Molecular Philogeography of Cacauhy (Theobroma speciosum) in the Brazilian Amazon

Filogeografia molecular de cacaú (Theobroma speciosum) na Amazônia brasileira

DOI:10.34117/bjdv6n12-417

Recebimento dos originais: 08/11/2020
Aceitação para publicação: 17/12/2020

Juliana de Freitas Encinas Dardengo
Doutorado em Biodiversidade e Biotecnologia (BIONORTE-MT)
Universidade do Estado de Mato Grosso (UNEMAT)
Endereço: Av. Perimetral Rogério Silva, Sn, 78.580-000 Alta Floresta, Mato Grosso, Brasil
E-mail: julianadardengo@outlook.com

Alex Souza Rodrigues
Mestrando em Genética e Melhoramento de Plantas pela Universidade Estadual do Norte Fluminense Darcy Ribeiro – UENF-Campos dos Goytacazes - RJ
E-mail: alexsouzarodrigues@outlook.com

Uéliton Alves de Oliveira
Programa de Pós-Graduação em Biotecnologia Vegetal
Universidade Estadual do Norte Fluminense Darcy Ribeiro – UENF-Campos dos Goytacazes - RJ
E-mail: uelitonalves2011@hotmail.com

Joameson Antunes Lima
Mestre em Genética e Melhoramento de Plantas (UNEMAT)
Doutorando no Programa de Pós-graduação em Genética e melhoramento de Plantas pela Universidade Estadual do Norte Fluminense Darcy Ribeiro
Endereço: Av. Alberto Lamego, 875 - Parque California, Campos dos Goytacazes – RJ
E-mail: anjoamerson@gmail.com

Magali Gonçalves Garcia
Doutorado em Genética e Melhoramento pela Universidade Federal de Viçosa (2010)
Universidade Federal do Pará (UFPAR)
Endereço: 030, R. Cel. José Porfírio - Recreio, Altamira – PA
E-mail: mgarcia.bio@gmail.com

Carolina Joana da Silva
Doutorado em Ecologia e Recursos Naturais pela Universidade Federal de São Carlos - UFSCAR (1990)
Universidade do Estado de Mato Grosso (UNEMAT)
Endereço: Av. Tancredo Neves, 1095 - Cavalhada II, Cáceres - MT, 78200-000
E-mail: ecopanta@terra.com.br
ABSTRACT
This study aims to determine the phylogeographic patterns in natural populations of Theobroma speciosum through the analysis of cpDNA and ITS sequence polymorphisms. We sampled individuals of T. speciosum distributed in 13 populations in the Brazilian Amazon to obtain data of DNA sequences of the chloroplastidic and nuclear genomes. The DNAsp program was used to estimate the number of haplotypes, haplotypic diversity, nucleotide diversity, and number of polymorphic sites. The values of genetic diversity estimated for populations from cpDNA present low values of haplotype and nucleotide diversity. In contrast, for the nuclear region, high values of haplotype and nucleotide diversity were found (except for three populations). The phylogeographic analyses based on the chloroplastidic and nuclear genome of T. speciosum evidenced the existence of 07 haplotypes for the cpDNA and 36 for the ITS, denoting a haplotype sharing, which means that T. speciosum does not present a pronounced but a generalist phylogeographic structure.

Keywords: conservation, diversity, Biogeography.

RESUMO
Estudos de filogeografia enfatizam os princípios e processos que governam a distribuição geográfica das linhagens genealógicas, principalmente em nível intraespecífico. Assim, este estudo tem como objetivo determinar os padrões filogeográficos em populações naturais de Theobroma speciosum por meio da análise de polimorfismos de sequências de cpDNA e ITS. Foram amostrados indivíduos de T. speciosum distribuídos em 13 populações na Amazônia brasileira, para obtenção de dados de sequências de DNA do genoma cloroplastídico e nuclear. O programa DNAsp foi utilizado para estimar o número de haplótipos (H), a diversidade haplotípica (Hd), a diversidade nucleotídica (pi), o número de sítios polimórficos (S). Os valores de diversidade genética estimados para as populações a partir de cpDNA apresentam valores baixos de diversidade haplotípica e nucleotídica, diferentemente, para a região nuclear foram encontrados altos valores de diversidade haplotípica e nucleotídica (exceto para MAC, NPR e HUM). As análises filogeográficas fundamentadas no genoma cloroplastídico e nuclear de T. speciosum evidenciam a existência de 07 haplótipos para o cpDNA e 36 para o ITS, denotando um compartilhamento de haplótipos, o que faz com que T. speciosum não apresente uma estrutura filogeográfica pronunciada e sim generalista.

