JSE Journal of Systematics and Evolution

doi: 10.1111/jse.13045

Research Article

A phylogenetic framework to study desirable traits in the wild relatives of *Theobroma cacao* (Malvaceae)

Ana M. Bossa-Castro[1*], Matheus Colli-Silva[2,3], José R. Pirani[2], Barbara A. Whitlock[4], Laura T. Morales Mancera[1], Natalia Contreras-Ortiz[5,6], Martha L. Cepeda-Hernández[7,8], Federica Di Palma[9,10], Martha Vives[1], and James E. Richardson[5,11,12,13]

[1]Departamento de Ciencias Biológicas, Facultad de Ciencias, Universidad de los Andes, Carrera 1 18A-12, Bogotá, Colombia
[2]Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, Rua do Matão 277, São Paulo 05508-090, São Paulo, Brazil
[3]Royal Botanic Gardens, Kew, Richmond, Surrey, UK
[4]Department of Biology, University of Miami, 1301 Memorial Drive, Coral Gables, Florida 33146, USA
[5]Tropical Diversity Section, Royal Botanic Garden Edinburgh, Edinburgh EH3 5NZ, United Kingdom
[6]Department of Molecular Plant Science, University of Edinburgh, Old College, South Bridge, Edinburgh EH8 9YL, UK
[7]Facultad de Ciencias, Universidad de los Andes, Carrera 1 18A-12, Bogotá, Colombia
[8]Corporación Corpogen, Carrera 4 20-41, Bogotá, Colombia
[9]School of Biological Sciences, University of East Anglia, Norwich NR4 7TU, UK
[10]Genome British Columbia, 575 W 8th Ave 400, Vancouver BC V5Z 0C4, Canada
[11]School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland
[12]Environmental Research Institute, University College Cork, Ellen Hutchins Building, Lee Road, Cork T23 XE10, Ireland
[13]Departamento de Biología, Facultad de Ciencias Naturales, Universidad del Rosario, Calle 12C, 6-25, Bogotá, Colombia
*Author for correspondence. E-mail: ana.bossa@alumni.colostate.edu
Received 26 September 2023; Accepted 27 November 2023

Abstract Crop wild relatives (CWRs) of cultivated species may provide a source of genetic variation that can contribute to improving product quantity and quality. To adequately use these potential resources, it is useful to understand how CWRs are related to the cultivated species and to each other to determine how key crop traits have evolved and discover potentially usable genetic information. The chocolate industry is expanding and yet is under threat from a variety of causes, including pathogens and climate change. *Theobroma cacao* L. (Malvaceae), the source of chocolate, is a representative of the tribe Theobromateae that consists of four genera and c. 40 species that began to diversify over 25 million years ago. The great diversity within the tribe suggests that its representatives could exhibit advantageous agronomic traits. In this study, we present the most taxonomically comprehensive phylogeny of Theobromateae to date. DNA sequence data from *WRKY* genes were assembled into a matrix that included 56 morphological characters and analyzed using a Bayesian approach. The inclusion of a morphological data set increased resolution and support for some branches of the phylogenetic tree. The evolutionary trajectory of selected morphological characters was reconstructed onto the phylogeny. This phylogeny provides a framework for the study of morphological and physiological trait evolution, which can facilitate the search for agronomically relevant traits.

Key words: cacao, crop wild relatives, *Herrania*, Malvaceae, morphological and molecular characters, phylogeny, *Theobroma*, trait evolution.

1 Introduction

1.1 Phylogenies and genetic diversity in crop improvement

Crops face a myriad of threats, including diseases and climate change. They often have limited genetic diversity due to the selection of a few individuals with desirable traits, making them more vulnerable to these threats (Hollingsworth et al., 2005; Maxted & Kell, 2009; Flint-Garcia, 2013; Mammadov et al., 2018). Therefore, it is crucial to assess the genetic diversity of crop species (e.g., Osorio-Guarín et al., 2017) and explore the potential use of genetic diversity from wild relatives (Cortés et al., 2022) to ensure food security. Crop wild relatives (CWRs) encompass close relatives of domesticated species used in agriculture (Maxted et al., 2006; Maxted & Kell, 2009). Leveraging the genetic diversity of CWRs offers the opportunity to expand the gene pool for crop improvement through traditional breeding methods or allele mining (Brozynska et al., 2016), as well as

through innovative technologies such as CRISPR-Cas9 (Zhang et al., 2014).

Crop improvement requires a fundamental understanding of the biology of the cultivated species, and phylogenies provide contextual insights into the evolutionary history of genes or gene families linked to traits of interest. Advancements in high-throughput genotyping and phenotyping methods have facilitated the unravelling of the genetic architecture underlying complex traits associated with abiotic and biotic stress, such as drought and disease resistance. For instance, genes like *DRO-1*, *ERECTA*, and other transcription regulators (*AP2/ERF*, *ZFPs*, *WRKY*, and *MYB*) have been characterized for their roles in drought tolerance in wheat (*Triticum* spp., Poaceae), and an understanding of those genes can contribute to enhancing its response to drought (Kulkarni et al., 2017). Genomic technologies have also been employed in wheat to identify genes associated with pathogen resistance (Periyannan et al., 2013; Saintenac et al., 2013). McElroy et al. (2018) conducted a comparison between genome-wide association studies (GWAS) and genomic selection (GS) in three cacao populations for predicting resistance to frosty pod rot and witches' broom disease, concluding that GS exhibited higher efficiency. While comprehensive genomic information allows screening for resistance in related species, understanding the evolutionary aspects of these traits would require the incorporation of a phylogeny.

The combination of phylogenies with knowledge about the medicinal uses of particular species can also assist in detecting clades or species groups that may contain valuable chemical compounds. By employing artificial intelligence models alongside phylogenies, drug discovery efforts can be accelerated by identifying species that may harbor molecules with high medicinal potential (Saslis-Lagoudakis et al., 2012). A similar approach can be applied to identify CWRs with agronomically important traits. This information can guide and facilitate the transfer of appropriate genes associated with these traits from wild relatives to target crops.

1.2 The cacao group as a case study

Chocolate, derived from *Theobroma cacao* L. beans, is a thriving industry projected to reach a value of US$190 billion by 2026 (Voora et al., 2019). The industry supports over five million farmers globally (Voora et al., 2019) and is facing an ever-increasing demand (Grilli et al., 2022). This industry faces various threats from pathogens, including black-pod rot (caused by *Phytophthora* spp.) and witches' broom and frosty pod rot (caused by species of *Moniliophthora* H.C. Evans, Stalpers, Samson & Benny), as well as emerging diseases like vascular streak die-back (caused by the basidiomycete *Ceratobasidium theobromae* (Talbot & Keane) Samuels & Keane) and the viral swollen shoot disease, found in Southeast Asia and West Africa, respectively (Marelli et al., 2019). The devastating impact of the 1980s–1990s witches' broom outbreak in Brazil particularly underscores the industry's vulnerability (Evans, 2007).

Climate change also poses a great risk to crops such as cacao in both native and cultivated ranges. Much needs to be learned about how plants might adapt to the abiotic stresses imposed by climate change (Anderson & Song, 2020). Historical studies have indicated that plants have been unable to adapt to changes even when those changes were occurring over large time scales. For instance, a lineage of Annonaceae, adapted to wet forests in Africa, contracted its range due to aridification that occurred over periods of millions of years (Couvreur et al., 2008). Crops like *Coffea arabica* L. (Rubiaceae) are already at risk as suitable cultivation areas shrink (Moat et al., 2017). Davis et al. (2012) also highlighted the threats posed by climate change to native populations of *C. arabica*. While cultivated cacao can tolerate less humid climates with irrigation, it generally performs poorly, even in short dry seasons, without ample shade and local humidity (Cuatrecasas, 1964). Läderach et al. (2013) highlight the potential impact of climate change on cacao in West Africa, where warming and reduced moisture are predicted up to 2100 (Anjos et al., 2021). Recent research in Brazil has predicted reduced areas for cacao cultivation under various climate change scenarios (Igawa et al., 2022). Inspiration comes in the form of studies on wild relatives of coffee. Davis et al. (2021) have studied the properties of *Coffea stenophylla* G. Don, a wild species from West Africa that is similar in terms of sensory profile to Arabica coffee, yet has the capacity to grow at higher temperatures. Species related to those that have been traditionally used to produce coffee, therefore, have the potential to replace or augment production from Arabica or to provide the genetic resources necessary to produce climate change resilient forms.

Another issue for the cacao industry is the low number of flowers that successfully set fruit that might be related, at least in part, to pollination success (Van Synghel et al., 2022). An understanding of the floral morphology of relatives of *T. cacao* might provide the information that would allow us to make pollination in cacao more efficient. An additional and more recent problem has arisen in the cacao industry, that is, market-imposed limits on the amount of the heavy metal cadmium permitted in chocolate products (European Commission, 2014). Cacao CWRs may have variants that are resistant to fungal pathogens, that may already have become better adapted to more extreme climates (González-Orozco et al., 2022), or that are low cadmium accumulators. CWRs may be good candidates for searching for the genetic apparatus that could be utilized to improve performance in cacao now and in the future (González-Orozco et al., 2020).

Theobroma cacao is a tree that grows to about 12 m in height. Its native range is the Neotropical rainforests, and it was originally domesticated at least 5500 years ago in south-eastern Ecuador and Montegrande, Jaen, Peru (Valdez, 2013; Ochoa, 2017; de laFuente deDiez Canseco, 2018; Olivera-Núñez, 2018; Zarrillo et al., 2018). Taxonomically, cacao is classified within the Malvaceae, subfamily Byttnerioideae and in the tribe Theobromateae, including four genera (Whitlock et al., 2001). Among these, *Theobroma* L. (23 species), *Herrania* Goudot (17 species), and *Guazuma* Adans. (three species) are native to the Neotropics (Colli-Silva & Pirani, 2020), whereas *Glossostemon* Desf. is found only in the deserts of the Middle East (Ali, 2020). *Guazuma ulmifolia* Lam. is found mostly in dry forests, whereas *Guazuma crinita* Mart., *Theobroma*, and *Herrania* are restricted to wet forests and are thus likely vulnerable to any form of climatic change that would

reduce rainfall. *Theobroma* was classified by Cuatrecasas (1964) into six sections: *T*. sect. *Andropetalum* Cuatrec., *T*. sect. *Glossopetalum* Bernoulli, *T*. sect. *Oreanthes* Bernoulli, *T*. sect. *Rhytidocarpus* Bernoulli, *T*. sect. *Telmatocarpus* Bernoulli, and *T*. sect. *Theobroma* that are distinguished based on morphological characters of branch growth, flowers, and fruits. Schultes (1958) recognized two sections of *Herrania*, *H*. sect. *Herrania* and *H*. sect. *Subcymbicalyx* R.E. Schult., based on sepal features. Finally, Freytag (1951) placed the species of the genus *Guazuma* into sections *G*. sect. *Gynophoriola* Freytag and *G*. sect. *Euguazuma* K. Schum., considering features of leaves, flowers, and their geographic distribution.

