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 Hugo Avendaño-Arrazate, Alfredo Vázquez-Ovando

About 4000 years ago cacao trees were domesticated in Mesoamerica and are still grown. In this study we analyzed the sequence variation of chloroplast DNA *trnH-psbA* intergenic spacer in 28 cacao trees from different farms in the Soconusco region in southern Mexico. Genetic relationships were established by two analysis approaches, based on geographical origin (five populations) and genetic origin (based on a previous study). In our results we identified six polymorphic sites, where five insertion / deletion (indels) type and one substitution were detected. We also found that the overall nucleotide diversity was low for both approaches (geographic = 0.003197; genetic = 0.003841), conversely was obtained moderate to high haplotype diversity (0.6610 and 0.7949), with ten and 12 haplotypes, respectively. The common haplotype (H2) for both networks involved cacao trees of 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-purifying from expansion of narrow populations. No genetic differentiation (AMOVA, $F_{ST} = 0$) was found, and the F_{ST} value (0.04339) of SAMOVA was not big enough to show moderate differentiation between populations. Only one population showed high frequency of haplotypes, thus it could be considered as an important reservoir of genetic material. The indels located in the intergenic spacer *trnH-psbA* of cacao trees could be useful for markers of DNA Barcoding development.

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

About 4000 years ago cacao trees were domesticated in Mesoamerica and are still grown. In this study we analyzed the sequence variation of chloroplast DNA *trnH-psbA* intergenic spacer in 28 cacao trees from different farms in the Soconusco region in southern Mexico. Genetic relationships were established by two analysis approaches, based on geographical origin (five populations) and genetic origin (based on a previous study). In our results we identified six polymorphic sites, where five insertion / deletion (indels) type and one substitution were detected. We also found that the overall nucleotide diversity was low for both approaches (geographic = 0.003197; genetic = 0.003841), conversely was obtained moderate to high haplotype diversity (0.6610 and 0.7949), with ten and 12 haplotypes, respectively. The common haplotype (H2) for both networks involved cacao trees of 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-purifying from expansion of narrow populations. No genetic differentiation (AMOVA, F_{ST} = 0) was found, and the F_{ST} value (0.04339) of SAMOVA was not big enough to show moderate differentiation between populations. Only one population showed high frequency of haplotypes, thus it could be considered as an important reservoir of genetic material. The indels located in the intergenic spacer *trnH-psbA* of cacao trees could be useful for markers of DNA barcoding development.

Keywords

Chloroplast DNA, Haplotype, Nucleotide diversity, indels, *trnH-psbA*

47

48 **Introduction**

49 The Neotropical cacao tree (*Theobroma cacao* L.), is widely cultivated in America, Africa and
50 Asia, it is considered an economical important crop because its seeds are used in the chocolate
51 industry (Wood & Lass, 1985). Based on morphological traits and the geographical origin, trees
52 are classified as; Criollo, Forastero and Trinitario (Cheesman, 1944). In Mesoamerica, the
53 Criollo cacao is being widely used as food for nearly 4000 years ago (De la Cruz *et al*., 1995;
54 Whitkus *et al*., 1998; Powis *et al*., 2011).
55 Based on simple sequence repeat (SSR) analysis, Motamayor *et al*. (2008) propose ten genetic
56 groups. But nowadays Criollo retains the identity as a separate group, while the other proposed
57 genetic groups comprising all trees from South America. In this region, it has reported the
58 highest genetic diversity of cacao trees.
59 On the other hand, genetic diversity of cacao in the South of Mexico was registered as moderate
60 to low in natural populations (Whitkus *et al*., 1998; by using RAPD markers), and cultivated
61 forms (Vazquez-Ovando *et al*., 2014 , by using microsatellite markers), but a wide diversity in
62 cacao pod and seed morphology was observed. In Soconusco farms (Chiapas, Mexico) Vazquez-
63 Ovando *et al*. (2014) found moderate to high allelic richness, however high levels of
64 homozygosity, they also reported the presence of trees sharing genetic identity with those
65 considered "Ancient Criollo" but also the presence of private alleles. These alleles may be
66 associated with commercial interest phenotypic traits, while preserving relation with other
67 polymorphic regions DNA.
68 The chloroplast DNA (cpDNA) and markers based on it, they are increasingly used for studies of
69 genetic population structure, evolution, gene flow, haplotype frequency and phylogenetic

relationships. Given its high conservation due to maternal uniparental inheritance, cpDNA is the main source of data for construction of phylogenetic relationships in plants (Shaw & Small, 2005). However, there are DNA regions that have variability, 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 further insight of the centre of origin of cultivated grape populations in Europe (Arroyo-Garcia *et al*., 2006) and to explain the diversity and structure population of cultivated Chinese cherry (Chen *et al*., 2013).

