

## Characterization and phylogenetic analysis of the complete plastid genome of *Theobroma bicolor* (Malvaceae) from Peru

Daniel Tineo<sup>a</sup>, Martha S. Calderon<sup>a,b</sup>, Jorge L. Maicelo<sup>a</sup>, Manuel Oliva<sup>a</sup>, Ángel F. Huamán-Pilco<sup>c</sup>, Oswaldo Ananco<sup>a</sup> and Danilo E. Bustamante<sup>a,b</sup>

<sup>a</sup>Instituto de Investigación para el Desarrollo Sustentable de Ceja de Selva (INDES-CES), Universidad Nacional Toribio Rodríguez de Mendoza, Chachapoyas, Peru; <sup>b</sup>Instituto de Investigación en Ingeniería Ambiental (INAM), Facultad de Ingeniería Civil y Ambiental (FICIAM), Universidad Nacional Toribio Rodríguez de Mendoza, Chachapoyas, Peru; <sup>c</sup>Departamento de Sanidad Vegetal, Facultad de Ciencias Agronómicas, Universidad de Chile, Santiago, Chile

### ABSTRACT

*Theobroma bicolor* Bonpl. 1806 is distributed in the Neotropics from southern Mexico to the Peruvian and Brazilian Amazon. High-throughput sequencing of *T. bicolor* from Peru (KUELAP2926) resulted in the assembly of its complete plastid genome (GenBank accession number OQ557154). The chloroplast genome of *T. bicolor* is A+T-rich (62.97%), having 160,317 bp in size and containing 130 genes; including a pair of inverted repeat regions (IRs) of 25,462 bp separated by a large single copy region (LSC) of 89,221 bp and a small single copy region (SSC) of 20,172 bp. This plastid genome is similar in length, content, and organization to other members of the genus *Theobroma*. Phylogenetic analyses of *T. bicolor* support its sistership to the clade comprising *T. cacao* and *T. grandiflorum*. This study may contribute valuable information to the phylogenetic relationships within the genus *Theobroma*.

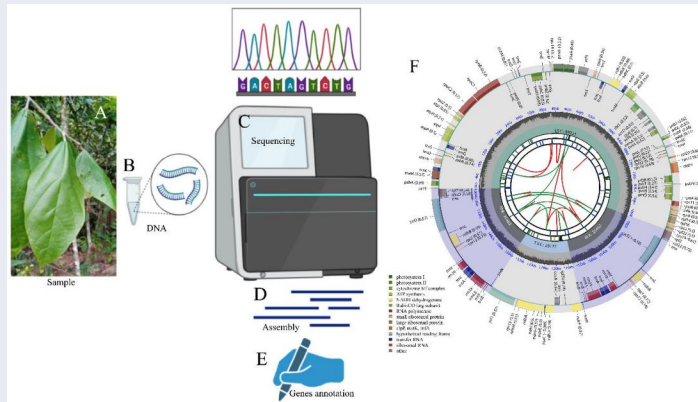
### ARTICLE HISTORY

Received 29 May 2023  
Accepted 21 January 2024

### KEYWORDS

Amazonas region;  
Malvaceae; Peru; plastid  
genome; *theobroma*



### GRAPHICAL ABSTRACT




## Introduction

The genus *Theobroma* L. belongs to the family Malvaceae and comprises 22 species (Bayer et al. 1999). One of this species, *Theobroma bicolor* has a wide distribution in the Neotropics, ranging from southern Mexico to the Peruvian and Brazilian Amazon (Aikpokpodion 2012; GBIF 2020). It is also considered to have significant economic potential (Ponce-Sánchez et al. 2021). In Colombia and Peru, *T. bicolor* is called as macambo or maraco (Barrera et al. 2006). The pulp of this species is used to make juices, jams and traditional beverages (Kufner and McNeil 2006; Gálvez-Marroquín et al. 2016). The seeds are