Whitlock & Baum (1999) were the first to perform a broad-scale molecular phylogenetic analysis of members of Theobromateae using 11 species of *Theobroma* and seven species of *Herrania*. They used the *vicilin* gene that codes for a protein present in seeds (Whitlock & Baum, 1999) with both genera recovered as monophyletic, but only *Herrania* was well supported. Whitlock et al. (2001) used plastid *ndh*F sequences that resulted in a polytomy of three *Theobroma* species with a clade of two *Herrania* species. Subsequently, Sousa Silva & Figueira (2004) used another seed protein gene (trypsin), sampling 11 *Theobroma* and three *Herrania* species, but no outgroup. Borrone et al. (2007) used the same species sample as Whitlock & Baum (1999) and five *WRKY* transcription factor loci. Genetic mapping of the *WRKY3, WRKY11*, and *WRKY14* indicated that they belonged to separate linkage groups in the nuclear genome (Borrone et al., 2004). Analyses of each of the *WRKY3* and *WRKY13* genes resulted in a paraphyletic *Theobroma*, with *Herrania* nested within it, whereas analyses of the other *WRKY* genes resolved both genera as monophyletic.

Hence, phylogenetic relationships between both genera, as well as within their respective sections, remain somewhat unclear and would benefit from the addition of more species. Acquiring specimens of all species through fieldwork is often challenging and expensive, particularly when species are only known from restricted localities, as is the case for Theobromateae. However, recent developments in the use of ancient DNA have opened the possibility of sourcing DNA from herbarium specimens (Sarkinen et al., 2012; Staats et al., 2013). Richardson et al. (2015) used data from Borrone et al. (2007) to place the evolution of the tribe in spatiotemporal context that also helps in understanding the evolutionary processes within the tribe. The *c*. 25-million-year timeframe for the evolution of Theobromateae, predominantly in the Neotropics, has seen many geological and climatic changes to which its representatives have either had to adapt to or have modified their geographic distributions to continue living under the same conditions. Although the genome of cacao itself has long been published (Argout et al., 2011, 2017; Motamayor et al., 2013), much research is still required to phenotypically and genotypically characterize its wild relatives.

In this study, we aimed to reconstruct the phylogeny of Theobromateae with as many species as possible using *WRKY* DNA sequences and morphological data. We characterize the evolutionary trajectory of morphological characters and provide a framework that permits the study of the evolution of key agronomic traits in cacao and its wild relatives.

2 Material and Methods

2.1 Tissue sampling, DNA extraction, and sequencing

Herbarium specimens from the genera *Glossostemon*, *Guazuma*, *Herrania*, and *Theobroma* were sampled from the following collections: ANDES, COL, NY, and F (Table 1). Approximately 20 mg per sample was collected and placed in a 2 mL tube with 30 mg of polyvinylpolypyrrolidone. Plant tissue was ground with three glass beads using a TissueLyser II (Qiagen, Valencia, CA, USA). DNA was extracted using a modified celyltrimethylammonium bromide-based extraction protocol (Doyle & Doyle, 1990), in which an extra chloroform-isoamyl alcohol step was added, as well as an extra wash. DNA was finally eluted in 50 μL 1× Tris-EDTA buffer. DNA yield and integrity were assessed using Nanodrop 2000 (Thermo Scientific, Waltham, MA, USA) and Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, USA), as well as 0.8% agarose gel electrophoresis. For samples with a 260/280 ratio higher than 2.0, RNase treatment was conducted by adding RNAse to the elution (ThermoFisher Scientific Inc., Carlsbad, CA, USA), incubating DNA at 37 °C for 30 min, according to manufacturer instructions.

Each *WRKY* locus was amplified separately from each template. The *WRKY* gene family of transcription factors is involved in many pathways of regulatory networks in many plant species (Eulgem et al., 2000; Borrone et al., 2007; Rushton et al., 2010), containing highly conserved DNA binding domains (which can depict deep phylogenetic relationships) interrupted by introns (which can carry variation at lower taxonomic levels). Their potential to unravel phylogenetic relationships has been successfully discussed and shown, specifically with *Theobroma* (Borrone et al., 2007). We used previously designed primers (Borrone et al., 2007) to amplify genomic regions adding a nested PCR approach, as first-round amplifications did not yield sufficient DNA for sequencing. New primers were designed for this approach using Primer-BLAST (Table 2) and cross-checking against the reference genome *T. cacao* Matina (Motamayor et al., 2013), downloaded from phytozome 13 website (Goodstein et al, 2012). The first PCR contained: 0.2 μM of the forward primer, 0.2 μM of the reverse primer, 2 mg/mL of bovine serum albumin, 12.5 μL of GoTaq Green Master Mix (Promega, Madison, WI, USA), and 200–250 ng of DNA in a total volume of 25 μL. Amplifications were conducted using a professional TRIO thermocycler (Analytik Jena, Jena, Germany) with the following conditions: 95 °C, 2 min; (95 °C, 30 s; 56–57 °C, 40 s; 72 °C, 60–120 s) × 30 cycles; 72 °C, 7 min; 4 °C, hold. The second PCR contained: 0.2 μM of the forward primer, 0.2 μM of the reverse primer, 1 μM of bovine serum albumin, 12.5 μL of GoTaq Green Master Mix (Promega), and 5 μL of the product of PCR1 in a total volume of 25 μL. As in the first PCR, amplifications were conducted using a thermocycler; the conditions were 95 °C, 2 min; (95 °C, 30 s; 56–57 °C, 40 s; 72 °C, 60–120 s) × 25 or 35 cycles; 72 °C, 7 min; 4 °C, hold. Amplification success was determined by agarose gel electrophoresis. The PCR products were cleaned using EXOSAP (GE Healthcare, Chicago, IL, USA). The samples were sequenced using Sanger technology in an ABI PRISM 3500 XL® sequencer (Applied Biosystems, Foster City, CA, USA) at the Gencore sequencing center (Universidad de los Andes, Bogotá, Colombia). We were unable to obtain

Table 1 Voucher specimens and GenBank accession numbers of the *WRKY* sequences used

Species	Voucher	Country/Year	Collection*	GenBank Accession Nos.				
				WRKY-03	*WRKY-11*	*WRKY-12*	*WRKY-13*	*WRKY-14*
Glossostemon bruguieri Desf.	J.S. Collenette 2411	Saudi Arabia, 1981	K	OR045314	_____	OR045328	_____	OR045339
Guazuma crinita Mart.	O.J.M. Samor s.n.	Brazil, 2014	NY	OR045315	_____	OR069740	OR091167	_____
G. longipedicellata Freytag	M. Carlson 651	El Salvador, 1946	F	_____	OR069739	_____	_____	OR045340
G. ulmifolia Lam.	B. Whitlock & M. Bowser 360	*sine loco*, s.d.	GH	EF640168	EF640237	EF640191	EF640214	EF640260
Herrania albiflora Goudot	TC01377	Costa Rica, s.d.	CATIE	EF640169	EF640238	EF640192	EF640215	EF640261
H. balaensis Preuss	A. Rimbach 20741	Ecuador, s.d.	F	_____	OR069742	OR045329	_____	OR045341
H. cuatrecasasiana García-Barr	B. Whitlock 312	*sine loco*, s.d.	GH	EF640170	EF640239	EF640193	EF640216	_____
H. dugandii García-Barr	R.E. Schultes 6038	*sine loco*, 1944	COL	_____	_____	OR045330	_____	OR045342
H. kanukuensis R.E. Schult.	S. Mori 24727	*sine loco*, s.d.	NY	EF640171	EF640240	EF640194	EF640217	EF640262
H. laciniifolia Goudot	D. Sanín et al. 4454	Colombia, 2011	COL	_____	_____	_____	_____	OR045343
H. lemniscata (Schomb.) R.E. Schult.	D. Clarke 1555	Guyana, 1996	NY	OR045316	OR069741	OR045331	OR091168	OR045344
H. mariae (Mart.) Goudot	F. Woytkowski 5130	Peru, 1958	F	OR045317	OR069743	_____	_____	_____
H. nitida (Poepp.) R.E. Schult.	TC01371	Costa Rica, s.d.	CATIE	EF640172	EF640241	EF640195	EF640218	EF640263
H. nycterodendron R.E. Schult.	B. Whitlock 315	*sine loco*, s.d.	GH	EF640173	EF640242	EF640196	EF640219	EF640264
H. pulcherrima Goudot	R. Jaramillo-Mejía 202	Colombia, 1987	F	OR045318	OR069744	OR045332	_____	OR045345
H. purpurea (Pitt.) R.E. Schult. (1)	B. Whitlock 318	*sine loco*, s.d.	GH	EF640174	EF640243	EF640197	EF640220	EF640265
H. purpurea (2)	T.B. Croat et al. 16838	Panama, 1971	F	OR045319	_____	_____	_____	_____
H. umbratica R.E. Schult.	TC01376	Costa Rica, s.d.	CATIE	EF640175	EF640244	EF640198	EF640221	EF640266
Theobroma angustifolium DC (1)	B. Whitlock 303	*sine loco*, s.d.	GH	EF640176	EF640245	EF640199	EF640222	EF640267
T. angustifolium (2)	TC01368	Costa Rica, s.d.	CATIE	EF640177	EF640246	EF640200	EF640223	EF640268
T. bernouillii Pittier	A. Lucas s.n.	Panama, 1949	F	OR045320	_____	_____	_____	_____
T. bicolor Humb. & Bonpl.	Hunter 1029	*sine loco*, s.d.	WIS	EF640178	EF640247	EF640201	EF640224	EF640269
T. cacao L. cv. TSH516 (1)	TC00157	Brazil, s.d.	CEPLAC	EF640179	EF640248	EF640202	EF640225	EF640270
T. cacao (2)	B. Whitlock 361	*sine loco*, s.d.	GH	EF640180	EF640249	EF640203	EF640226	EF640271
T. cacao (3)	B. Whitlock s.n.	*sine loco*, s.d.	GH	EF640181	EF640250	EF640204	EF640227	EF640272
T. cacao (4)	C. González et al. s.n.	Colombia, 2019	ANDES	_____	_____	_____	_____	OR045346
T. chocoense Cuatrec.	B. Whitlock 356	*sine loco*, s.d.	GH	EF640182	EF640251	EF640205	EF640228	EF640273
T. cirmolinae Cuatrec.	V.M. Patiño 241	Colombia, 1963	COL	OR045321	_____	_____	_____	OR045347
T. gileri Cuatrec.	B. Whitlock 301	*sine loco*, s.d.	GH	EF640183	EF640252	EF640206	EF640229	EF640274
T. glaucum H. Karst.	s.c. [SEF] 8525	Ecuador, 1982	NY	OR045322	_____	OR045333	OR091169	OR045348
T. grandiflorum (Willd. ex Spreng.) K. Schum.	B. Whitlock 305	*sine loco*, s.d.	GH	EF640184	EF640253	EF640207	EF640230	EF640275
T. hylaeum Cuatrec.	H. Pittier 4194	Panama 1911	NY	OR045323	_____	OR045334	_____	OR045349
T. mammosum Cuatrec. & J. Léon	TC01367	Costa Rica, s.d.	CATIE	EF640185	EF640254	EF640208	EF640231	EF640276
T. microcarpum Mart. (1)	B. Whitlock 302	*sine loco*, s.d.	GH	EF640186	EF640255	EF640209	EF640232	EF640277
T. microcarpum (2)	TC01369	Costa Rica, s.d.	CATIE	EF640187	EF640256	EF640210	EF640233	EF640278
T. obovatum Klotzsch ex Bernoulli	B.A. Krukoff 1668	Brazil, 1931	F	OR045324	OR069745	OR045335	_____	OR045350
T. simiarum Donn. Sm. (1)	B. Whitlock 321	*sine loco*, s.d.	GH	EF640188	EF640257	EF640211	EF640234	EF640279

