Morphometrics and Phylogenomics of Coca (*Erythroxylum* spp.) Illuminate Its Reticulate Evolution, With Implications for Taxonomy

Natalia A.S. Przelomska [1,2,3,*,†] Rudy A. Diaz [2,†] Fabio Andrés Ávila [4] Gustavo A. Ballen [5,6] Rocío Cortés-B. [7] Logan Kistler [3] Daniel H. Chitwood [8,9] Martha Charitonidou [10] Susanne S. Renner [11] Oscar A. Pérez-Escobar [2,*,‡] and Alexandre Antonelli [2,12,13,*,‡]

[1]School of Biological Sciences, University of Portsmouth, Portsmouth PO1 2DY, UK
[2]Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AE, UK
[3]Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington DC 20560, USA
[4]The New York Botanical Garden, New York, NY 10458, USA
[5]Instituto de Biociências, Universidade Estadual Paulista, Botucatu, São Paulo, Brazil
[6]School of Biological and Behavioural Sciences, Queen Mary University of London, London E1 4NS, UK
[7]Herbario Forestal Universidad Distrital, Campus El Vivero, CR 5E 15-82 Bogotá, Colombia
[8]Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA
[9]Department of Computational Mathematics, Science & Engineering, Michigan State University, East Lansing, MI 48824, USA
[10]Department of Biological Applications and Technology, University of Ioannina, 45110 Ioannina, Greece
[11]Department of Biology, Washington University, Saint Louis, MO 63130, USA
[12]Gothenburg Global Biodiversity Centre, Department of Biological and Environmental Sciences, University of Gothenburg, SE 41319 Göteborg, Sweden
[13]Department of Biology, University of Oxford, Oxford OX1 3RB, UK
†Lead authors.
‡Senior authors.

***Corresponding authors:** E-mails: natalia.przelomska@port.ac.uk; o.perez-escobar@kew.org; a.antonelli@kew.org.
Associate editor: Aida Ouangraoua

Abstract

South American coca (*Erythroxylum coca* and *E. novogranatense*) has been a keystone crop for many Andean and Amazonian communities for at least 8,000 years. However, over the last half-century, global demand for its alkaloid cocaine has driven intensive agriculture of this plant and placed it in the center of armed conflict and deforestation. To monitor the changing landscape of coca plantations, the United Nations Office on Drugs and Crime collects annual data on their areas of cultivation. However, attempts to delineate areas in which different varieties are grown have failed due to limitations around identification. In the absence of flowers, identification relies on leaf morphology, yet the extent to which this is reflected in taxonomy is uncertain. Here, we analyze the consistency of the current naming system of coca and its four closest wild relatives (the "coca clade"), using morphometrics, phylogenomics, molecular clocks, and population genomics. We include name-bearing type specimens of coca's closest wild relatives *E. gracilipes* and *E. cataractarum*. Morphometrics of 342 digitized herbarium specimens show that leaf shape and size fail to reliably discriminate between species and varieties. However, the statistical analyses illuminate that rounder and more obovate leaves of certain varieties could be associated with the subtle domestication syndrome of coca. Our phylogenomic data indicate extensive gene flow involving *E. gracilipes* which, combined with morphometrics, supports *E. gracilipes* being retained as a single species. Establishing a robust evolutionary-taxonomic framework for the coca clade will facilitate the development of cost-effective genotyping methods to support reliable identification.

***Key words:* coca, *Erythroxylum*, leaf crops, morphometrics, phylogenomics, taxonomy.**

Received: August 08, 2023. **Revised:** May 01, 2024. **Accepted:** May 10, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Open Access

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/ ... June 2025

Introduction

The expansion of humans into South America some 15,000 to 13,500 years ago (Rothhammer and Dillehay 2009) was accompanied by the adoption, and sometimes domestication of native plants into human agroecosystems. One of the earliest known and most widely cultivated crops is coca (Rury and Plowman 1983), taxonomically circumscribed as *Erythroxylum coca* Lam. and *E. novogranatense* (D. Morris) Hieron (Schulz 1907; Plowman 1979), whose history of use traces back at least 8,000 years (Dillehay et al. 2010). Coca was predominantly involved in the cultural evolution of societies owing to its medicinal, stimulatory, and cultural properties (Schultes 1979; Plowman 1984), and many Andean and Amazonian communities today remain reliant on this plant (Naranjo 1979; Cristancho and Vining 2004; Azevedo 2021; Vergara et al. 2022).

Despite many ethnobotanical and physiological studies, the taxonomic boundaries between cultivated varieties and their wild relatives in the genus *Erythroxylum* (family Erythroxylaceae) are poorly defined. The implications of an inadequate classification system of coca are pertinent to plantations, many of which supply local populations with leaves for medicinal and culinary uses (Restrepo et al. 2019), while others are designated for the extraction of the plant's coveted alkaloid cocaine. Focused coca eradication to hamper drug trafficking, cocaine-linked deforestation, and armed conflict have become associated with the latter type of plantation in the last century (Rincón-Ruiz and Kallis 2013; Dávalos et al. 2016, 2021; Negret et al. 2019). While eradication programs have targeted many cocaine-producing plantations, they inevitably pose a threat to the Indigenous biocultural diversity of coca in the process, due to geographical proximity. In Colombia alone, 10% of illegal plantations occur in Indigenous territories (UNODC 2023), and in Peru, the major cocaine production area in the valley of the Apurímac, Ene, and Mantaro Rivers (known as VRAEM in Spanish), overlaps with Indigenous lands (Watson and Arce 2024). The United Nations Office on Drugs and Crime (UNODC) has been monitoring areas of coca cultivation annually since 1999 yet has limited tools for discriminating species or varieties. The development of new, reliable plant identification tools would therefore be highly valuable for better understanding the complex landscape of coca growing across South America.

The genus *Erythroxylum* P. Browne consists of over 270 species, of which about three-quarters are native to the American tropics (Plowman and Hensold 2004; Jara-Muñoz et al. 2022). There are four varieties of cultivated cocas—two each in the species *Erythroxylum coca* (var. *coca* and var. *ipadu* Plowman) and *E. novogranatense* (var. *novogranatense* and var. *truxillense* [Rusby] Plowman). All of them have largely allopatric distributions in northwestern South America (Bohm et al. 1982; our Fig. 1). The most widely cultivated species is *E. coca* (Huánuco coca). Its variety *coca* is native to wet montane forests of the eastern Andean slopes of Peru and Bolivia, whereas variety *ipadu* (Amazonian coca) is grown across the lowland Amazon basin. The less widely cultivated *E. novogranatense* has historically been grown in the dry valleys of the Cordilleras and the Sierra Nevada de Santa Marta (Bohm et al. 1982), but has also recently been found in the Pacific region (Chocó and Cauca) (UNODC 2016). Its variety *truxillense* (also known as "Trujillo coca") is cultivated in arid regions of northwestern Peru for traditional use and is a flavoring and stimulant additive to the soft drink *Coca Cola*© (Plowman 1986; Gootenberg 2003).

Attribution of species status to the two broad types of cultivated coca (*Erythroxylum coca* and *E. novogranatense*) was undoubtedly influenced by their ethnobotanical significance, and differences between them were discerned through morphology, chemotaxonomy, and reproductive systems (Rury 1981; Bohm et al. 1982; Plowman and Rivier 1983; Plowman 1986). A better understanding of the relationships among lineages of cultivated crops can now be gained with molecular markers (Viruel et al. 2021). Recent population-level genomic work (White et al. 2021) has suggested that cultivated coca is polyphyletic, with wild *E. gracilipes* Peyr. inferred as the closest living relative of both coca species.

As a leaf crop, selective pressures may have affected coca leaf shape and size to the point of these becoming taxonomically informative; leaf morphology can be part of the domestication syndrome (Galindo Bonilla and Fernández-Alonso 2010; Arias et al. 2021). Indeed, leaf characteristics have been crucial in the formal taxonomy of coca (Plowman 1979, 1982), and are therefore used for coca identification in monitoring surveys (e.g. UNODC 2012). The leaves of cultivated coca are thought to be distinguishable from leaves of closely related, and often sympatric wild *Erythroxylum* species by being smaller, rounder, and softer (Rury 1981; White et al. 2021). However, inter-grading variation across the two cultivated coca species and their wild relatives is pervasive (Rury and Plowman 1983), and phenotypic plasticity additionally renders leaf morphology-based identification of individual coca leaves problematic (Rury 1981; Rury and Plowman 1983). To date, rigorous statistical analyses are lacking.

Here, we investigate genetic relationships in the coca clade, assess their degree of correspondence with currently accepted taxa, and examine the discriminatory power of leaf morphology in identifying species and varieties using digitized herbarium specimens. We employ Gaussian mixture models (GMMs) to infer probabilistic morphometric clusters and assess their overlap with the currently accepted taxa. We then infer population-level nuclear and plastid phylogenies for the coca clade through a hybrid approach of genome skimming herbarium specimens and mining published *Erythroxylum* target capture genomic datasets. To test for gene flow among taxa, we apply phylogenetic network analysis. Finally, we apply population genomic tests to delimit population groups of coca and molecular-clock models to estimate lineage divergence times.

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

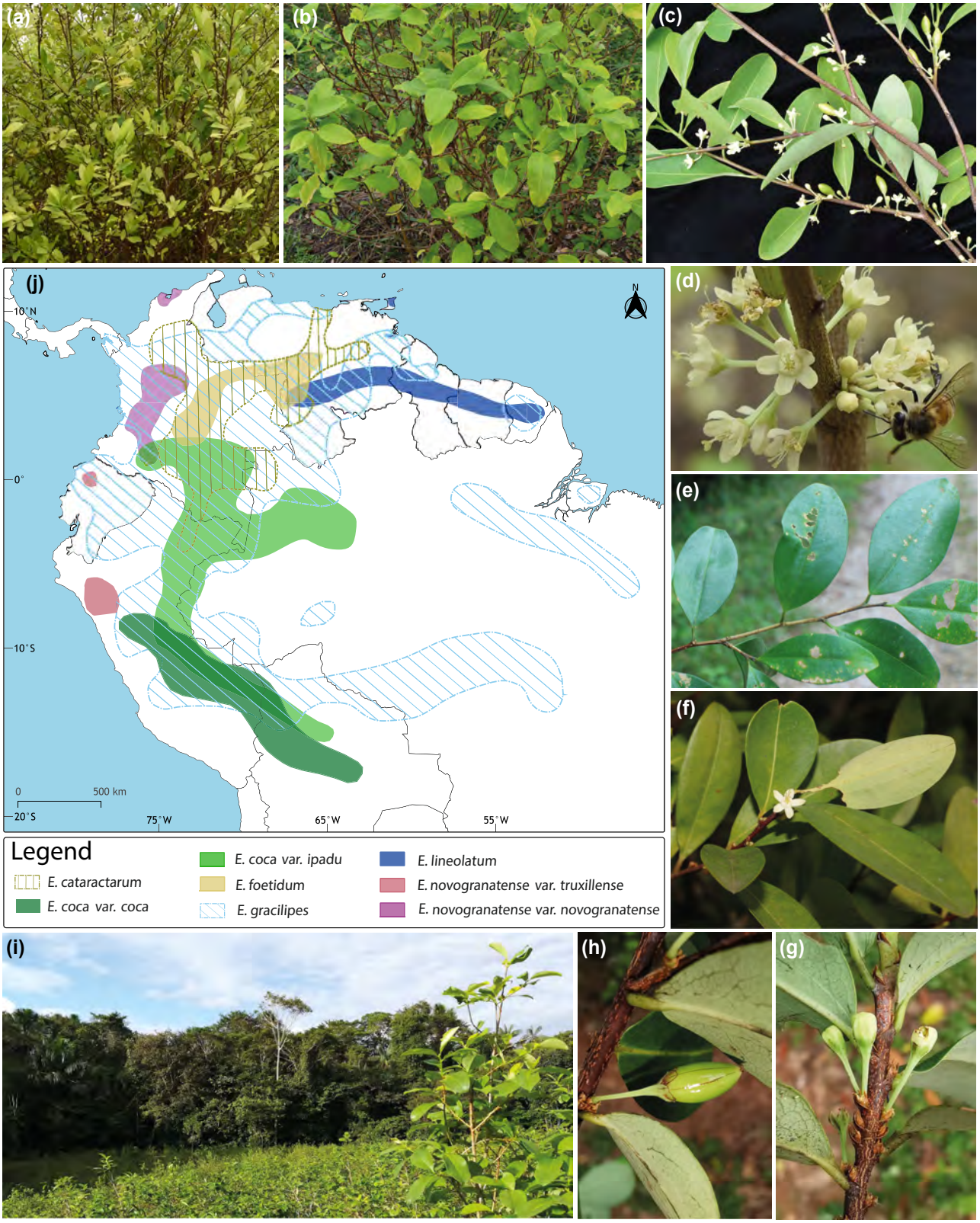

Fig. 1. Morphology and geographical distribution of taxa in the coca clade. a) Bush of *Erythroxylum novogranatense*. b) Bush of *E. coca* var. *coca*. c) Leaves, flowers, immature flowers, and immature fruit of *E. coca* var. *ipadu*. d) Flowers of *E. coca* var. *ipadu* with potential pollinator. e) Leaves of *E. gracilipes*. f) Leaves of *E. cataractarum*. g) Immature flowers of *E. cataractarum*. h) Immature fruit of *E. cataractarum*. i) Illegal coca plantation established in the Amazonian forests of Putumayo department, Colombia. j) geographical distributions of cultivated and wild relative *Erythroxylum taxa*. Photos: Rocío Cortés-B. (a, b, c, d), William Ariza (e, f), and José Aguilar Cano (g, h, i).