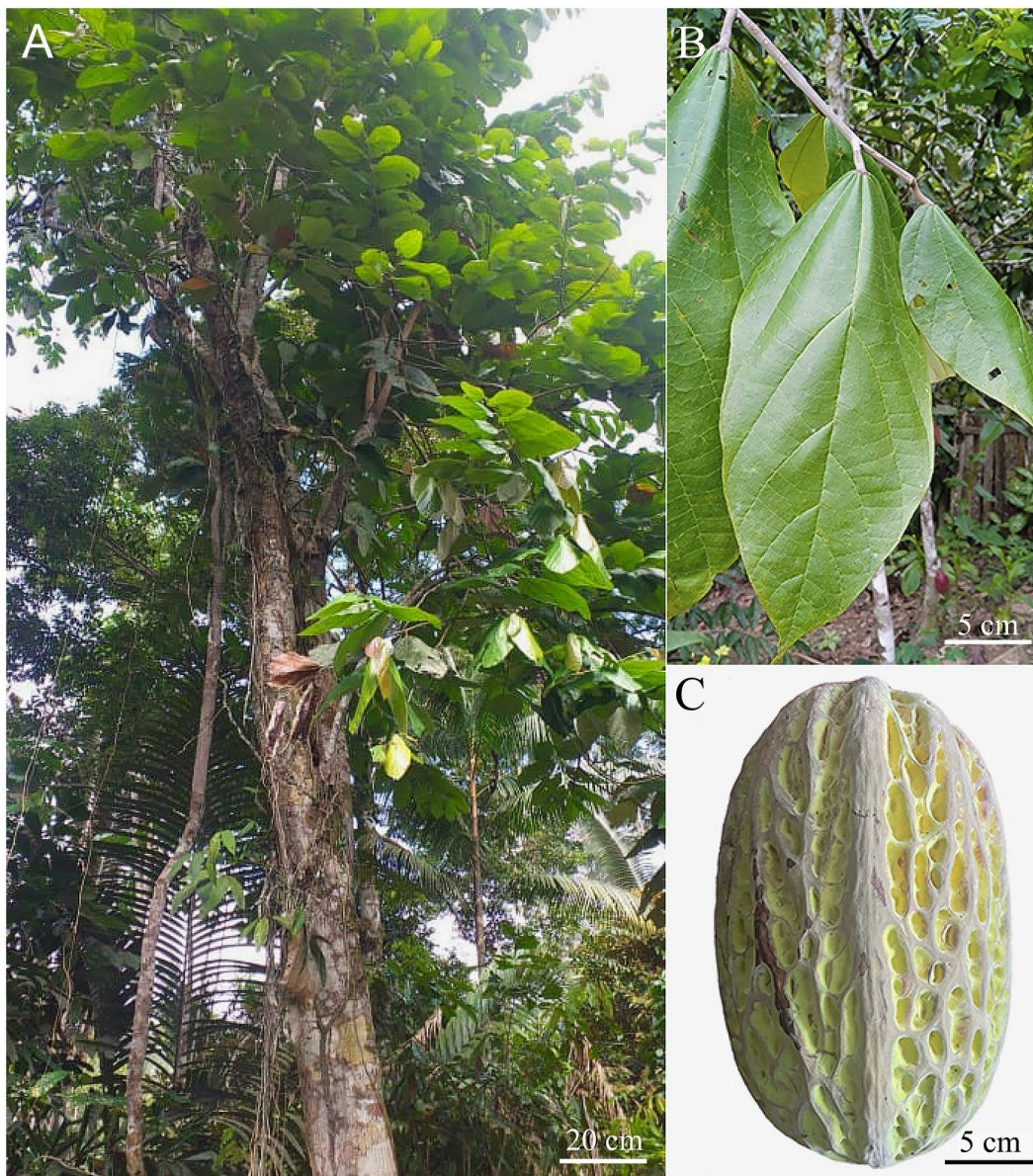
used as snacks and are used to make a type of chocolate called bacalate (Mantilla et al. 2008; González and Torres 2010), which has good quality butter containing 42.0% of fatty acids (Torres et al. 2002). *T. bicolor* beans have lower phenol levels compared to *T. cacao*, which has the potential to improve the flavor and nutritional quality of cocoa products (Febrianto and Zhu 2022). Additionally, *T. bicolor* is resistant to pests and diseases. For instance, *Moniliophthora roreri* and *Moniliophthora perniciosa* commonly attack *T. cacao*, but do not generate strong affection on *T. bicolor* (Lagneaux et al. 2021). These special traits in *T. bicolor* offer great potential in

**CONTACT** Danilo E. Bustamante  [danilo.bustamante@untrm.edu.pe](mailto:danilo.bustamante@untrm.edu.pe)  Instituto de Investigación para el Desarrollo Sustentable de Ceja de Selva (INDES-CES), Universidad Nacional Toribio Rodríguez de Mendoza, Chachapoyas, Peru

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/23802359.2024.2310134>.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.