The cpDNA intergenic spacer mostly used is *trnH-psbA,* which has showed high variability and besides useful to elucidate genetic relationships at the intraspecific level (Azuma *et al*., 2001; Hamilton, Braverman & Soria-Hernanz, 2003). The *trnH-psbA* sequenced region of ten cacao accessions deposited in the NCBI database, expose only one haplotype (Kane *et al*., 2012), while Jansen *et al*. (2011) showed the presence of polymorphic sites, which set 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 report 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 by some authors (Bieniek, Mizianty & Szklarczyk, 2015; Whitlock, Hale & Groff, 2010), because the mechanism by indels are generated remains unclear. However, indels are informative character states, since genetic variability analyzed using polymorphism due to indels or substitutions can be studied without distinction (Nei, 1987) therefore they are used as markers. Moreover, when included in diversity or phylogenetic analysis the discriminant power between species is

93 enhanced (Raymúndez *et al*., 2002; Hamilton, Braverman & Soria-Hernanz, 2003; Kress &
94 Erickson, 2007; Sun *et al*., 2012) and even between conspecific individuals (Pérez-Jiménez *et al*.,
95 2013). Therefore, the aim of this study was to evaluate the genetic diversity and to describe the
96 relationship between individuals of *Theobroma cacao* L. Criollo type of the Soconusco region
97 (Chiapas, Mexico) by using the variation of chloroplast DNA revealed by *trnH-psbA* spacer
98 sequence.
99

100 **Material & methods**

101 Plan material and sample collection

102 A total of 45 samples cacao were including in this study. 38 trees were sequencing and analyzed,
103 and seven sequences accessions were taken from GenBank as references. From 38 trees
104 sequenced, 28 were selected from plantations of Soconusco (Chiapas, Mexico) based on a
105 previous characterization (Vazquez-Ovando *et al*., 2014) using ten SSR´s molecular markers.
106 The individuals were selected based on traits of fruits (pod) and seeds, which resembled those of
107 the Criollo variety. Pods were elongated, deeply grooved, pointed at the pod end, had a lumpy
108 surface with a warty appearance outside, white or slightly pigmented seeds, and sweet mucilage.
109 In agree with that reported by Vazquez-Ovando *et al*. (2014) the individuals were classified as:
110 12 trees with high Criollo ancestry, ten non Criollo group (hereinafter called Forastero) and six
111 admixtures (hereinafter called hybrid) (Table 1). Additionally, ten accessions were sequenced
112 and included as references: two Forastero variety (Catongo and EET 399), one Trinitario variety
113 (RIM 24) and seven wild Criollo (one collected in the Lacandon rainforest [SL01] and six
114 obtained from germplasm of the Instituto Nacional de Investigaciones Forestales, Agrícolas y
115 Pecuarias, México [Yaxcabá, Xocen, Lacandón 06 Lacandón 28, Lagarto and Carmelo]) (Table

1). *Theobroma bicolor* was used as outgroup. The average age of the trees was 30 year. Leaves were sampled and placed in plastic bags, taken to the laboratory (4°C) and they stored at -20 °C until processing.

DNA extraction, amplification and sequencing

Total DNA extraction was performed by modifying the method described by Doyle & Doyle (1990). Leaves were washed with sterile water and 70% ethyl alcohol. Approximately 200 mg cacao leaves were ground with liquid nitrogen with 60 mg polyvinyl pyrrolidone and 1 mL 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 precipitation 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 Milli-Q water) was checked by 0.8% agarose electrophoresis and quantified by spectrophotometry at 260 nm (Jenway, Genova), and the purity was inferred by the 260/280 and 260/230 absorbance ratios.