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

Results

Leaf Size is Insufficient for Identifying the Cultivated Species and Varieties of Coca

We find that leaf size metrics have limited power to discriminate between varieties of cultivated coca and between cultivated cocas and their wild relatives. In our principal component analysis (PCA) on linear metrics, the taxa do not appear as distinct clusters. Nevertheless, PC1 does segregate most *E. gracilipes* individuals from the remaining taxa (Fig. 2a) due to the greater leaf area (loadings for area, width, and length: 0.590, 0.574, and 0.572). Leaf area is likewise significantly greater in *E. foetidum* compared to the remaining taxa (*t*-test, $P < 2.2\times10^{-4}$) (Fig. 2a, supplementary figs. S1 and S4, Supplementary Material online). PC2 can be attributed to the leaf length-to-width ratio (loadings: −0.73 and 0.68 respectively, loading of area: 0.04). We note the taxonomic signal for length-to-width ratio, supported to an extent by diverging slopes for different taxonomic groups on a locally estimated scatterplot smoothing plot (supplementary fig. S5, Supplementary Material online). *Erythroxylum novogranatense* var. *truxillense* has longer, narrower leaves than those observed in *E. novogranatense* var. *novogranatense* (Fig. 2). *Erythroxylum coca* in turn has slightly larger leaves (Fig. 2, supplementary fig. S6, Supplementary Material online). However, the varietal groups of *E. novogranatense* and *E. coca*, collectively with *E. cataractarum*, exhibit a large proportion of overlap in leaf size morphospace (Fig. 2a), precluding statistically detectable taxon identification. The *E. cataractarum* type specimen appears near the extreme end of its morphospace (Fig. 2a), and closer to the centroid of *E. n novogranatense* var. *truxillense*'s distribution.

The GMMs with the highest level of statistical support model three clusters for linear metric data, with consistent support even after downsampling to 50 leaves per taxon and inclusion of samples growing either outside of

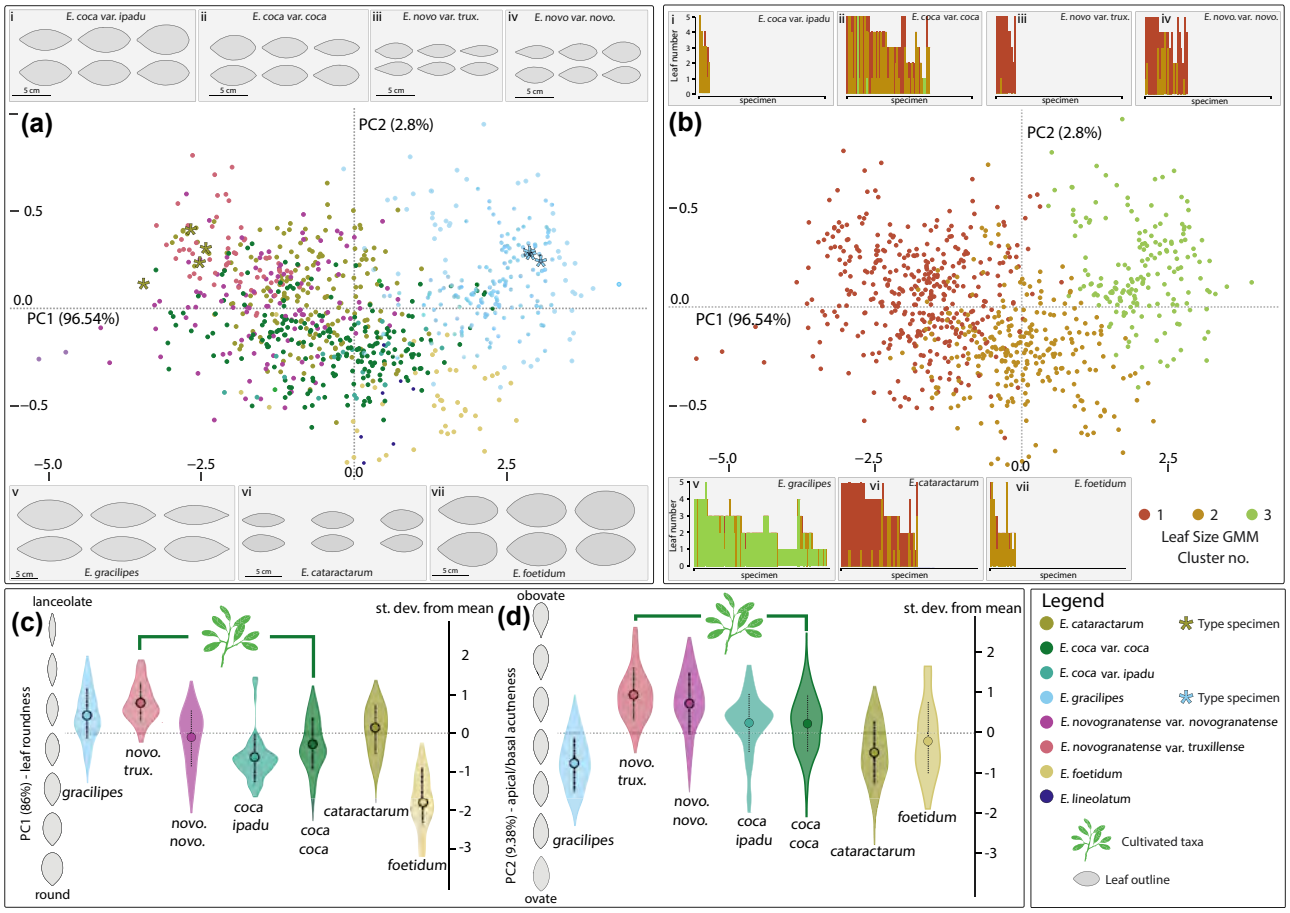

Fig. 2. Summary of leaf morphometric analyses, encompassing leaf size and leaf shape, based on principal dataset (261 specimens, 844 leaves total). a) Morphospace representing PCA based on linear metrics, colored by verified pre-defined taxonomic identifications. Insets are eigenleaves computed for each taxonomic subset of the data, reconstructed using the reverse Fourier transform for the mean, and ± 1.5 standard deviations from the mean, on the first two PCs of shape variation within a species. b) Morphospace representing PCA based on linear metrics, colored by assignment resulting from the GMM clustering model. Insets are bar charts representing the number of leaves sampled, and their individual cluster assignments, grouped by pre-defined taxonomic identifications. c) Plots of the distribution of values along the first PC of the EFA PCA, representing leaf roundness, grouped by taxonomic identification. d) Plots of the distribution of values along the second PC of the EFA PCA, representing acuteness at the base or apex, grouped by taxonomic identification. *E. lineolatum* was omitted from the EFA shape analyses due to limited sample size.

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

their native South American range or in cultivation (supplementary figs. S1 to S3, Supplementary Material online). The clearest element of congruence in terms of GMM-inferred groups and taxonomy is Cluster 3, which exhibits a high degree of overlap with the data taxonomically determined as *E. gracilipes* (Fig. 2a and b.v). Cluster 2 overlaps well with leaves taxonomically assigned to *E. foetidum* and *E. coca* var. *ipadu* and is the most frequently assigned cluster for leaves of *E. coca* var. *coca* (Fig. 2b.i, ii, vii), whereas Cluster 1 is assigned to most *E. novogranatense* var. *truxillense* and is the most common classification for individuals of var. *novogranatense* and *E. cataractarum* (Fig. 2b.iii, iv, vi). Importantly, these clustering methods based on linear metrics fail to distinguish the three wild relative species consistently and confidently from cultivated cocas. Furthermore, linear metric-based clustering was discordant with the taxonomic groupings overall (Rand index, full dataset = 0.691, Rand index, cultivated cocas only = 0.632; supplementary table S6, Supplementary Material online).

Leaf Shape Morphometrics Illuminate Traits That Define Cultivated Species

Using a reverse Fourier transform, we directly visualize outlines of leaf shapes using the statistical framework of the PCA into which our raw leaf outline data were input (Fig. 2a; supplementary figs. S1E, F, S2E, F, and S3E, F, Supplementary Material online), revealing the two most defining leaf shape traits (Fig. 2a, c and d, supplementary figs. S1 to S3 and S8, Supplementary Material online), as discussed below. PC1 describes leaf shape from obovate/orbicular to lanceolate, or "roundness" of the leaves (Fig. 2ai–vii and c). *Erythroxylum foetidum* has the roundest leaves, followed by *E. coca* var. *ipadu* and *E. coca* var. *coca* which do not differ (*t*-test $P = 0.096$; supplementary table S5a, Supplementary Material online). *E. novogranatense* var. *novogranatense* is not highly distinctive in roundness from the two *E. coca* varieties (*t*-test $P > 0.01$ for each comparison; supplementary table S5a, Supplementary Material online) whereas *E. novogranatense* var. *truxillense* is very different, having the most lanceolate leaves (Fig. 2d; supplementary table S5a, Supplementary Material online). PC2 describes obovate versus ovate shape of the leaves, or more specifically "acuteness at the base or apex" (Fig. 2ai–vii and d), where wild taxa (*E. gracilipes*, *E. cataractarum*, and *E. foetidum*) have more ovate leaves than the cultivated species grouped together (*t*-test, $P < 0.01$ for all comparisons; supplementary table S5b, Supplementary Material online). The cultivated varieties are better characterized by obovate leaf shape, especially *E. novogranatense* var. *truxillense* (Fig. 2d). There is no difference of high significance ($P < 0.01$) between varieties and notably, *E. gracilipes* is only highly significantly different from one variety: *E. novogranatense* var. *truxillense* (supplementary table S5c, Supplementary Material online). As with the linear metric data, our GMM of the highest statistical support clusters the data into three groups (consistent after downsampling), albeit with differences in group composition. Aspects of congruence between shape and size GMM clusters (Fig. 2b, supplementary fig. S8, Supplementary Material online) are as follows: one of the modeled groups (Cluster 3) is assigned almost exclusively to leaves from *E. gracilipes* individuals (supplementary fig. S8, Supplementary Material online). The principal group to which leaves of *E. foetidum* are assigned (Cluster 2; supplementary fig. S8, Supplementary Material online) is in turn distinct from this. Cluster 1 is prevalent, dominating in all cultivated taxa and *E. cataractarum*, but also common in *E. foetidum* and *E. gracilipes* (supplementary fig. S8, Supplementary Material online). The principal coordinate analysis (PCoA) based on mean vectors of per taxon elliptical Fourier analysis (EFA) data likewise shows how all cultivated varieties form a close cluster (supplementary fig. S9, Supplementary Material online). Overall, there is less uniformity of taxa when grouped using shape data compared to when grouped based on linear metric data (supplementary tables S7 and S6, Supplementary Material online respectively), and an even greater mismatch of GMM-inferred clusters to the taxonomy (Rand index, full dataset = 0.555, Rand index, cultivated cocas only = 0.425; supplementary table S7, Supplementary Material online).

Erythroxylum gracilipes Gene Flow is Pervasive and Supported by Hybrid Edges

The uniparental plastid phylogeny and the 326-gene nuclear phylogeny both underscore the complex genetic structure of paraphyletic *E. gracilipes* (Fig. 3a to d), and each genomic source places *E. foetidum* along with specimen Spruce 3725 as sister to the coca clade. The remaining portions of the respective phylogenies exhibit incongruence, especially in the placement of the ten *E. gracilipes* specimens, including the Kew (Index Herbariorum code K) isotype of *E. gracilipes* (Spruce 3068; supplementary table S2, Supplementary Material online). This sample is encompassed in the poorly supported *E. gracilipes* + *E. coca* clade in the nuclear tree and has an unresolved position (50% likelihood bootstrap support (LBS)) within the well-supported *E. novogranatense* + *E. cataractarum* clade in the plastid tree. The plastid PCA based on genotype likelihoods (GLs) (Fig. 3d) additionally demonstrates that the *E. gracilipes* isotype plastid shares more genetic variation with *E. gracilipes* samples than with plastids of other taxa.