Palavras-chave: Diversidade genética, Amazônia, Biogeografia, cpDNA.
1 INTRODUCTION

The Amazon rainforest presents an area of approximately 6.5 million km² in Northern South America (Sivam, 2012). Recent data from the diversity of trees in the Amazon point to the existence of 11,676 species distributed in 1225 genus and 140 families (Steege et al., 2016).

There is currently considerable interest in reconstructing the evolutionary history of species in order to determine their origins and their reactions to environmental changes (Dutech et al., 2003). Phylogeography is an area of study that emphasizes the principles and processes that govern the geographical distribution of genealogical lineages, mainly at the intraspecific level or among related species (Avise, 2000). Molecular phylogeography studies seek to examine molecular data from contemporary natural populations in order to search for evolutionary traces left in the genome by historical processes, thereby contributing to the enhancement of biogeographic hypotheses that may explain the current geographical distributions of these species and populations.

The family Malvaceae presents fruit species of great economic importance, including the genus Theobroma L., distributed from Mexico to the limits of the Amazon rainforest (Santos et al, 2020). The genus Theobroma presents 22 species, all belonging to Tropical America (Cuatrecasas, 1964), 8 of which are found in the Brazilian Amazon, such as the species *Theobroma speciosum* Willd. ex Spreng., which is found in a primary forest of non-flooded land and presents medium-sized trees. According to Lorenzi (2000), each plant of *T. speciosum* produces a large quantity of fruits that are very much appreciated by medium-sized mammals. These animals are natural dispersers of the species and generally consume the fruits and discard the seeds when they are still in the tree, which contributes to the occurrence of aggregate seed shadows in the immediate vicinity of the maternal plants, resulting in a spatial aggregation of individuals sharing a recent common ancestor (Dardengo et al., 2017).

In this study, we used molecular phylogeography tools for the analysis of DNA sequence data from the chloroplastidic and nuclear genomes of the *T. speciosum* species in order to determine the phylogeographic relationships among the natural populations, considering the aggregate distribution pattern found in previous studies with the species.

2 MATERIALS AND METHODS

2.1 SAMPLING STRATEGY AND DNA EXTRACTION

This study analyzed specimens of *T. speciosum*, collected from 13 populations in the Brazilian Amazon (Figure 1, Table 1). The sampling sites were chosen as to cover most of the species’ geographical range. To define the vegetation formations in which the species occurred, we consulted the Terrestrial Ecoregions of the World from the World Wildlife Fund (Olson et al., 2001). The habitat
classifications for Brazil from the Brazilian Institute of Geography and Statistics (IBGE, 1997) provided details about the vegetation on a local scale.

Figure 1. Geographic distribution of 13 populations of *Theobroma speciosum* and their associated vegetation formations. Terrestrial ecoregions: NT0707, Guyana savanna; NT0135, Madeira-Tapajós moist forests; NT0157, Purus-Madeira moist forests; NT0166, South-west Amazon moist forests; NT0168, Tapajós-Xingu moist forests; NT0170, Tocantins/Pindaré moist forests; NT0173, Uatuma-Trombetas moist forests; and NT0180, Xingu-Tocantins-Araguaia moist forests.
Table 1. *Theobroma speciosum* populations with their codes, geographic coordinates, ecoregions, and number of specimens sampled for cpDNA and ITS analyses.