The cpDNA amplification of trnH-psbA intergenic spacer was conducted by using forward primer 5'-CGCGCATGGTGGATTCACAATCC-3' and reverse primer 5'-GTTATGCATGAACGTAATGCTC-3' (Shaw & Small, 2005). PCR conditions was performed following to described by Shaw & Small (2005) with changes in the concentration of MgCl2 (2mM) and using the average value of the temperature melting reported. PCR was performed in a 25 µL reaction mixture, containing 100 ng genomic DNA, 4 µL of 10x PCR buffer ViBuffer A (VivantisTM), 1 µL of MgCl2 (50 mM), 0.5 µL of dNTP Mix (10mM, Promega), 0.05 mM of each primer and 2.5 U of Taq DNA polymerase (VivantisTM). Following one cycle of 5 min at

94 °C, 35 PCR cycles of 30 s at 94 °C, 30 s at 53 °C, 1 min at 72 °C and 10 min at 72 °C final extension were performed in a thermal cycler TC3000 (Techne, Cambridge, UK). 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 the Image Lab (v. 4.0.1, Bio-Rad Laboratories) and integrating GeneRuler™ 100 bp DNA Ladder Plus (Fermentas®) as a molecular weight marker.

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

Sequence alignment and data analysis

The sequence quality was checked and electropherograms edited by using BioEdit © (Hall, 1999). Sequences were limited at the ends to avoid the presence of variable sites due to artifacts sequencing by polymerase (approx. 40 bp) and aligned with ClustalW 1.81 (Thompson, Higgins & Gibson, 1994). Visual inspection and manual edition of sequences for confirming variable sites was performed. We used two different analytical approaches, based on the geographic origin and the genetic origin of the samples (Table 1). In both approaches molecular diversity indices; the number of segregating sites (S), number of haplotypes, haplotype diversity (Hd) and nucleotide diversity (πd) were estimated following Nei (1987) in DnaSP© 5.1 (Rozas *et al*., 2010).

159 In order to infer evolutionary relationships at the intraspecific level, we evaluated network
160 building. The method used was median-joining (MD) based on parsimony criteria (Bandelt,
161 Forster & Röhl, 1999; Polzin & Daneshmand, 2012), which performed with the software
162 Network© 4.6.1.3.

163 Analysis of molecular variance (AMOVA), pairwise *Fst* values as well as the statistics of
164 molecular variances F_{CT} (test by permuting individuals within populations), F_{ST} (test by
165 permuting genotypes among populations but within groups) and F_{SC} (test by permuting
166 genotypes among groups) were estimated using the Arlequin© version 3.0 (Excoffier, Laval &
167 Schneider, 2005). Significance was evaluated by 99 999 random permutations of sequences. In
168 order to determine whether sample sites clustered on a population level, a spatial analysis of
169 variance (SAMOVA) was conducted (Dupanloup, Schneider & Excoffier, 2002), using
170 haplotype data and geographic co-ordinates of each of the 5 sample sites. The SAMOVA was
171 run for K = 2–5 putative populations to determine the maximum F_{ST} value, the highest
172 proportion of differences between populations due to genetic variation.

173 The neutral evolution of chloroplast DNA was evaluated to examine whether any population had
174 experienced historical demographic changes using test Tajima´s D (Tajima, 1989) by using
175 Arlequin© version 3.0 (Excoffier, Laval & Schneider, 2005). We evaluated for geographical
176 approach, overall as well as populations.

177 When the analysis was conducted for genetic origin approach, seven accessions from NCBI
178 database as reference were included: Matina-06 (HQ336404.2), Criollo-22 (JQ228379.1)
179 Amelonado (JQ228380.1) Scavina (JQ228382.1), ICS-01 (JQ228381.1), ICS-06 (JQ228383.1),
180 ICS-39 (JQ228387.1).

181

182 **Results**

183 Sequence characterization and genetic diversity

184 The sequences of *trnH-psbA* intergenic spacer in 45 samples *Theobroma cacao* (Table 1) were
185 aligned with a consensus length of 526 bp, of which six polymorphic sites (Table 2), including a
186 substitution of T (A) at position 134, five indels (Figure 1) and six segregating sites were
187 detected. This resulted in 12 haplotypes, of which four were singletons represented by a unique
188 sequence in the sample (Table 2). The nucleotide composition of the fragment revealed AT-rich
189 (A+T, 75.52%).

190 The geographic approach analysis revealed overall the average values of haplotype diversity
191 (Hd) and nucleotide diversity (μd) were 0.6610 and 0.0031, respectively (Table 3). 10 haplotypes
192 identified, the most frequent haplotype (H2) was shared by 19 trees of the seven geographic
193 populations formed *a priori* (Table 2). Four trees that belong to Population 1 (one tree),
194 Population 3 (one tree) and Population 5 (two trees) were the second most common haplotype
195 (H1). Overall 50% haplotypes were singletons (Table 2). The analysis showed that most genetic
196 diversity was found in Population 4 (Mazatán), with the highest values for all indices; also it
197 included 50% of the haplotypes identified. The other populations maintained moderate Hd and
198 low πd with similar values each population (Table 3). Yucatán and Selva Lacandona populations
199 (wild) they exhibited Hd 1 and 0 respectively, although these data are influenced by the low
200 number reference individuals.

