelsy F, Primikirios N, Risovannaya V, Roubelakis-Angelakis
391 KA, Snoussi H, Sotiri P, Tamhankar S, This P, Troshin L, Malpica JM, Lefort F,
392 Martinez-Zapater JM. 2006. Multiple origins of cultivated grapevine (*Vitis vinifera* L.
393 ssp. *sativa*) based on chloroplast DNA polymorphisms. *Molecular Ecology* 15:3707-
394 3714.

395 Avendaño-Arrazate CH, Ogata-Aguilar N, Gallardo-Méndez RA, Mendoza-López A, Aguirre-
396 Medina JF, Sandoval-Esquivez A. 2010. *Cacao Diversidad en México*. Publicación
397 Especial No. 1. Instituto de Investigaciones Forestales, Agrícolas y Pecuarias. Centro de
398 Investigación Pacífico Sur. Campo Experimental Rosario Izapa. Tuxtla Chico, Chiapas,
399 México, 86 pp.

400 Avise CJ. 2000. *Phylogeography: the history and formation of species*. Harvard University
401 Press. Cambridge, Massachusetts, Londres. 228 pp.

402 Azuma H, García-Franco JG, Rico-Gray V, Thien LB. 2001. Molecular phylogeny of
403 themagnoliaceae: the biogeography of tropical and temperate disjunctions. *American*
404 *Journal of Botany* 88(12): 2275–2285.

405 Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific
406 phylogenies. *Molecular Biology and Evolution* 16(1):37-48.

407 Bieniek W, Mizianty M, Szklarczyk M. 2015. Sequence variation at the three chloroplast loci
408 (*matK*, *rbcL*, *trnH-psbA*) in the Triticeae tribe (Poaceae): comments on the relationships

409 and utility in DNA barcoding of selected species. *Plant Systematics and Evolution*
410 301:1275–1286.
411 Chen T, Wang X-R, Tang H-R, Chen Q, Huang X-J, Chen J. 2013. Genetic diversity and
412 population structure of Chinese cherry revealed by chloroplast DNA *trn* Q-*rps* 16
413 intergenic spacers variation. *Genetic Resources and Crop Evolution* 60(6):1859-1871.
414 Cheesman E. 1944. Notes on the nomenclature, classification and possible relationships of cacao
415 populations. *Tropical Agriculture* 21:144-159.
416 Crandall KA, Templeton AR. 1993. Empirical test of some predictions from coalescent theory
417 with applications to intraspecific phylogeny reconstruction. *Genetics* 134(3):959-969.
418 De la Cruz M, Whitkus R, Gómez-Pompa A, Mota-Bravo L. 1995. Origins of cacao cultivation.
419 *Nature* 375:542-543.
420 De Sahagún B. 2009. *Historia general de las cosas de la Nueva España II*. Editorial Dastin
421 Export, México. Libro tercero, cap. III y XII.
422 Doyle JJ, Doyle JL. 1990. A rapid total DNA preparation procedure for fresh plant tissue. *Focus*
423 12:13-15.
424 Dupanloup I, Schneider S, Excoffier LG. 2002. A simulated annealing approach to define the
425 genetic structure of populations. *Molecular Ecology* 11: 2571-2581.
426 Excoffier L, Laval G, Schneider S. 2005. Arlequin Ver. 3.0: an integrated software package for
427 population genetics data analysis. *Evolutionary Bioinformatics Online* 1:47-50.
428 Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis
429 program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98.

Hamilton MB, Braverman JM, Soria-Hernanz DF. 2003. Patterns and relative rates of nucleotide and insertion/deletion evolution at six chloroplast intergenic regions in new world species of the Lecythidaceae. *Molecular Biolology and Evolution* 20(10):1710–1721.

Hedrick PW. 2005. *Genetics of populations*. Jones and Bartlett Publishers, Sudbury, MA, EUA. 737 pp.

Ingvarsson PK, Ribstein S, Taylor DR. 2003. Molecular evolution of insertions and deletion in the chloroplast genome of *Silene*. *Molecular Biology and Evolution* 20(11):1737-1740.

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. *Molecular Biolology and Evolution* 28(1):835–847.