Continued

Table 1 Continued

Species	Voucher	Country/Year	Collection*	GenBank Accession Nos.				
				WRKY-03	*WRKY-11*	*WRKY-12*	*WRKY-13*	*WRKY-14*
T. simiarum (2)	P.A. Teunissen 16785	Suriname, 1981	NY					OR045351
T. speciosum Willd. ex Spreng	Hunter 1033	*sine loco*, s.d.	WIS	EF640189	EF640258	EF640212	EF640235	EF640280
T. sinuosum Pav.	J. Jaramillo et al. 13459	Ecuador, 1990	NY	OR045325		OR045336	OR091170	OR045352
T. subincanum Mart.	C.C. Berg et al. P19856	Brazil, 1973	NY	OR045326	OR069746	OR045337	OR091171	OR045353
T. sylvestre Mart.	J.M. Pires et al. 51263	Brazil, 1961	F	OR045327		OR045338		OR045354
T. velutinum Benoist	S. Mori 24731	*sine loco*, s.d.	NY	EF640190	EF640259	EF640213	EF640236	EF640281

*Botanical collections abbreviations: Royal Botanic Gardens, Kew (K), New York Botanical Garden (NY), Field Museum (F), Harvard University Herbaria (GH), Tropical Agricultural Research and Training Center (CATIE), Herbario Nacional Colombiano (COL), Wisconsin State Herbarium (WIS), Herbário Centro de Pesquisas do Cacau (CEPLAC/CEPEC), Museo de Historia Natural C.J. Marinkelle (ANDES).

sequence data from the following taxa because of low-quality DNA: *Theobroma stipulatum* Cuatrec., *Theobroma canumanense* Pires & Fróes ex Cuatrec., *Theobroma nemorale* Cuatrec., *Herrania breviligulata* R.E. Schult., *Herrania kofanorum* R.E. Schult., *Herrania camargoana* R.E. Schult., and *Herrania tomentella* R.E. Schult.

2.2 Morphological matrix

The construction of the morphological matrix was based on a comprehensive review of the literature (Freytag, 1951; Schultes, 1958; Cuatrecasas, 1964; Ali, 2020) and on an examination of preserved and spirit specimen collections from around the world. We particularly focused on botanical collections A, COAH, COL, ECON, F, FMB, HUA, GH, INPA, K, L, MG, MO, NY, RB, U, US, and WAG (herbarium acronyms following Thiers, 2022 [continuously updated]), along with supplementary data from living collections and plants gathered during field expeditions in Amazonia, as well as information on species occurrences found in other preserved specimen collections (Colli-Silva et al., 2023). This comprehensive survey enabled the establishment of a morphological data set comprising 56 characters (Table 3). The morphological traits considered crucial for specific delimitations included branch and leaf architecture, indumentum, as well as the position and number of flowers per inflorescence. Additionally, we considered corolla, calyx, and androecium variation, including morphology, color, indumentum, and dimensions, as well as fruit shape and indumentum, seed germination, and wood anatomy.

Character coding was derived from various sources, including the works of Ali (2020), Freytag (1951), Schultes (1958), Cuatrecasas (1964), and recent specimens that were not available to those authors. For leaf size and dimensions, we utilized the leaf blade area classes described in Table 3, following Ellis et al. (2009). Regarding petal structure, we adopted terminologies used by Schultes (1958) and Cuatrecasas (1964) in *Theobroma*, *Herrania*, and *Guazuma*, in which the petals are described as consisting of two parts connected by a narrow joint. The lower part comprises the petal "claw" or "hood" and is cucullate and erect, with 1-3-7 prominent ridges on the adaxial surface that we refer to as guidelines. The upper portion of the petal, referred to as the "ligule" in this study, is membranous, flat, and linear at the base, sometimes expanding toward the apex. Not all species of *Theobroma* have a prominent ligule. Petals are flat in *Glossostemon*, and lack the distinct claw, ligule, and joint between the two that is seen in other Theobromateae.

Interpretation of the androecium morphology of Theobromateae and other Byttnerioideae varies. In *Theobroma*, *Herrania*, and *Guazuma*, dithecate anthers occur in antipetalous groups, borne on filaments with varying degrees of connation. The filaments of these antipetalous groups alternate with staminodes, together forming a short staminal tube. Here, we consider each dithecate anther to represent a stamen. *Theobroma cacao* thus has 10 stamens, in five antipetalous pairs, with filaments of each pair connate for most of their length. This arrangement is referred to as two-antheriferous. Alternatives include an androecium of 15 or 20 stamens in five antipetalous pairs that are referred to as three- or four-antheriferous (triads or tetrads), respectively.

Table 2 Primers used to amplify *WRKY* genes

Locus name	Gene identifier*	Primer name	Nested	Sequence (5′–3′)	Tm (°C)	Expected size (bp)	Source
TcWRKY03	Thecc.02G338600	FPW12-3		TCCTTACCCAAGGTAATGCCCCTG	57.9	649	Borrone et al. (2007)
		RPW12-3		TGCTTACGGACGTTGCATCCT	56.5		Borrone et al. (2007)
		FPW12-3				736	
		RPW12-3-2	External	CTAGTTTTGGCAGCTGGTACGTC			This study
TcWRKY11	Thecc.05G199200	Tc11pF		GGTAGTGAATATCCAAGAAGC	56.8	1055	Borrone et al. (2007)
		Tc11pR		ACAGGACATCCAGGAGTTG	60.1		Borrone et al. (2007)
		Tc11pF2	Internal	CAGGGATATTAAGCTTGGGTC		795	This study
		Tc11pR					
TcWRKY12	Thecc.07G022800	Tc12pF		ACGCATCCTAATTGTGAAGTG	61.5	893	Borrone et al. (2007)
		Tc12pR		TTTTCTAACAGGGCAACCG	62.5		Borrone et al. (2007)
		Tc12pF				982	
		Tc12pR2	External	CTGGTCCTTGCAGTAGGTAC			This study
TcWRKY13	Thecc.05G051600	Tc13pF		AAGCAAGTGAAAGGAAGTGAG	60.1	1165	Borrone et al. (2007)
		Tc13pR		TGAAAGCTCTTGGATCATCCGATGC	72.7		Borrone et al. (2007)
		Tc13pF2	Internal	CTGAAATTGTCTACAAGGGTG		1062	This study
		Tc13pR					
TcWRKY14	Thecc.01G176500	Tc14pF		GCCAAGGGAAATCCATGTC	64.2	481	Borrone et al. (2007)
		Tc14pR		GGATTGTTCGGTCTTCTGC	62.2		Borrone et al. (2007)
		Tc14pF				581	
		Tc14pR2	External	CCTGATAGTAGCATTCGTGC			This study

*Gene identifiers according to *Theobroma cacao* Matina reference genome (Motamayor et al., 2013).

Table 3 Morphological characters and character states used for ancestral character state reconstructions