201 Meanwhile when the data analysis was based on the genetic origin, the highest Hd (0.93) was
202 found in the hybrid group (Table 3). In contrast, the Trinitario-reference group had the lowest

value Hd (0.5). The πd was low (0.0025 to 0005) for all groups, similar to another approach. Forastero-reference and Trinitario-reference groups did not present singletons (Table 3). Sequences from the NCBI database were grouped into one haplotype (H12), except HQ336404.2 it grouped in H11 whit EET399, that corresponding to Forastero-reference group.

Intraspecific relationship haplotype

Figure 2 shows the haplotype network built with data from geographic approach (a) and genetic approach (b). Both networks show a star arrangement. The general base has a common haplotype for the two networks (H2) that included cacao trees from all geographic populations (a) and four of six genetic groups (b). From this common haplotype (H2) derive a unique set of intermediate haplotypes and interconnected by one or two mutational steps, in both networks. H4-H6 haplotypes were farthest from the central clade, i.e. haplotypes newly created (Figure 2). Haplotypes H3 and H8-H10 were singletons.

Population genetic structure

The analysis of molecular variance (AMOVA) was not significant and with a value *Fst*=0; while in the spatial analysis of molecular variance (SAMOVA) it was found that the value of *K*=2 extends the *Fst* to 0.04339 (Table 4), generating two clusters; the first contained only the Population 4 (Mazatán) and the second cluster grouped the other geographic populations (Table 4).

Neutrality tests showed non-significant value in the Tajima´s D, except for the 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).

224

225 **Discussion**

226 In this study was found high haplotype variation in chloroplast DNA cacao trees grown in the
227 Soconusco region. No found inversions, transitions or transversions which reported as commons
228 in other plants (Whitlock, Hale & Groff, 2010; Zeng *et al*., 2012). However, we found
229 polymorphism, type insertions or deletions (indels) in three poly-A regions (Figure 1). This
230 agrees with that reported by Jansen *et al*. (2011), in HQ336404.2 accession and support the
231 affirmation Aldrich *et al*. (1988) that indels are a presumably common feature in the region *trnH-*
232 *psbA*. For the data analysis, we included the indels as informative character states, being as the
233 high interspecific divergence of the region spacer allow even be used as a marker of DNA
234 Barcoding (Kress & Erickson, 2007). Molecular diversity indices found in the present study have
235 similarity to the results of Zeng *et al*. (2012) using the same intergenic spacer, which revealed 11
236 haplotypes for 35 samples of *Thinopyrum intermedium*, low nucleotide diversity ($\pi d = 0.00473$)
237 and moderately high haplotype diversity (Hd = 0.7331) (our results for geographic populations
238 were $\pi d = 0.003197$, Hd = 0.6610). The results of these authors further support the use of one
239 intergenic spacer to reveal nucleotide polymorphism.

240 Our results of the haplotype diversity are contrary, those reported by Vazquez-Ovando *et al*.
241 (2014) who reported low genetic diversity when the study conducted with individuals of the
242 same region (in particular the Population 4 Mazatán) using microsatellite markers. However, the
243 low nucleotide diversity found in this study is supported by the low genetic variability found
244 with microsatellites. Individuals included in both studies showed great phenotypic pod variability
245 resemble to Criollo-type (eg. different degrees of roughness, color, deeply grooved). This could

246 be revealing greater association between morphological variability cacao pod with reported
247 allelic richness (Vazquez-Ovando *et al*., 2014) and variability of haplotypes found in our study.
248 The number of haplotypes was higher than polymorphic sites (Table 2) this feature is associated
249 with ancestral species that have diverged enough, accumulating mutations among different
250 haplotypes (Roger, 1995). Population 7 (Selva Lacandona) exhibited no haplotype diversity (Hd
251 = 0), however the haplotype (H2) located in this population is considered the common ancestor
252 due to it share all populations (Figure 2A). On the contrary, the two individuals belong the
253 Population 6 (Yucatán), which exhibited each other different haplotypes (H2 and H9),
254 interrelated by only a mutational step (Figure 2A). This shows that in the trees belonging to
255 Yucatán population, an individual eventually descended from other of this region where the
256 Maya people grown cacao.