| Populations (acronym)         | Location       | Ecoregions     | DNA Analyses | ITS         |
|------------------------------|----------------|----------------|--------------|-------------|
|                              |                |                | cpDNA N | Pi | Hd | N | Pi | Hd |
| Alta Floresta-MT (AF)        | S56°05'24" W9°52'33,6" | NT0135         | 14 | 0 | 0 | 14 | 0.007 | 0.967 |
| Cujubim-RO (CUJ)             | S62°35'24" W9°21'46,8" | NT0135         | 2  | 0 | 0 | 5  | 0.006 | 0.7  |
| Nova Bandeirantes e Apiacás – MT (NBA) | S57°48'38" W09°50'59" | NT0135         | 8  | 0 | 0 | 9  | 0.003 | 0.694 |
| Parque Nacional do Juruena – MT (JUR) | S58°12'20" W07°23'56" | NT0135         | 3  | 0.0018 | 1 | 3  | 0.007 | 1  |
| Porto Velho-RO (POR)         | S63°53'60" W8°34'43,2" | NT0157         | 3  | 0 | 0 | 3  | 0.006 | 1  |
| Humaitá-AM (HUM)             | S63°01'12" W7°30'21,6" | NT0157         | 4  | 0 | 0 | 3  | 0    | 0  |
| Rio Branco-AC (RBC)          | S67°48'36" W09°58'30" | NT0166         | 8  | 0.0004 | 0.536 | 7 | 0.007 | 0.09 |
| Altamira-PA (ALT)            | S52°12'36" W3°12'10,8" | NT0168         | 4  | 0 | 0 | 4  | 0.002 | 0.5  |
| Novo Progresso-PA (NPR)      | S55°22'48" W07°8'52,8" | NT0168         | 3  | 0.0006 | 0.667 | 3  | 0    | 0  |
| Aurora do Pará-PA (ARI)      | S47°33'36" W02°8'24" | NT0170         | 5  | 0.0003 | 0.286 | 5  | 0.002 | 0.4  |
| Manaus-AM (MAN)              | S60°01'12" W03°06'7,2" | NT0173         | 5  | 0.0043 | 0.4  | 4  | 0.002 | 0.833 |
| Piçarra-PA (PIC)             | S48°52'12" W06°26'16,8" | NT0180        | 4  | 0.0013 | 0.833 | 7  | 0.005 | 0.714 |
| Macapá-AP (MAC)              | S51°04'12" W0°02'19,9" | NT0707        | 4  | 0 | 0 | 4  | 0    | 0  |
| TOTAL                        |                |                | 67 |     |   | 71 |     |     |
Leaf samples were transported to the laboratory while still fresh and then kept at -80 °C; alternatively, they were dried immediately using silica gel and kept at room temperature until the DNA was extracted. Total genomic DNA was extracted following the procedure described in Doyle and Doyle (1990). Genomic DNA from all specimens was archived in our laboratory, at the State University of Mato Grosso, Brazil.

2.2 ASSEMBLY OF THE CPDNA DATASET

The \textit{trnT–trnL} spacer was amplified using the primer combination A2 described by Cronn et al. (2002) and aC (5”-CGTAGCGTCTACCGATTTCG-3”), complementary to the primer C described by Taberlet et al. (1991). This primer pair amplified the intergenic spacer between \textit{trnT} (UGU) and the \textit{trnL} (UAA) 5” exon. The second was performed with primers C and D and amplified the \textit{trnL} (UAA) intron.

PCR reactions and amplifications were carried out according to the following protocol described by Garcia et al. (2011). PCR products were cleaned using ExoSAP IT (USB), using a ratio of 9μl reaction to 3μl enzyme. Sequencing was performed by Macrogen Inc., South Korea (www.macrogen.com) using the same primers as in the PCR reactions. Sequences were imported into Sequencher for editing.

2.3 ASSEMBLY OF ITS DATASETS

The entire internal transcribed spacer (ITS) region of the nuclear 18S-26S ribosomal RNA genes was amplified according to standard PCR protocols of Garcia et al. (2011), using primer ITS4 described by White et al. (1990) and primer ITS.LEU described by Baum et al. (1998). PCR products were cleaned using ExoSAP IT (USB), at the rate of 9μl reaction to 3μl enzyme. Sequencing was performed by Macrogen Inc., South Korea (www.macrogen.com), using the same primers as in the PCR reactions. Sequences were imported into the program Sequencher 4.8 (Gene Codes) for editing. Complete sequence alignments were performed with the introduction of gaps to compensate for the presence of insertion/deletions (indels).