1. Lifeform: (0) arboreal; (1) shrubby-herbaceous
2. Branching architecture: (0) monopodial; (1) sympodial
3. Stipules: (0) caducous; (1) persistent
4. Stipule shape: (0) linear to subulate; (1) elliptic, broadened
5. Stipule length: (0) up to 1.5 cm long; (1) more than 1.5 cm long
6. Leaf blade division: (0) simple; (1) palmately-compound
7. Petiole length: (0) up to 3 cm long; (1) 3–36 cm long; (2) more than 36 cm long
8. Ratio petiole-midrib width: (0) >1.0; (1) ~1.0
9. Pulvinus at the petiole apex: (0) absent; (1) present
10. Leaflet attachment (when applicable): (0) sessile; (1) petiolulate
11. Leaf(-let) area size class (Webb, 1959; Ellis et al., 2009): (0) leptophyll (<25 mm^2); (1) nanophyll (25–225 mm^2); (2) microphyll (225–2025 mm^2); (3) notophyll (2025–4500 mm^2); (4) mesophyll (4500–18 225 mm^2); (5) macrophyll (18 225–164 025 mm^2); (6) megaphyll (>164 025 mm^2)
12. Leaf blade shape: (0) elliptic-oblong; (1) ovate; (2) obovate; (3) linear; (4) lanceolate; (5) special
13. Leaf(-let) blade basal width: (0) symmetric; (1) asymmetric
14. Leaf(-let) blade basal extension: (0) symmetric; (1) asymmetric
15. Leaf(-let) blade lobation: (0) absent; (1) present
16. Leaf(-let) blade margin: (0) untoothed; (1) partially toothed; (2) entirely toothed
17. Leaf(-let) blade base shape: (0) straight or cuneate; (1) concave; (2) convex; (3) decurrent; (4) cordate
18. Times the acumen is longer than the leaf length: (0) <10×; (1) 10–40×; (2) >40×
19. Number of leaf basal veins: (0) 1–2; (1) 3 or more
20. Leaf(-let) primary vein framework: (0) craspedodromous; (1) eucamptodromous; (2) brochidodromous
21. Leaf(-let) margin outreach extension: (0) inconspicuous; (1) conspicuous, <2 mm long; (2) conspicuous, ≥2 mm long
22. Major secondary vein spacing: (0) regular; (1) irregular; (2) decreasing proximally; (3) gradually increasing proximally; (4) abruptly increasing proximally
23. Major secondary attachment to midvein: (0) decurrent; (1) proximal secondaries decurrent; (2) excurrent; (3) deflected.
24. Intersecondary vein proximal course: (0) parallel to major secondaries; (1) perpendicular to midvein.
25. Leaf(-let) blade indumentum on abaxial surface: (0) glabrous; (1) slightly pubescent; (2) densely pubescent
26. Leaf(-let) blade indumentum composition on abaxial surface: (0) homotrichous, stellate trichomes only; (1) heterotrichous, simple and stellate trichomes; (2) heterotrichous, two layers of stellate trichomes of different sizes; (3) glabrous
27. Texture on abaxial surface of the leaf(-let) blade: (0) velvety; (1) asperous
28. Inflorescence location: (0) cauliflorous; (1) ramiflorous; (2) terminal
29. Flower number per inflorescence: (0) 1–3-flowered; (1) 4–10-flowered; (2) many-flowered
30. Calyx color externally: (0) yellowish-green; (1) ochraceous to ferrugineous; (2) crimson; (3) purple
31. Sepal number: (0) 5; (1) 3
32. Calyx inclusion: (0) cupuliform; (1) patelliform
33. Petal shape: (0) cucullate at base; (1) flat throughout
34. Petal hood guidelines: (0) 1–2; (1) 3; (2) 4–6; (3) 7 more
35. Corolla indumentum: (0) glabrous; (1) sparsely pubescent; (2) densely pubescent
36. Petal color: (0) yellow; (1) pink to light red; (2) crimson; (3) purple; (4) cream
37. Petal claw–ligule ratio: (0) ligule smaller than the claw; (1) ligule almost the same size of the claw; (2) claw up to 2× longer than the claw; (3) ligule 2–10× longer than the claw; (4) ligule more than 10× longer than the claw.
38. Petal ligule shape: (0) linear and filiform; (1) widened and expanded
39. Petal ligule margin: (0) untoothed; (1) retuse at the apex; (2) sinuose
40. Petal ligule bifid: (0) absent; (1) present
41. Staminode shape: (0) filiform; (1) elliptic or ovate; (2) lanceolate
42. Staminode color: (0) yellow; (1) pink to light red; (2) crimson; (3) purple; (4) cream
43. Stamen grouping: (0) dyads only; (1) triads only; (2) tetrads only; (3) dyads and triads, alternating; (4) dyads and tetrads, alternating; (5) solitary and dyads, alternating; (6) groups of four free stamens on either side of each staminode
44. Fruit shape: (0) spheroidal; (1) oblong to ellipsoid; (2) conical
45. Fruit indumentum: (0) glabrous; (1) slightly pubescent; (1) densely pubescent
46. Epicarp surface: (0) muricate; (1) irregularly tuberculate-rugose; (2) slightly ridged; (3) deeply ridged; (4) spiculose; (5) smooth; (6) feathery
47. Number of fruit ridges: (0) 5; (1) 10; (2) undefined; (3) no ridges
48. Fruit basal constriction: (0) absent; (1) present
49. Fruit apex: (0) acute; (1) obtuse; (2) reflex; (3) straight; (4) rounded; (5) truncate
50. Fruit color when ripe: (0) ochraceous to ferruginous; (1) yellowish-green; (2) brownish-black; (3) crimson to purple
51. Fruit pericarp: (0) woody; (1) slightly pulpose; (2) heavily pulpose
52. Cotyledon color: (0) whitish to yellowish-green; (1) purple to dark-brown
53. Germination: (0) hypogeal; (1) epigeal
54. Growth ring distinctiveness: (0) distinct; (1) non-distinct
55. Wood vessel porosity: (0) diffuse-porous; (1) ring-porous
56. Wood vessel grouping: (0) exclusively solitary; (1) solitary or grouped in multiples of 2–3; (2) grouped in multiples of 4 more

Triads are present in most species of *Theobroma* and *Guazuma*, whereas *Herrania* can have tetrads or dyads or triads alternating with tetrads. *Glossostemon* has 40 stamens with groupings of four stamens with unfused filaments on each side of, and alternating with, staminodes. This interpretation of stamen fits the variation that can be seen across Byttnerioideae, which includes taxa with a wide range of stamen/anther number and fusion of filaments, including many taxa with completely free stamens. We coded this variation as a single character with the following states: dyads only; triads only; tetrads only; dyads and triads, alternating; dyads and tetrads, alternating; solitary and dyads, alternating; groups of four free stamens on either side of each staminode.

2.3 Sequence alignment, phylogenetic analysis, and ancestral state reconstructions

DNA sequences were processed using Geneious v. 10.1.3 (Kearse et al., 2012), where they were assembled, edited, and concatenated. The alignment was performed automatically

using MAFFT, followed by manual adjustments using BioEdit (Hall, 1999) to ensure accuracy. Accession numbers for the resulting 54 new sequences can be found in Table 1, and their data have been submitted to the GenBank database under accession numbers OR045314-OR045354, OR069739-OR069746, and OR091167-OR091171. Additionally, previously published *WRKY* gene sequences from Borrone et al. (2007) were included in our study (Table 1). Out of 44 species of Theobromateae, this study generated new molecular data for 17 species. For the phylogenetic analyses, *Guazuma* species were used as the outgroup, as previous studies have indicated it to be the sister group to all other Theobromateae (Whitlock et al., 2001). Outgroups from other tribes were not used due to challenges in aligning their *WRKY* sequences unambiguously with our data. We coded 11 binary state phylogenetically informative insertion/deletion (indel) characters derived from the DNA sequence matrix. Indels were identified through visual inspection of the aligned molecular matrix to identify regions with long insertions or deletions. These were counted as one insertion/deletion (query-insertion, 0/1) and included as an additional categorical data type along with the morphological matrix.

The phylogenetic analyses were conducted using two data sets: (i) a combination of *WRKY* gene sequence data, indels, and morphological data (total evidence) and (ii) *WRKY* gene sequence data and indels, with the exclusion of morphological data. The purpose of the latter analysis was to determine how much more statistical support for nodes was provided by the inclusion of morphological data.

Bayesian inference were performed in BEAST 2.2.0 (Bouckaert et al., 2014), with the inclusion of the "morphmodels" v. 1.1.4 package to allow the inclusion of categorical (morphological and indel) data. The best molecular substitution model for each partition was identified using the jModeltest (Guindon & Gascuel, 2003; Darriba et al., 2012) as follows: GTR (*WRKY*-3), TrN + G (*WRKY*-11 and *WRKY*-13), HKY + G (*WRKY*-12), and HKY (*WRKY*-14). The substitution model used for the morphological and indel characters was GTR, as this is forced by the morpho-models package on BEAST. For details on evolutionary model descriptions, see Bouckaert et al. (2014). The uncorrelated lognormal relaxed molecular clock (Drummond et al., 2006) was implemented, assuming each branch would have its own independent evolutionary rate. This approach was chosen as the most appropriate to use in BEAST as one of the related groups, *Glossostemon*, shows substantial long branches in comparison to its relatives (Richardson et al., 2015; Hernández-Gutiérrez & Magallón, 2019).

Markov chain-specific parameters were set as follows: 50 000 000 generations, with four runs and four chains, sampling the Markov chain every 10 000 times. Tree posterior probabilities (pps) were obtained, and the intervals of credibility values were set as a fraction of the highest pps. Clade support was then represented by pp values, with nodes with 0.50–0.79 indicating weak support, 0.80–0.89 indicating regular support, and values higher than 0.9 indicating strong support (see Swenson et al., 2008). We evaluated posterior distributions using Tracer v. 3.2.6 (Rambaut et al., 2018). We discarded a fraction of 0.25 trees as burn-in to reach stationarity, and the remaining trees were saved and used to build a maximum clade credibility (MCC) tree. Mesquite v. 3.70 (Maddison & Maddison, 2021) was used to reconstruct the ancestral states of morphological characters on the MCC tree (Table S1) under a likelihood criterion, aiming to maximize the probability of arriving at the observed states for a given node or terminal, given the branch lengths provided by the tree obtained from the Bayesian analysis. The characters which we chose to focus on here were those that we considered could influence drought tolerance and pollination. These included indumentum of vegetative and reproductive organs and floral characters such as claw guidelines, petal claw–ligule length ratios, and staminode shape.

3 Results

The final data matrix comprised 42 terminals, including one species of *Glossostemon*, three out of three species of *Guazuma*, 13 out of 17 species of *Herrania*, and 19 out of 23 species of *Theobroma*. The matrix included a total of 3866 characters, with 3795 corresponding to DNA sequence data sets, with variation in base substitutions, and 67 corresponding to categorical data (morphological matrix and indel region information). The aligned matrix is available as supplementary material (Table S1). The molecular data included five regions of the *WRKY* genes with varying sizes after alignment (*WRKY*-3 with 635 bp; *WRKY*-11 with 968 bp, *WRKY*-12 with 770 bp, *WRKY*-13 with 1074 bp, and *WRKY*-14 with 344 bp). Among the categorical data, 60 characters were derived from morphological data (obtained from Table 3), with an average percentage of missing data at 11.6% for all morphological characters. Additionally, 11 informative indels were identified from the aligned molecular data matrix.

Bayesian analysis of total evidence resulted in an MCC tree that revealed that *Herrania* is nested within *Theobroma* with high pp values ranging from 0.85 to 1.0 in all nodes of the phylogeny (Fig. 1). The sections of *Theobroma*, as recognized by Cuatrecasas (1964), emerged as monophyletic, except for *Theobroma* sect. *Glossopetalum*, in which *Theobroma mammosum* Cuatrec. & J. León from *Theobroma* sect. *Andropetalum* was nested. All the sections had 1.0 pp support. In *Herrania*, *Theobroma* sect. *Subcymbicalyx*, as established by Schultes (1958), was found to be paraphyletic, with *Herrania* sect. *Herrania* nested within it as monophyletic with 1.0 pp. The outgroup, *Guazuma*, is monophyletic, with *Guazuma crinita* being sister to a strongly supported monophyletic group consisting of the other two species in the genus.

