ORIGINAL RESEARCH

Human influence on the distribution of cacao: insights from remote sensing and biogeography

Matheus Colli-Silva[1,2] · James Edward Richardson[3,4,5,6] · Antonio Figueira[7] · José Rubens Pirani[1]

Abstract

Cacao (*Theobroma cacao*, Malvaceae) is an important tree crop in Africa and in the Americas. Current genomic evidence suggests that its original range in Tropical Americas was smaller than its current distribution and that human-mediated dispersal occurred before European colonization. This includes regions like Mesoamerica and Eastern Amazonia where cacao is supposedly naturally occurring. In this study, we utilize remote sensing and land use data to examine the influence of human activities on cacao-growing regions and explore patterns between cacao distribution and anthropized areas. By evaluating nearly nine thousand preserved specimen collections, we worked with a comprehensive occurrence dataset that considers taxonomy and distribution. We then analyzed remote sensing images of specimen locations and compared land use profiles of regions into which cacao was introduced with documented native areas. Our findings revealed a clear association between anthropized areas and cacao specimens, with the majority located in areas strongly affected by human activities. Conversely, regions closer to the proposed native range of cacao exhibit less human impact. These results, while accounting for sampling bias, reinforce the idea that humans may have played a significant role in cacao's dispersal, even in parts of the Amazon where its native status remains uncertain. The discussion on cacao's native range and identification of introduced areas hold implications for jurisdiction, access to genetic resources, and conservation efforts. Additionally, it is relevant to debates surrounding the repatriation of genetic data of economically important crops. Understanding the historical human influence on cacao's distribution is crucial for addressing issues of crop improvement, conservation, and sustainable use.

Keywords Amazonian crops · Biogeography · Center of diversity · Cocoa · Domestication · Geoprocessing

---

Communicated by Daniel Sanchez Mata.

Extended author information available on the last page of the article

Published online: 10 February 2024

⛎ Springer

Introduction

Among the vast array of more than 7,000 edible plants, over 400 species are considered as major food crops (Ulian et al. 2020), ranging from minimal alterations to their wild phenotype to the development of distinct cultivars or cultigens with specific features advantageous to humans. In fact, human communities have actively managed and selected plants since the early Holocene (Clement et al. 2010; Watling et al. 2018), favoring traits that enhance edibility, productivity, nutritional value, or other desirable characteristics. While comprehensive syntheses exist for the geographical origins of selected crops such as maize (Wang et al. 2017), tobacco (Duke et al. 2021), or rice (Sweeney and McCouch 2007; Gutaker et al. 2020), the geographic history of many other important crops remains elusive.

One such crop with a rich history is cacao (*Theobroma cacao L.,* Malvaceae), which has been cultivated and utilized by Meso-American societies centuries before Columbus arrival in 1492, primarily for its use in a bitter drink (Cuatrecasas 1964; Bletter and Daly 2006). Today, cacao is cultivated for its fermented seeds, which are essential for chocolate production. Cacao, known for its shade tolerance, thrives in various agroforestry scenarios, whether it's under thinned forests typical of areas in the Brazilian Atlantic coast (*cabruca* system), temporary shading amidst food crops, or in the presence of introduced tree species for definitive shading (Sambuishi et al. 2012; Gama-Rodrigues et al. 2021). The amalgamation of cacao with both woody (*e.g.*, *Erythrina spp.*, *Hevea spp.*) and non-woody species (*e.g.*, banana, cassava) exemplifies the compatibility and sustainability of multistrata production systems (Gama-Rodrigues et al. 2021). Cacao-based agroforestry systems are pivotal for sustainable development in emerging countries, especially in South America and Mesoamerica (Zequeira-Larios et al. 2021). These systems emulate the attributes of natural forests and mitigate human pressure on the original forest cover in cacao regions.

Beyond plantations, cacao, along with other related species, thrives in lowland rainforests of the Americas. Traditionally assigned to two related genera, *Theobroma L.* (22-23 species, Cuatrecasas 1964) and *Herrania* Goudot (17 species; Schultes 1958), these understory trees bear fruits that are typically known as "*cacao,*" "*cupuí,*" "*cacaorana*" or similar, with cacao (*T. cacao*) being the most widely recognized species. It should be noted that most of these species are primarily native to the Neotropical region (Colli-Silva et al. 2023a). Very few species, however, are cultivated or extensively used by humans, whose which include e.g. *T. grandiflorum* (Willd. ex Spreng.) K. Schum. and *T. bicolor* Humb & Bonpl. (Cuatrecasas 1964; Bletter and Daly 2006).

The history of cacao is more complex than previously assumed. The species would have been originated around ten million years ago (Richardson et al. 2015), but the role humans played in establishing its current broad distribution across the Tropics is not entirely clear. Genomic studies have revealed that cacao's domestication involved the introduction of Ecuadorian varieties into Mesoamerica, likely facilitated by indigenous populations (Cornejo et al. 2018). Furthermore, archaeological findings provided insights into the consumption of cacao in present-day Ecuador over 5,000 years ago (Zarillo et al. 2018), emphasizing the intricate history of cacao's cultivation and utilization in the Americas and underscoring the role played by indigenous societies in its dispersal and consumption in areas where it is found growing today, even in a seemingly "wild" condition. Historical evidence suggests that cacao was introduced into the South American Atlantic coastal forests in the eighteenth century, from where it spread to West Africa during the period of European colonization (Soria 1970; Motamayor et al. 2003).