2.4 PHYLOGENETIC ANALYSES

DnaSP v5 software (Librado et al., 2009) was used to estimate the haplotype number (H), haplotipic diversity (Hd), nucleotidic diversity (pi), number of polimorphic sites (S), and the gene diversity and nucleotide diversity (Nei, 1973) indices at the population level. Gene genealogies were inferred using the Median Joining (MJ) network method (Bandelt et al., 1999) as
implemented in NETWORK 4.5.0.2 (Fluxus Technology Ltd) using default parameters. The program was run such that indels were considered to be a fifth character state and coded such that each indel, regardless of its size, was considered to be a single mutation.

3 RESULTS

3.1 CPDNA DATASET

Sequencing of the target trnT-trnL spacer of the chloroplast genome yielded two segments of sequence. We aligned a total of 577 bases for fragment AB and 545 bases for fragment CD. The assembly of these two sets of sequences resulted in a final chloroplast dataset that was 1122 bases long. Sequence analysis revealed a total of 07 haplotypes and 17 polymorphic sites (Figure 2), 16 of which were base substitutions and 01 of which was indel. Segment CD showed lower levels of polymorphism, containing 4 base substitutions.

Figure 2. Sequence alignment of the variable sites in the cpDNA trnT–trnL spacer that define the 07 haplotypes of *Theobroma speciosum*. Each fragment spans 1122 bases. Dots indicate similarity to haplotype A and hyphens indicate gaps. Numbers on top indicate the nucleotide position with haplotype A acting as the reference sequence during alignment. The number of occurrences of each haplotype (#) is indicated on the right.

3.2 NUCLEAR ITS DATASET

Our dataset for the nuclear ITS region consisted of aligned sequences of 534 bases. There were 36 ITS haplotypes among the 71 ribosomal sequences (Figure 3). Sequence alignment revealed the presence of 33 polymorphic sites that were characterized by either indels or base
substitutions, 29 of which were base substitutions, 3 of which were 1-bp indels, and 1 was 16-bp indel.

Figure 3. Sequence alignment of the variable sites in the nuclear ITS that defines the 36 ITS haplotypes of *Theobroma speciosum*. Each fragment spans 534 bases. Dots indicate similarity to haplotype 1 and hyphens indicate gaps. Numbers on top indicate the nucleotide position with haplotype 1, acting as the reference sequence during alignment. The number of occurrences of each haplotype among direct sequences (#) is indicated on the right.
3.3 GENEALOGICAL AND GEOGRAPHIC RELATIONSHIPS AMONG LINEAGES

Genetic structure was assessed with MJ networks, which were constructed for the cpDNA and nuclear ITS datasets separately. In both networks, the haplotypes were organized around the most frequent ones, with the least common haplotypes being found at the periphery of the network. The genetic structure together with the geographic distributions of the cpDNA haplotypes (cpDNA) (Figure 4) and ITS haplotypes (Figure 5) showed that there was a clear sharing of haplotypes among the regions.

The MJ network that was constructed with the chloroplastidic DNA dataset presents 7 cpDNA haplotypes, which were organized around the high frequency haplotype A (found in 56 of the 67 specimens analyzed) (Fig. 4). Based on the pattern generated in the analysis of the chloroplastidic genome, it is possible to observe that most populations presented a single haplotype (Haplotype A - POR, CUJ, NBA, AF, AUR, ALT, MAC, and HUM), with only 3 populations presenting unique haplotypes (MAN, JUR, and PIC).

The MJ network that was constructed with the nuclear dataset presents 36 ITS haplotypes, which were organized around the high frequency haplotypes 19 and 12 (found in 5 populations - AF, NBA, POR, PIC, RBC, and ALT) (Fig. 4). The populations HUM, MAC, and NPR present just one haplotype each and the haplotypes 30 and 36 were exclusives from the populations HUM and MAC, respectively. ITS haplotypes from Manaus were separated from the others through 8 intermediate missing haplotypes. In the RBC population we found 5 different haplotypes, 1 of which (19) was separated from the others through 5 intermediate missing haplotypes.