Regarding *E. coca*, the plastid tree suggests closely related but phylogenetically distinct *coca* and *ipadu* varieties nested within a clade of *E. gracilipes* + *E. coca* (Fig. 3b). In the nuclear dataset, the multispecies coalescent (MSC) ASTRAL species tree indicates that gene-tree incongruence is substantial (normalized quartet score: 0.657). Furthermore, for any given bipartition within this *E. gracilipes* + *E. coca* clade, there is a mean total of 12 gene trees supporting the bipartition with strong support and

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

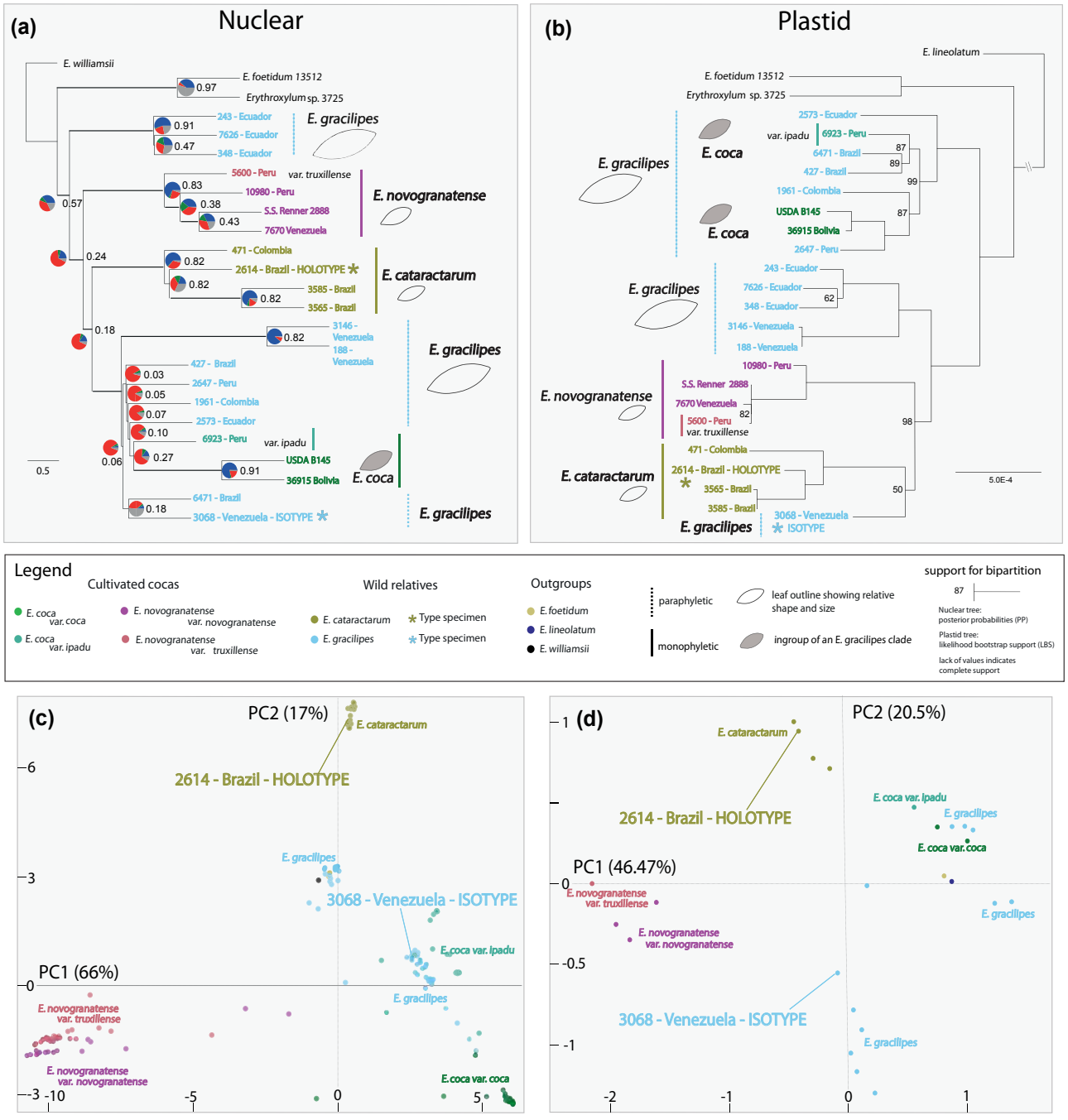

Fig. 3. Comparative phylogenomic and population genomic analyses of coca and wild relatives. a) ASTRAL summary species tree topology derived from 326 ML phylogenies built from nuclear genes from 25 samples of *Erythroxylum*, with percent quartet support indicated and gene concordance factors presented. Pie charts at nodes indicate the proportion of gene trees that support (blue), or conflict with (green, red) the species tree. Here, the green proportion of the pie chart indicates the proportion of gene trees recovering the second most frequent alternative bipartition to the species tree, and red—any other alternative conflicting bipartition. Gray represents non-informative gene trees. b) ML tree reconstructed from plastomes sequenced in 25 samples of *Erythroxylum*, with bootstrap support after 500 iterations indicated and bootstrap support indicated. c) Nuclear genomic PCA based on 13,643 GLs computed from 18 sequenced and 155 data-mined accessions of *Erythroxylum* samples. Sequenced type specimens for *E. cataractarum* and *E. gracilipes were* highlighted. d) Plastid PCA based on 4,664 GLs computed from 18 sequenced and 9 data-mined accessions of 25 *Erythroxylum* samples. Sequenced type specimens *for E. cataractarum* and *E. gracilipes are* highlighted in (c) and (d).

94 gene trees conflicting with strong support (Fig. 3a) (mean gene concordance factor (gCF)= 0.11, mean internode certainty (IC) = 0.08, supplementary fig. S11A and B, Supplementary Material online). Within this clade, a monophyletic grouping of *ipadu* and *coca* varieties has moderate support (32 supporting trees, 78 conflicting trees, gCF = 0.27, IC = 0.28), but includes a highly supported sister status of the two *coca* samples (gCF = 0.91, IC = 0.86). In contrast, *E. cataractarum* and *E. novogranatense* are well-supported, monophyletic groups, each

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

with 100% LBS in the plastid phylogeny (Fig. 3b) and characterized by high gene-tree support values in the ASTRAL tree (*E. cataractarum* gCF = 0.82, IC = 0.86; *E. novogranatense* gCF = 0.83, IC = 0.87). Similar values were obtained for the concatenated RAxML species tree (supplementary fig. S12A and B, Supplementary Material online).

The phylogenetic network analysis (Fig. 4c), conducted on the same set of 25 samples to gain a broader picture of the evolutionary history of the clade, provides support to the hypothesis of gene flow involving *E. gracilipes*. The best network suggests two reticulations. The first is highly supported (LBS = 98, credible intervals for major and minor edges 0.62 to 0.80 and 0.20 to 0.37 respectively) and originates in the same clade as the early diverging, monophyletic Ecuadorian *E. gracilipes* clade, with the *E. gracilipes* + *E. coca* clade as the recipient (inheritance proportion = 0.70). The second most frequent hybridization edge detected (credible intervals for major and minor edges 0.67 to 0.90 and 0.10 to 0.33 respectively) sits within the *E. gracilipes* + *E. coca* clade, and likewise involves a monophyletic *E. gracilipes* outgroup supplying genetic material to an ingroup including the cultivated lineages (inheritance proportion = 0.83). An alternative recipient position with lower support was recovered for this second hybridization. The main source of uncertainty in this hybridization event likely lies with the recipient branch, on the basis that the donor branch has a higher combined support (combined bootstrap support considering all alternative recipients = 76). However, since the orientation of gene flow relies on support of the minor edge in the hybridization event, and is concordant here, we consider the gene flow direction to be highly supported.

Bioculturally Important Varieties Differ in Degree of Intraspecific Genetic Differentiation

Our nuclear PCA analyses based on nuclear GLs called in the 18 sequenced specimens and 155 data-mined samples combined (hereafter, the "merged dataset") show genetic structure within *E. gracilipes*. The most substantial proportion of its diversity manifests as a cluster containing the isotype of *E. gracilipes* (Spruce 3068) (supplementary table S2, Supplementary Material online), corresponding to "gracilipes1" sensu White et al. (2021) (Fig. 3c). This group, along with a secondary *E. gracilipes* and a single *E. cataractarum* cluster, are observed in both the merged dataset (Fig. 3c) and in our wild relatives-focused sampling from Kew specimens only (supplementary fig. S13, Supplementary Material online). The merged dataset PCA (Fig. 3c) exhibits the largest proportion of the variation as being driven by genetic differentiation between *E. novogranatense* and *E. coca* var. *coca*. Within *E. novogranatense*, the varieties *novogranatense* and *truxillense* do form separate clusters, but with more subtle differentiation. In comparison, the varieties of *E. coca*: var. *coca*, and var. *ipadu* are clearly distinct. Unsupervised clustering predicted the highest values of ΔK for scenarios of genetic structuring where *E. novogranatense* (both varieties), *E. coca* var. *coca*, *E. gracilipes* (containing *E. coca* var. *ipadu*), and *E. cataractarum* resolve as three or four broadly separate genetic entities (Fig. 4b, supplementary figs. S14 and S15, Supplementary Material online). The isotype-defined *E. gracilipes* group becomes distinct at $K = 5$ (Fig. 4b) and *E. coca* var. *ipadu* at $K = 6$ (supplementary fig. S14, Supplementary Material online). Only under a model of eight hypothetical populations does *E. novogranatense* var. *novogranatense* emerge as a cluster distinct from *E. novogranatense* var. *truxillense* (supplementary fig. S14, Supplementary Material online).

Lineage Divergence of Cultivated Cocas Long Precedes the Peopling of South America

We inferred a time-calibrated phylogeny using a Bayesian MSC approach to estimate ages of divergence within this clade. In our population-focused approach, we combined a species tree relaxed molecular-clock model with information on population membership determined using NGSadmix (Skotte et al. 2013). From this, we reliably reconstructed a species tree for the coca species and varieties and inferred absolute ages of their divergence from closely related wild relatives in the presence of topological discordance (Ogilvie et al. 2017). The results imply that the coca clade diversified 2.2 million years ago (Ma) (± 1 Ma, posterior probability (PP) = 1.0) (Fig. 4a). The cultivated cocas and their wild-living relatives shared a common ancestor ~1.6 Ma ago (± 700 thousand years (Kyrs), PP = 1.0). The divergence of the cultivated *E. novogranatense* and *E. coca* var. *coca* lineages from *E. cataractarum* and *E. gracilipes* (1,400 to 800 Kyrs, PP = 0.93 and 800 to 400 Kyrs, PP = 1.0 respectively) precedes the peopling of the American tropics (~15.5 Ka; Dillehay et al. 2008; Rothhammer and Dillehay 2009; Prates et al. 2020) by hundreds of millennia.

Discussion

Establishing a robust taxonomy is vital for research into plants that are medicinally, nutritionally, or culturally valuable to humans (e.g. Pironon et al. 2024). Plant classification at this scale is often non-trivial, since it involves elucidating the plants' evolutionary trajectories within a wider pool of genetic diversity comprising crop wild relatives, hybrids, and semi-domesticated forms (Pellicer et al. 2018; Pérez-Escobar et al. 2021, 2022; Simon et al. 2022). For coca, its naming system has a further layer of relevance, being directly linked to the existence of legal frameworks, which on one side aim to hamper trafficking, but on the other hold the possibility of promoting opportunities for local communities that depend upon coca cultivation for traditional uses.

Evolution and Reticulation of Lineages in the Coca Clade

Our new plastid tree for the coca clade corroborates the hypothesis of White et al. (2021) proposing *E. gracilipes*