Exploring the depths of Amazonian primary forests reveals a notable presence of cacao plants within areas exhibiting diverse levels of human impact, encompassing abandoned farms, degraded lands, and seemingly untouched dense forests. This distribution pattern implies that the historical native range of cacao might have been more restricted in the past due to its presumably limited natural dispersal abilities (as described e.g., in Cuatrecasas 1964). Human intervention has played a pivotal role in the introduction, selection, and hybridization of cacao populations, contributing to the development of present-day cultivars (Cornejo et al. 2018), akin to the processes observed in other crops in the Amazon (see Clement et al. 2015; Levis et al. 2017). Consequently, cacao trees demonstrate adaptability across a spectrum of environments ranging from anthropized areas to primary or secondary forests.

Therefore, investigating the influence of human activity on the geographic distribution of cacao not only enriches our comprehension of its original habitat but also holds implications for discussing genetic resources, enhancing crop development, conservation efforts, and discussions concerning the retention or retrieval of genetic data.

The objective of this study is to assess the impact of human influence on the distribution of cacao by comparing areas identified as native ranges with introduced areas. To achieve this, we compiled a comprehensive occurrence dataset by evaluating preserved specimen collections to better allocate the native ranges and introduced areas according to several criteria. Remote sensing images were obtained for locations where cacao specimens are found, and land use profiles were compared between introduced areas and the hypothesized center of origin of the species. We aim to provide insights into the role of human influence in the current distribution of cacao and discuss its potential implications for various aspects, including jurisdiction, access to genetic resources, conservation, and repatriation of genetic data. By doing so, we intend to contribute to both policymaking and academia, offering valuable information and novel perspectives on how cacao's geographic distribution should be interpreted.

Material and methods

Literature survey and study area

A comprehensive literature survey was conducted encompassing studies that discuss the origin, distribution, and dispersal of cacao before and after human influence. This survey included classic botanical monographs of *Theobroma* by Bernoulli (1869), Schumann (1886), and Ducke (1925, 1940), as well as the most recent taxonomic treatment available for the genus by Cuatrecasas (1964). In addition, agronomic and historic literature was consulted to understand the association between known cultivars and the botanical circumscriptions of the species, which helped with formulating hypotheses regarding the origin and dispersal of cacao. Relevant works consulted in this regard included Morris (1882), Preuss (1901), van Hall (1914), Cheesman (1927, 1929, 1932, 1944), Pittier (1924), Pittier and Chevalier (1925), Pittier et al. (1926), Pound (1938, 1945), Ciferri and Ciferri (1957), Schultes (1984), Figueira et al. (1994), and Bartley (2005). Furthermore, studies that employed genomic data to delimit the origin and distribution of cacao were also reviewed. These studies, consulted for their insights, included Laurent et al. (1994), N'Goran et al. (1994), Motamayor et al. (2002; 2008), Motamayor and Lanaud (2002),

Thomas et al. (2012), Clement et al. (2015), Lachenaud and Motamayor (2017), Osorio-Guarin et al. (2017), Cornejo et al. (2018), Zarillo et al. (2018), and Fouet et al. (2022).

This survey was important because it allowed us to achieve a more accurate determination of the specific regions where *T. cacao* occurs as wild, and it served as the foundation for defining major and minor regions of interest (Fig. 1) that were used for downstream analyses. The major areas were categorized as follows: (1) Areas of late introduction, where human introduction after the Pre-Columbian era (i.e., before Christopher Columbus arrived in 1492 in the Americas) is well-documented in the literature; (2) Potential early introduction sites, where cacao may have been introduced during the Pre-Columbian era; (3) Areas of early introduction, where human introduction during the Pre-Columbian era is certain based on the literature; and (4) Potential native area of *T. cacao* based on the most recent evidence compiled in here (namely Bartley 2005; Thomas et al. 2012; Clement et al. 2015; Cornejo et al. 2018; Fouet et al. 2022). To achieve a more detailed resolution for Tropical Americas, we further subdivided the region into smaller scales using the biogeographical delimitations proposed by Morrone (2014) (Fig. 1).

Cacao relies on mammals, like rodents and primates, and birds for dispersal in nature (Cuatrecasas 1964; Silva et al. 2010), and it has dispersion limitations due to fruit characteristics, which have affected its distribution. In this sense, it is important to make clear that the term "introduced specimen" used in this study refers to any specimen that was intentionally introduced by humans at a specific time and location, whereas the term "native area" refers to the region where only wild specimens were naturally dispersed, without human intervention.

Specimen occurrence data

The primary occurrence dataset for *T. cacao* used in this study was compiled as part of a larger dataset for all *Theobroma* and *Herrania* species (Colli-Silva et al. 2023a). This dataset was constructed through an extensive literature survey and incorporated data obtained from the GBIF repository (Global Biodiversity Information Facility; GBIF.org 2020). GBIF-mobilized data underwent rigorous review processes, including georeferencing procedures and thorough taxonomic revision of nearly nine thousand preserved specimen collections of *Theobroma* and *Herrania* species (see Colli-Silva et al. 2023a).

For this study, only preserved specimens of *T. cacao* were extracted from the larger dataset, and we excluded records with the same geographic location. This decision was made because preserved specimen collections provide more reliable and accurate geographic data compared to human observations, photographs, or other sources of information. The occurrence information derived from these preserved specimens has been included as Supplementary Information and can be found in Appendix S1. This data set incorporated a total of 637 locations (unique geographic point occurrences).