In contrast, the MCC tree resulting from the analysis of *WRKY* gene sequence data and indels only (Fig. S1) showed reduced resolution and support for several branches compared to the total evidence tree (Fig. 1). For instance, the total evidence tree (Fig. 1) fully resolved the relationships between *Theobroma simiarum* Donn. Sm., *Theobroma chocoense* Cuatrec., *Theobroma hylaeum* Cuatrec., and *Theobroma cirmolinae* Cuatrec., whereas these relationships remained unresolved without the inclusion of morphological data (Fig. S1). We focus our discussion on the total evidence tree in Fig. 1 because it had a better total summed value of

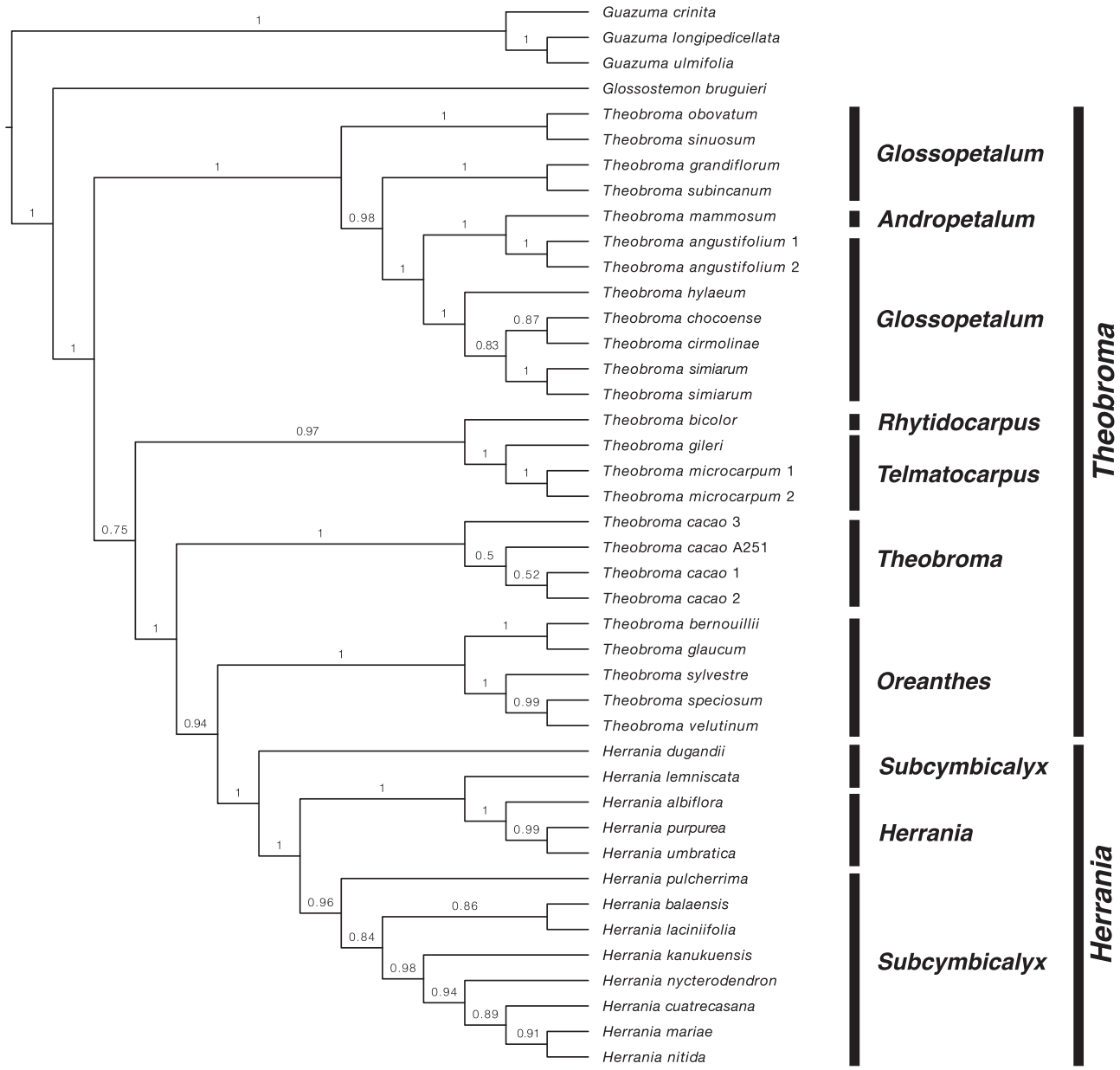

Fig. 1. Maximum clade credibility tree of Theobromateae using *WRKY* gene sequence data (including indel characters) and morphological data (total evidence analysis). Posterior probabilities indicated above each branch.

support, that is, a higher summed MCC value (36.10) compared to the tree that excluded morphological data (Fig. S1; 35.25).

The size of the colored segments in the pie charts placed on nodes in Fig. 2 indicates the likelihood of the state at that node based on a maximum likelihood estimation. The ancestral state reconstructions of selected morphological characters are indicated at all nodes on the total evidence tree (Figs. 2A–2F). The morphological characters that have most transitions and are homoplasious are petal claw–ligule ratio and fruit surface type. Ancestral states of crown nodes of sections have lower probability for petal claw–ligule ratios and fruit surface type due to their homoplasious nature (Figs. 2B, 2F). The crown nodes of *Guazuma*, *T.* sect. *Theobroma*, and *T.* sect. *Oreanthes* all have three claw guidelines as the most probable state, whereas in *T.* sections *Herrania* and *Glossopetalum* it is four to six or seven to more (Fig. 2A). The crown node of *Theobroma* and of all sections of the genus has leaf abaxial surfaces that have pubescence in the form of a single layer of stellate trichomes with the exception of *T.* sect. *Theobroma* that has glabrous leaf undersurfaces (Fig. 2C). Independent apomorphies in the form of two layers of stellate trichomes are found in *Theobroma velutinum*, *T. simiarum*, *T. grandiflorum*, and *Glossostemon* (Fig. 2C). The likely state of staminodes at the crown nodes differs among various sections of

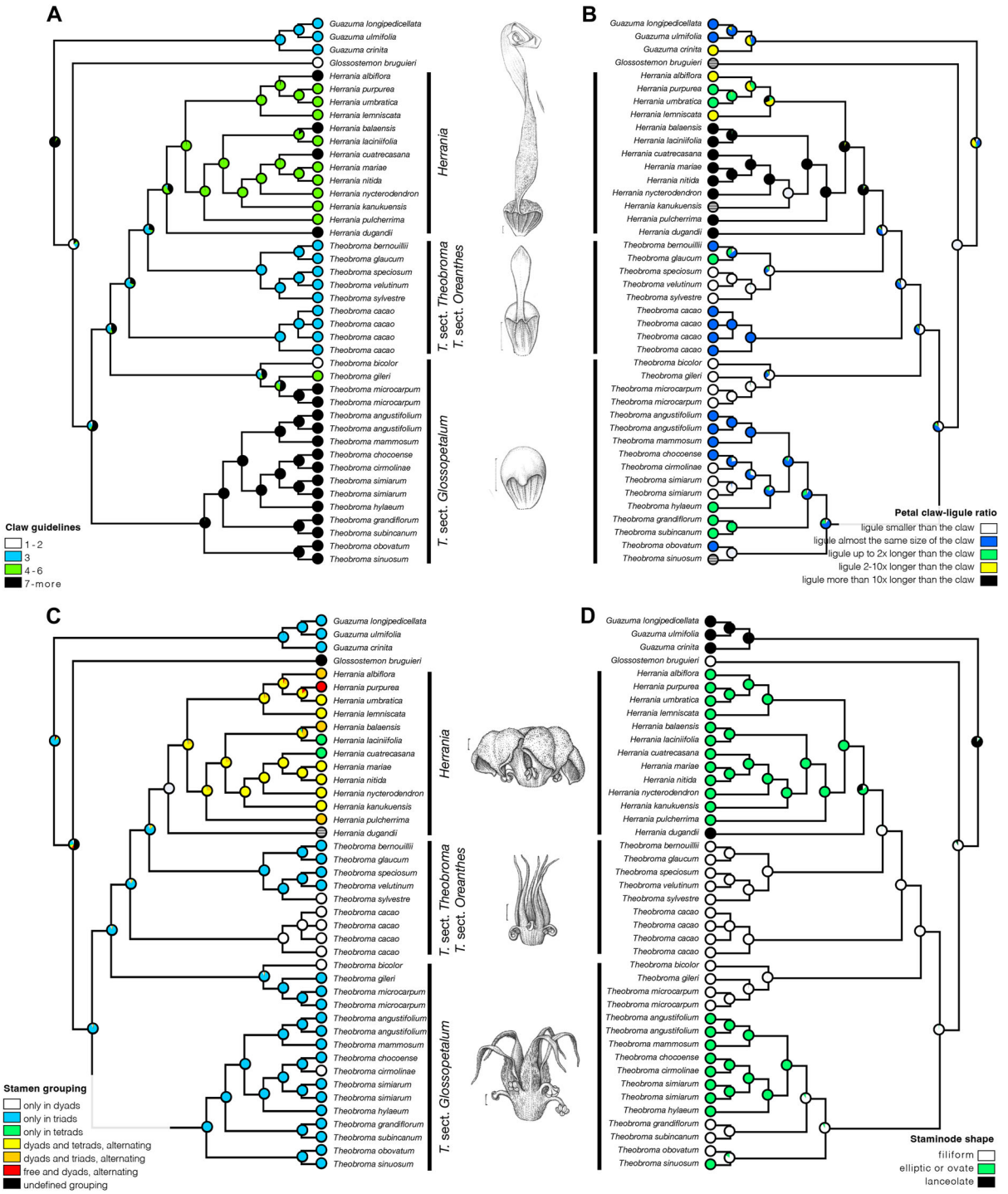

Fig. 2. Ancestral state reconstructions of key morphological characters plotted on the maximum clade credibility tree derived from the total evidence analysis. **A**, Claw guidelines. **B**, Ligule–claw ratio. **C**, Stamen grouping—number of anthers per filament. **D**, Staminode shape. **E**, Fruit indumentum. **F**, Fruit surface. Pie charts indicate the most likely character state for each corresponding node. For a clearer understanding of character variation among different groups of *Theobroma*, selected species were illustrated. The illustrations were created by Klei Sousa. The illustrated species in each figure are as follows: (A, B) (from top to bottom): *Herrania breviligulata* (*Herrania* sect. *Subcymbicalyx* sensu Schultes (1958)), *Theobroma cacao* (*Theobroma* sect. *Theobroma* sensu Cuatrecasas (1964)), and *Theobroma microcarpum* (*Theobroma* sect. *Telmatocarpus* sensu Cuatrecasas (1964)). (C, D) (from top to bottom): *Herrania purpurea* (*Herrania* sect. *Herrania* sensu Schultes (1958)), *T. cacao*, and *Theobroma grandiflorum* (*Theobroma* sect. *Glossopetalum*, sensu Cuatrecasas (1964)). (E, F) (from top to bottom): *Herrania mariae* (*Herrania* sect. *Subcymbicalyx* sensu Schultes (1958)), *T. cacao*, and *T. grandiflorum*.