257 Low levels nucleotide polymorphism could be explained by rapid population expansion events in
258 its distribution range, whereas high haplotype diversity may be due to continuous introduction of
259 individuals from different locations. In populations recently introduced or expanded from a small
260 number of founders, would have a common haplotype shared by most individuals and many rare
261 haplotypes connected to the main by few independent mutations (Slatkin & Hudson, 1991;
262 Avise, 2000) as in the present study (Figure 2). A similar argument is explained when using
263 microsatellite markers (Vazquez-Ovando *et al*., 2014).

264 The relative low variability in cacao cultivated populations is also supported by the lack of
265 neutrality revealed by the global test Tajima. Specifically in the Population 4 (Mazatán) the
266 negative value of Tajima´s D (-0.93302), could be related to an event "bottleneck" which would
267 indicate population expansion, not natural because of it is cultivated populations. It is reported
268 that in the past occurred unclearly events (disease, volcanic eruptions or other natural events)

269 they may have caused the almost complete disappearance of populations established by the
270 people in the Mesoamerican region (De Sahagún, 2009, *Codex Florentino*). A process rapid
271 expansion due to recolonization in populations, probable introduction of other varieties of trees
272 cacao not native in the region, would subject the populations to event bottleneck in very recent
273 periods. However, these are presumptive weak to infer the population history by use a single
274 locus. The bottleneck event could also be related to the loss of alleles (haplotypes), mainly rare
275 alleles, which is much greater than the loss of genetic variance *per se*. Although these rare alleles
276 contribute little to the total genetic variability, can provide unique responses against challenges
277 evolutionary as found in this study a high number of unique haplotypes (3 haplotypes in
278 population 4). The presence of both common and rare haplotypes can be understood by a
279 directional-purifying process selection or expansion events from small populations (Hedrick,
280 2005). H3 and H6-H10 haplotypes (cultivated populations) are singletons, agree with Crandall &
281 Templeton (1993) the singletons located in this study were connected with haplotypes from the
282 same population. Population 4 (Mazatán), shows the highest haplotype diversity, which makes
283 this population an important reservoir of genetic material at the level of chloroplast, and possibly
284 phenotypic, as also it is the abundance of morphology in pods seen in this population.

285 Overall, cacao trees with high ancestry was located in the center of haplotype network, this
286 supported by coalescence theory that predicts the ancient haplotype should be the most common
287 and most distributed among populations. In concordance derived haplotypes would be less
288 frequent, and in many cases private; these would be located in regions for cacao cultivated
289 populations latest. H4 and H5 are haplotypes perhaps recently created because of they are
290 located at the ends of the network, which may be due to germplasm exchange with traits of
291 interest of cacao farmers. These anthropogenic activities perhaps had a strong impact on the

292 levels of variation observed in cpDNA sequences, which explains the observed no
293 differentiation. In addition, migration over long distances by the exchange by farmers
294 contributed to the colonization of new regions founded by few individuals, establishing different
295 alleles by mutation and genetic drift.

296 Related to the genetic origin MAJH02 and Carmelo individuals are located in the haplotypes 4
297 and 6 respectively (Figure 2B). They possibly belong to "hybrid" group rather than the Criollo.
298 But also they are contenders for Modern Criollo group i.e. individuals classified as Criollo but
299 which might have been introgressed with Forastero genes (Motamayor *et al*., 2002) and preserve
300 phenotypic traits of Criollo ancient.

301 Furthermore the value found for Fst = 0.00 by AMOVA reveals that all molecular variance is
302 within populations. Indeed the F_{ST} value of SAMOVA (Table 4) is not enough to show at least
303 moderate differentiation between populations ($Fst \geq 0.05$). This provides some explanations
304 regarding the demographic history of *T. cacao* trees, indicating that populations formed *a priori*
305 have experienced gene flow, which results in population homogenization. Spatial analysis
306 reveals highest differentiation between groups when K=2 is tested; meanwhile a K=3 (Fst =
307 0.00088) grouped trees of Yucatán, Selva and Cacahoatán in the same genetic population. This is
308 unusual; being that geographic distance is longer among the three localities and may be
309 associated with distribution of trees in the past, i.e. the ancestral haplotype (H2) grouped
310 individuals of Selva and that for one mutational step it originated the individuals of Yucatán,
311 which in turn originated at individuals of Cacahoatán by the same event (network haplotype by
312 genetic origin, Figure 2B). Following this criterion, the H3 has a greater correspondence with the
313 Criollo genotype, rather than it reported previously as hybrid (Vazquez-Ovando *et al*., 2014).