Acquisition of remote sensing data

All downstream analyses were conducted in R v. 4.2.1 and Python v. 3.10.2 environments (van Rossum and Drake 1995; R Core Team 2021). Satellite images were obtained iteratively from the Sentinel-2 collection within the Google Earth Engine API platform (Gorelick et al. 2017) in Python. The search was restricted to images captured between January 1, 2020, and January 1, 2022. Briefly, Sentinel-2 images provide multispectral surface
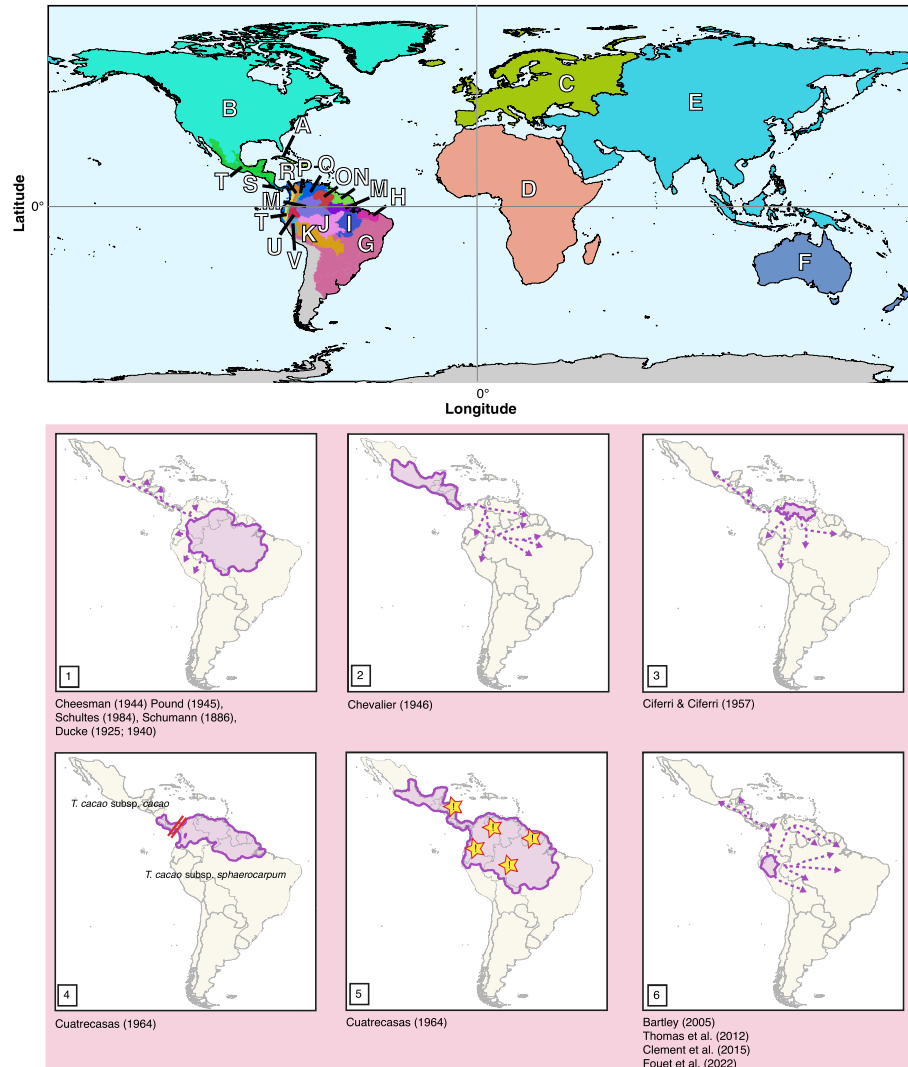
Fig. 1 Summary of the main regions considered in this study and the presumed scenarios of the origin of *T. cacao* and its dispersal through the Neotropics, as proposed by various authors (detailed in Methods, "Literature survey and study area" section). Dispersal events are represented by purple arrows. In Scenario 4 (Cuatrecasas 1964), the red lines indicate the emergence of the Panama isthmus that facilitated overland dispersal of terrestrial organisms. In Scenario 5 (Cuatrecasas 1964), the yellow "!" stars represent the occurrence of several mutations that, according to these hypotheses, would have originated in various cacao morphotypes. Minor areas (sensu Morrone 2014) considered in this study were grouped into major regions based on the origin of the specimens, as described in Methods ("Literature survey and study area" section). The major regions include: (1) Unequivocal late introduction: Antilles (A), Canada and United States (B), Europe (C), Africa (D), Asia (E), Oceania and Pacific islands (F), South American Atlantic coast (G). (2) Potential early introduction: Pará province (H), Xingu-Tapajós province (I), Madeira province (J), Rondônia province (K), Imerí province (L), Roraima province (M), Guianan Lowlands province (N), Pantepui province (O), Paramo province (P). (3) Unequivocal early introduction: Pacific dominion (Q), Magdalena province (R), Puntarenas-Chiriquí province (S), Mesoamerican dominion (T). (4) Potential native area: Napo province (U), Ucayali province (V)

reflectance data with bands in the visible and near-infrared regions of the electromagnetic spectrum at a resolution of 10 m.