**Figure 1.** Morphology of *Theobroma bicolor* (KUELAP-2926). (A) Habit showing tree with branched clusters. (B) Complex leaf. (C) Immature fruits oblong-ellipsoid in shape. All images have been obtained from Oswaldo Ananco from the province of Condorcanqui, Amazonas region.

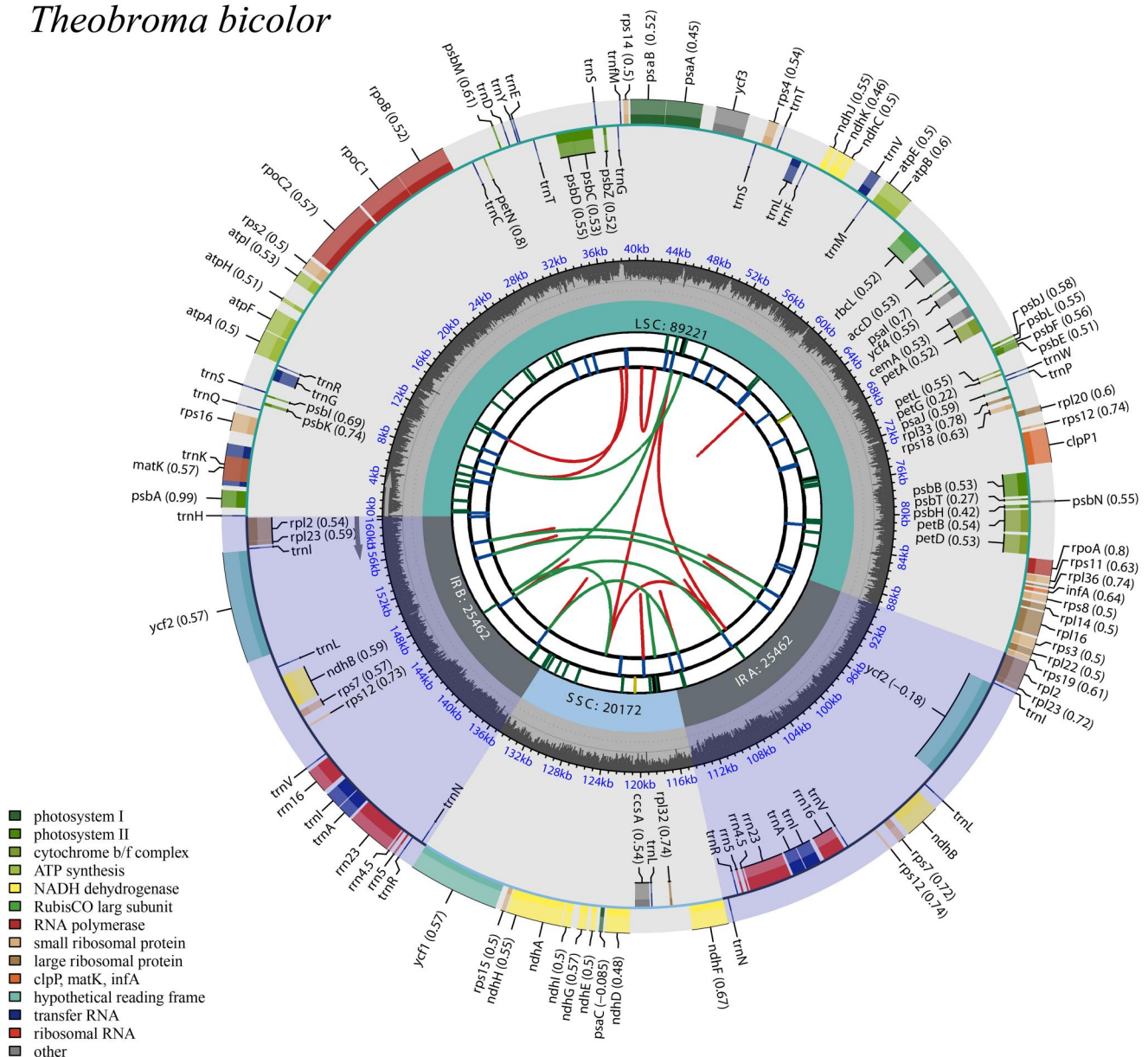
the genetic improvement of *T. cacao* and also might enhance the flavor and nutritional quality of *T. cacao* derivatives such as chocolate, butter and cocoa powder (Torres et al. 2002; Febrianto and Zhu 2022). However, few studies related to its chemical characterization, post-harvesting management, and genomic analyses are available in comparison to its relative *T. cacao* (González et al. 2016; Ponce-Sánchez et al. 2021). It was confirmed that the plastidial genome has useful information for topics such as population genetics, evolution, and biotechnology due to its maternal inheritance (Rogalski et al. 2015). Accordingly, the aim of this study was decoding the complete chloroplast genome of *T. bicolor* from Peru using high throughput sequencing technology.