Although the populations showed some differences in the distribution of genetic and, consequently, haplotypic variability, for the nuclear genome the most evident pattern observed was marked by the existence of some populations with unique haplotypes and some shared between adjacent regions.
Figure 4. A: Median-joining networks depicting the relationships among haplotypes of *Theobroma speciosum* based on cpDNA trnT–trnL. Circles represent cpDNA haplotypes (coded with letters) and size is proportional to the relative frequencies. The number of substitutions is indicated with bars when more than one. Each black rhombus indicates an intermediate missing haplotype. B: Geographic distribution of the 07 cpDNA haplotypes of *Theobroma speciosum*. Each circle represents a population; a color within a circle denotes the cpDNA haplotype. Population codes are as given in Table 1.
Figure 5. A: Median-joining networks depicting the relationships among haplotypes of *Theobroma speciosum* based on nuclear ITS. Circles represent ITS haplotypes (coded with numbers) and size is proportional to the relative frequencies. The number of substitutions is indicated with bars when more than one. Each black rhombus indicates an intermediate missing haplotype. B: Geographic distribution of the 36 ITS haplotypes of *Theobroma speciosum*. Each circle represents a population; a color within a circle denotes the ITS haplotype. Population codes are as given in Table 1.
4 DISCUSSION

The phylogeographic analyses based on the chloroplastidic and nuclear genome of *T. speciosum* evidenced the existence of 07 haplotypes for the cpDNA and 36 for the ITS, denoting a sharing of haplotypes, which causes a generalist phylogeographic structure in *T. speciosum*, as well as that shown for *Curatella americana* L. (Canuto, 2011).

A clear separation was observed between the Manaus lineage and the other populations, which shared only one haplotype of cpDNA (haplotypes A) and no haplotype of ITS with the rest of the populations sampled. In addition, the lineage of Manaus is separated from the other populations by 12 and 08 mutation steps (for cpDNA and ITS, respectively), and in the ITS network intermediate haplotypes (hypothetical) separate the lineage from the others. According to Avise (2000), the presence of spatially restricted lineages separated by a considerable amount of mutational steps distinguishes the ancestral allopatric lineages in a gene tree.

The cause of the phylogeographic distance between the ancestral lineages of Manaus and the others cannot be justified only with the results of this research, and more populations should be sampled in the next studies with the species, mainly in the western region of Amazonia.

The comparison between haplotype diversity and nucleotide diversity is a simple way to infer the demographic history of a group (Avise, 2000). The values of haplotype and nucleotide diversity estimated for populations from cpDNA were low, and this pattern allows inferring a rapid population growth from an ancestral population, with sufficient time for recovery of the variation by mutation, but not enough time to accumulate large differences in the sequences (Avise, 2000). Differently, for the nuclear region high values of haplotypic and nucleotide diversity were found (except for MAC, NPR, and HUM). This result is expected, since the analyses evidenced the existence of 7 haplotypes for cpDNA and 36 for ITS.

For ITS data, three populations (MAC, MAN, and HUM) presented exclusive haplotypes that were isolated from the others, and for cpDNA they presented the haplotype A shared with the other populations. According to McCauley (1995), these characteristics are due to the fact that organellar genes have a more distinct population structure than nuclear genes because of differences in effective population size.

In plants, the gene flow can be promoted by the dispersion of seeds and pollen. In most angiosperms the cpDNA is maternally inherited and can move only through the dispersion of the seeds, so the migration rates of cpDNA in angiosperms are much lower than the migration rates of nuclear genes that can move through the dispersion of seeds and pollen (McCauley, 1995). There are no studies
showing the maternal inheritance of chloroplastidic DNA for *T. speciosum*; however, for *T. cacao* this characteristic was evidenced by Laurent et al. (1993).

5 CONCLUSIONS

The phylogeographic analyses based on the chloroplastidic and nuclear genome of *T. speciosum* evidenced a sharing of haplotypes, denoting a generalist phylogeographic structure of *T. speciosum* populations. To improve the information about the evolutionary history of *T. speciosum*, studies with modeling and simulations from the fossil record and phylogeographic studies with populations of *T. speciosum* from the western part of the Brazilian Amazon, as well as from other countries, are needed.
REFERENCES

Avise JC. Phylogeography: the history and formation of species. First Edition. Cambridge: Harvard Univ. Press; 2000.