For each of the 637 locations, the most suitable image for the area of interest was selected, ensuring that the chosen images had cloud coverage of less than 10%. The selected images were then reprojected to the Universal Transverse Mercator zone, and the Normalized Difference Vegetation Index (NDVI) was calculated for a buffer area standardized as 500×500 m^2 centered around each point location. The NDVI was selected because of its wide and intuitive usage in vegetation analyses (Rouse et al. 1974) which we consider appropriate for this exploratory analysis. In short, NDVI is computed as the ratio of the difference to the sum of reflectance values in the red and near-infrared regions (Rouse et al. 1974). NDVI values range from -1 to 1 and are used to classify vegetations into categories, distinguishing dense forests from sparse vegetations, grasslands, water bodies, barren lands, and built-up areas.

NDVI values were reclassified into the following categories, based on an classification based on NDVI threshold values, defined as folows: (1) water bodies (NDVI < 0); (2) barren lands and built-up areas ($0 \leq NDVI < 0.18$); (3) grasslands and agricultural lands ($0.18 \leq NDVI < 0.27$); (4) sparse vegetations ($0.27 \leq NDVI < 0.36$); and (5) dense forests (NDVI ≥ 0.36). In this study, "areas of human influence" refer to regions showing indications of anthropogenic presence or influence, such as roads, deforested areas, agricultural lands, or other areas that exhibit signatures detectable through standard remote sensing analyses. Areas of human influence were standardized based on NDVI values between 0 and 0.27.

NDVI profile analyses

NDVI profiles were generated for each specific point location (for all 637 locations) and then contrasted across the various areas delineated in "Literature survey and study area" section. Several key variables were collected for each location, including: (i) the median NDVI value of the site, (ii) the proportional occurrence of dense vegetations, and (iii) the proportional occurrence of areas influenced by human activity. This analysis involved buffer extraction and spatial data manipulation carried out using R packages "raster" v. 2.0-12 (Hijmans 2023), "sp" v. 1.5-0 (Pebesma and Bivand 2005; Bivand et al. 2013), and "sf" v. 1.0-8 (Pebesma and Bivand 2023). The metrics were calculated from all pixels included inside the buffer taken for each site.

In order to discern statistical differences, we investigated whether the NDVI scores within explicitly designated regions differed from the other major areas outlined in "Literature survey and study area" section. The assessment of variances between these classes employed a Kolmogorov–Smirnov test (Ks-test) with a significance threshold set at 0.05. The statistical evaluation was conducted using the "dgof" v. 1.4 package (Arnold and Emerson 2011) within R.

Addressing potential sampling biases

In theory, all specimen collections are often biased towards areas that are easily accessible and closer to regions of human influence (Oliveira et al. 2016). To ensure that our results account for these potential biases, we specifically addressed two concerns associated with our dataset. By addressing these biases associated with accessibility and anthropic

influence, we aimed to enhance the robustness of our findings and provide a more comprehensive understanding of the observed patterns in our empirical data.

First, to address potential sampling bias in our cacao specimens, we conducted additional investigations on related species within the *Theobroma*/*Herrania* genera. Using the same dataset of Colli-Silva et al. (2023a), we replicated image acquisition and data analysis, focusing on these wild cacao relatives, excluding *T. cacao*. This expanded perspective, though still limited, allowed us to discern distribution patterns beyond our primary target species. Supplementary Information (Appendix S1) contains the dataset used for this analysis. Employing a Ks-test, we compared the sampling patterns of these wild cacao relatives with our main dataset. A significant disparity would suggest that cacao collections deviate from the sampling patterns observed in other *Theobroma* species, thereby negating the influence of sampling bias.

Second, we aimed to address the concern that sampling tends to be concentrated in more accessible areas with higher anthropic influence, while less accessible areas with lower anthropic influence are underrepresented. To account for this potential bias, we conducted a randomization procedure by randomly swapping occurrence points in our dataset 1000 times, while maintaining the geographic range of our original records (defined as a 25 km buffer around all records). For each replicate, we obtained satellite images for the randomized points and calculated the same metrics as with the empirical data. Subsequently, we compared the results of the replicated scenarios with the empirical dataset to assess their statistical similarity. If the replicated scenarios significantly differed from the empirical dataset, it would suggest that factors other than chance influence our cacao collections, which aligns with the objectives of this study. The files containing the replicates are available as Supplementary Information (Appendix S1, S2).

Results

We found distinct patterns when associating cacao geographic distribution and land use profiles, as identified through NDVI classification. Notably, we observed significant differences between areas introduced to cacao cultivation after the pre-Columbian era and those closer to the suggested native cacao region (Table 1). Regions where cacao was introduced post–pre-Columbian era (like the South American Atlantic coast, North America, and other overseas locations), showed a higher occurrence of cacao specimens closer to areas affected by human activities and fewer occurrences near dense forests (Table 1; Fig. 2). Conversely, cacao specimens from regions closer to the suggested native area, as indicated by literature (such as Napo and Ucayali biogeographical provinces), were primarily situated away from human-influenced areas (Table 1; Fig. 2). Regions where early pre-Columbian introduction of cacao is observed, like Eastern Amazonia and Mesoamerica, displayed intermediate values for the measured variables (Table 1; Fig. 2).