## Materials and methods

The specimen of *T. bicolor* was collected in March 2022. This specimen was an adult flowering plant, approximately 30 years

old (Figure 1(A)). Tissue samples of approximately 50 mm<sup>2</sup> were taken from leaf tips (Figure 1(B)) for genomic analyses and placed in prelabelled 1.5 mL Safelock Eppendorf tubes with the addition of silica gel. In addition, samples of leaves (Figure 1(B)), flowers and fruits (Figure 1(C)) of *T. bicolor* were collected (Figure 1) and deposited at the herbarium of Universidad Nacional Toribio Rodríguez de Mendoza (KUELAP, <https://www.untrm.edu.pe>, Curator Eli Pariente, email: [eli.pariante@untrm.edu.pe](mailto:eli.pariante@untrm.edu.pe)) under the voucher number KUELAP2926 (collected by Oswaldo Ananco from Condorcanqui, Peruvian Amazonas region;  $-4.245242\text{ S}$ ,  $-77.722025\text{ W}$ ).

DNA genomic was extracted from *T. bicolor* (Specimen Voucher- KUELAP-2926) using the kit NucleoSpin (Macherey-Nagel, Düren, Germany) following the manufacturer's instructions. The 150 bp PE Illumina library construction and sequencing was performed using the HiSeq 6000 platform by MacroGen (Seoul, South Korea). The genome was assembled using default de novo settings in MEGAHIT (Li et al. 2015)

*Theobroma bicolor*

**Figure 2.** Schematic map of overall features of the chloroplast genome of *Theobroma bicolor*. The map contains six tracks in default. From the center outward, the first track shows the dispersed repeats connected with arcs. The second track shows the long tandem repeats as short bars. The third track shows the short tandem repeats or microsatellite sequences as short bars. The small single-copy (SSC), inverted repeat (IRs), and large single-copy (LSC) regions are shown on the fourth track. The GC content along the genome is plotted on the fifth track. The genes are shown on the sixth track. The optional codon usage bias is displayed in the parenthesis after the gene name. Genes are coded by their functional classification. The transcription directions for the inner and outer genes are clockwise and anticlockwise, respectively. The functional classification of the genes is shown in the bottom left corner.

and Geneious Prime 2023.2 (<https://www.geneious.com>) to close gaps. Sequencing depth and coverage was calculated following Yang et al. (2023). The genes were annotated manually using blastx, NCBI ORFfinder, and tRNAscan-SE 2.0 (Lowe and Chan 2016). The *T. bicolor* plastid genome was aligned to other plastomes using MAFFT (Katoh and Standley 2013). The Maximum Likelihood and Bayesian Inference phylogenetic tree was constructed using IQ-TREE v.2.2.0 software (Trifinopoulos et al. 2016) with 1500 ultrafast bootstrap replicates and the test model (-m TEST) (Minh et al. 2020). The tree was visualized with TreeDyn 198.3 at Phylogeny.fr (Dereeper et al. 2008).