Conclusions

Indels located in the chloroplast DNA *trnH-psbA* spacer region of cacao trees could allow the development of genetic markers barcode. The molecular analysis of nucleotide diversity showed low diversity, but high haplotype diversity, which may be due to events bottlenecks populations, confirmed with negative Tajima's D and haplotype network in a star arrangement. It also allowed identifying ten different haplotypes (trees grown) of which H3 and H6-H10 resulted singletons because they are not associated with other cacao grown or with those reported in the molecular databases. The presence of these haplotypes, accompanied by the low number of mutational steps that groups might suggest a very short evolutionary history or events of disappearance-expanding populations of southern Mexico. A geographical population (Pop 4 Mazatán) was located high frequency haplotypes, which makes this zone an important reservoir of genetic material at the level of chloroplast, and possibly phenotypic, since they were also observed in this population abundance of morphology in pods. The genetic differentiation between populations was zero, by suggesting that gene flow homogenized populations.

Acknowledgements

To Nancy Gálvez- Reyes for his advice on data analysis and comments on the manuscript.

References

334 Aldrich J, Cherney BW, Merlin E, Christopherson L. 1988. The role of insertion/deletions in the
335 evolution of the intergenic region between *psbA* and *trnH* in the chloroplast genome.
336 Current Genetics 14:137-146.
337 Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López MA, Arnold C, et al. 2006. Multiple
338 origins of cultivated grapevine (*Vitis vinifera* L. ssp. sativa) based on chloroplast DNA
339 polymorphisms. Molecular Ecology 15:3707-3714.
340 Avise, C. J. 2000. Phylogeography: the history and formation of species. Harvard University
341 Press. Cambridge, Massachusetts, Londres. 228 pp.
342 Azuma H, García-Franco JG, Rico-Gray V, Thien LB. 2001. Molecular phylogeny of
343 themagnoliaceae: the biogeography of tropical and temperate disjunctions. American
344 Journal of Botany 88(12): 2275–2285.
345 Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific
346 phylogenies. Molecular Biology and Evolution 16 (1):37-48.
347 Bieniek W, Mizianty M, Szklarczyk M. 2015. Sequence variation at the three chloroplast loci
348 (*matK*, *rbcL*, *trnH-psbA*) in the Triticeae tribe (Poaceae): comments on the relationships
349 and utility in DNA barcoding of selected species. Plant Systematics and Evolution
350 301:1275–1286.
351 Chen T, Wang X-R, Tang H-R, Chen Q, Huang X-J, Chen J. 2013. Genetic diversity and
352 population structure of Chinese cherry revealed by chloroplast DNA *trn* Q-*rps* 16
353 intergenic spacers variation. Genetic Resources and Crop Evolution 60(6)1859-1871.
354 Cheesman E. 1944. Notes on the nomenclature, classification and possible relationships of cacao
355 populations. Tropical Agriculture 21:144-159.

356 Crandall KA y Templeton AR. 1993. Empirical test of some predictions from coalescent theory
357 with applications to intraspecific phylogeny reconstruction. Genetics 134(3): 959-969.
358 De la Cruz M, Whitkus R, Gómez-Pompa A, Mota-Bravo L. 1995. Origins of cacao cultivation.
359 Nature 375:542-543.
360 De Sahagún B. 2009. Historia general de las cosas de la Nuev aEspaña II. Editorial Dastin
361 Export, México. Libro tercero, cap. III y XII.
362 Doyle JJ, Doyle JL. 1990. A rapid total DNA preparation procedure for fresh plant tissue. Focus
363 12:13-15.
364 Dupanloup I, Schneider S, Excoffier LG. 2002. A simulated annealing approach to define the
365 genetic structure of populations. Molecular Ecology 11: 2571-2581.
366 Excoffier L, Laval G, Schneider S. 2005. Arlequin Ver. 3.0: an integrated software package for
367 population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.
368 Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis
369 program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95-98.
370 Hamilton MB, Braverman JM, Soria-Hernanz DF. 2003. Patterns and relative rates of nucleotide
371 and insertion/deletion evolution at six chloroplast intergenic regions in new world species
372 of the Lecythidaceae. Molecular Biolology and Evolution 20(10):1710–1721.
373 Hedrick PW. 2005. Genetics of populations. Jones and Bartlett Publishers, Sudbury, MA, EUA.
374 737 pp.
375 Jansen RK, Saski C, Lee SB, Hansen AK, Daniell H. 2011. Complete plastid genome sequences
376 of three rosids (*Castanea*, *Prunus*, *Theobroma*): Evidence for at least two independent
377 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, et al. 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.