In our study, we performed two separate analyses to investigate potential sampling biases associated with our cacao specimens (see "Addressing potential sampling biases" section). Firstly, we examined the possibility of sampling bias by considering that other
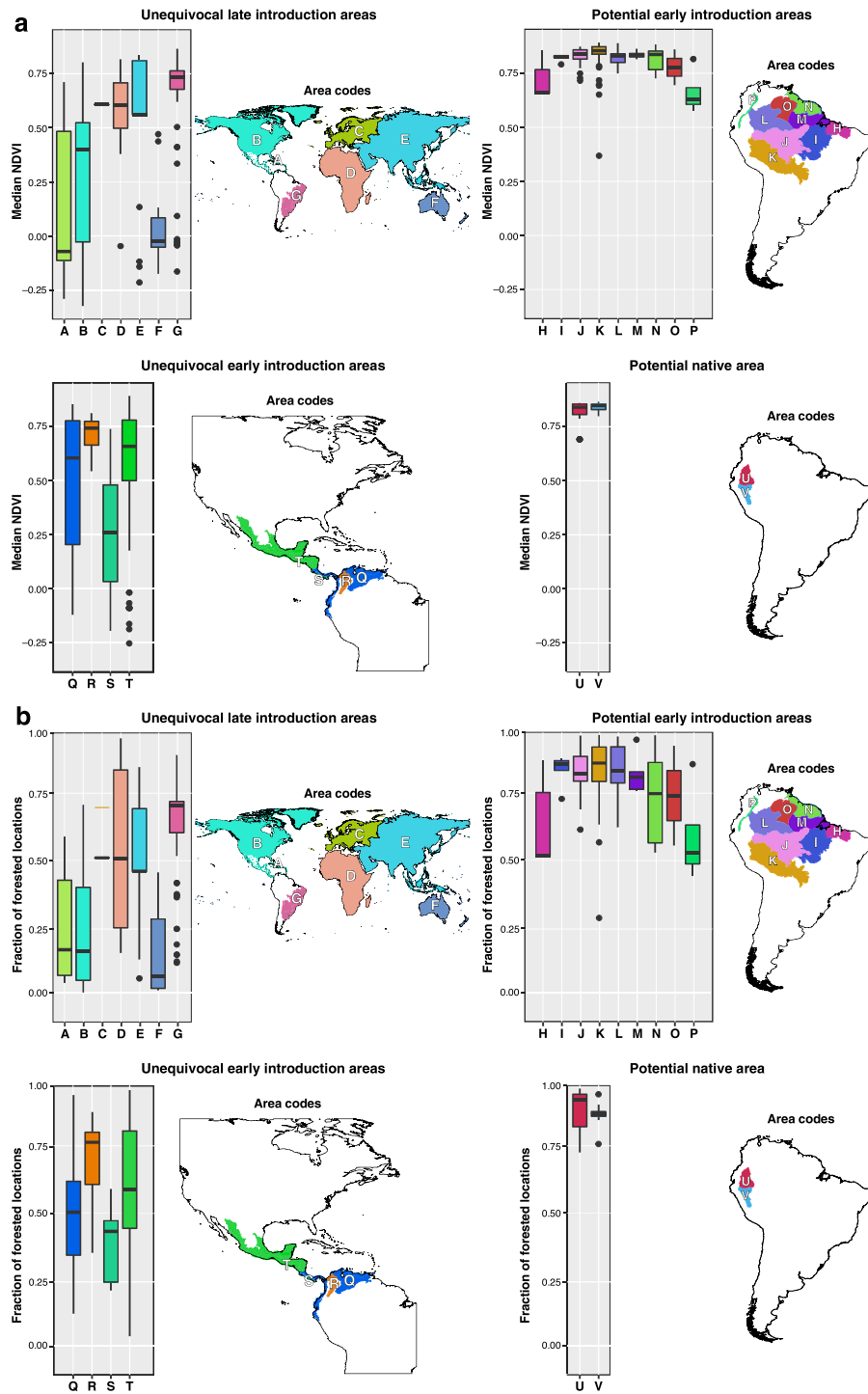
Table 1 Differences of selected variables of land use profiles based on point occurrence locations of cacao specimens in the globe

Origin area	Median NDVI	Relative frequency of areas of dense forests	Relative frequency of areas of human influence
Unequivocal late introduction	0.487 (±0.344)	0.490 (±0.273)	0.049 (±0.067)
Unequivocal early introduction	0.374 (±0.354)	0.541 (±0.213)	0.036 (±0.041)
Potential early introduction	0.787 (±0.156)	0.837 (±0.134)	0.022 (±0.025)
Potential native area	0.754 (±0.236)	0.883 (±0.071)	0.020 (±0.013)

Average values plus/minus standard deviation are presented for the three variables considered here: median NDVI for the locations, relative frequency of areas of dense forests and relative frequency of areas of human influence. Origin areas were separated as described in Methods (see "Literature survey and study area" section in Methods), and minor biogeographical regions were depicted according to Morrone (2014). Areas of dense forests include reclassified pixels with NDVI > 0.36, and areas of human influence include reclassified pixels with $0 < NDVI \leq 0.27$ (see "NDVI profile analyses" section in Methods for more details). For all cases, pairwise differences among areas were statistically significant (p-values < 0.05). See Appendix S2 for p-values for all pairwise comparisons