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

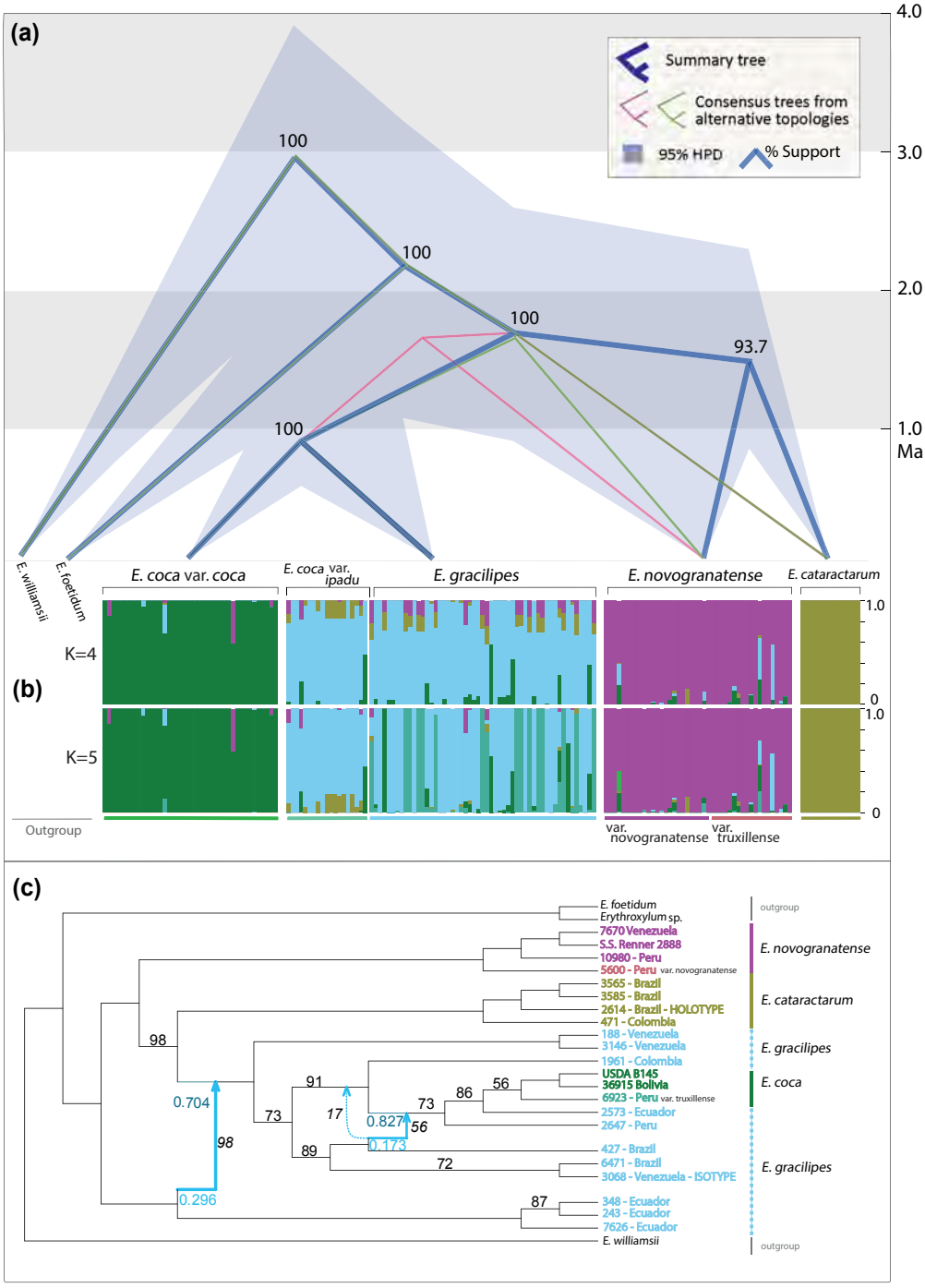

Fig. 4. Absolute times of divergence of lineages in the coca clade based on populations. a) Chronogram representing dates of divergence within the clade currently including *E. gracilipes*, *E. cataractarum*, *E. coca* and *E. novogranatense*. *E. foetidum* and *E. williamsii* serve as successive outgroups. Consensus topology shown in dark blue. Alternative topologies in pink and green. 95% highest posterior density intervals of absolute ages indicated as shaded area and percentage support provided at each node. b) NGSadmix population structure inference showing per-individual ancestry of sequenced and data-mined accessions of 173 *Erythroxylum* specimens at highly supported values of $K = 4$ and $K = 5$. Full results of per-individual ancestry inference for varying values of K provided in Supplementary Material online. c) Phylogenetic network demonstrating waves of gene flow ("hybrid events"). Numbers on branches represent bootstrap values lower than 100. Numbers in italics are bootstrap values for the hybridizations. Any alternative recipient positions for a hybridization event shown as a dotted line. Inheritance proportion values are presented for the optimal network. Colored numbers represent the inheritance proportions for the major (dark blue) and minor (light blue) edges, depicting their point estimates (see main text for credible intervals).

as a clade within which *E. cataractarum*, *E. novogranatense*, and *E. coca* are nested. Our population genomic and phylogenomic analyses which build upon this study illustrate extensive genetic structure characterizing *E. gracilipes*, a shrub proposed as the wild progenitor of all described varieties of cultivated coca (Macbride 1949; White et al. 2021). It is unsurprising that a plant species with a broad geographic distribution should show non-discrete clustering

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

(Dodsworth et al. 2021) and *E. gracilipes* is indeed distributed widely—in wet biomes across tropical South America (White et al. 2019) (Fig. 1j). Counteracting this structure, we find prevalent gene flow between lineages currently considered as *E. gracilipes*, involving both early diverging clades and the admixed *E. gracilipes* + *E. coca* clade. This differs to the findings of White et al. (2021) who conducted maximum likelihood (ML) Treemix analysis (Pickrell and Pritchard 2012) to conclude that none of the three most significant waves of gene flow in the coca clade involved the poorly resolved *E. gracilipes* + *E. coca* clade, but instead featured *E. cataractarum* as donor or recipient of genetic material. Treemix can be biased when applied to a dataset that includes multiple admixed populations and this method warrants extra support in most scenarios (Lipson et al. 2013, 2020). Future analyses using more genomic markers will allow for the re-evaluation of all gene flow events inferred for the coca clade thus far.

The dominant *E. gracilipes* group, including the Spruce 3068 isotype, was previously identified as "gracilipes1" (White et al. 2021), who proposed to retain it as a genetically constricted species of *E. gracilipes*, while re-classifying other pockets of genetic variation within *E. gracilipes* as separate species. While there is a clear rationale for this proposition, we believe there are reasons for considering an alternative approach. First, there is insufficient evidence that the *E. gracilipes* population groups are independently evolving lineages, given the shallowness of many branches on the phylogenomic trees and degree of nuclear gene-tree conflict first indicated by White et al. (2021) and supported by gene flow patterns detected in this study. As a second argument, expressed phenotypes remain highly relevant to botanical classification (Wells et al. 2022) not least due to their applicability in ecological contexts such as functional diversity (Sultan 2000). Rigorous phenotypic analysis, facilitated by advances in the field of morphometrics (Christodoulou et al. 2020) here supports a high degree of consistency in terms of the large, mucronate to acuminate leaves of *E. gracilipes*. Combined with dependable characters used for taxonomy mainly based on lanceolate and coriaceous leaves, and free styles in brevistylar and longistylar flowers, this still supports the retention of a single species. Splitting of *E. gracilipes* could also have consequences for interpretation of the evolution of the domesticated *E. novogranatense* and *E. coca* lineages, reinforcing the idea of domestication events isolated in space and time. Such a model is incongruent with the theory of a protracted timescale of human selection on coca that spans a variety of habitats (Plowman 1984), now well-supported by the landscape-level domestication paradigm, built on evidence of extensive human-mediated genetic exchange (Allaby et al. 2022; Fuller et al. 2022).

We deduce that the gene pool ancestral to *E. gracilipes* encompassed rich standing genetic variation, from which lineages of *E. novogranatense*, *E. cataractarum*, and *E. coca* var. *coca* also emerged, probably in response to novel environmental and ecological conditions. Our molecular dating results showing clade divergence of the lineages of species currently regarded to be cultivated in the mid to late Pleistocene suggest that these had already undergone adaptation to ecological conditions distinct from those of their sister wild-living, forest-dwelling relatives; at the point of being selected, these populations were feasibly pre-adapted to a human environment in which they then experienced anthropogenic selection. The species *E. cataractarum* was mostly shaped by adaptation to drier, high-altitude habitats of the eastern Andean slopes (White et al. 2021). Despite limited sampling, genomic data support monophyly of this group. There is some ethnobotanical evidence of the use of *E. cataractarum* (Schultes 1981), including the inscription on the K-type specimen itself "Ipadú das cachocinas" ("ipadu" being an Indigenous name for coca widely used in Brazil; Plowman 1979). *Erythroxylum cataractarum* is not documented as a cultivated variety of coca but, considering the high degree of leaf morphological overlap with cultivated cocas, its use cannot be ruled out (Plowman and Rivier 1983).

Erythroxylum novogranatense and *E. coca* var. *coca* have been taxonomically proposed as "neo-species" resulting from long-term human cultivation and selection (Rury 1981), and the population genomic results confirm the distinctness of these taxa sensu *lato*. We urge caution in labeling these "independently evolving monophyletic lineages" (White et al. 2021), since this is not supported by the evidence for ancient gene flow we detected across datasets and overlooks the diffuse nature of domestication (Allaby et al. 2008; Gross and Olsen 2010). On the other hand, we agree that the differentiation is appreciable, phylogenomically supported, and one argument for retaining their respective species designations. The differentiation could have been driven in part by breeding system: the majority of *Erythroxylum* spp. are reportedly heterostylous (White et al. 2019), with *E. coca* var. *coca* and *E. novogranatense* exhibiting self-incompatibility (Ganders 1979). Heterostyly has previously been linked to elevated genetic differentiation between Neotropical species of *Erythroxylum* (Abarca et al. 2008). In contrast, *E. coca* var. *ipadu*, which is the only variety of coca that is predominantly out-crossing (Plowman 1979) exhibits genetic identity linked to that of the isotype-defined *E. gracilipes* group. This reinforces our phylogenomic conclusion—of a close genetic relationship of *E. coca* var. *ipadu* to this population of the wild *E. gracilipes*, which could have implications for the future taxonomic status of this variety.

Distinct Leaf Phenotypes in the Coca Clade Revealed by Morphometrics

Our leaf morphometric dataset provides evolutionary insights that complement the phylogenomic inferences, supporting multiple origins of coca as a leaf crop. Our most significant finding in this context is that of leaf size distinctness between wild *E. gracilipes* and smaller-leaved domesticated *E. coca* and *E. novogranatense*. *E. coca* var. *ipadu* is not excluded from this pattern, despite its close genetic affinity to *E. gracilipes*. Our data also show a tendency for

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

rounder leaves in *E. coca* compared to wild species. Hence, we provide statistical evidence that partly supports cultivated coca leaves being smaller and rounder than those of their wild progenitor (White et al. 2021), a hypothesis built upon in-depth physiological studies of stomatal density and vein morphology (Rury 1981). A possible scenario partially supported by our results is that cultivated cocas are the result of evolutionarily distinct adaptations from wild forest progenitors to novel, open habitats with greater exposure to sunlight. Those habitats likely arose as a consequence of global temperature decline and climatic fluctuations following the onset of Pleistocene glaciations at c. 2.6 Ma, which led to widespread contraction of rainforests and expansion of drier habitats (Silva et al. 2018). The species currently regarded as domesticated then needed to invest fewer resources into leaf expansion than they would in the moist forest habitat of *E. gracilipes*. Experimental work has revealed latent plasticity in leaf size within all described varieties of cultivated coca—manifested not only in smaller "sun leaves" and larger "shade leaves" not only in their native neotropical setting, but also when cultivated under glass at temperate latitudes (Rury and Plowman 1983). Leaf morphology is generally evolutionarily labile in plants and thus readily adaptable to new microhabitats or biomes (Spriggs et al. 2018). We propose that the consistently smaller leaves in cultivated cocas are the result of genetic canalization (Flatt 2005; Piperno 2017), whereby phenotypic variation in the cultivated species has largely become developmentally constrained to small leaf size. A significant shift in environmental conditions, including escape from and survival outside of human cultivation, might over time unlock persistent cryptic genetic variation (Flatt 2005), here responsible for the large-leaved *E. gracilipes* phenotype. *Erythroxylum cataractarum* is a highly valuable control taxon to benchmark this theory of leaf shape evolution. Having presumably circumvented the selective pressures associated with human cultivation, it has nonetheless attained a leaf size statistically indistinguishable from that of cultivated cocas.

Another possible domestication syndrome trait is acuteness at the leaf base (lending an obovate shape), which contrasts to the frequently observed apical acuteness of wild *E. gracilipes*. Given the sheer volume of leaves harvested by certain Indigenous communities to supply their daily needs of almost constant coca chewing (Schultes 1981), it is plausible that reorganization of the leaf structure to be amenable to human leaf picking could have given these coca bushes an adaptive advantage. Interestingly, the EFA distils basal acuteness as a trait occurring in cultivated coca varieties but not in *E. cataractarum*. Finally, the length-to-width ratio of coca leaves could warrant further study; this has been pinpointed e.g. as the primary source of variation between accessions and species of apple, with a heritable basis (Migicovsky et al. 2018).

Studies of other plants cultivated for their fruit (Chitwood et al. 2013; Chitwood and Otoni 2017; Klein et al. 2017) or roots (Gupta et al. 2020) have shown that leaf shape, as inferred through an EFA framework, is highly heritable. EFA has been used to demonstrate that morphological groupings are consistent with species boundaries (Andrade et al. 2008; Sayıncı et al. 2015; Klein et al. 2017), but this is not consistently true across plant groups, and particularly not for domesticates occupying ecosystems more diverse than that of a modern agricultural setting (Soares et al. 2011; Nascimento et al. 2021), which is the case for traditionally cultivated coca.

Taxonomic Implications

Synthesizing the evidence from phylogenomics, gene flow analyses, phenotypic plasticity in coca leaf shape, and the limited discriminatory power of other characters such as leaf venation and floral anatomy (Plowman 1979; Rury 1981), a phylogenetic species concept could be applied here to lump *E. gracilipes*, *E. coca*, *E. novogranatense*, and *E. cataractarum* into a single species. Nevertheless, the insights from population genomics and molecular dating do support the identity of long-recognized, and distinct cocas. As such, we propose that the names of these are retained for the time being, so as not to compromise their cultural significance. A future prospect could entail re-classifying *E. coca* var. *ipadu*, *E. coca* var. *coca*, *E. novogranatense* var. *novogranatense*, and *E. novogranatense* var. *truxillense* as varieties of equal standing within a more complex species, to more easily accommodate any new varieties of coca that are described using genomic, ethnobotanical, and metabolomic evidence.

Conclusion

The challenging, subtle nature of morphological differences between cultivated cocas and their closest wild relatives first underscored by Rury (1981) supports the failure of our linear metric and EFA analyses to reliably classify coca leaves by current taxonomy. Previous attempts to discriminate Colombian coca varieties by leaf morphology have been similarly inconclusive (Galindo Bonilla and Fernández-Alonso 2010; Rodríguez Zapata 2015). Yet, the analyses introduced here did yield statistically supported insights regarding leaf phenotypes which can be leveraged for further research of this clade. The phylogenomic evidence base we assembled can serve as a platform for further studies interrogating the complex evolutionary trajectory of lineages in the coca clade and for potential taxonomic revisions. We also hope that it will contribute to the development of sensitive genomic methods to identify species and varieties of coca and to highlight the importance of safeguarding portions of its diversity which are under the stewardship of Indigenous communities.