Pérez-Jiménez M, Besnard G, Dorado G, Hernandez P. 2013. Varietal tracing of virgin olive oils based on plastid DNA variation profiling. PLoS ONE 8(8):e70507.

Polzin T, Daneshmand SV. 2012. NETWORK 4.6.1.3 Fluxus Technology Ltd. All rights reserved. Steiner (MP) algorithm.

Powis T, Cyphers A, Gaikwad N, Grivetti L, Cheong K. 2011. Cacao use and the San Lorenzo Olmec. Proceedings of the National Academy of Sciences of United States of America 108(21): 8595-8600.

Raymúndez MB, Mathez J, Xena de Enrech N, Dubuisson JY. 2002. Coding of insertion–deletion events of the chloroplastic intergene *atp-rbcL* for the phylogeny of the Valerianeae tribe (Valerianaceae). Comptes Rendus Biologies 325: 131–139.

400 Roger RA. 1995. Genetic evidence for Pleistocene population explosion. Evolution 49(4): 608-
401 615.
402 Rozas J, Librado P, Sánchez-Del Barrio JC, Messeguer X, Rozas R. 2010. DNA Sequence
403 Polymorphism, Ver. 5.10.1 Universidad de Barceolona. http://www.ub.edu/dnasp/.
404 Shaw J, Small RL. 2005. Chloroplast DNA phylogeny and phylogeography of the North
405 American plums (*Prunus* subgenus *Prunus* section *Prunocerasus*, Rosaceae). American
406 Journal of Botany 92: 2011–2030.
407 Shaw J, Lickey EB, Edward E, Schilling EE, Small RL. 2007. Comparison of whole chloroplast
408 genome sequences to choose noncoding regions for phylogenetic studies in angiosperms:
409 the tortoise and the hare III. American Journal of Botany 94(3):275–288. 2007.
410 Sun XQ, Zhu YJ, Guo JL, Peng B, Bai MM, et al. 2012. DNA Barcoding the dioscorea in china,
411 a vital group in the evolution of monocotyledon: use of *matK* gene for species
412 discrimination. PLoS ONE 7(2): e32057.
413 Slatkin M, RR Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable
414 and exponentially growing populations. Genetics 129, 555-562.
415 Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA
416 polymorphism. Genetics 123:585-595.
417 Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensivity of
418 progressive multiple sequence alignment through sequence weighting, position-specific
419 gap penalties and weight matrix choise. Nucleic Acids Research 22: 4673-4680.
420 Vázquez-Ovando JA, Molina-Freaner F, Nuñez-Farfán J, Ovando-Medina I, Salvador-Figueroa
421 M. 2014. Genetic identification of *Theobroma cacao* L. trees with high Criollo ancestry
422 in Soconusco, Chiapas, Mexico. Genetic and Molecular Research 13 (4):10404-14.

423 Wang FY, Gong X, Hu CM, Hap G. 2008. Phytogeography of an alpine species *Primula*
424 *secundiflora* inferred from the chloroplast DNA sequence variation. Journal of
425 Systematics and Evolution 46:13-22.
426 Whitkus R, de la Cruz M, Mota-Bravo L, Gómez-Pompa A. 1998. Genetic diversity and
427 relationships of cacao (*Theobroma cacao* L.) in southern Mexico. Theoretical and
428 Applied Genetics 96(1-2): 621-627.
429 Whitlock BA, Hale AM, Groff PA. 2010. Intraspecific inversions pose a challenge for the *trnH-*
430 *psbA* plant DNA barcode. PLoS ONE 5(7):e11533.
431 Wood GAR, Lass RA. 1985. Cacao. Tropical Agriculture Series, 4ta ed. Blackwell Sciencie
432 Logman Group Ltd. Nueva York.
433 Zeng J, Fan X, Sha LNn, Kang HY, Zhang HQ, Liu J, Wang XL, Zhou YH, Yang RW. 2012.
434 Nucleotide polymorphism pattern and multiple maternal origin in *Thinopyrum*
435 *intermedium* inferred by *trnH-psbA* sequences. Biologia Plantarum 56 (2):254-260.