Kane N, Sveinsson S, Dempewolf H, Yang JY, Zhang D, M Engels JM, Cronk Q. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* 99(2):320-329.

Kress WJ, Erickson DL. 2007. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *PloS ONE* 2:e508.

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, Lachenaud P, da Silva e Mota JW, Loor R, Kuhn DN, Brown JS, Schnell RJ. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L). *PLoS ONE* 3(10):e3311.

Nei M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, Nueva York, pp 512.

452 Pérez-Jiménez M, Besnard G, Dorado G, Hernandez P. 2013. Varietal tracing of virgin olive oils
453 based on plastid DNA variation profiling. *PLoS ONE* 8(8):e70507.
454 Polzin T, Daneshmand SV. 2012. NETWORK 4.6.1.3 Fluxus Technology Ltd. All rights
455 reserved. Steiner (MP) algorithm.
456 Powis T, Cyphers A, Gaikwad N, Grivetti L, Cheong K. 2011. Cacao use and the San Lorenzo
457 Olmec. *Proceedings of the National Academy of Sciences of United States of America*
458 108(21):8595-8600.
459 Raymúndez MB, Mathez J, Xena de Enrech N, Dubuisson JY. 2002. Coding of insertion–
460 deletion events of the chloroplastic intergene *atp-rbcL* for the phylogeny of the
461 Valerianeae tribe (Valerianaceae). *Comptes Rendus Biologies* 325:131–139.
462 Roger RA. 1995. Genetic evidence for Pleistocene population explosion. *Evolution* 49(4):608-
463 615.
464 Rozas J, Librado P, Sánchez-Del Barrio JC, Messeguer X, Rozas R. 2010. DNA Sequence
465 Polymorphism, Ver. 5.10.1 Universidad de Barcelona. http://www.ub.edu/dnasp/
466 (accessed 20 july 2015).
467 Shaw J, Small RL. 2005. Chloroplast DNA phylogeny and phylogeography of the North
468 American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). *American*
469 *Journal of Botany* 92:2011–2030.
470 Shaw J, Lickey EB, Edward E, Schilling EE, Small RL. 2007. Comparison of whole chloroplast
471 genome sequences to choose noncoding regions for phylogenetic studies in angiosperms:
472 the tortoise and the hare III. *American Journal of Botany* 94(3):275–288. 2007.

473 Sun XQ, Zhu YJ, Guo JL, Peng B, Bai MM, Hang YY. 2012. DNA Barcoding the dioscorea in
474 china, a vital group in the evolution of monocotyledon: use of *matK* gene for species
475 discrimination. *PLoS ONE* 7(2):e32057.
476 Slatkin M, RR Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable
477 and exponentially growing populations. *Genetics* 129:555-562.
478 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
479 polymorphism. *Genetics* 123:585-595.
480 Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensivity of
481 progressive multiple sequence alignment through sequence weighting, position-specific
482 gap penalties and weight matrix choise. *Nucleic Acids Research* 22:4673-4680.
483 Vázquez-Ovando JA, Molina-Freaner F, Nuñez-Farfán J, Ovando-Medina I, Salvador-Figueroa
484 M. 2014. Genetic identification of *Theobroma cacao* L. trees with high Criollo ancestry
485 in Soconusco, Chiapas, Mexico. *Genetic and Molecular Research* 13(4):10404-14.
486 Wang FY, Gong X, Hu CM, Hap G. 2008. Phytogeography of an alpine species *Primula*
487 *secundiflora* inferred from the chloroplast DNA sequence variation. *Journal of*
488 *Systematics and Evolution* 46:13-22.
489 Whitkus R, de la Cruz M, Mota-Bravo L, Gómez-Pompa A. 1998. Genetic diversity and
490 relationships of cacao (*Theobroma cacao* L.) in southern Mexico. *Theoretical and*
491 *Applied Genetics* 96(1-2):621-627.
492 Whitlock BA, Hale AM, Groff PA. 2010. Intraspecific inversions pose a challenge for the *trnH-*
493 *psbA* plant DNA barcode. *PLoS ONE* 5(7):e11533.