Materials and Methods

Leaf Morphometrics

Image Preparation

A total of 1,163 leaf outlines were extracted from 342 digital herbarium specimens, representing individuals collected

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

in the wild (species: *E. gracilipes* Peyr., *E. cataractarum* Spruce ex. Peyr, *E. lineolatum* DC., and *E. foetidum* Plowman), as well as plants cultivated in the neotropics or under glass at temperate latitudes (*Erythroxylum coca* Lam. (var. *coca* and var. *ipadu*) and *E. novogranatense* (D. Morris) Hieron (var. *novogranatense* and var. *truxillense* [Rusby] Plowman) (supplementary table S1, Supplementary Material online). Species identities of the specimens were recorded from herbarium labels. We obtained taxonomic identities directly from specimens and compared them with specimens cited in local monographs (Pineda 2016), morphological characters proposed by White et al. (2021; supplementary table S1, Supplementary Material online), type specimens, protologs, and geographical distributions. Based on these resources, we excluded any candidate specimens that could not be confidently assigned to any *Erythroxylum* species of interest to this study and re-assigned the taxon for several cultivated samples ($n = 14$). A balanced sampling for each specimen was always expected as input, but the number of leaves on each specimen that were both intact and fully developed was variable. Hence, we set an upper limit of five leaves and a lower limit of two leaves per specimen. A leaf was considered fully developed if it was roughly the same size as the largest leaf on the sheet and showed shape characteristics consistent with botanical descriptions. Leaves were sampled only if they were pressed flat and intact from the base to the apex.

Image features that obstructed leaf contours (e.g. herbarium tape, twigs, and cracks) were manually removed with a digital paintbrush and images were segmented to isolate selected leaves. Outlines were extracted as coordinates using the "DiaOutline" software (Wishkerman and Hamilton 2018), implemented in R. Digital noise was removed from the outlines using the "coo.smooth" function in the R package "Momocs" v.1.3 (Bonhomme et al. 2014). A total of 200 equidistant, non-homologous pseudo-landmarks were sampled from each outline ("coo.sample" function in "Momocs"; Bonhomme et al. 2014). Two additional landmarks were manually defined at homologous positions on the base and apex of every leaf ("def_ldk" function in "Momocs"; Bonhomme et al. 2014).

Calculating Linear Metrics and Shape Variables

Outlines were standardized to 100 pixels per centimeters and linear metrics (length, width, and area) were recorded for each leaf using the "coo_toolbox" in "Momocs" (Bonhomme et al. 2014). To generate quantitative shape variables, leaf outlines were decomposed by elliptical Fourier analysis ("efourier" function in "Momocs"; Kuhl and Giardina 1982; Bonhomme et al. 2014). This type of outline analysis relies on the principle of Fourier series to express an outline as the sum of simpler trigonometric functions. The frequencies of each harmonic in the series are described by four scalar coefficients. These coefficients can be treated statistically as homologous, quantitative variables (Zelditch et al. 2004; Claude 2008). The minimum number of harmonics necessary for ample shape reconstruction was estimated through estimation of harmonic power (Lestrel 1997; Bonhomme et al. 2014). We computed 14 harmonics, representing a cumulative harmonic power near 100%. Because the coefficients of an elliptic Fourier descriptor are not invariant in size, rotation, shift, and starting point of chain-coding about a contour (Yoshioka et al. 2004), we standardized the Fourier coefficients prior to this shape analysis. Outlines were normalized with Bookstein baseline superimposition to the two homologous landmarks defined in outline preparation ("fgProcrustes" function in "Momocs"; Friess and Baylac 2003; Bonhomme et al. 2014). The dataset was passed through a second round of normalization, as part of the default "efourier" function in "Momocs".

Leaves with aberrant shapes heavily skew dimensionality reduction, thus we removed these from the dataset prior to further analyses. Aberrant leaf shape outliers were identified by modeling the shape variation for each species as normally distributed, with a confidence level of $1e^{-3}$ ("which_out" function in "Momocs"; Bonhomme et al. 2014). In total, 15 leaves with aberrant shapes were excluded. Upon inspection, aberrant contours were mostly due to distortion from the drying process.

Morphometrical Statistical Analyses

Shape and size are separate aspects of an organism's morphology, and natural patterns can be obscured if an investigation should confound the two (Christodoulou et al. 2020). Therefore, we treated linear metrics and shape variables separately in statistical analyses. To define morphological spaces for respective linear metric and shape cluster analyses, data dimensionality was reduced via PCA (Claude 2008). To visualize the distribution of the most common leaf shapes within a species, we performed PCA separately for each taxonomic subset of Fourier data, after which eigen-leaves were reconstructed using the reverse Fourier transform for the mean, and +/− 1.5 standard deviations from the mean, on the first two PCs of shape variation within a species ("PCcontrib" function in "Momocs"; Bonhomme et al. 2014). To assess the distribution of our taxon-binned leaf data along the first two axes of variation, corresponding to defining leaf traits, we constructed eigen-leaves for the first two PCs of the shape—EFA PCA, encompassing 95.4% of the variation combined (supplementary fig. S1E, Supplementary Material online).

To investigate the question of whether underlying structure in this morphometric dataset reflects current taxonomic boundaries, variation within this dataset was examined without a priori identifications. We inferred morphological groups probabilistically using model-based clustering using GMMs (Bouveyron and Brunet-Saumard 2014; Bouveyron et al. 2019). This method of cluster analysis assumes that continuously distributed data can be described as a mixture (weighted average) of G multivariate normal distributions (clusters). The aim is to identify the number of groups (G) and the geometric properties of their densities; different model parameters correspond to different hypotheses of group structure (Bouveyron et al. 2019).

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

Model selection was performed with the expectation-maximization (EM) algorithm, which estimates model parameters by maximum likelihood (Dempster et al. 1977). Measures of empirical support are based on the Bayesian information criterion (BIC; (Schwarz 1978), wherein Δ BIC expresses the gain in the explanatory power of the model when an additional group is considered by the algorithm. Based on Δ BIC, an EEV geometrically-constrained clustering model (ellipsoidal, equal volume, and shape) was selected for Fourier data and an EVE model (ellipsoidal, equal volume, and orientation) was selected for linear metrics. Each leaf was assigned to a cluster by soft classifications—expectations of the assignments under the applied probability model. GMMs were fitted using the "mclust" v.5.0 package (Scrucca et al. 2016), with parameters set to model two to fifteen groups. If leaves are well-classified by a given model, the conditional probability that a leaf belongs to one of the postulated groups should be close to 1. To statistically quantify equivalence between taxonomic clusters and GMM-inferred clusters, we applied the Rand index, whose value indicates the degree of correspondence of two different clustering outcomes from 0 = complete mismatch 1 = to perfect match (Rand 1971). We also applied the adjusted Rand index which corrects for chance (Hubert and Arabie 1985), carrying out all analyses in the "Fossil" package (Vavrek 2011) implemented in R.

Noisy variables in high-dimensional data are known to degrade the performance of model-based clustering, so variable selection is crucial (Bouveyron et al. 2019). Only three morphological variables were recorded for leaf size, so data were relatively low-dimensional. As such, all three principal components (PC) of their log-transformation were used in cluster analysis. Although Fourier descriptors of leaf shape are more complex in dimension, experimental exploration of a modeling approach to model selection (Maugis et al. 2009) did not provide valuable insight. Therefore, we retained the four PCs explaining most of the variation, deeming these most useful in defining group structure following previous studies (Sneath and Sokal 1975; Ezard et al. 2010). To visualize clustering for each leaf on a given herbarium specimen, bar charts were created to represent the number of leaves sampled, and their individual cluster assignments (Fig. 2b, supplementary figs. S7 and S8, Supplementary Material online).

Canonical Variants Analysis

To visualize relationships between the a priori taxonomically defined groups in morphological space, we also computed mean vectors for the shape variables. Euclidean distances between these were calculated using the R package "rdist" (Blaser 2020) and the output distance matrix was used for PcoA ("pcoa" implemented in R package "ape").

Dataset Filtering and Downsampling

Finally, we carried out four iterations of filtering our full dataset of 1,163 digitized, taxonomically verified leaf outlines from 342 specimens. The primary purpose was to produce a dataset excluding those samples assigned dubious IDs after the taxonomic evaluation and those which were cultivated outside of South America. This dataset consisted of 844 leaf outlines from 261 specimens. We reran the analyses using the filtered dataset for our main interpretation of the results. The second filtering iteration had the goal of demonstrating that there was no bias due to different sample sizes for different taxa. Therefore, for each pre-assigned taxonomic group, we downsampled to a random subset of 50 leaves (excluding *E. lineolatum*, for which this quota was not reached), resulting in a dataset of 335 leaves from 167 samples. The final two iterations involved retaining from our main interpretation dataset only the cultivated coca species with a main goal of computing the degree of correspondence between the taxonomic groups and GMM clusters. One of the iterations comprised the full dataset of *E. coca* and *E. novogranatense* individuals (404 samples total, 106 specimens) and the other is a subset of these, including only one leaf per specimen (97 samples total).

Genomics

Taxon Sampling and Genomic Data Mining

Our taxon sampling builds upon previous phylogenomic and population genomic studies of the coca clade (White et al. 2019, 2021). We generated new data from 18 accessions of coca and its closest wild relatives, following the current taxonomy (Plowman 1982; White et al. 2019, 2021) and spanning a large proportion of the geographical distribution (supplementary fig. S10, Supplementary Material online). Our sampling included 12 specimens of *E. gracilipes* (including a Kew (K) isotype), four of *E. cataractarum* (including a K holotype), one specimen of *E. lineolatum*, and one specimen of *E. foetidum* (to serve as outgroups, based on White et al. 2019), as well as one specimen of *E. n. novogranatense* (S.S. Renner 2888) (supplementary table S2, Supplementary Material online). For each sample, we weighed out 0.2 to 0.3 g of leaf tissue and pulverized this using steel ball bearings in a SPEX sample prep tissue homogenizer (SPEX Inc, NJ, USA). We extracted genomic DNA following a CTAB protocol, with the addition of 2% v/v 2-mercaptoethanol (Doyle and Doyle 1990). DNA extracts were purified using a bead clean up method with a 2:1 ratio of Ampure XP beads (Beckman Coulter, USA) to DNA eluate. DNA extracts were quantified using a Quantus fluorometer (Promega, USA) and the degree of DNA fragmentation was assessed using a 4200 TapeStation system (Agilent Technologies, USA). We conducted library preparation using a NEBNext Ultra II DNA Library Preparation Kit according to the manufacturer's protocol. Sequencing of DNA libraries was carried out on an Illumina HiSeq platform with a paired-end 150 bp configuration, by GeneWiz (South Plainfield, USA). To complement our *Erythroxylum* genomic dataset, we mined read data for wild relatives and cultivated species of coca from NCBIs sequence read archive (SRA) repository. This comprised raw sequence data from 155 herbarium specimens from which DNA had

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

been extracted and enriched for 427 nuclear genes via target capture with a custom-designed set of RNA probes (White et al. 2019, 2021) (hereafter: "mined dataset") (supplementary table S3, Supplementary Material online). This brought the total of samples used for some downstream analyses to 173.

Processing of Raw Read Data and Alignment to Target sequences

We trimmed raw reads (both sequenced and data-mined) using AdapterRemoval v.2.1 (Schubert et al. 2016). With the goal of retrieving nuclear genomic information, the trimmed reads were mapped against a set of low-copy nuclear genes previously selected to infer phylogenetic relationships in *Erythroxylum* (White et al. 2019). The reads were mapped using Bowtie v.2.3.4.1., after which they were realigned around indels using GATK v.3.8.1 and filtered for duplicates using picard-tools, all steps being executed within the pipeline "PALEOMIX" v.1.2.13 (Schubert et al. 2014). To obtain plastid alignments, the trimmed data was also mapped to a *E. novogranatense* plastid genome from NCBI (NC_030601) using the PALEOMIX pipeline and settings identical to those above.

Plastid and Nuclear Phylogenomic Analysis

To investigate phylogenetic relationships between recognized wild relatives and cultivated taxa of coca, we aimed to produce, for the first time, a plastid phylogeny for this clade, complemented by a nuclear phylogeny. To this end, we reconstructed sequences from reads mapped both to the full plastid genome and to the low-copy nuclear genes. We opted for a pseudohaploid approach to sequence generation owing to the low read depth across the nuclear genes—a consequence of our genome skimming approach.

For the plastome, using our data from 18 samples mapped to the *E. novogranatense* reference, we generated pseudohaploid sequences using ANGSD v.0.930 (Korneliussen et al. 2014) sampling the most common base (-doFasta 2), setting a minimum mapping quality of 25, a minimum base quality of 25 (loosening stringency to account for taxonomic breadth of handled samples and sample degradation respectively) and minimum depth of 10 (afforded by the high depth of recovered plastome data). We also set a threshold of genome completeness to at least 90% retain a sample. Using the mined dataset, we reconstructed sequences using the same procedure as with our data, with one difference of requiring a minimum depth of 3 to retain a site. As a result, we successfully retrieved plastomes of nine data-mined samples. The low rate of full plastome recovery from the mined data is unsurprising, given the authors' experimental aim of nuclear gene target enrichment, in which the majority of organellar DNA will be purged in laboratory steps. To build a phylogenetic representation of relationships between these 27 plastid genomes, we ran RAxML v.8.2.12 (Stamatakis 2014) under the rapid bootstrap analysis mode, with the GTR substitution model, the GAMMA model of rate heterogeneity (Yang 1994) and 500 bootstrap iterations.