Fig. 2 **a** Median NDVI (Normalized Difference Vegetation Index) values of the areas where cacao specimens are found, categorized based on the major and minor regions defined in this study (see "Literature survey and study area" section in Methods). **b** Relative frequency of forested areas in the regions where cacao specimens are found. **c** Relative frequency of areas of human influence (as defined in "Acquisition of remote sensing data" section. in Methods), according to the major and minor areas defined for this work (see "Literature survey and study area" section. in Methods). Areas are defined as follows: (1) Unequivocal late introduction: Antilles (A), Canada and United States (B), Europe (C), Africa (D), Asia (E), Oceania and Pacific islands (F), South American Atlantic coast (G); (2) Potential early introduction: Pará province (H), Xingu-Tapajós province (I), Madeira province (J), Rondônia province (K), Imerí province (L), Roraima province (M), Guianan Lowlands province (N), Pantepui province (O), Paramo province (P); (3) Unequivocal early introduction: Pacific dominion (Q), Magdalena province (R), Puntarenas-Chiriquí province (S), Mesoamerican dominion (T); (4) Potential native area: Napo province (U), Ucayali province (V)

relatives from the same genus might exhibit a similar biased pattern. Secondly, we aimed to address the concern that, by default, botanical sampling usually is concentrated in more accessible areas with higher anthropic influence, while less accessible areas with lower anthropic influence are underrepresented. Our findings indicate that these identified biases do not solely explain the observed distribution patterns for our data. Statistical differences were found in over 98% of the replicates generated for this study, compared to the empirical data (p-values < 0.05; Appendix S2 in Supplementary Information). Moreover, when comparing the cacao dataset with its wild relatives, the median NDVI, frequency of forested areas, and frequency of anthropized areas showed statistical differences in most regions. These results indicate that factors beyond chance or the biases addressed in our study contribute to the observed distribution patterns of cacao specimens. All p-values and occurrence datasets can be found in the Supplementary Information (Appendices S1 and S2).
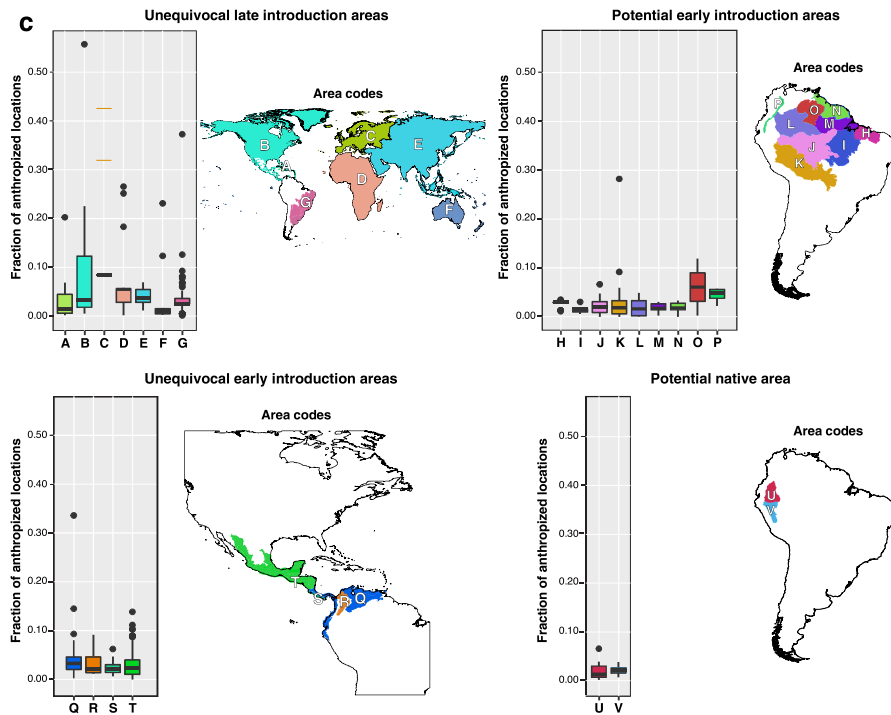
a

Unequivocal late introduction areas

Potential early introduction areas

Unequivocal early introduction areas

Potential native area

b

Unequivocal late introduction areas

Potential early introduction areas

Unequivocal early introduction areas

Potential native area

Fig. 2 (continued)

Discussion

Human impacts on cacao dispersal

Our results are consistent with the following scenario outlined in Fig. 2: origin of *Theobroma cacao* with a native range in areas U and V (this has been suggested as the native range based on genetic diversity studies, e.g. Thomas et al. 2012; early introduction into Northern South America and Central America (areas Q, R, S and T); potential early introduction to Eastern Amazonia and the Guiana (areas I-P); late introduction to Eastern Brazil and tropical areas outside of the Americas (areas G and A-F, respectively). This scenario is closest to that outlined in Fig. 1 (Map 6).

We observed a solid presence of cacao specimens in areas strongly influenced by human activities, providing support for the idea that human intervention would have played a significant role in cacao dispersal to various regions. This raises questions about the true native status of cacao in some areas. Bartley (2005) outline possible pathways of cacao's dispersal in areas in the African and Asian tropics, where cacao was spread from Mesoamerica to the Philippines in the seventeenth century, and to Africa from Amazonian varieties in the nineteenth century. Plantations in these continents were likely established based on very few individuals, and they may exhibit low genetic diversity. In the American Tropics, while cacao occurrence is often associated with human-impacted areas, recent literature identifies the native range of cacao as the primary forests of Western Amazonia (Thomas et al. 2012; Clement et al. 2015; Cornejo

et al. 2018; Fouet et al. 2022). This suggests that, compared to other areas potentially influenced by human introduction, these regions harbor a higher abundance of wild cacao specimens within primary forests, far away from human settlements or urban areas. Alternatively, it implies that these findings might be attributed to the actions of indigenous populations who cultivated/stimulated cacao plantations in these regions prior to modern settlements.