494 Wolfe AD, Randle CP. 2004. Recombination, heteroplasmy, haplotype polymorphism, and
495 paralogy in plastid genes: implications for plant molecular systematics. *Systematic*
496 *Botany* 29:1011–1020.
497 Wood GAR. 2001. Consumption and manufacture. In: Wood GAR, Lass RA (eds). *Cocoa*. 4th
498 ed. Blackwell Science Ltd. Oxford, UK pp. 587-597.
499 Yamane K, Yano K, Kawahara T. 2006. Pattern and rate of indel evolution inferred from whole
500 chloroplast intergenic regions in sugarcane, maize and rice. *DNA Research* 13:197-204.
501 Yang JY, Scascitelli M, Motilal LA, Sveinsson S, Engels JMM, Kane NC, Dempewolf H, Zhang
502 D, Maharaj K, Cronk QCB. 2013. Complex origin of Trinitario-type *Theobroma cacao*
503 (Malvaceae) from Trinidad and Tobago revealed using plastid genomics. *Tree Genetics &*
504 *Genomes* 9(3):829-840.
505 Yang JY, Motilal LA, Dempewolf H, Maharaj K, Cronk QC. 2011. Chloroplast microsatellite
506 primers for cacao (*Theobroma cacao*). *American Journal of Botany* 98(12):e372–e374.
507 Zeng J, Fan X, Sha LN, Kang HY, Zhang HQ, Liu J, Wang XL, Yang RW, Zhou YH. 2012.
508 Nucleotide polymorphism pattern and multiple maternal origin in *Thinopyrum*
509 *intermedium* inferred by *trnH-psbA* sequences. *Biologia Plantarum* 56(2):254-260.

1

Location of indels (blue arrows) and the transversion (red arrow) in a sequenced fragment of the chloroplast DNA *trnH-psbA* intergenic spacer from *Theobroma cacao* trees. See Table 1 for sample details.