To build a nuclear phylogeny that would correspond to our plastid tree, we aimed for identical sampling of the 27 specimens. We used our dataset of K samples aligned to the low-copy nuclear genes (see *Processing of raw read data and alignment to target sequences*), where 15 of the 18 samples yielded sufficient data for inclusion. We also included data-mined samples of *E. gracilipes*, *E. cataractarum*, *E. novogranatense*, and *E. coca* from which we had retrieved "by-catch" plastid data and used in the plastid tree (supplementary table S4a, Supplementary Material online). Finally, we data-mined accessions of *E. foetidum* ($n = 1$), *E. lineolatum* ($n = 1$), and *E. williamsii* Standl. ex Plowman ($n = 1$) from the NCBI repository as outgroups (based on the phylogenetic hypothesis of White et al. 2019) (supplementary table S4b, Supplementary Material online). Notably, a different sample of *E. foetidum* was used in the nuclear phylogeny, due to insufficient retrieval of nuclear genes from our genome skimmed *E. foetidum* 13,512 samples that had been used in the plastid phylogeny. For the nuclear alignments, we generated pseudohaploid sequences, by sampling a random allele at each site using ANGSD (-doFasta 1), requiring a minimum depth of 3, a minimum quality score of 25, and a minimum mapping quality of 25. Those samples for which at least 134 genes were genotyped to at least 80% of their total length were taken forward for phylogenomic analysis. Two samples were retained as outgroups—one *E. foetidum* sample as a first outgroup and one *E. williamsii* as the second outgroup. *Erythroxylum williamsii* was deemed an equivalent alternative on the basis of being as related to the ingroup as *E. lineolatum* (White et al. 2019), of which no samples passed our "minimum number of genes" threshold. Due to the thresholds, the final number of samples in the phylogeny was 25 (and the plastid phylogeny was downsampled accordingly). A total of 326 genes, qualifying based on their retention in at least 75% of the samples, were used in the final nuclear alignment. For each gene, we computed a gene tree using RAxML v.8.2.12 with the rapid bootstrap analysis mode, a GTR + GAMMA model of substitution and 500 bootstrap replicates. The resulting 326 gene trees were summarized into a multispecies coalescent tree phylogeny using gene-tree reconciliation in ASTRAL-III v.5.7.8 (Zhang et al. 2018). We did not collapse any of the branches since the lowest posterior probability support value was 0.4. To assess incongruence between the 326 gene trees used to make the species tree, we began by conducting a bipartition analysis with PhyParts (Smith et al. 2018). This method, which leverages added precision gained by using rooted gene trees as input, was used to summarize at each node the conflicting, concordant, and unique bipartitions with respect to the ASTRAL species tree topology. We visualized the output using a script that plots pie charts (Smith et al. 2015).

To further interrogate gene-tree support for species tree, we computed metrics which have complementary explanatory power: gene concordance factors (gCF; Baum

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

2007) and internode certainty (IC, ICA). For every bipartition of the tree, the gene concordance factor is the percentage of decisive trees that contain that bipartition (Minh et al. 2020), whereas IC is a quantified degree of certainty for individual bipartitions which considers the frequencies of the most frequent specifically conflicting bipartition (Salichos and Rokas 2013). ICA in turn considers all gene trees with a decreasing logarithmic weight (Salichos et al. 2014). We computed these metrics using a custom script designed to consider well-supported bipartitions on the basis of gene-tree bootstrap values in gene trees with variable support among branches (Simon et al. 2022). Finally, using fasta alignments, we also computed site concordance factors (Minh et al. 2020), which measure the percentage of decisive sites supporting a branch in the reference tree, using IQ-TREE2 (Minh et al. 2020).

As a supplementary exploration of the nuclear phylogeny, we concatenated alignments of the 326 genes and constructed an ML tree using RAxML v.8.2.12 with the GTRCAT model of substitution, to produce a reference tree against which we interrogated the ML gene trees. We also computed support metrics gCF, IC, and ICA for this tree using the same method as above. Finally, we plotted all output trees in FigTree v.1.4.4 (Rambaut 2016).

Population Genomic Analysis

To examine the shared variance between samples in our dataset of coca and its wild relatives ($n = 18$) and to explore patterns of clustering and genetic identity of the type specimens in the context of published data from wild relatives and cultivated coca (the "merged dataset," our data plus the mined dataset, $n = 173$), we conducted population genomic analyses based on GLs. We called the GLs using ANGSD v.0.930 (Korneliussen et al. 2014), setting a minimum *P*-value of $1e^{-6}$ to call a variant, a minimum quality score of 25, a minimum mapping quality of 25, and -remove_bads 1 to exclude any possible leftover duplicate or failed reads. We used the matrix of per-sample genotype likelihoods to carry out principal component analysis using PCangsd v.1.02 (Meisner and Albrechtsen 2018) with 10,000 iterations and requiring a minimum of 50% samples to be genotyped at any site. We also set a minimum minor allele frequency (MAF) of 0.2 (our dataset, $n = 18$) and an MAF of 0.05 (merged dataset, $n = 173$).

Furthermore, we modeled population structure for the merged dataset using "NGSadmix" v.32 (Skotte et al. 2013). We prepared the dataset by filtering for missingness of individuals at a site (tolerating up to 50% missingness) and setting an MAF threshold of 0.05. A total of 13,643 filtered sites were analyzed with default parameters for a maximum of 2,000 EM iterations. We modeled the per-individual ancestries assuming from 2 to 10 ancestral populations ($K = 2$ to 10), running ten iterations of the analysis per value of K with different random seeds. We computed the theoretical best value of K, using the Evanno method (Evanno et al. 2005), as implemented in CLUMPAK (Kopelman et al. 2015), identifying the highest values of ΔK.

To complement our phylogenomic analysis of plastids, we also computed a PCA based on genetic variation within the dataset of 18 sequenced and nine data-mined coca clade plastids. The goal of this was to explore any alternative insights on data clustering suggested by a phylogeny-free representation of the genetic variance. We ran PCangsd on this dataset, setting an MAF of 0.1, analyzing a total of 4,664 sites.

Molecular-Clock Dating Analyses

The time of origin and diversification of the tropical American lineages of *Erythroxylum* remain elusive. The first and only inference of absolute ages of *Erythroxylum* was conducted by White (2019), who relied on a penalized likelihood approach implemented on a phylogram derived from a concatenated supermatrix of 544 nuclear genes plus a fossil constraint applied to the stem node of the stem lineage of the tropical American *Erythroxylum* and secondary calibrations. However, the cultivated cocas were not included in this analysis, and *E. williamsii*, a taxon deemed sister to *E. lineolatum* and the coca clade by White et al. (2019), shared a common ancestor with *E. cataractarum* and *E. gracilipes* ~20 Ma. Because estimation of absolute ages based on concatenated supermatrices are known to be prone to produce erroneous branch lengths in the presence of incomplete lineage sorting (Ogilvie et al. 2017), here we opted to derive ages of divergence using a Bayesian Multispecies Coalescent approach that is known to reliably estimate calibrated time trees, even in the presence of gene-tree conflict. This approach uses multiple gene alignments, age prior calibrations, and prior knowledge of population memberships as input (Ogilvie et al. 2017; Yan et al. 2022). To obtain ages of divergence in the coca clade, we first estimated the root-to-tip tree variance, concordance, and length of our 326 maximum likelihood gene trees for 25 specimens (the same as input into species tree inference through gene-tree reconciliation in ASTRAL—see *Plastid and nuclear phylogenomic analysis*), using SortaDate v.1.0 (Smith et al. 2018). We assigned a priori population memberships from the results of NGSadmix analysis, for $K = 4$. We then selected the 20 most clock-like (i.e. lowest root-to-tip variance), congruent gene alignments that represented the panel of 25 samples included in the ASTRAL-III species tree analysis. This subset of alignments was imported into BEAUTi v.2.6 (Bouckaert et al. 2019) as unlinked partitions using the following priors: (*a*) a weakly informative secondary calibration applied to the root of the tree (i.e. the MRCA of *E. williamsii*, *E. lineolatum*, and the coca clade) modeled by an exponential distribution with lambda equal to 0.1498, so that 95% of the density is between 0 Ma and 20 Ma, and the mean 6.60 Ma; (b) an uncorrelated log-normal relaxed clock with a log-normal prior on the mean rate with parameters log-mean −6.909 and log standard deviation 1.1746, so that the mean rate with 95% of the density is between 0.0001 and 0.001 substitutions/site/Ma, a mutation rate for angiosperms (Lu et al. 2021); (c) a ploidy level of 2 (option "autosomal nuclear") as recommended for diploid

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

organisms and following the known ploidy levels of the cultivated cocas (Rodríguez Zapata 2015); (d) a Coalescent Constant Population tree model with a mean population size of 1.0 and a non-informative prior of 1/X (as recommended by Drummond and Bouckaert (2015) for datasets that contain population-level samplings); (e) a theoretical number of populations (K) of four, following the clusters produced by NGSadmix and that attained a high delta likelihood. We used the function findParams of the tbea package (Ballen and Reinales 2024) implemented in R (R Core Team 2021), to find the parameter values that best describe the probabilistic expectations in the priors. We ran the molecular-clock analyses for 1 billion generations, sampling every 50,000 states and ensuring that all parameters converged by attained effective sample sizes (ESS) > 200. Additionally, we ran three independent analyses and examined the posterior marginal distribution for each parameter, to check whether convergence was attained beyond within-change convergence measures such as ESS. We found that the three independent runs arrived at essentially the same posterior distributions for all the parameters in the model. Lastly, to ensure that all priors implemented were informative, we conducted one independent analysis where sampling was drawn from the priors, which revealed that indeed, our prior parameters were informative. Parameter and tree summaries were generated and visualized using Tracer v1.7.2 (Rambaut and Drummond 2013) and Treeannotator v.1.0 (available at https://beast.community/treeannotator).

Phylogenetic Networks

To explicitly estimate sources and directionality of gene flow within the coca clade, we made use of the 326 loci which comprise the same genes as used in our phylogenomic tree reconstructions (see *Plastid and nuclear phylogenomic analysis*), with which we inferred phylogenetic networks using a pseudolikelihood approach in species networks applying quartets (SNaQ) (Solís-Lemus and Ané 2016). This was implemented in the Julia package PhyloNetworks (Solís-Lemus et al. 2017). The input for SNaQ normally comprises gene-tree point estimates, however, we chose to use posterior gene tree densities estimated via Bayesian inference to account for topological uncertainty. Gene-tree posterior densities were estimated for each locus using MrBayes (Huelsenbeck and Ronquist 2001). After assessing topological convergence using the standard deviation of split frequencies < 0.05 (SDSF, Nylander et al. 2008), we formatted the trees to plain newick and subsampled the posterior gene-tree sample (a procedure demanded by memory constraints) using phyx (Brown et al. 2017). PhyloNetworks was then used for calculation of concordance factors (Solís-Lemus et al. 2017) using the composite tree sample. Network estimation was carried out using SNaQ with the same initial tree used in divergence time estimation and considering the number of hybridizations to be $h = 0$ to 3. Each analysis included 20 independent runs to aid optimization via more thorough exploration of parameter space. We applied the heuristic criterion of gradient stabilization to decide how many hybridizations to allow in the network (Solís-Lemus and Ané 2016; Tiley et al. 2023). Finally, a bootstrap analysis was carried out with 100 replicates, each running 20 parallel searches. Credible intervals for the inheritance proportion in minor and major hybrid edges were constructed from the quantiles 0.025 and 0.975 of their bootstrap samples. Phylogenetic networks were plotted using the package PhyloPlots, available at https://github.com/cecileane/PhyloPlots.jl.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Sue Zmartzy from RBG Kew for curatorial assistance with the herbarium specimens and Eve Lucas for consultation on knowledge pertaining to type specimens and taxonomic practice. We thank the laboratory staff at the Jodrell, and particularly Robyn Cowan for support enabling wet lab work. We are grateful to Carly Cowell for consultation on policy. A subset of computational (phylogenomic) analyses performed for this paper was conducted on the Smithsonian High Performance Cluster, Smithsonian Institution: https://doi.org/10.25572/SIHPC. A.A. acknowledges support from the Swedish Research Council (2019-05191), the Swedish Foundation for Strategic Environmental Research MISTRA (Project BioPath), and the Kew Foundation. O.A.P.E. is supported by the Sainsbury Orchid Fellowship at the Royal Botanic Gardens Kew and the Swiss Orchid Foundation. G.A.B. was supported through a FAPESP postdoctoral fellowship (#2023/07838-1) and a PDRA position funded by the BBSRC (grant BB/T01282X/1 awarded to M. dos Reis). G.A.B. thanks C. Solís-Lemus for her feedback on phylogenetic network analysis. Finally, we thank D. M. White for helpful discussions and three anonymous referees for their constructive input.

Author Contributions

O.A.P.E. and F.A.A. conceived the study, with further contributions from A.A., R.A.D., and N.A.S.P. A.A. provided financial support. F.A.A. conducted taxonomic evaluations prior to specimen analysis. R.A.D. conducted morphometric analyses, with mentorship from D.C. and further contributions from N.A.S.P. The initial morphometric analyses were conducted by R.A.D. for his MSc thesis (Queen Mary University of London and RBG Kew). N.A.S.P. conducted wet lab work. N.A.S.P., O.A.P.E., and L.K. conducted phylogenomic analyses. N.A.S.P. and O.A.P.E. conducted population genomic analyses. G.A.B. and O.A.P.E. carried out divergence time estimation analysis. G.A.B. conducted phylogenetic network analysis. S.S.R. provided samples and advised on taxonomy. R.C-B shared taxonomic expertise

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

and ideas for discussion. O.A.P.E. and N.A.S.P. produced figures, with contributions from R.A.D. and M.C. N.A.S.P. wrote the manuscript, with contributions from R.A.D., O.A.P.E., A.A., and F.A.A. All co-authors read and approved the final manuscript.