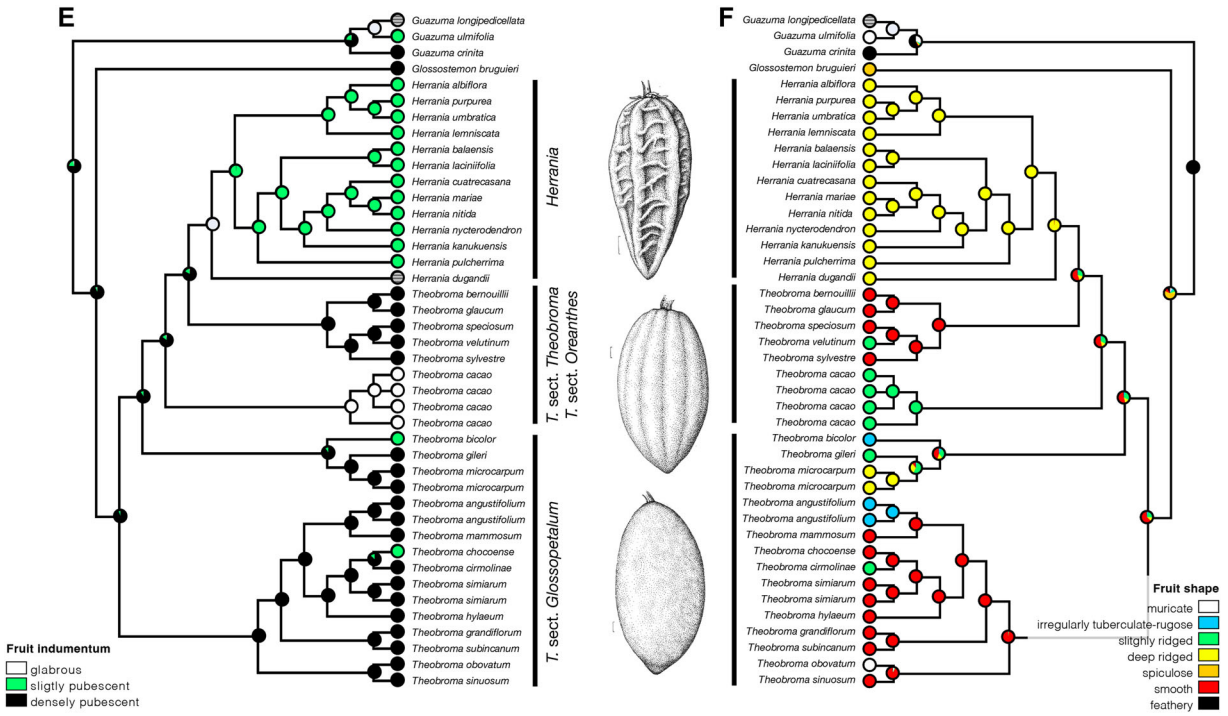

Fig. 2. Continued

Theobroma. In most sections, the staminodes are expected to be filiform, with the exception of *Theobroma* sect. *Herrania*, where they are more likely to be elliptic or ovate. *Guazuma*, on the other hand, has lanceolate staminodes (Fig. 2D). *Theobroma* sect. *Theobroma* is the only one that has fruits that are glabrous reconstructed as the state at its crown node. All other sections and *Guazuma* are either slightly or densely pubescent (Fig. 2E).

4 Discussion

4.1 Phylogenetic relationships

In this study, we present the most comprehensive phylogeny of Theobromateae published to date, offering valuable insights into morphology and key crop traits (Figs. 1, 2). Our analyses strongly support the non-monophyly of *Theobroma*, with *Herrania* nested within it, confirming the suggestions from previous studies such as Borrone et al. (2007) (Fig. 1). *Herrania* had traditionally been distinguished from *Theobroma* based on its palmately compound leaves, in contrast to the simple leaves of the latter. Additionally, *Herrania* species typically exhibit petal ligules longer than those found in *Theobroma* species. Despite these distinctions, several morphological similarities between the two genera, such as lifeform, inflorescence type, and fruit characteristics, would support the inclusion of *Herrania* within *Theobroma*. Both genera are characterized as understory trees with a cymose inflorescence, often in the form of a monochasium or dichasium with reduced branches. Their fruits are baccate, usually spherical, conical, or cylindrical, featuring a mesocarp composed of two portions, with an outer fibrous layer and an inner pulpose layer, each one separated by a sclerenchyma ring. Additionally, they share the same chromosome number, $2n = 20$ (Azevedo et al., 2017). With the close relationship between *Herrania* and *Theobroma* that we describe here, the potential of *Herrania* as a valuable source of genetic resources for cacao improvement should not be overlooked. The phylogenetic relationships inferred for *Guazuma* in our study support the earlier proposals made by Freytag (1951) based on morphology and geographical distribution. Our analysis places *G. crinita* as sister to a well-supported clade comprising the other two species, *G. longipedicellata* and *G. ulmifolia*.

4.2 Evolution of drought tolerance-related traits

A key feature with respect to the ability to tolerate drought is pubescence. Hairy leaves are a common feature of plants in arid environments as they have been shown to reduce leaf absorptance resulting in a reduced heat load and consequently lower leaf temperatures and transpiration rates (Ehleringer & Mooney, 1978). Loss of water through transpiration may also be decreased by the formation of a boundary layer of unperturbed air, maintained by hairs, that slows the rate of gas exchange. Our reconstructions indicate that the earliest diverging lineages of Theobromateae have abaxial surfaces of leaves with a single layer of stellate trichomes (Fig. 2C) that may be considered the ancestral state for the tribe as a whole and for most of the sections of *Theobroma*. *Theobroma* sect. *Theobroma* in contrast has the abaxial surface of the leaves glabrous as an apomorphy, whereas a few species have more dense pubescence as an apomorphy (*T. velutinum*, *T. simiarum*, and *T. subincanum*).

Theobroma cacao is the only species that has glabrous fruits (Fig. 2E). Calyces that are tomentose on the outer surface with abundant stellate, ochraceous, or ferruginous hairs are found in early diverging lineages of *Theobroma* but are glabrous in sections *Theobroma* and *Telmatocarpus*. Thus, pubescence appears to be a plesiomorphic feature within Theobromateae. Perhaps the conditions under which *T. cacao* originated were so humid that pubescence was not necessary to prevent water loss, and the recovery of pubescence could improve the heat/drought tolerance of *T. cacao*.

Compound leaves are believed to have arisen numerous times among several groups of angiosperms, including Malvaceae, often with reversions back to the simple leaves, suggesting that the conversion between simple and compound leaves can be attained with relative ease (Champagne & Sinha, 2004). Developmental studies have shown that simple and compound leaves, as well as leaflets and serrations, are regulated by distinct ontogenetic programs (e.g., Efroni et al., 2010), but which factors might have driven the evolution of these different leaf types among several families are still a matter of speculation.

The apomorphic condition of the compound leaves in *Herrania* might have advantages over the plesiomorphic simple-leaved form in *Theobroma*. The explanation of reduction in wind resistance due to reduced surface area does not seem adequate given that, in the case of *Theobroma* and *Herrania*, both simple and compound leaved forms occupy similar habitats with little variation in wind. Another possible explanation is that reducing surface area also reduces water loss, meaning that species of *Herrania* may be better adapted to drier conditions than the simple leaved Theobromateae. A comparison of the climatic conditions under which compound and simple leaved forms grow would be necessary to determine whether the former is found in conditions of lower precipitation. It is possible that a "compound leaved *Theobroma*" may perform better in drier conditions, in terms of vegetative growth and reproduction, than the simple leaved form and could, therefore, be more resilient to drought. Interestingly, *Herrania*, like many understory trees, has a combination of monopodial trunk and palmate leaves, whereas simple leaved *Theobroma* exhibit sympodial growth. These different growth architectures, described by Hallé et al. (1978), may have different physiological characteristics, with the monopodial form being more efficient in growth, or energy acquisition, and use than the sympodial form.

Further investigations into the morphology and physiology of *Guazuma* could lead to a better understanding of drought tolerance in Theobromateae. *Guazuma crinita* is found in rainforests, whereas *G. ulmifolia* is found predominantly in dry forests of Latin America. The adaptations to drier conditions exhibited by *G. ulmifolia* could include particular states in some of the characters we investigated here such as growth ring distinctiveness, wood vessel porosity, and wood vessel grouping that might be related to differences in water conductivity. Variations in these features have been shown to influence the degree of cavitation that has a severe impact on water conductivity (Taneda & Sperry, 2008). Conducting a more extensive comparative study on the wood anatomy of both species, as well as that of *Theobroma* and *Herrania*, may provide insights into some of the factors that confer drought tolerance on *G. ulmifolia*, while its related species appear to thrive in more humid conditions.

4.3 Evolution of traits related to pollination biology

We found that petal ligules shorter or the same size as claws is a plesiomorphic condition within the tribe based on our character reconstructions of the early diverging nodes (Fig. 2B). Similarly, the triangular or filiform staminodes of *Guazuma* and *Glossostemon*, respectively, are reconstructed as plesiomorphic (Fig. 2D). Transitions in floral morphological evolution include development of elliptic or ovate petaloid staminodes, larger particularly in width, from ancestral triangular/filiform forms (Fig. 2D). *Guazuma*, the sister genus of the remainder of the tribe (Whitlock et al., 2001; Richardson et al., 2015), has triangulate staminodes. *Glossostemon* and *Theobroma* sect. *Theobroma*, *T.* sect. *Telmatocarpus*, and *T.* sect. *Rhytidocarpus* have filiform staminodes, whereas elliptic or ovate, petaloid forms have evolved on two occasions independently in *T.* sect. *Glossopetalum* and in *Herrania* (Fig. 2D).

The number of guidelines on the petal claw or hood seems to be an informative feature with few independent evolutionary transitions to similar states, for example, four to six guidelines are found in *T. gileri* and *Herrania* (Fig. 2A). *Theobroma* sect. *Glossopetalum* has seven guidelines or more per claw, *Herrania* has four to six and sections *Theobroma* and *Oreanthes* have three. Each of the sections of *Theobroma* and *Herrania* have little variation in petal color. *Theobroma* sections *Oreanthes* and *Glossopetalum* have smooth fruits, *Glossostemon* has spiculose whereas *Theobroma* sections *Theobroma*, *Telmatocarpus*, and *Herrania* have ridged fruits (Fig. 2F). These types of epicarp surface seem to be correlated with degree of fruit pubescence (Fig. 2E). Cuatrecasas (1964) speculated that the sections of *Theobroma* with pseudoterminal growth (*T.* sect. *Glossopetalum*), may be an ancestral condition than the ones which have lost the axillary buds of the jorquette branches, necessitating lateral shoots to continue growing (as in *T.* sections *Oreanthes*, *Rhytidocarpus*, and *Theobroma*).

Ligules longer than claws have evolved independently in, for example, *Herrania*, *T. grandiflorum* and *T. subincanum*, *T. hylaeum* and *T. glaucum*. A transition to the apomorphic state that include an increase in length of petal ligules relative to petal claws (Fig. 2B) and an increase in staminode width to more petaloid forms (Fig. 2D) could be related in some way to different pollination syndromes. Pollination in *Theobroma* seems to be undertaken predominantly by small-bodied midges in the families Ceratopogonidae and Cecidomyiidae (Soria, 1970; Entwistle, 1972; Bystrak & Wirth, 1978; Winder, 1978; Alvim, 1984; Toledo-Hernández et al., 2017). In *Herrania*, pollination is thought to be mainly by stout-bodied Diptera of the family Phoridae (Young, 1984). Schultes (1958) has suggested that variations in pollen morphology between *Theobroma* and *Herrania* might be associated with their distinct pollinators. Whether these different pollinators play some selective role in causing the differences in floral morphology is unclear. Studies on each group would be required to determine whether particular

floral character states result in different pollinators, differences in pollination success and fruit set. This information could in turn be used to try to improve pollination success in cacao.