Theobroma species rely on mammals, such as rodents and primates, for fruit dispersion (van Hall 1914; Cuatrecasas 1964). Limited dispersal of cacao is evident due to certain species characteristics, including indehiscent fruits, flower self-incompatibility, short pollination distances, or high rates of vegetative propagation (Silva et al. 2010; Thomas et al. 2012; Levis et al. 2017). Additionally, genetic bottlenecks have been observed in introduced populations of cacao in Mesoamerica, and there is a lack of palynological records of the species in Mexico and Eastern Amazonia before the Holocene (Clement et al. 2010; Thomas et al. 2012; Osorio-Guarin et al. 2017; Cornejo et al. 2018). Furthermore, Bartley (2005) suggested that cacao might first have been used for its pulp by indigenous people and this may have aided its dispersion, as they took fruits on their migrations into the forest, ate the pulp and spat out the seeds. A similar trend is observed within a related species, *T. grandiflorum* or *cupuaçu* (Colli-Silva et al. 2023b). These factors indicate that wild cacao populations may have faced barriers to expand their geographic distributions over ecological time, which contrasts with the wide distribution of cacao seen today.

Interestingly, most of cacao's genetic diversity is concentrated in the border areas of northeastern Peru, northern Bolivia, southwestern Colombia, western Brazil (Acre state), and eastern Ecuador (Motamayor et al. 2002; Thomas et al. 2012; Clement et al. 2015; Cornejo et al. 2018). This specific region is recognized as a biogeographical area of endemism (sensu Morrone 2014), delimited by the Ucayali and Napo rivers and by the Andes to the west. These rivers likely played a crucial role in the diversification of various species (Silva and Oren 1996; Hubert et al. 2007; Harvey et al. 2014; Dumont et al. 1990; Kreft et al. 2004; Morrone 2014). Particularly, the change of these river courses has been discussed as a potential factor shaping species diversification in this region (Tuomisto and Ruokolainen 1997). Cacao may have also been affected by this, when considering its natural history before and human influence.

The extent to which ancient Amazonian societies reshaped the region's landscapes remains a topic of intense debate. For instance, Levis et al. (2017) uncovered a significant link between archaeological sites and the occurrence of certain plant species. Their research revealed that domesticated species were five times more likely to dominate in these areas compared to non-domesticated ones. This trend was also consistent across the Amazon basin, with forests surrounding archaeological sites showing higher abundance and diversity of domesticated plant species. These findings underscore the substantial impact of historical plant domestication by Amazonian indigenous groups on the structure of tree communities. Clement (1989) provided a review shedding light on numerous other Amazonian crops, belonging to different botanical families, that might have undergone similar processes. The collective and increasing body of evidence from different disciplines emphasizes the legacy of ancient Amazonian peoples' influence on the region's flora, amplifying the significance of historical plant domestication in shaping the Amazonian landscape. Our research aligns with this narrative, offering further insights that contribute to understanding this historical legacy.

Implications for genetic resource repatriation

Our findings shed light on the fact that many areas traditionally considered as the "native" range of cacao may actually consist of introduced populations that were established before or after the pre-Columbian era and that may contain specimens that have spontaneously grown and persisted outside its native range. If this is correct, there can be significant implications for issues related to jurisdiction and access to genetic resources for crop improvement and conservation of cacao. In this sense, it would be crucial to better characterize germplasm accessions that have contributed significantly to our understanding of cacao diversity, as suggested by many authors (Bartley 2005; End et al. 2010; Laliberté 2012; Malhotra and Apshara 2014).

The principle of sovereign rights of a country over the genetic resources of plant species native to its territory is well-established in international law (Correa 1995) and widely recognized by international bodies such as the Food and Agriculture Organization (FAO) and the Convention on Biological Diversity (CDB). According to FAO, national governments have the authority to regulate access to genetic resources, which is subject to national legislation. Hence, each country should possess the right and jurisdiction over the resources native to its territory. Disputes concerning genetic resource rights and patents for *Theobroma* species have already arisen in some countries, such as those for the Peruvian cultivars "Chuncho" and "Cacao Amazonas Peru" (INDECOPI 2016), or the Brazilian *cupuaçu* (*T. grandiflorum*; see Rezende and Ribeiro 2009 and Colli-Silva et al. 2023b). Determining the origins and natural history of *T. cacao* before and after humans, and understanding its subsequent introductions rely on the biogeographical context summarized here, with different scenarios indicating specific countries as the native range of the species.

Based on our results (Fig. 2a–c) and on the existing evidence in the literature, the native range of cacao would be limited to areas in Ecuador, Colombia, Peru, and perhaps the westernmost part of Brazil (Acre and Western Amazonas states). Consequently, many areas within Brazil, as well as the Guianas, might harbor specimens that are not authentically native but rather cultivated or naturalized specimens in regions significantly impacted by human activities. However, this assertion requires cautious interpretation and should be further evaluated, given evidence from extensive surveys, such as those conducted in the Guianas, which describe populations that contradict this notion (Lachenaud et al. 2004; Lachenaud and Zhang 2008).

Further limitations in our methodology also require attention. Our study utilized a broad range of satellite images covering various regions and collected over a significant timeframe (see "Acquisition of remote sensing data" section of Methods). The choice of images across such diverse dates can significantly affect the NDVI values. For instance, a forest might display higher NDVI values in the wet season and lower values in the dry season. This variation could explain the high standard deviation of NDVI patterns even in nearby areas, especially in highly seasonal ecosystems, as seen in previous research on Amazonia (Silva et al. 2013). However, in our study, this issue is consistent across all images, spreading this bias evenly throughout the datasets. Moreover, limitations linked to the NDVI extend to instrumental factors, including uncertainties in satellite navigation, fluctuations in the satellite's local crossing time, and sensor degradation (Santos and Negri 1997). While potential correlations between human activities and environmental factors could be confounding, Levis et al. (2017) demonstrated that human influence alone explains roughly half of the variation in the abundance of domesticated species in certain regions.