Data Availability

All high-throughput sequencing files are archived in the NCBI SRA database under the accession number (PRJNA1117374). Morphometric datasets are available at: 10.6084/m9.figshare.25709913.

References

Abarca CA, Martínez-Bauer A, Molina-Freaner F, Domínguez CA. The genetic consequences of evolving two sexes: the genetic structure of distylous and dioecious species of *Erythroxylum*. *Evol Ecol Res*. 2008:**10**(2):281–293.

Allaby RG, Fuller DQ, Brown TA. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc Natl Acad Sci U S A*. 2008:**105**(37):13982–13986. https://doi.org/10.1073/pnas.0803780105.

Allaby RG, Stevens CJ, Kistler L, Fuller DQ. Emerging evidence of plant domestication as a landscape-level process. *Trends Ecol Evol*. 2022:**37**(3):268–279. https://doi.org/10.1016/j.tree.2021.11.002.

Andrade IM, Mayo SJ, Kirkup D, Van den Berg C. Comparative morphology of populations of *Monstera* Adans. (Araceae) from natural forest fragments in Northeast Brazil using elliptic Fourier Analysis of leaf outlines. *Kew Bull*. 2008:**63**(2):193–211. https://doi.org/10.1007/s12225-008-9032-z.

Arias T, Niederhuth CE, McSteen P, Pires JC. The molecular basis of kale domestication: transcriptional profiling of developing leaves provides new insights into the evolution of a *Brassica oleracea* vegetative morphotype. *Front Plant Sci*. 2021:**12**:637115. https://doi.org/10.3389/fpls.2021.637115.

Azevedo DL. Pátu: o "pó da memória" dos conhecedores ye'pamasa. *Mundo Amazónico*. 2021:**12**(2):136–152. https://doi.org/10.15446/ma.v12n2.87541.

Ballen GA, Reinales S. tbea: tools for pre- and post-processing in Bayesian Evolutionary Analyses; 2024. https://www.biorxiv.org/content/10.1101/2024.06.18.599561v1.

Baum DA. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*. 2007:**56**(2):417–426. https://doi.org/10.1002/tax.562013.

Blaser N. rdist: calculate pairwise distances. R package version 0.0.5; 2020 [accessed 2022 Sep 6]. https://CRAN.R-project.org/package=rdist.

Bohm BA, Ganders FR, Plowman T. Biosystematics and evolution of cultivated coca (Erythroxylaceae). *Syst Bot*. 1982:**7**(2):121–133. https://doi.org/10.2307/2418321.

Bonhomme V, Picq S, Gaucherel C, Claude J. Momocs: outline analysis using R. *J Stat Softw*. 2014:**56**(13):1–24. https://doi.org/10.18637/jss.v056.i13.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, Maio D, *et al.* BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 2019:**15**(4):e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: a review. *Comput Stat Data Anal*. 2014:**71**:52–78. https://doi.org/10.1016/j.csda.2012.12.008.

Bouveyron C, Celeux G, Murphy TB, Raftery AE. *Model-based clustering and classification for data science: with applications in R*. Cambridge: Cambridge University Press; 2019.

Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. *Bioinformatics*. 2017:**33**(12):1886–1888. https://doi.org/10.1093/bioinformatics/btx063.

Chitwood DH, Kumar R, Headland LR, Ranjan A, Covington MF, Ichihashi Y, Fulop D, Jiménez-Gómez JM, Peng J, Maloof JN, *et al.* A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. *Plant Cell*. 2013:**25**(7):2465–2481. https://doi.org/10.1105/tpc.113.112391.

Chitwood DH, Otoni WC. Morphometric analysis of *Passiflora* leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. *GigaScience*. 2017:**6**(1):1–13. https://doi.org/10.1093/gigascience/giw008.

Christodoulou MD, Clark JY, Culham A. The Cinderella discipline: morphometrics and their use in botanical classification. *Bot J Linn Soc*. 2020:**194**(4):385–396. https://doi.org/10.1093/botlinnean/boaa055.

Claude J. *Morphometrics with R*. New York: Springer-Verlag; 2008.

Cristancho S, Vining J. Culturally defined keystone species. *Hum Ecol Rev*. 2004:**11**(2):153–164. https://www.jstor.org/stable/24707675.

Dávalos LM, Davalos E, Holmes J, Tucker C, Armenteras D. Forests, coca, and conflict: grass frontier dynamics and deforestation in the Amazon-Andes. *J Illicit Econ Dev*. 2021:**3**(1):74–96. https://doi.org/10.31389/jied.87.

Dávalos LM, Sanchez KM, Armenteras D. Deforestation and coca cultivation rooted in twentieth-century development projects. *BioScience*. 2016:**66**(11):974–982. https://doi.org/10.1093/biosci/biw118.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Methodol*. 1977:**39**(1):1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Dillehay TD, Ramírez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD. Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science*. 2008:**320**(5877):784–786. https://doi.org/10.1126/science.1156533.

Dillehay TD, Rossen J, Ugent D, Karathanasis A, Vásquez V, Netherly PJ. Early holocene coca chewing in northern Peru. *Antiquity*. 2010:**84**(326):939–953. https://doi.org/10.1017/S0003598X00067004.

Dodsworth S, Christenhusz MJM, Conran JG, Guignard MS, Knapp S, Struebig M, Leitch AR, Chase MW. Extensive plastid-nuclear discordance in a recent radiation of *Nicotiana* section *Suaveolentes* (Solanaceae). *Bot J Linn Soc*. 2021:**193**(4):546–559. https://doi.org/10.1093/botlinnean/boaa024.

Doyle JJ, Doyle JL. Isolation of plant DNA from fresh tissue. *Focus*. 1990:**12**:13–15.

Drummond AJ, Bouckaert RR. *Bayesian evolutionary analysis with BEAST*. Cambridge: Cambridge University Press; 2015.

Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*. 2005:**14**(8):2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x.

Ezard HGT, Pearson PN, Purvis A. Algorithmic approaches to aid species' delimitation in multidimensional morphospace. *BMC Ecol Evol*. 2010:**10**:175. https://doi.org/10.1186/1471-2148-10-175.

Flatt T. The evolutionary genetics of canalization. *Q Rev Biol*. 2005:**80**(3):287–316. https://doi.org/10.1086/432265.

Friess M, Baylac M. Exploring artificial cranial deformation using elliptic Fourier analysis of procrustes aligned outlines. *Am J Phys Anthropol*. 2003:**122**(1):11–22. https://doi.org/10.1002/ajpa.10286.

Fuller DQ, Denham T, Kistler L, Stevens C, Larson G, Bogaard A, Allaby R. Progress in domestication research: explaining expanded empirical observations. *Q Sci Rev*. 2022:**296**:107737. https://doi.org/10.1016/j.quascirev.2022.107737.

Galindo-Bonilla A, Fernández-Alonso JL. Plantas de coca en Colombia. Discusión crítica sobre la taxonomía de las especies cultivadas del género *Erythroxylum* P. Browne. *Revista De La*

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

Academia Colombiana De Ciencias Exactas, Físicas Y Naturales. 2010:**34**(133):455–465.
Ganders FR. Heterostyly in *Erythroxylum coca* (Erythroxylaceae). *Bot J Linn Soc*. 1979:**78**(1):11–20. https://doi.org/10.1111/j.1095-8339.1979.tb02182.x.
Gootenberg P. Between coca and cocaine: a century or more of US-Peruvian drug paradoxes, 1860–1980. *Hisp Am Hist Rev*. 2003:**83**(1):119–150. https://doi.org/10.1215/00182168-83-1-119.
Gross BL, Olsen KM. Genetic perspectives on crop domestication. *Trends Plant Sci*. 2010:**15**(9):529–537. https://doi.org/10.1016/j.tplants.2010.05.008.
Gupta S, Rosenthal DM, Stinchcombe JR, Baucom RS. The remarkable morphological diversity of leaf shape in sweet potato (*Ipomoea batatas*): the influence of genetics, environment, and G×E. *New Phytol*. 2020:**225**(5):2183–2195. https://doi.org/10.1111/nph.16286.
Hubert L, Arabie P. Comparing partitions. *J Classif*. 1985:**2**(1):193–218. https://doi.org/10.1007/BF01908075.
Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 2001:**17**(8):754–755. https://doi.org/10.1093/bioinformatics/17.8.754.
Jara-Muñoz OA, White DM, Rivera-Díaz O. Morphological and molecular evidence support elevating *Erythroxylum macrophyllum* var. *savannarum* (Erythroxylaceae) to specific status. *Syst Bot*. 2022:**47**(2):467–476. https://doi.org/10.1600/036364422X16512572274990.
Klein LL, Caito M, Chapnick C, Kitchen C, O'Hanlon R, Chitwood DH, Miller AJ. Digital morphometrics of two north American grapevines (*Vitis*: Vitaceae) quantifies leaf variation between species, within species, and among individuals. *Front Plant Sci*. 2017:**8**:373. https://doi.org/10.3389/fpls.2017.00373.
Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour*. 2015:**15**(5):1179–1191. https://doi.org/10.1111/1755-0998.12387.
Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014:**15**(1):356. https://doi.org/10.1186/s12859-014-0356-4.
Kuhl FP, Giardina CR. Elliptic Fourier features of a closed contour. *Comput Graph Image Proccess*. 1982:**18**(3):236–258. https://doi.org/10.1016/0146-664X(82)90034-X
Lestrel PE. *Fourier descriptors and their applications in biology*. Cambridge: Cambridge University Press; 1997. p. 632–647.
Lipson M. Applying f4-statistics and admixture graphs: theory and examples. *Mol Ecol Resour*. 2020:**20**(6):1658–1667. https://doi.org/10.1111/1755-0998.13230.
Lipson M, Loh PR, Levin A, Reich D, Patterson N, Berger B. Efficient moment-based inference of admixture parameters and sources of gene flow. *Mol Biol Evol*. 2013:**30**(8):1788–1802. https://doi.org/10.1093/molbev/mst099.
Lu Z, Cui J, Wang L, Teng N, Zhang S, Lam H-M, Zhu Y, Xiao S, Ke W, Lin J, *et al.* Genome-wide DNA mutations in Arabidopsis plants after multigenerational exposure to high temperatures. *Genome Biol*. 2021:**22**(1):160. https://doi.org/10.1186/s13059-021-02381-4.
Macbride JF. *Erythroxylaceae. Flora of Peru*. Chicago (IL), USA: Field Museum of Natural History; 1949. p. 632–647.
Maugis C, Celeux G, Martin-Magniette M-L. Variable selection for clustering with Gaussian mixture models. *Biometrics*. 2009:**65**(3):701–709. https://doi.org/10.1111/j.1541-0420.2008.01160.x.
Meisner J, Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*. 2018:**210**(2):719–731. https://doi.org/10.1534/genetics.118.301336.
Migicovsky Z, Li M, Chitwood DH, Myles S. Morphometrics reveals complex and heritable apple leaf shapes. *Front Plant Sci*. 2018:**8**:2185. https://doi.org/10.3389/fpls.2017.02185.
Minh BQ, Hahn MW, Lanfear R. New methods to calculate concordance factors for phylogenomic datasets. *Mol Biol Evol*. 2020:**37**(9):2727–2733. https://doi.org/10.1093/molbev/msaa106.
Naranjo P. Hallucinogenic plant use and related indigenous belief systems in the Ecuadorian Amazon. *J Ethnopharmacol*. 1979:**1**(2):121–145. https://doi.org/10.1016/0378-8741(79)90003-5.
Nascimento MGP, Mayo SJ, de Andrade IM. Distinguishing the Brazilian mangrove species *Avicennia germinans* and *A. schaueriana* (Acanthaceae) by elliptic Fourier analysis of leaf shape. *Feddes Repertorium*. 2021:**132**(2):77–107. https://doi.org/10.1002/fedr.202000025.
Negret PJ, Sonter L, Watson JEM, Possingham HP, Jones KR, Suarez C, Ochoa-Quintero JM, Maron M. Emerging evidence that armed conflict and coca cultivation influence deforestation patterns. *Biol Conserv*. 2019:**239**:108176. https://doi.org/10.1016/j.biocon.2019.07.021.
Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*. 2008:**24**(4):581–583. https://doi.org/10.1093/bioinformatics/btm388.
Ogilvie HA, Bouckaert RR, Drummond AJ. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol Biol Evol*. 2017:**34**(8):2101–2114. https://doi.org/10.1093/molbev/msx126.
Pellicer J, Saslis-Lagoudakis CH, Carrió E, Ernst M, Garnatje T, Grace OM, Gras A, Mumbrú M, Vallès J, Vitales D, *et al.* A phylogenetic road map to antimalarial *Artemisia* species. *J Ethnopharmacol*. 2018:**225**:1–9. https://doi.org/10.1016/j.jep.2018.06.030.
Pérez-Escobar OA, Bellot S, Przelomska NAS, Flowers JM, Nesbitt M, Ryan P, Gutaker RM, Gros-Balthazard M, Wells T, Kuhnhäuser BG, *et al.* Molecular clocks and archeogenomics of a late period Egyptian date palm leaf reveal introgression from wild relatives and add timestamps on the domestication. *Mol Biol Evol*. 2021:**38**(10):4475–4492. https://doi.org/10.1093/molbev/msab188.
Pérez-Escobar OA, Tusso S, Przelomska NAS, Wu S, Ryan P, Nesbitt M, Silber MV, Preick M, Fei Z, Hofreiter M, *et al.* Genome sequencing of up to 6,000-year-old Citrullus seeds reveals use of a bitter-fleshed species prior to watermelon domestication. *Mol Biol Evol*. 2022:**39**(8):msac168. https://doi.org/10.1093/molbev/msac168.
Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. 2012:**8**(11):e1002967. https://doi.org/10.1371/journal.pgen.1002967.
Pineda YAM. *Revisión taxonómica de Erythroxylum (Erythroxylaceae) para Colombia. Tesis de pregrado, Biología*. Bogotá, Colombia: Universidad de los Andes; 2016.
Piperno DR. Assessing elements of an extended evolutionary synthesis for plant domestication and agricultural origin research. *Proc Natl Acad Sci U S A*. 2017:**114**(25):6429–6437. https://doi.org/10.1073/pnas.1703658114.
Pironon S, Ondo I, Diazgranados M, Allkin R, Baquero AC, Cámara-Leret R, Cantiero C, Dennehy-Carr Z, Govaerts R, Hargreaves S, *et al.* The global distribution of plants used by humans. *Science*. 2024:**383**(6680):293–297. https://doi.org/10.1126/science.adg8028.
Plowman T. The identity of Amazonian and Trujillo Coca. *Bot Mus Lealf Harv Univ*. 1979:**27**(1--2):45–68. https://doi.org/10.5962/p.295220.
Plowman T. The identification of coca (*Erythroxylum* species): 1860–1910. *Bot J Linn Soc*. 1982:**84**(4):329–353. https://doi.org/10.1111/j.1095-8339.1982.tb00368.x.
Plowman T. The ethnobotany of Coca (*Erythroxylum* spp., Erythroxylaceae). *Adv Econ Bot*. 1984:**1**:62–111.
Plowman T. Coca chewing and the botanical origins of coca (*Erythroxylum* spp.) in Latin America. In: Pacini D, Franquemont C, editors. *Coca and cocaine: effects on people and policy in Latin America*. Cambridge (MA): Cultural Survival, Inc. and LASP, Cornell University; 1986. p. 5–34.
Plowman T, Hensold N. Names, types, and distribution of neotropical species of *Erythroxylum* (Erythroxylaceae). *Brittonia*. 2004:**56**(1):1–53. https://doi.org/10.1663/0007-196X(2004)056[0001:NTADON]2.0.CO;2.