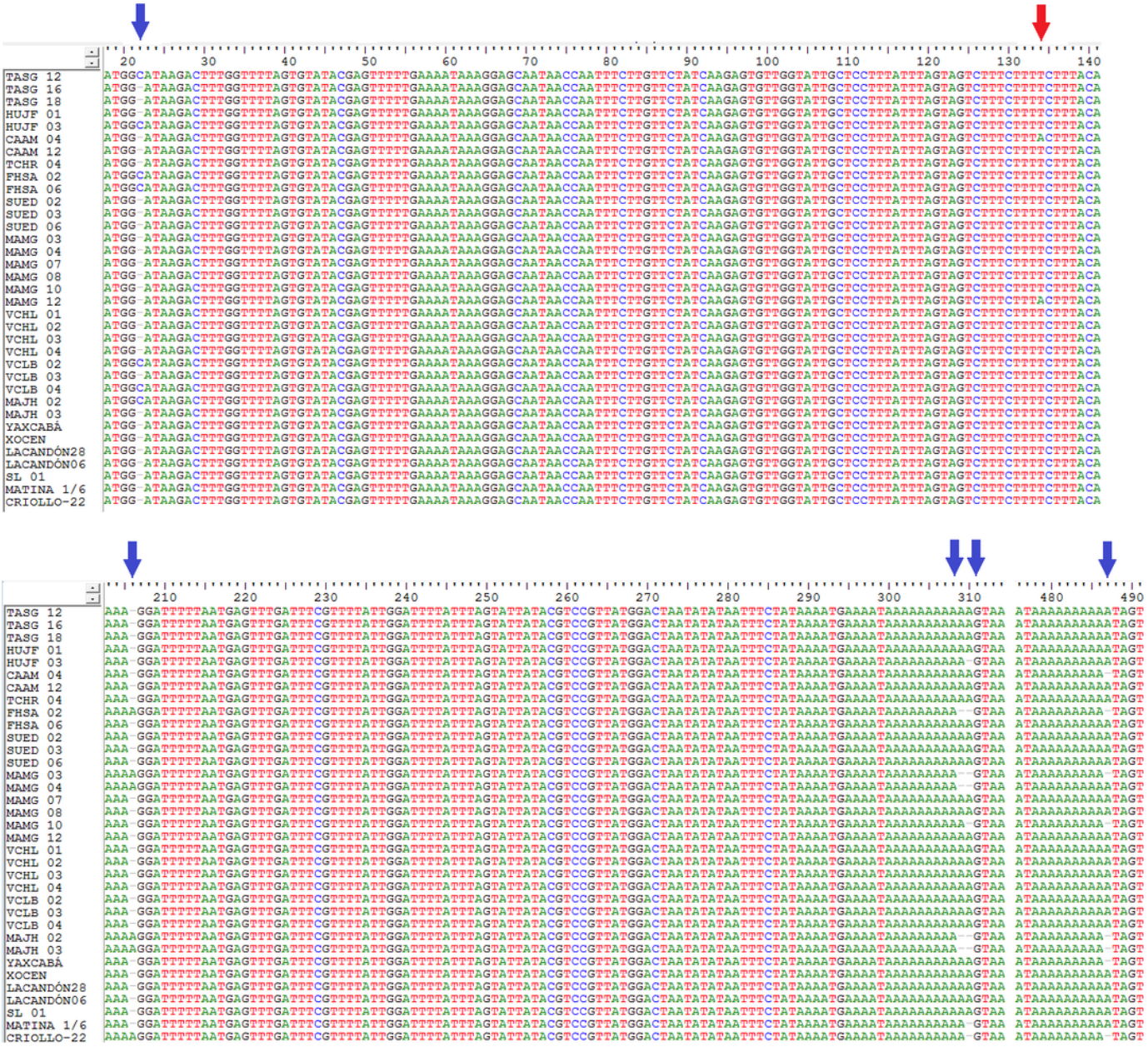
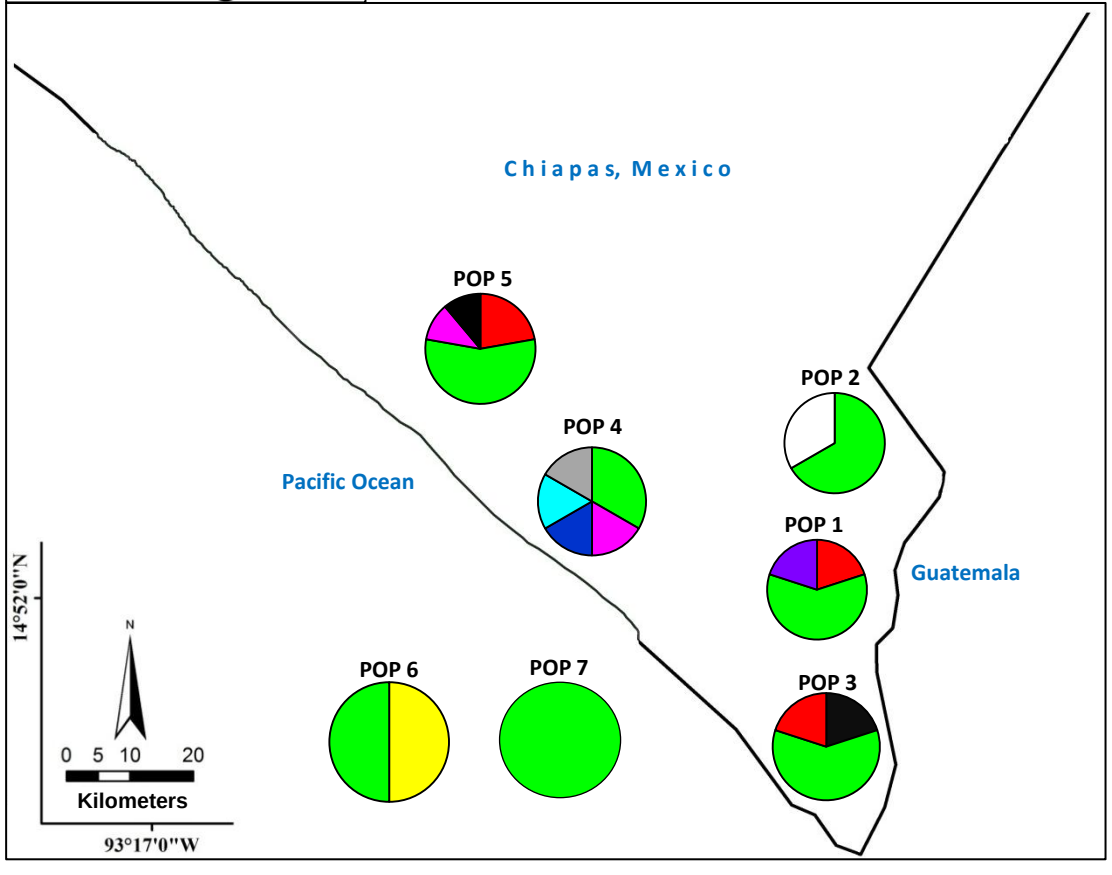