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

Baum DA, Small RL, Wendel JF. Biogeography and floral evolution of baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematic Biology* 1998; 47:181–207.

Canuto JZ. Phylogeography of *Curatella americana* L. (Dilleniaceae): an arboreal species of the Amazonian savannas and Brazil [dissertation]. Central Research Institute of the Amazon - INPA, Manaus; 2011.

Cronn RC, Small RL, Haselkorn T, Wendel JF. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany* 2002; 89: 707-725.

Cuatrecasas J. A taxonomic revision of the genus *Theobroma*. *Contributions from U.S. National Herbarium*. 1964; 35: 379-607.

Dardengo JFE, Rossi AAB, Silva CJ, Silveira M. Spatial structure of *Theobroma subincanum* Mart. and *Theobroma speciosum* Will. ex. Spreng. in the Parque Nacional do Juruena, Mato Grosso state, Brazil. *Revista Árvore*. 2017; 41(1), e: 410101.

Doyle JJ, Doyle JL. Isolation of plant DNA from fresh tissue. *Focus*. 1990; 12: 1315.

Dutech C, Maggia L, Tardy C, Joly HI, Jarne P. Tracking a genetic signal of extinction-recolonization events in a neotropical tree species: *Vouacapoua americana* Aublet in French Guiana. *Evolution*. 2003; 57(12): 2753-2764.

Garcia MG, Silva RS, Carniello MA, Veldman JW, Rossi AAB, De Oliveira LO. Molecular evidence of cryptic speciation, historical range expansion, and recent intraspecific hybridization in the Neotropical seasonal forest tree *Cedrela fissilis* (Meliaceae). *Molecular Phylogenetic Evolution*, 2011; 61 (3): 639-649.

IBGE - INSTITUTO DE GEOGRAFIA E ESTATÍSTICA. Recursos Naturais e Meio Ambiente: uma visão do Brasil. Departamento de Recursos Naturais e Estudos Ambientais. Second edition. Rio de Janeiro: IBGE; 1997.

Laurent V, Risterucci AM, Lanaud C. Choroplast and mitocontrial DNA diversity in *Theobroma cacao*. *Theoretical and applied genetics*. 1993; 87:81-88.

Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009; 25:1451-1452.
Lorenzi H. Weeds of Brazil: terrestrial, aquatic, parasitic and toxic. First Edition. Nova Odessa: Instituto Plantarum; 2000.

Mccauley DF. The use of chloroplast DNA polymorphism in studies of gene flow in plants. *Trends Ecology Evolution*, 1995; 10: 198-202.

Nei M. Analysis of Gene Diversity in Subdivided Populations. *Proc. Nat. Ac. Sci. PNAS*, USA. 1973; 70(12): 3321-3323.

Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GVN, Underwood EC, et al. Terrestrial ecoregions of the world: a new map of life on earth. *BioScience*. 2001; 51, 933-938.

SANTOS, J. A. R.; LOBÃO M. S.; ARAUJO, R. M. G. Uso da dendrocronologia para avaliar a produção anual de frutos da espécie *Theobroma cacao* L. no sudoeste da amazônia. *Brazilian Journal of Development*. 6(7): 51001-51011. DOI:10.34117/bjdv6n7-662

SIVAM - SISTEMA DE VIGILÂNCIA DA AMAZÔNIA. Amazônia Legal, 2012. Accessed in 20 mai. 2020. Available from: <HTTP://www.sivam.gov.br/AMAZONIA/apres1.htm>.

Steege HT, Vaessen RW, Cárdenas-López D, Sabatier D, Antonelli A, Oliveira SM, et al. The Discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. *Scientific Reports* 2016; 6: 29-49.

Taberlet P, Gielly L, Pautou G, Bouvet J. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 1991; 17:1105-1109.

White TJ, Bruns TD, Lee S, Taylor J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ. editors. PCR protocols, a guide to methods and applications. San Francisco: Academic Press; 1990.