4.4 Taxonomic implications

Cuatrecasas' sections of *Theobroma* emerged as monophyletic except for *T.* sect. *Glossopetalum* that has *Theobroma mammosum* of *T.* sect. *Andropetalum* nested within. In contrast to Cuatrecasas' (1964) sections of *Theobroma*, the sections of *Herrania* defined by Schultes (1958) do not appear to be monophyletic and some re-circumscription may be necessary and will be undertaken in a separate manuscript (Colli-Silva et al., in preparation). Inclusion of morphological data increased clade resolution and support values, for example, the support of the clade including sections *Telmatocarpus* and *Rhytidocarpus*, the placement of *T. mammosum* within *T. sect. Glossopetalum*, and especially the paraphyly of *Theobroma* as well as the relationships among species of *Herrania*. This was particularly important in resolving relationships between the species within sections and showed the consistency of the definition of many sections by previous authors. For example, *Theobroma* sect. *Glossopetalum*, which has 13 species, had poorly resolved relationships between species with only molecular data, but much better resolution with total evidence (Figs. 1 and S1). The addition of more sequence data should increase resolution and support within these clades.

4.5 Final remarks

Our study has clarified relationships amongst nearly all species of the tribe Theobromateae. We have characterized the evolutionary trajectory of morphological characters within the group and provided a framework for understanding evolution in the tribe. The traits that are relevant to drought tolerance, disease resistance, and floral biology are complex and an understanding of their function and development needs to be integrated into that of the whole plant. Our study represents a first step in understanding the evolution of those traits. The various morphologies of species in Theobromateae may mean that they perform better in certain respects than cacao itself under particular conditions. Greater pubescence or compound leaves might mean greater tolerance to arid conditions, and if so these features could be introduced into cacao. Harris et al. (in preparation) have characterized the climatic envelope of each species of Theobromateae providing an indication of which of them are more tolerant to arid conditions. Knowledge of the precise mechanisms of drought tolerance in those species might assist with producing a drought-tolerant cacao. Alteration in form of the cacao flower might attract a wider range of pollinators, improving fruit set. Mournet et al. (2020) have characterized some aspects of the response of *T. grandiflorum* to witches' broom infection, but more studies need to be done on other taxa in the group to characterize their degree of resistance to this and other diseases. Our phylogeny may be used to assist with the search for these agronomically important traits. It may also be used to guide the search for novel products. For example, the pulp of *T. grandiflorum* is used to produce flavorings and fruit juices. Our results indicate that this species is related to a group of other species whose similar morphological traits may make them also suitable for producing similar products. Genomics studies are required to fully understand how the genetic diversity of CWRs may be utilized in the cacao industry. Many of these cacao CWRs are under-collected and likely threatened with extinction. To safeguard the potential future use of the genetic diversity of CWRs within Theobromateae it is vital that wild populations and their genetic diversity be conserved, preferably in situ, together with all the organisms that they have developed interactions with over the course of millions of years.

Acknowledgements

This work was primarily funded by the UK Research and Innovation (UKRI) Global Challenges Research Fund (GCRF) GROW Colombia grant via the UK's Biotechnology and Biological Sciences Research Council (BB/P028098/1) awarded to F.D.P. This study was partially financed by the following Brazilian sources: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)—Finance Code 001—and by the São Paulo Research Foundation (FAPESP; Grant IDs 2020/01375-1, 2020/10206-9 and 2021/08635-1), with M.C-S. and J.R.P. as grantees. We thank New York Botanical Garden (NY), Field Museum (F), Herbario Nacional Colombiano (COL), Museo de Historia Natural C.J. Marinkelle (ANDES), Royal Botanic Gardens Kew (K), Instituto Amazónico de Investigaciones Científicas SINCHI (COAH), Harvard University Herbaria (GH and ECON), Instituto de Investigación de Recursos Biológicos Alexander von Humboldt (FMB), Universidad de Antioquia (HUA), Instituto Nacional de Pesquisas da Amazônia (INPA), Naturalis Biodiversity Center collections (L, U, and WAG), Museu Paraense Emilio Goeldi (MG), Missouri Botanical Garden (MO), Jardim Botânico do Rio de Janeiro (RB), and Smithsonian Institution (US) for allowing our team to review specimens and collect samples for genomic analyses. We thank New York Botanical Garden—Pfizer Plant Research Laboratory for allowing our team to perform the DNA extractions of specimens. N.C. was funded by a PhD scholarship from the Darwin Trust of Edinburgh.

Author Contributions

A.M.B-C. performed the research, collected samples, and wrote the manuscript. J.E.R. conceived the project, designed the experiments, and wrote the manuscript. M.C-S. produced the morphological matrix, performed phylogenetic analyses, and assisted with manuscript preparation. J.R.P. assisted with manuscript preparation. B.A.W. collected and contributed data and assisted with manuscript preparation. L.T.M.M. collected data and assisted with manuscript preparation. N.C. collected samples and contributed data. M.C-H., F.D.P., and M.V. contributed to project conception and administration, assisted with manuscript preparation and funding acquisition.

References

Ali ZAA. 2020. Taxonomic study of *Glossostemon bruguieri* Desf. (Malvaceae) in Iraq. *Plant Archives* 20: 926–929.

Alvim PDE. 1984. Flowers of cocoa. *Cocoa Growers' Bulletin* 35: 23–31.

Anderson JT, Song B-H. 2020. Plant adaptation to climate change—Where are we? *Journal of Systematics and Evolution* 58: 533–545.

Anjos LJS, Barreiros de Souza E, Amaral CT, Igawa TK, Mann de Toledo P. 2021. Future projections for terrestrial biomes indicate widespread warming and moisture reduction in forests up to 2100 in South America. *Global Ecology and Conservation* 25: e01441.

Argout X, Martin G, Droc G, Fouet O, Labadie K, Rivals E, Aury JM, Lanaud C. 2017. The cacao Criollo genome v2.0: An improved version of the genome for genetic and functional genomic studies. *BMC Genomics* 18: 730.

Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C. 2011. The genome of *Theobroma cacao*. *Nature Genetics* 43: 101–108.

Azevedo DSR, Gustavo S, Lemos LSL, Lopes UV, Patrocinio NGRBP, Alves RM, Marcellino LH, Clement D, Micheli F, Gramacho KP. 2017. Genome size, cytogenetic data and transferability of EST-SSRs markers in wild and cultivated species of the genus *Theobroma* L. (Byttnerioideae, Malvaceae). *PLoS One* 12: e0170799.

Borrone JW, Kuhn DN, Schnell RJ. 2004. Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*. *Theoretical and Applied Genetics* 109: 495–507.

Borrone JW, Meerow AW, Kuhn DN, Whitlock BA, Schnell RJ. 2007. The potential of the *WRKY* gene family for phylogenetic reconstruction: An example from the Malvaceae. *Molecular Phylogenetics and Evolution* 44: 1141–1154.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10: e1003537.

Brozynska M, Furtado A, Henry RJ. 2016. Genomics of crop wild relatives: Expanding the gene pool for crop improvement. *Plant Biotechnology Journal* 14: 1070–1085.

Bystrak PG, Wirth WW. 1978. The North American species of *Forcipomyia*, subgenus *Euprojoannisia* (Diptera: Ceratopogonidae). *U.S. Department of Agriculture Technical Bulletin* no. 1591.

Champagne C, Sinha N. 2004. Compound leaves: Equal to the sum of their parts? *Development*. 131: 4401–4412.

Colli-Silva M, Pirani JR. 2020. Estimating bioregions and under-collected areas in South America by revisiting Byttnerioideae, Helicteroideae and Sterculioideae (Malvaceae) occurrence data. *Flora* 271: 151688.

Colli-Silva M, Richardson JR, Pirani JR. 2023. A taxonomic dataset of preserved specimen occurrences of *Theobroma* and *Herrania* (Malvaceae, Byttnerioideae) stored in 2020. *Biodiversity Data Journal* 11: e99646.

Cortés AJ, Cornille A, Yockteng R. 2022. Evolutionary genetics of crop-wild complexes. *Genes* 13: 1.

Couvreur TLP, Chatrou LW, Sosef MSM, Richardson JE. 2008. Molecular phylogenetics reveal multiple tertiary vicariance origins of African rain forest trees. *BMC Evolutionary Biology* 6: 54–63.

Cuatrecasas J. 1964. Cacao and its allies: A taxonomic revision of the genus *Theobroma*. *Contributions from the United States National Herbarium* 35: 379–614.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: More models, new heuristics and parallel computing. *Nature Methods* 9: 772.

Davis AP, Gole TW, Baena S, Moat J. 2012. The impact of climate change on indigenous Arabica coffee (*Coffea arabica*): Predicting future trends and identifying priorities. *PLoS One* 7: e47981.

Davis AP, Mieulet D, Moat J, Sarmu D, Haggar J. 2021. Arabica-like flavour in a heat-tolerant wild coffee species. *Nature Plants* 7: 413–418.

de laFuente deDiez Canseco L ed. 2018. *Cacao, the treasure of the Amazon*. Lima: Fondo Editorial USIL.

Doyle JJ, Doyle JL. 1990. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin of the Botanical Society of America* 19: 11–15.

Drummond AJ, Ho SYW, Philips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 4(5): e88.

Efroni I, Eshed Y, Lifshitz E. 2010. Morphogenesis of simple and compound leaves: A critical review. *The Plant Cell* 22: 1019–1032.

Ehleringer JR, Mooney HA. 1978. Leaf hairs: Effects on physiological activity and adaptive value to a desert shrub. *Oecologia* 37: 183–200.

Ellis B, Daly DC, Hickey LJ, Mitchell JD, Johnson KR, Wilf P, Wing SL. 2009. *Manual of leaf architecture*. Ithaca: Cornell University Press.

Entwistle PF. 1972. *Pests of cocoa*. London: Longmans.

Eugem T, Rushton PJ, Robatzek S, Somssich IE. 2000. The SRKY superfamily of plant transcription factors. *Trends in Plant Science* 5(5): 199–206.

European Commission. 2014. Commission Regulation (EU) No 488/2014 of 12 May 2014 amending Regulation (EC) No 1881/2006 as regards maximum levels of cadmium in foodstuffs Text with EEA relevance. *Official Journal of the European Union* 138: 75–79.

Evans HC. 2007. Cacao diseases-the trilogy revisited. *Phytopathology* 12: 1640–1643.

Flint-Garcia SA. 2013. Genetics and consequences of crop domestication. *Journal of Agricultural and Food Chemistry* 61: 8267–8276.