Figure 2(on next page)

Median joining network for chloroplast DNA *trnH-psbA* intergenic spacer haplotypes of *Theobroma cacao* trees from Soconusco, Mexico, and the outgroup haplotype (*Theobroma bicolor*).

The map indicates the geographic distribution of the haplotypes. The colored portions represent the proportions of the same haplotype occurring in each sampling locality. Trees employed as the references (Pop 6 and Pop 7) are shown outside the map. The population code and details are shown in Table 1.



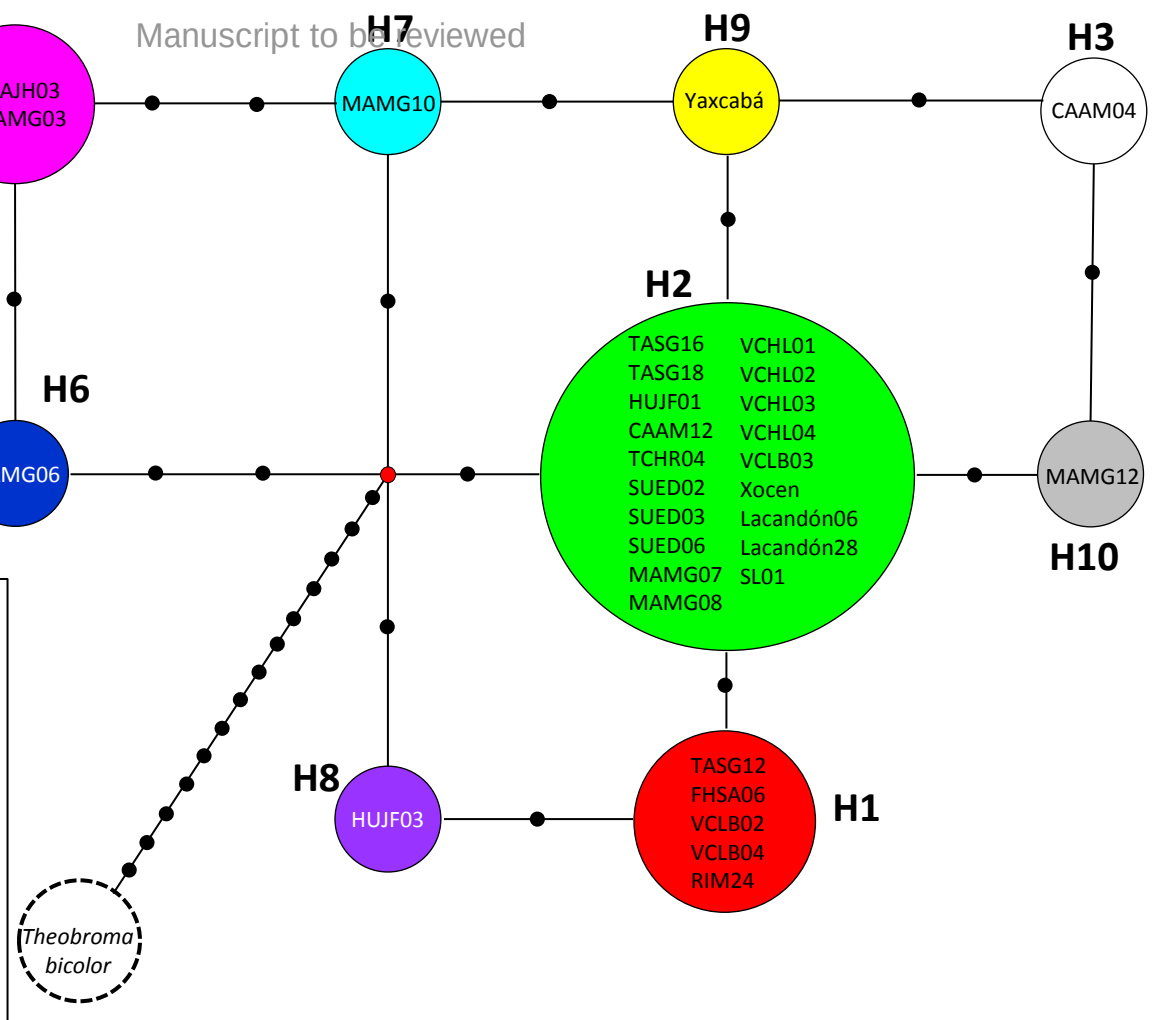


Figure 3(on next page)

Median joining network for the chloroplast DNA *trnH-psbA* intergenic spacer haplotypes of *Theobroma cacao* trees cultivated in Soconusco, Mexico, and the reference accessions.

The circle sizes are proportional to the haplotype frequencies, and the color represents the proportions of the same haplotype occurring in each genetic group. For genetic group details see Table 1.