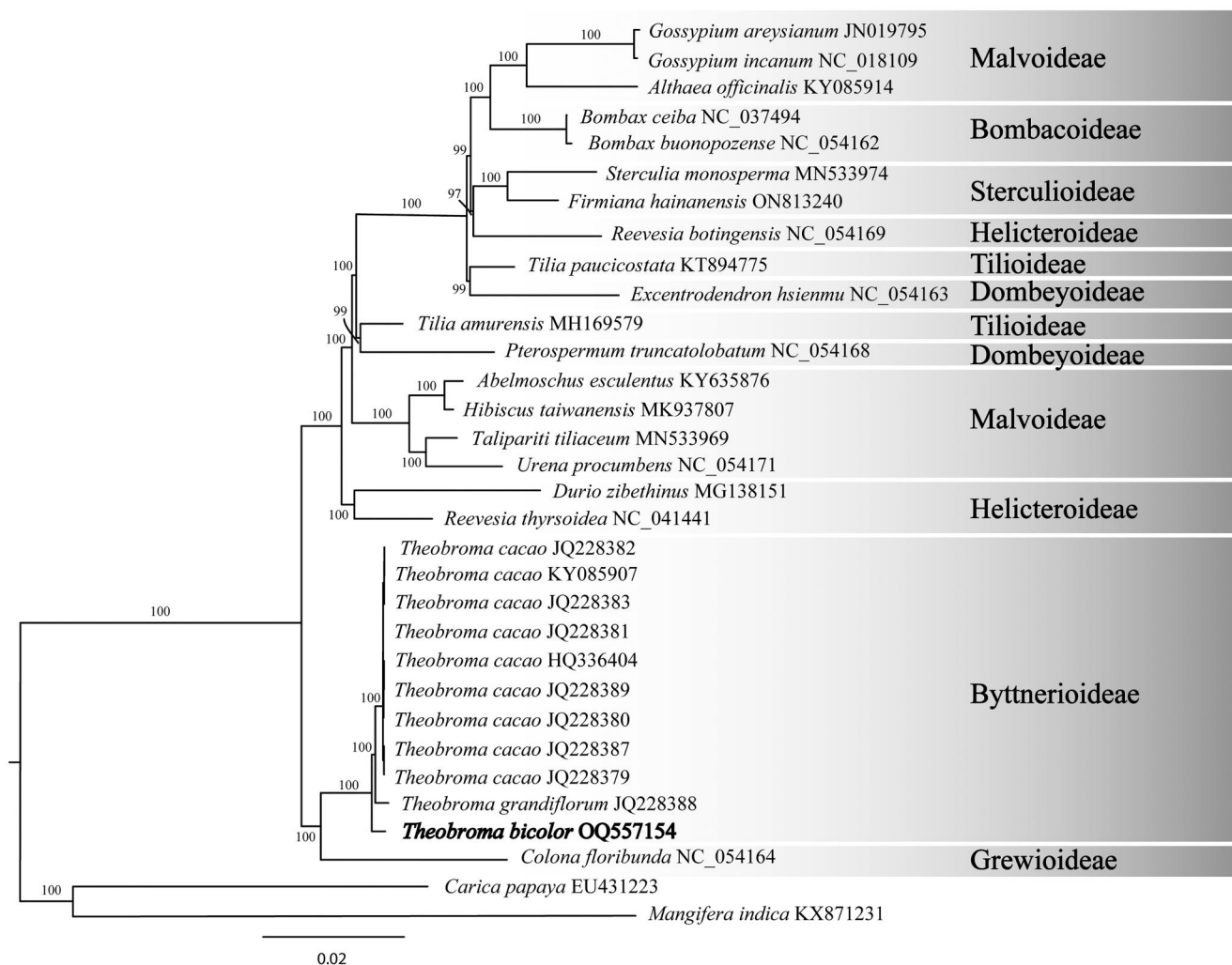
## Results

### Genome organization and composition

The plastid genome of *T. bicolor* is 160,317 bp in length and contains 130 genes (Figure 2), including a pair of inverted repeat regions (IRs) of 25,462 bp separated by a large single-copy region (LSC) of 89,221 bp and a small single copy (SSC) region of 20,172 bp. The maximum, minimum and average sequencing depth was 1926, 29, and 121.63, respectively (Figure S1). This plastid genome is A + T rich (62.97%), including 25 ribosomal proteins, 37 tRNA (*trnA*, *trnG*, *trnN*, and *trnT* occur in duplicates; *trnS*, *trnR* and *trnV* occur in triplicates,