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

Plowman T, Rivier L. Cocaine and cinnamoylcocaine content of *Erythroxylum* species. *Ann Bot*. 1983:**51**(5):641–659. https://doi.org/10.1093/oxfordjournals.aob.a086511.

Prates L, Politis GG, Perez SI. Rapid radiation of humans in South America after the last glacial maximum: a radiocarbon-based study. *PLoS One*. 2020:**15**(7):e0236023. https://doi.org/10.1371/journal.pone.0236023.

Rambaut A. *Figtree—tree figure drawing tool, version 1.4.3*. United Kingdom: Institute of Evolutionary Biology, University of Edinburgh; 2016.

Rambaut A, Drummond AJ. Tracer v1.6; 2013 [accessed 2023 Apr 29]. http://tree.bio.ed.ac.uk/software/tracer/.

Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc*. 1971:**66**(336):846–850. https://doi.org/10.1080/01621459.1971.10482356.

R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021. https://www.R-project.org/.

Restrepo DA, Saenz E, Jara-Muñoz OA, Calixto-Botía IF, Rodríguez-Suárez S, Zuleta P, Chavez BG, Sanchez JA, D'Auria JC. *Erythroxylum* in focus: an interdisciplinary review of an overlooked genus. *Molecules*. 2019:**24**(20):3788. https://doi.org/10.3390/molecules24203788.

Rincón-Ruiz A, Kallis G. Caught in the middle, Colombia's war on drugs and its effects on forest and people. *Geoforum*. 2013:**46**:60–78. https://doi.org/10.1016/j.geoforum.2012.12.009.

Rodríguez Zapata FV. Genome size and descriptors of leaf morphology as indicators of hybridization in Colombian cultigens of coca *Erythroxylum* spp. [master's thesis]. Bogotá, Colombia: Universidad de Los Andes; 2015.

Rothhammer F, Dillehay TD. The late Pleistocene colonization of South America: an interdisciplinary perspective. *Ann Hum Genet*. 2009:**73**(Pt 5):540–549. https://doi.org/10.1111/j.1469-1809.2009.00537.x.

Rury PM. Systematic anatomy of *Erythroxylum* P. Browne: practical and evolutionary implications for the cultivated cocas. *J Ethnopharmacol*. 1981:**3**(2-3):229–263. https://doi.org/10.1016/0378-8741(81)90056-8.

Rury PM, Plowman T. Morphological studies of archeological and recent coca leaves (*Erythroxylum* spp.). *Bot Mus Lealf Harv Univ*. 1983:**29**(4):297–341. https://doi.org/10.5962/p.168665.

Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*. 2013:**497**(7449):327–331. https://doi.org/10.1038/nature12130.

Salichos L, Stamatakis A, Rokas A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol*. 2014:**31**(5):1261–1271. https://doi.org/10.1093/molbev/msu061.

Sayıncı B, Kara M, Ercişli S, Duyar Ö, Ertürk Y. Elliptic Fourier analysis for shape distinction of turkish hazelnut cultivars. *Erwerbs-Obstbau*. 2015:**57**(1):1–11. https://doi.org/10.1007/s10341-014-0221-7.

Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, *et al*. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc*. 2014:**9**(5):1056–1082. https://doi.org/10.1038/nprot.2014.063.

Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes*. 2016:**9**:88. https://doi.org/10.1186/s13104-016-1900-2.

Schultes RE. The Amazonia as a source of new economic plants. *Econ Bot*. 1979:**33**(3):259–266. https://doi.org/10.1007/BF02858251.

Schultes RE. Coca in the northwest Amazon. *J Ethnopharmacol*. 1981:**3**(2-3):173–194. https://doi.org/10.1016/0378-8741(81)90053-2.

Schulz OE. *Erythroxylaceae*. Vol. 29. Leipzig: Engelmann; 1907.

Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978:**6**(2):461–464. https://doi.org/10.1214/aos/1176344136.

Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J*. 2016:**8**(1):289–317. https://doi.org/10.32614/RJ-2016-021.

Silva GAR, Antonelli A, Lendel A, Moraes EDM, Manfrin MH. The impact of early Quaternary climate change on the diversification and population dynamics of a South American cactus species. *J Biogeogr*. 2018:**45**(1):76–88. https://doi.org/10.1111/jbi.13107.

Simon MF, Mendoza Flores JM, Liu H-L, Martins MLL, Drovetski SV, Przelomska NAS, Loiselle H, Cavalcanti TB, Inglis PW, Mueller NG, *et al*. Phylogenomic analysis points to a South American origin of *Manihot* and illuminates the primary gene pool of cassava. *New Phytol*. 2022:**233**(1):534–545. https://doi.org/10.1111/nph.17743.

Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. *Genetics*. 2013:**195**(3):693–702. https://doi.org/10.1534/genetics.113.154138.

Smith SA, Brown JW, Walker JF. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. *PLoS One*. 2018:**13**(5):e0197433. https://doi.org/10.1371/journal.pone.0197433.

Smith SA, Moore MJ, Brown JW, Yang Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol Biol*. 2015:**15**:150. https://doi.org/10.1186/s12862-015-0423-0.

Sneath PHA, Sokal RR. Numerical taxonomy. The principles and practice of numerical classification. *Q Rev Biol*. 1975:**50**(4):525–526. https://doi.org/10.1086/408956.

Soares MLC, Mayo SJ, Gribel R, Kirkup D. Elliptic Fourier analysis of leaf outlines in five species of *Heteropsis* (Araceae) from the reserva florestal adolpho ducke, manaus, amazonas, Brazil. *Kew Bull*. 2011:**66**(3):463–470. https://doi.org/10.1007/s12225-011-9290-z.

Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet*. 2016:**12**(3):e1005896. https://doi.org/10.1371/journal.pgen.1005896.

Solís-Lemus C, Bastide P, Ané C. PhyloNetworks: a package for phylogenetic networks. *Mol Biol Evol*. 2017:**34**(12):3292–3298. https://doi.org/10.1093/molbev/msx235.

Spriggs EL, Schmerler SB, Edwards EJ, Donoghue MJ. Leaf form evolution in *Viburnum* parallels variation within individual plants. *Am Nat*. 2018:**191**(2):235–249. https://doi.org/10.1086/695337.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014:**30**(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Sultan SE. Phenotypic plasticity for plant development, function and life history. *Trends Plant Sci*. 2000:**5**(12):537–542. https://doi.org/10.1016/S1360-1385(00)01797-0.

Tiley GP, Flouri T, Jiao X, Poelstra JW, Xu B, Zhu T, Rannala B, Yoder AD, Yang Z. Estimation of species divergence times in presence of cross-species gene flow. *Syst Biol*. 2023:**72**(4):820–836. https://doi.org/10.1093/sysbio/syad015.

UNODC. Censo de cultivos de coca 2012—Colombia; 2012 [accessed 2023 Aug 6]. https://www.unodc.org/documents/colombia/2013/Agosto/censo_de_cultivos_de_coca_2012_BR.pdf.

UNODC. Colombia. Survey of territories affected by illicit crops—2016; 2016 [accessed 2023 Aug 6]. https://www.unodc.org/documents/crop-monitoring/Colombia/Colombia_Coca_survey_2016_English_web.pdf.

UNODC. Colombia. Monitoreo de los territorios con presencia de cultivos de coca 2022; 2023 [accessed 2024 Feb 17]. https://www.unodc.org/documents/colombia/2023/septiembre-9/INFORME_MONITOREO_DE_TERRITORIOS_CON_PRESENCIA_DE_CULTIVOS_DE_COCA_2022.pdf.

Vavrek MJ. Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontol Electron*. 2011:**14**(1):16.

Vergara H, Becking M, Posada V, Troyano D, Baptiste Ballera B, Len Pulido J. Estrategia de investigación regional sobre los usos y potencialidades de la hoja de coca: un instrumento para incidir en

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025

las políticas públicas agrícolas, alimenticias y farmacéuticas en Colombia. 2022. https://doi.org/10.57793/9789587566789.
Viruel J, Kantar MB, Gargiulo R, Hesketh-Prichard P, Leong N, Cockel C, Forest F, Gravendeel B, Pérez-Barrales R, Leitch IJ, *et al*. Crop wild phylorelatives (CWPs): phylogenetic distance, cytogenetic compatibility and breeding system data enable estimation of crop wild relative gene pool classification. *Bot J Linn Soc*. 2021:**195**(1):1–33. https://doi.org/10.1093/botlinnean/boaa064.
Watson A, Arce M. Indigenous mobilization and territorial ordering in the Amazon. *Polit Geogr*. 2024:**108**:103013. https://doi.org/10.1016/j.polgeo.2023.103013.
Wells T, Carruthers T, Muñoz-Rodríguez P, Sumadijaya A, Wood JRI, Scotland RW. Species as a heuristic: reconciling theory and practice. *Syst Biol*. 2022:**71**(5):1233–1243. https://doi.org/10.1093/sysbio/syab087.
White D. Biogeography, diversification, and domestication in the coca family (Erythroxylaceae). *Doctoral dissertation*. University of Illinois at Chicago; 2019.
White DM, Huang JP, Jara-Muñoz OA, Madriñán S, Ree RH, Mason-Gamer RJ. The origins of coca: museum genomics reveals multiple independent domestications from progenitor *Erythroxylum gracilipes*. *Syst Biol*. 2021:**70**(1):1–13. https://doi.org/10.1093/sysbio/syaa074.
White DM, Islam MB, Mason-Gamer RJ. Phylogenetic inference in section Archerythroxylum informs taxonomy, biogeography, and the domestication of coca (*Erythroxylum* species). *Am J Bot*. 2019:**106**(1):154–165. https://doi.org/10.1002/ajb2.1224.
Wishkerman A, Hamilton PB. Shape outline extraction software (DiaOutline) for elliptic Fourier analysis application in morphometric studies. *App Plant Sci*. 2018:**6**(12):e01204. https://doi.org/10.1002/aps3.1204.
Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Syst Biol*. 2022:**71**(2):367–381. https://doi.org/10.1093/sysbio/syab056.
Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 1994:**39**(3):306–314. https://doi.org/10.1007/BF00160154.
Yoshioka Y, Iwata H, Ohsawa R, Ninomiya S. Analysis of petal shape variation of *Primula sieboldii* by elliptic Fourier descriptors and principal component analysis. *Ann Bot*. 2004:**94**(5):657–664. https://doi.org/10.1093/aob/mch190.
Zelditch ML, Swiderski DL, Sheets HD, Fink WL. *Geometric morphometrics for biologists: a primer*. San Diego (CA): Elsevier Academic Press; 2004.
Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*. 2018:**19**(Suppl 6):153. https://doi.org/10.1186/s12859-018-2129-y.

Downloaded from https://academic.oup.com/mbe/article/41/7/msae114/7708214 by guest on 01 June 2025