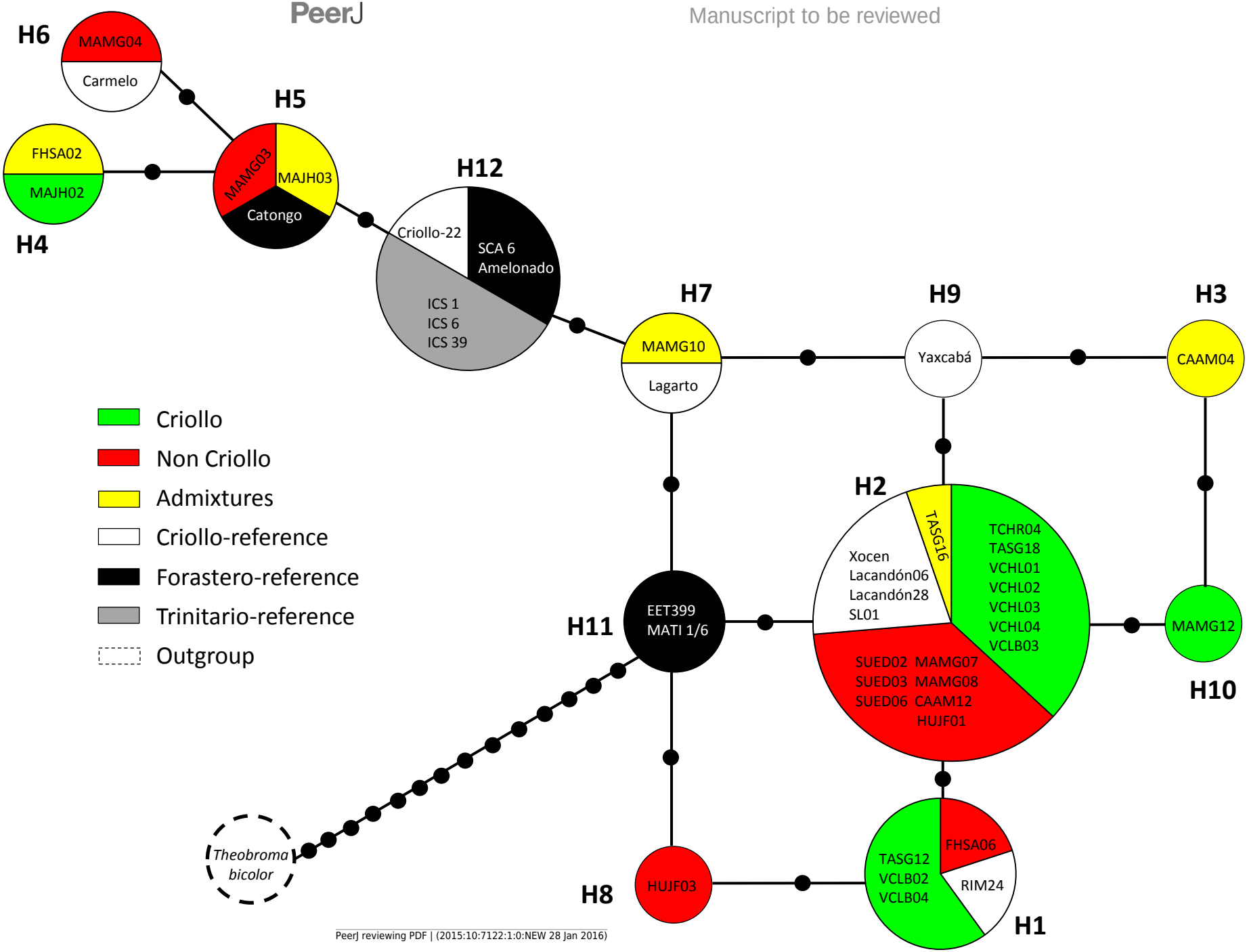

H6
MAMG04
Carmelo
H5
FHSA02
MAJH02
H4
MAMG03
MAJH03
Catongo
H12
Criollo-22
SCA 6
Amelonado
ICS 1
ICS 6
ICS 39
H7
MAMG10
Lagarto
H9
Yaxcabá
H3
CAAM04
Criolo
Non Criollo
Admixtures
Criollo-reference
Forastero-reference
Trinitario-reference
Outgroup
H2
TASG16
Xocen
Lacandón06
Lacandón28
SL01
TCHR04
TASG18
VCHL01
VCHL02
VCHL03
VCHL04
VCLB03
SUED02 MAMG07
SUED03 MAMG08
SUED06 CAAM12
HUJF01
H11
EET399
MATI 1/6
MAMG12
H10
Theobroma bicolor
H8
HUJF03
TASG12
VCLB02
VCLB04
FHSA06
RIM24
H1

Table 1(on next page)

*A priori** population grouping and genetic classification of the analyzed *Theobroma cacao* trees.

For populations 1-5 (from farms in Soconusco, Mexico) genetic clustering was based on membership to the Criollo group (%) described by Vázquez Ovando *et al*. (2014) using SSR markers. For the reference trees** (populations 6-9), the genetic grouping was suggested by Avendaño-Arrazate *et al*. (2010) and the database accessions (ICGD).

Pop*	Coordinates latitude (N)/ longitude (W)	Criollo (n=20)	Non Criollo (n=16)	Admixtures (n=9)
1	14°59’28’’N, 92°26’44’’W (Huehuetán) 14°52’55’’N, 92°21’42’’W (Tapachula)	TASG12 (93%) TASG18 (95%)	HUJF01 (9%) HUJF03 (2%)	TASG16 (86%)