Freytag GF. 1951. A revision of the genus *Guazuma*. *Ceiba* 1: 193–225.

González-Orozco CE, Galán AAS, Ramos PE, Yockteng R. 2020. Exploring the diversity and distribution of crop wild relatives of cacao (*Theobroma cacao* L.) in Colombia. *Genetic Resources and Crop Evolution* 67: 2071–2085.

González-Orozco CE, Porcel M, Rodriguez-Medina C, Yockteng R. 2022. Extreme climate refugia: A case study of wild relatives of cacao (*Theobroma cacao*) in Colombia. *Biodiversity and Conservation* 31: 161–182.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* 40(D1): D1178–D1186.

Grilli G, Cantillo T, Ferrini S, Richardson JE, Turner K, Di Maria C, Erazo J, Azcárate J, Di Palma F. 2022. Perspectives on a bioeconomy development path for Colombia. Cacao farming for peace building and rural development. *Report 3. GROW Colombia*

Project Series. Norwich: GROW Colombia Project UKRI GCRF Grant BB/P028098/1.

Guindon O, Gascuel S. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

Hall TA. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41: 95–98.

Halle F, Oldeman RAA, Tomlinson PB. 1978. *Tropical Trees and Forests: an architectural analysis*. Berlin: Springer-Verlag. 463.

Hernández-Gutiérrez R, Magallón S. 2019. The timing of Malvales evolution: Incorporating its extensive fossil record to inform about lineage diversification. *Molecular Phylogenetics and Evolution* 140: a106606.

Hollingsworth PM, Dawson IK, Goodall-Copestake WP, Richardson JE, Weber JC, Sotelo Montes C, Pennington RT. 2005. Do farmers reduce genetic diversity when they domesticate tropical trees? A case study from Amazonia. *Molecular Ecology* 14: 497–501.

Igawa TK, Toledo PMD, Anjos LJS. 2022. Climate change could reduce and spatially reconfigure cocoa cultivation in the Brazilian Amazon by 2050. *PLoS One* 17: e0262729.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Mentjies P, Drummond A. 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.

Kulkarni M, Soolanayakanahally R, Ogawa S, Uga Y, Selvaraj MG, Kagale S. 2017. Drought response in wheat: Key genes and regulatory mechanisms controlling root system architecture and transpiration efficiency. *Frontiers in Chemistry* 5: 106.

Läderach P, Martinez-Valle A, Schroth G, Castro N. 2013. Predicting the future climatic suitability for cocoa farming of the world's leading producer countries, Ghana and Côte d'Ivoire. *Climatic Change* 119: 841–854.

Maddison WP, Maddison DR. 2021. Mesquite: A modular system for evolutionary analysis. Version 3.70 [online]. Available from http://www.mesquiteproject.org. [accessed 19 December 2023].

Mammadov J, Buyyarapu R, Guttikonda SK, Parliament K, Abdurakhmonov IY, Kumpatla SP. 2018. Wild relatives of maize, rice, cotton, and soybean: Treasure troves for tolerance to biotic and abiotic stresses. *Frontiers in Plant Science* 9: a886.

Marelli JP, Guest DI, Bayley BA, Evans HC, Brown JK, Junaid M, Barreto RW, Lisboa DO, Puig AS. 2019. Chocolate under threat from old and new diseases. *Phytopathology* 109: 1331–1343.

Maxted N, Ford-Lloyd BV, Jury S, Kell S, Scholten M. 2006. Towards a definition of a crop wild relative. *Biodiversity and Conservation* 15: 2673–2685.

Maxted N, Kell S. 2009. Establishment of a global network for the *in situ* conservation of crop wild relatives: Status and needs. *Commission on Genetic Resources for Food and Agriculture*. FAO.

McElroy MS, Navarro AJR, Mustiga G, Stack C, Gezan S, Peña G, Sarabia W, Saquicela D, Sotomayor I, Douglas GM, Migicovsky Z, Amores F, Tarqui O, Myles S, Motamayor JC. 2018. Prediction of cacao (*Theobroma cacao*) resistance to *Moniliophthora* spp. diseases via genome-wide association analysis and genomic selection. *Frontiers in Plant Science* 9: 343.

Moat J, Williams J, Baena S, Wilkinson T, Gole TW, Challa ZK, Demissew S, Davis AP. 2017. Resilience potential of the Ethiopian coffee sector under climate change. *Nature Plants* 3: 17081.

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, Bustamante CD, Schnell RJ, Main D, Gilbert D, Parida L, Kuhn DN. 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* 14: r53.

Mournet P, Beviláqua de Albuquerque PS, Alves RM, Oliveira Silva-Werneck J, Rivallan R, Marcellino LH, Clément D. 2020. A reference high-density genetic map of *Theobroma grandiflorum* (Willd. ex Spreng) and QTL detection for resistance to witches' broom disease (*Moniliophthora perniciosa*). *Tree Genetics and Genomes* 16: a89.

Ochoa R. 2017. Jaén y la cultura Marañón [VIDEO]. Available from https://larepublica.pe/domingo/1147164-montegrandey-la-cultura-maraNOn/

Olivera-Núñez Q. 2018. *Jaén: Arqueología y turismo*. Perú: Municipalidad Provincial de Jaén.

Osorio-Guarín JA, Berdugo-Cely J, Coronado RA, Zapata YP, Quintero C, Gallego-Sánchez G, Yockteng R. 2017. Colombia a source of cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. *Frontiers in Plant Science* 8: 1994.

Periyannan S, Moore J, Ayliffe M, Bansal U, Wang X, Huang L, Deal K, Luo M, Kong X, Bariana H, Mago R, McIntosh R, Dodds P, Dvorak D, Lagudah E. 2013. The gene Sr33, an ortholog of barley Mla genes, encodes resistance to wheat stem rust race Ug99. *Science* 341: 786–788.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67: 901–904.

Richardson JE, Whitlock BA, Meerow AW, Madriñán S. 2015. The age of chocolate: A diversification history of *Theobroma* and Malvaceae. *Frontiers in Ecology and Evolution* 3: 120.

Rushton PJ, Somssich IE, Ringler P, Shen QJ. 2010. *WRKY* transcription factors. *Trends in Plant Science* 15(5): 247–258.

Saintenac C, Zhang W, Salcedo A, Rouse MN, Trick HN, Akhunov E, Dubcovsky J. 2013. Identification of wheat gene Sr35 that confers resistance to Ug99 stem rust race group. *Science* 341: 783–786.

Sarkinen T, Staats M, Richardson JE, Cowan R, Bakker FT. 2012. How to open the treasure chest: Optimizing DNA extraction from herbarium specimens. *PLoS One* 7(8): e43808.

Saslis-Lagoudakis CH, Savolainen V, Williamson EM, Forest F, Wagstaff SJ, Baral SR, Watson MF, Pendry CA, Hawkins JA. 2012. Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences USA* 109: 15835–15840.

Schultes RE. 1958. A synopsis of the genus *Herrania*. *Journal of the Arnold Arboretum* 34: 217–278.

Soria SDJ. 1970. Studies on *Forcipomyia* spp. Midges (Diptera: Ceratopogonidae) related to the pollination of *Theobroma cacao* L. Doctoral Dissertation. Madison: University of Wisconsin.

Sousa Silva CR, Figueira A. 2004. Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. *Plant Systematics and Evolution*, 250(1–2): 93–104.

Staats M, Erkens RHJ, van de Vossenberg B, Wieringa JJ, Kraaijeveld K, Geml J, Richardson JE, Bakker FT. 2013. Exploring genomic treasure troves: Whole-genome sequencing of herbarium and insect museum specimens. *PLoS One* 8(7): e69189.

Swenson U, Richardson JE, Bartish IV. 2008. Multi-gene phylogeny of the pantropical subfamily Chrysophylloideae (Sapotaceae): Evidence of generic polyphyly and extensive morphological homoplasy. *Cladistics* 24: 1006–1031.

Taneda H, Sperry JS. 2008. A case-study of water transport in co-occurring ring- versus diffuse-porous trees: Contrasts in water-status, conducting capacity, cavitation and vessel refilling. *Tree Physiology* 28: 1641–1651.

Thiers B. 2022. Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/science/ih/. [accessed 19 December 2023].

Toledo-Hernández M, Wanger TC, Tscharntke T. 2017. Neglected pollinators: Can enhanced pollination services improve cocoa yields? A review. *Agriculture, Ecosystems and Environment* 247: 137–148.

Valdez FX. 2013. *Arqueología Amazonica: Las Civilizaciones Ocultas del Bosque Tropical*. Quito, Ecuador: IRD Éditions. 395.

Vansynghel J, Ocampo-Ariza C, Maas B, Martin EA, Thomas E, Hanf-Dressler T, Schumacher NC, Ulloque-Samatelo C, Tscharntke T, Steffan-Dewenter I. 2022. Cacao flower visitation: Low pollen deposition, low fruit set and dominance of herbivores. *Ecological Solutions and Evidence* 3: e12140.

Voora V, Bermúdez S, Larrea C. 2019. *Global market report: Cocoa*. Winnipeg, Canada: International Institute for Sustainable Development (IISD). 12.

Webb LJ. 1959. A physiognomic classification of Australian rain forests. *Journal of Ecology* 47: 551–570.

Whitlock BA, Baum DA. 1999. Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene vicilin. *Systematic Botany* 24: 128–138.

Whitlock BA, Bayer C, Baum DA. 2001. Phylogenetic relationships and floral evolution of the Byttnerioideae ("Sterculiaceae" or Malvaceae s.l.) based on sequences of the chloroplast gene, ndhF. *Systematic Botany* 26: 420–437.

Winder JA. 1978. Cocoa flower Diptera: Their identity, pollinating activity and breeding sites. *Proceedings of the National Academy of Sciences USA* 24: 5–18.

Young AM. 1984. Mechanisms of pollination by Phoridae (Diptera) in some *Herrania* species (Sterculiaceae) in Costa Rica. *Proceedings of the Entomological Society of Washington* 86: 503–518.

Zarrillo S, Gaikwad N, Lanaud C, Powis T, Viot C, Lesur I, Fouet O, Argout X, Guichoux E, Salin F, Solorzano RL, Bouchez O, Vignes H, Severts P, Hurtado J, Yepez A, Grivetti L, Blake M, Valdez F. 2018. The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. *Nature Ecology and Evolution* 2: 1879–1888.

Zhang F, Wen Y, Guo X. 2014. CRISPR/Cas9 for genome editing: Progress, implications and challenges. *Human Molecular Genetics* 23: R40–R46.

Supplementary Material

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.13045/suppinfo:

Fig. S1. Maximum clade credibility tree of Theobromateae based on analysis of *WRKY* gene sequence data and indels only.

Table S1. Morphological matrix and indication of indel regions for Theobromateae species (formatted as a NEXUS file). Full descriptions of the characters and character states are found on the main text (Table 3).