**Table 1.** Characteristics of plastid genome among species of *T. bicolor*, *T. grandiflorum* and *T. cacao*.

Characteristics	<i>T. bicolor</i> (OQ557154)	<i>T. grandiflorum</i> (JQ228388)	<i>T. cacao</i> (HQ336404)
Size (base pair; bp)	160,317	160,619	160,604
LSC length (bp)	89,221	89,333	89,395
SSC length (bp)	20,172	20,194	20,187
IR length (bp)	25,462	25,546	25,511
Number of genes	130	130	130
Protein-coding genes	85	85	85
tRNA genes	37	37	37
rRNA genes	8	8	8
Duplicate genes	17	17	17
Total (%)	37.02	36.80	36.90
LSC (%)	34.88	34.70	34.70
SSC (%)	31.35	31.10	31.20
GC content			
IR (%)	42.86	42.90	43.00
CDS (%)	37.90	37.90	37.90
rRNA (%)	55.54	55.50	55.50
tRNA (%)	52.94	52.90	53.00
Protein coding part (CDS) (% bp)	48.79	48.96	49.08

**Figure 3.** Maximum likelihood phylogram of *Theobroma bicolor* (OQ557154) and related genera. These genera are grouped in their subfamilies. For instance, the subfamily Grewioideae is based on the genus *Grewia* (Narkthai & Chantaranonthai, 2020); whereas the subfamily Byttnerioideae is based on the genus *Byttneria* (Colli-Silva & Pirani 2020). Numbers along branches are RaxML bootstrap supports based on 1500 replicates. The legend below represents the scale for nucleotide substitutions.

*trnI* and *trnL* occur in quadruplicate), 21 photosystem I and II, six *ycf*, seven cytochrome b/f complex, six ATP synthase, four RNA polymerase, eight rRNA, and 16 other genes (*accD*, *cemA*, *ccsA*, *ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*, *psbN*). Additionally, 13 cis-splicing

(*atpF*, *clpP1*, *ndhA*, *ndhB*, *petB*, *petD*, *rpoC1*, *rpl2*, *rpl16*, *rps16*, *ycf1*) and one trans-splicing genes were identified (*rps12*) (Figures S2 and S3). Fifty-four of the 130 genes are transcribed on the forward strand and the remaining 76 are coded on the reverse strand (Figure 2; Table 1).

## Phylogenetic analysis

Highly supported nodes in the phylogenetic analysis of plastid genome sequences of *T. bicolor* and other species of Malvales confirmed that *T. bicolor* is sister to the clade comprising *T. grandiflorum* and *T. cacao*. These three species are grouped in the subfamily Byttnerioideae, which is in a sister relationship to the subfamily Grewioideae.

## Discussion

Twenty species have been accepted in the genus *Theobroma* (Govaerts et al. 2021) and this study has assembled the third complete plastid of genome for this genus. The plastid genome of *T. bicolor* is highly conserved in length, content, and organization to other species currently assigned to *Theobroma* (Figure 3, Figure S4). Although the plastid genome of *T. bicolor* is slightly shorter (160,317 bp) than *T. cacao* (160,604 bp) and *T. grandiflorum* (160,619 bp) (Abdullah et al. 2020), the total GC content is slightly higher in *T. bicolor*. This increase is justified by the higher percentages of GC content in LSC, SSC, and rRNA. Conversely, it was suggested that GC content is substantially higher in IR genes (Yang et al. 2022). Phylogenomic analysis of the *T. bicolor* plastid genome fully resolved it in a clade sister to *T. grandiflorum* and *T. cacao*. This evolutionary relationship is congruent to recent multilocus phylogenies among *Theobroma* species (Niu et al. 2019; Abdullah et al. 2020). Pairwise genetic distances of the plastid genomes of *T. bicolor* and *T. grandiflorum* are 0.37%; whereas *T. bicolor* and *T. cacao* are 0.32%. These interspecific differences are useful for establishing a genetic threshold to delimit species in the genus *Theobroma* (Kuhn et al. 2010; Liu et al. 2018). Conclusively, this contribution provides a useful resource for conservation and a valuable framework to fully resolve the phylogenetic relationships of the genus *Theobroma*.