2	14°56’41’’N, 92°09’59’’W (Tuxtla Chico) 14°59’53’’N, 92°10’44’’W (Cacahotán)	TCHR04 (98%)	CAAM12 (1%)	CAAM04 (53%)
3	14°47’31’’N, 92°11’11’’W (Frontera Hidalgo) 14°38’27’’N, 92°13’47’’W (Suchiate)		FHSA06 (1%) SUED02 (2%) SUED03 (1%) SUED06 (1%)	FHSA02 (36%)
4	14°48’56’’N, 92°29’06’’W (Mazatán)	MAMG12 (98%)	MAMG03 (2%) MAMG04 (1%) MAMG07 (1%) MAMG08 (9%)	MAMG10 (24%)
5	15°28’07’’N, 92°48’42’’W (Mapastepec) 15°10’31’’N, 92°38’06’’W (Villa Comaltitlán) 15°11’17’’N, 92°36’55’’W (Villa Comaltitlán)	MAJH02 (96%) VCHL01 (97%) VCHL02 (96%) VCHL03 (97%) VCHL04 (97%) VCLB02 (97%) VCLB03 (98%) VCLB04 (98%)		MAJH03 (63%)
6**	20°32’29.25’’N, 88°50’35.82’’W (Yucatán)	Yaxcabá Xocen		
7**	16°06’42.92’’N, 90°56’31.28’’W (Selva Lacandona)	Lacandón 06 Lacandón 28 SL01		
8**	INIFAP (Several)	Lagarto Carmelo	Catongo EET 399	RIM 24
9**	Accessions (ICGD)	Criollo 22	SCA 6 (MIA 29885) Amelonado (TARS 16542) MATI 1/6	ICS 1 (TARS 16656) ICS 6 (TARS 16658) ICS 39 (TARS 16664)
ICGD= International Cocoa Germplasm Database; TARS= Tropical Agriculture Research Station; INIFAP= Instituto de Investigaciones Agrícolas y Pecuarias; RIM= Rosario Izapa Mexico.				