Table 1(on next page)

Theobroma cacao trees, geographical coordinates and Genbank accesions. Genetic origin (Criollo, Non Criollo and admixtures) based on Vázquez-Ovando et al. (2014). *Geographic population *a priori*, **reference tree

1

Pop*	Coordinates latitude (N)/ longitude (W)	Criollo (n=20)	Non Criollo (n=15)	Admixtures (n=10)
1	14°59’28’’N, 92°26’44’’W (Huehuetán) 14°52’55’’N, 92°21’42’’W (Tapachula)	HUJF01 TASG12	HUJF03	TASG16 TASG18
2	14°56’41’’N, 92°09’59’’W (Tuxtla Chico) 14°59’53’’N, 92°10’44’’W (Cacahotán)	TCHR04	CAAM12	CAAM04
3	14°47’31’’N, 92°11’11’’W (Frontera Hidalgo) 14°38’27’’N, 92°13’47’’W (Suchiate)		FHSA06 SUED02 SUED03 SUED06	FHSA02
4	14°48’56’’N, 92°29’06’’W (Mazatán)	MAMG12	MAMG03 MAMG04 MAMG07 MAMG08	MAMG10
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 VCHL01 VCHL02 VCHL03 VCHL04 VCLB02 VCLB03 VCLB04		MAJH03
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 EET399	Rim24
9**	GenBank	Criollo-22	Scavina Amelonado Matina-06	ICS-01 ICS-06 ICS-39

2

Table 2(on next page)

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

1

Haplotype	Polymorphic site						Populations								
	22	134	206	309	310	487	Pop 1	Pop 2	Pop 3	Pop 4	Pop 5	Pop 6	Pop 7	Pop 8	Pop 9
H1	C	T	-	A	A	A	1		1		2				
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				
H6	-	T	A	-	-	A				1					
H7	-	T	-	A	-	-				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 cacao from Soconusco (Chiapas, Mexico) grouped by geographic approach (Pop) and genetic origin approach

1

Pop	Locality	N	S	Sn	H	Hd ± de	πd ± ED
1	Huehuetán, Tapachula	5	2	1	3	0.7000±0.218	0.001905±0.00177
2	Cacahoatán, Tuxtla Chico	3	2	1	2	0.6667±0.3143	0.002545±0.002614
3	Frontera Hidalgo, Suchiate	5	5	0	3	0.7000±0.2184	0.004183±0.003223
4	Mazatán	6	5	3	5	0.9333±0.121	0.004825±0.00347
5	Mapastepec, Villa Comaltitlán	9	5	0	4	0.6944±0.1470	0.003908±0.002735
6	Yucatán	2	1	1	2	1.0000±0.500	0.001908±0.00269
7	Selva Lacandona	3	0	0	1	0	0
Total		3	--	6	--	0.6610±0.089	0.003197±0.00213
Genetic origin approach							
“Criollo”		1	6	1	4	0.6364±0.127	0.002506±0.00188
“Forastero”		1	5	1	5	0.6667±0.163	0.003295±0.00236
“Hybrid”		6	6	1	5	0.9333±0.121	0.005450±0.00383
Criollo-reference[a]		8	4	1	5	0.7857±0.150	0.003333±0.00245
Forastero-reference[a]		5	3	0	3	0.8000±0.164	0.003053±0.00251
Trinitario-reference[a]		4	4	0	2	0.5000±0.265	0.003802±0.00319
Total		4	--	4	--	0.7949±0.052	0.003841±0.00244

N=Samples sizes, S=Number of segregating, Sn=Singletons, H=Number of haplotypes, Hd=Haplotype diversity, πd=Nucleotide diversity. [a]Including sequences GenBank (Criollo-reference n=1, Forastero-reference n=3, Trinitario-reference n=3).

2

Table 4(on next page)

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

1

Source of variation	df	SS	VC	Variation (%)	Fixation indices	*P* value
Among groups	1	1.613	0.12816	13.98	F_{SC}= -0.11146	0.73412
Among populations within groups	5	2.507	-0.08789	-9.59	F_{ST}= 0.04393	0.00684
Within populations	26	22.789	0.8765	95.61	F_{CT}= 0.1398	0.14956
Total	32	26.909	0.91677			
df= degrees of freedom, SS=Sum of squares, VC=Variance components.						

2

3

1

Location of indels (blue arrows) in a fragment of chloroplast DNA *trnH-psbA* intergenic spacer of *Theobroma cacao* trees. See Table 1 for details of the samples

2

Haplotypes frequency in each geographical population (A) and contribution of genetic groups to the haplotype network (B). Built with Network© 4.6.13 by Median Joining method.