## Authors contributions

DT, DEB, and MSC designed the research. DT, OA, JLM, and MO generated data for analysis. DEB, MSC, DT, AFHP, and OA analyzed the data and conducted experiments. DEB, MSC, JLM, and MO supervised the research. DEB, MSC, DT, AFHP, OA, JLM, and MO drafted the work and revised it critically. All authors agree to be accountable for all aspects of the work. All authors have read and approved the manuscript.

## Ethics statement

Permit for sample collection and scientific research of wild flora (AUT-IFL-2020-0051-MINAGRI-SERFOR-DGGSPFFS-DGSPF) was provided by Servicio Nacional Forestal y de Fauna Silvestre (SERFOR).

## Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

## Funding

This research was funded by INDES-CES/UNTRM throughout Project CUI No. [2252878] 'FISIOBVEG'. This study was also funded by Vicerrectorado de Investigación de la Universidad Nacional Toribio Rodríguez de Mendoza [DITT-2023-BM]. It was also partially funded by CONCYTEC under Project MiCroResi [PE501079652-2022-PROCIENCIA].

## Data availability statement

The genome sequence data that support the findings of this study are openly available in GenBank of NCBI at <https://www.ncbi.nlm.nih.gov/> under the accession no. OQ557154. The associated BioProject, BioSample, and SRA numbers are PRJNA944261, SAMN33745908, and SRR23852829, respectively.

## References

- Abdullah, Waseem S, Mirza B, Ahmed I, Waheed MT. 2020. Comparative analyses of chloroplast genomes of *Theobroma cacao* and *Theobroma grandiflorum*. *Biologia*. 75(5): 761–771. doi:10.2478/s11756-019-00388-8.
- Aikpokpodion P. 2012. Phenology of flowering in cacao (*Theobroma cacao*) and its related species in Nigeria. *Afr J Agric Res*. 7:3395–3402.
- Barrera JA, Hernández MS, Vargas G, Martínez O, Melgarejo LM, Casas AE, Zambrano JE, Bedoya CD. 2006. Caracterización del crecimiento y desarrollo vegetativo de especies promisorias del género *Theobroma* bajo condiciones de la Amazonia colombiana. 1st edn. Bogotá, Colombia: *Theobroma*; p. 65–100.
- Bayer C, Fay MF, Bruijn AY, Savolainen V, Morton CM, Kubitzki K, Alverson WS, Chase MW. 1999. Support for an expanded family concept of Malvaceae within a circumscribed order Malvales: a combined analysis of plastid atpB and rbcL DNA sequences. *Bot J Linn Soc*. 129(4):267–303. doi:10.1111/j.1095-8339.1999.tb00505.x.
- Colli-Silva M, Pirani JR. 2020. Estimating bioregions and undercollected areas in South America by revisiting Byttnerioideae, Helicteroideae and Sterculioideae (Malvaceae) occurrence data. *Flora*. 271:151688. doi:10.1016/j.flora.2020.151688.
- Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, et al. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 36: W465–W469. doi:10.1093/nar/gkn180.
- Febrianto NA, Zhu F. 2022. Comparison of bioactive components and flavor volatiles of diverse cocoa genotypes of *Theobroma grandiflorum*, *Theobroma bicolor*, *Theobroma subincanum* and *Theobroma cacao*. *Food Res Int*. 161:111764. doi:10.1016/j.foodres.2022.111764.
- Gálvez-Marroquín L, Reyes-Reyes A, Avendaño-Arrazate C, Hernández-Gómez E, Díaz-Fuentes VH, Mendoza-López A. 2016. Pataxte (*Theobroma bicolor* Humb. & Bonpl.): especie subutilizada en México. *Agroproductividad*. 9:41–47.
- GBIF. 2020. *Theobroma bicolor* Humb. & Bonpl. in GBIF Secretariat. Global Biodiversity Information Facility Backbone Taxonomy. Checklist dataset doi:10.15468/dl.vjk22s%0A. Accessed via GBIF.org
- González AA, Moncada J, Idarraga A, Rosenberg M, Cardona CA. 2016. Potential of the amazonian exotic fruit for biorefineries: the *Theobroma bicolor* (Makambo) case. *Ind Crops Prod*. 86:58–67. doi:10.1016/j.indcrop.2016.02.015.
- González A, Torres G. 2010. Manual Cultivo de Macambo. 1st ed. Perú: IIAP.
- Govaerts R, Nic Lughadha E, Black N, Turner R, Paton A. 2021. The world checklist of vascular plants, a continuously updated resource for exploring global plant diversity. *Sci Data*. 8(1):215. doi:10.1038/s41597-021-00997-6.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780. doi:10.1093/molbev/mst010.
- Kufer J, McNeil C. 2006. The jaguar tree (*Theobroma bicolor* Bomp.). A cultural history of Cacao. In: *Chocolate in Mesoamerica*, Cameron MC. Gainesville: University Press of Florida; p. 90–104.
- Kuhn DN, Figueira A, Lopes U, Motamayor JC, Meerow AW, Cariaga K, Freeman B, Livingstone DS, Schnell RJ. 2010. Evaluating *Theobroma grandiflorum* for comparative genomic studies with *Theobroma cacao*. *Tree Genet. Genomes*. 6(5):783–792. doi:10.1007/s11295-010-0291-0.
- Lagneaux E, Andreotti F, Neher CM. 2021. Cacao, copoazu and macambo: Exploring *Theobroma* diversity in smallholder agroforestry systems of the Peruvian Amazon. *Agroforest Syst*. 95(7):1359–1368. doi:10.1007/s10457-021-00610-0.
- Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast-single-node solution for large and complex metagenomics assembly