1

Table 2(on next page)

Nucleotide polymorphic sites and cpDNA haplotypes in cacao populations based on variation in the intergenic *trnH-psbA* spacer region.

1

Haplotype	Polymorphic site						Populations (for details see Table 1)								
	22	134	206	309	310	487	Pop1	Pop2	Pop3	Pop4	Pop5	Pop6	Pop7	Pop8	Pop9
H1	C	T	-	A	A	A	1		1		2			1	
H2	-	T	-	A	A	A	3	2	3	2	5	1	3		
H3	-	A	-	A	A	-		1							
H4	C	T	A	-	-	-			1		1				
H5	-	T	A	-	-	-				1	1			1	
H6	-	T	A	-	-	A				1				1	
H7	-	T	-	A	-	-				1				1	
H8	C	T	-	A	-	A	1								
H9	-	T	-	A	A	-						1			
H10	-	A	-	A	A	A				1					
H11	-	T	-	A	-	A								1	1
H12	-	T	A	A	-	-									6

2

Table 3(on next page)

Genetic diversity in cacaos from Soconusco (Chiapas, Mexico) grouped by the geographic approach (Pop) and genetic origin approach.

1

Pop	Locality	N	S	Sn	H	Hd ± sd	πd ± sd
1	Huehuetán,	5	2	1	3	0.70±0.21	0.0019±0.0017
2	Cacahoatán, Tuxtla Chico	3	2	1	2	0.67±0.31	0.0026±0.0026
3	Frontera Hidalgo, Suchiate	5	5	0	3	0.70±0.21	0.0042±0.0032
4	Mazatán	6	5	3	5	0.93±0.12	0.0048±0.0035
5	Mapastepec, Villa Comaltitlán	9	5	0	4	0.69±0.14	0.0039±0.0027
6	Yucatán	2	1	1	2	1.00±0.50	0.0019±0.0027
7	Selva Lacandona	3	0	0	1	0	0
Total		33	--	6	--	0.66±0.08	0.0032±0.0021
Genetic origin approach*							
“Criollo”		12	6	1	4	0.64±0.13	0.0025±0.0019
“Non Criollo”		11	5	1	5	0.62±0.16	0.0030±0.0021
“Admixtures”		5	5	1	5	1.00±0.12	0.0060±0.0041
Criollo-reference[a]		8	4	1	5	0.79±0.15	0.0033±0.0025
Forastero-reference[a]		5	3	0	3	0.80±0.16	0.0031±0.0025
Trinitario-reference[a]		4	4	0	2	0.50±0.27	0.0038±0.0032
Total		45	--	4	--	0.80±0.05	0.0038±0.0024

N=Samples sizes, S=Number of segregating, Sn=Singletons, H=Number of haplotypes, Hd=Haplotype diversity, πd=Nucleotide diversity. sd=standard deviation. [a]Including sequences GenBank (Criollo-reference n=1, Forastero-reference n=3, Trinitario-reference n=3). *Classification based on membership (>90%) to Criollo type, see Table 1 (Vázquez-Ovando *et al*. 2014).

2

Table 4(on next page)

Spatial analysis of molecular variance (K = 2) for cacao populations and the statistical analysis of molecular variance fixation indices corresponding to the groups.

1

Source of variation	df	SS	VC	Variation (%)	Fixation indices	*P* value
Among groups	1	1.61	0.1282	13.98	F_{SC}= -0.1115	0.7341
Among populations within groups	5	2.51	-0.0879	-9.59	F_{ST}= 0.0439	0.0068
Within populations	26	22.80	0.8765	95.61	F_{CT}= 0.1398	0.1496
Total	32	26.91	0.9168			
df= degrees of freedom, SS=Sum of squares, VC=Variance components.						

2

3