- via succinct de Bruijn graph. *Bioinformatics*. 31(10):1674–1676. doi:10.1093/bioinformatics/btv033.
- Liu J, Niu YF, Ni SB, He XY, Zheng C, Liu ZY, Cai HH, Shi C. 2018. The whole chloroplast genome sequence of *Macadamia tetraphylla* (Proteaceae). *Mitochondrial DNA B Resour*. 3(2):1276–1277. doi:10.1080/23802359.2018.1532836.
- Lowe TM, Chan PP. 2016. tRNAscan-SE on-line: search and contextual analysis of transfer RNA Genes. *Nucleic Acids Res*. 44(W1):W54–W57. doi:10.1093/nar/gkw413.
- Mantilla L, Piñeres R, Fonseca D. 2008. Colombia Frutas de la amazonia. Bogotá, Colombia: Instituto Sinchi.
- Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 37(5):1530–1534. doi:10.1093/molbev/msaa015.
- Narkthai P, Chantaranonthai P. 2020. Taxonomic notes on the subfamily Grewioideae (Malvaceae) in Thailand. *Thai Forest Bull*. 48(1):72–76. doi:10.20531/tfb.2020.48.1.12.
- Niu YF, Ni SB, Liu J. 2019. The complete chloroplast genome of *Theobroma grandiflorum*, an important tropical crop. *Mitochondrial DNA B Resour*. 4(2):4157–4158. doi:10.1080/23802359.2019.1693291.
- Ponce-Sánchez J, Zurita-Benavides MG, Peñuela MC. 2021. Reproductive ecology of white cacao (*Theobroma bicolor* Humb. & Bonpl.) in Ecuador, western Amazonia: floral visitors and the impact of fungus and mistletoe on fruit production. *Braz J Bot*. 44:479–489.
- Rogalski M, do Nascimento VL, Fraga HP, Guerra MP. 2015. Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. *Front Plant Sci*. 6: 586. doi:10.3389/fpls.2015.00586.
- Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res*. 44(W1):W232–W235. doi:10.1093/nar/gkw256.
- Torres DEG, Assunção D, Mancini P, Torres RP, Mancini-Filho J. 2002. Antioxidant activity of macambo (*Theobroma bicolor* L.) extracts. *Eur J Lipid Sci Technol*. 104(5):278–281. doi:10.1002/1438-9312(200205)104:5<278::AID-EJLT278>3.0.CO;2-K.
- Yang N, Jingling L, Chang Z, Chang L. 2023. Generating sequencing depth and coverage map for organelle genomes. *Protocols io*. doi:10.17504/protocols.io.4r3l27jxg1y/v1.
- Yang T, Sahu SK, Yang L, Liu Y, Mu W, Liu X, Strube ML, Liu H, Zhong B. 2022. Comparative analyses of 3,654 plastid genomes unravel insights into evolutionary dynamics and phylogenetic discordance of green plants. *Front Plant Sci*. 13:808156. doi:10.3389/fpls.2022.808156.