Further research is imperative to accurately trace the origins of cacao and other wild crop species and validate the various biogeographical hypotheses outlined in our study, accouting for both the history before and after human arrival in the Americas. Levis et al. (2017) raised a significant question regarding the association between domesticated species and archaeological sites: Did humans enrich forests with domesticated species, or did they settle near naturally rich forests? Our approach cannot prove causation, but given additional supporting evidence, the former scenario appears more plausible. To facilitate such investigations, several crucial steps should be taken. Firstly, there should be a substantial increase in collecting new germplasm accessions from wild cacao populations in underrepresented areas (Sereno et al. 2006; Zhang et al. 2016). Secondly, exploring the morphological variability of *T. cacao* is necessary to identify potential characteristics that could define genetic clusters as distinct varieties (Motamayor et al. 2008). Lastly, historical biogeographical studies employing various analyses are essential to trace the origin of *Theobroma*-related species in South America, particularly in the Amazon basin. Additionally, considering the biogeography of species related to cacao, like endophytic or pathogenic fungi (Hanada et al. 2010), can offer insights into the geographic history of cacao, possibly indicating co-evolution with *T. cacao* or its relatives.

Declarations

Competing interests The authors declare no relevant financial or non-financial interests to disclose.

References

Arnold TB, Emerson JW (2011) Nonparametric goodness-of-fit tests for discrete null distributions. R J 3:34. https://doi.org/10.32614/RJ-2011-016

Bartley BGD (2005) The genetic diversity of cacao and its utilization. Cabi, Wallingford

Bernoulli KCG (1869) Neue Denkschriften der Allg. Schweizerischen Gesellschaft für die Gesammten Naturwissenschaften, vol 24, Band 3, 1–15. Zurich, 1871. Reprint issued 1869

Bivand RS, Pebesma E, Gómez-Rubio V (2013) Visualising spatial data. In: Bivand RS, Pebesma E, Gómez-Rubio V (eds) Applied spatial data analysis with R. Springer, New York, pp 59–82

Bletter N, Daly DC (2006) Cacao and its relatives in South America. In: McNeil P (ed) Chocolate in Mesoamerica: a cultural history of Cacao. Univ Press of Florida, Gainesville, pp 31–68

Cheesman EE (1927) Notes on the nomenclature, classification and possible relationships of cacao populations. Trop Agric 21:144–159

Cheesman EE (1929) Cacao note: the hybrid nature of Forastero. Trop Agric 6:174

Cheesman EE (1932) The economic botany of Cacao. Trop Agric 9:1–16

Cheesman EE (1944) Notes on the nomenclature, classification and possible relationships of cacao populations. Trop Agric 21:144–159

Ciferri R, Ciferri F (1957) The evolution of cultivated cacao. Int J Org Evol 11:381–397

Clement CR (1989) A center of crop genetic diversity in western Amazonia. Bioscience 39:624–631

Clement CR, De Cristo-Araújo M, Coppens D'Eeckenbrugge G, Alves Pereira A, Picanço-Rodrigues D (2010) Origin and domestication of native Amazonian crops. Diversity 2:72–106

Clement CR, Denevan WM, Heckenberger MJ, Junqueira AB, Neves, EG, Teixeira WG, Woods WI (2015) The domestication of Amazonia before European conquest. Proc R Soc B282: 20150813. https://doi.org/10.1098/rspb.2015.0813

Colli-Silva M, Richardson JE, Pirani JR (2023a) A taxonomic dataset of preserved specimen occurrences of *Theobroma* and *Herrania* (Malvaceae, Byttnerioideae) stored in 2020. Biodivers Data J. https://doi.org/10.3897/bdj.11.e99646

Colli-Silva M, Richardson JE, Neves EG, Watling J, Figueira A, Pirani JR (2023b) Domestication of the Amazonian fruit tree *cupuaçu* may have stretched over the past 8000 years. Commun Earth Environ. https://doi.org/10.1038/s43247-023-01066-z.

Cornejo OE, Yee MC, Dominguez V, Andrews M, Sockell A, Strandberg E, Livingstone D, Stack C, Romero A, Umaharan P, Royaert S, Tawari NR, Ng P, Gutierrez O, Phillips W, Mockaitis K, Bustamante CD, Motamayor JC (2018) Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. Commun Biol. https://doi.org/10.1038/s42003-018-0168-6

Correa CM (1995) Sovereign and property rights over plant genetic resources. Agric Hum Values 12:58–79

Cuatrecasas J (1964) Cacao and its allies: a taxonomic revision of the genus *Theobroma*. Contr US Natl Herb 35:379–614

da Silva JMC, Oren DC (1996) Application of parsimony analysis of endemicity in Amazonian biogeography: an example with primates. Biol J Linnean Soc 59:427–437. https://doi.org/10.1111/j.1095-8312.1996.tb01475.x

Ducke A (1925) Plantes nouvelles ou peu connues de la région amazonienne; Genre *Theobroma* L. Arch Jard Bot Rio De Janeiro 4:130–133

Ducke A (1940) As espécies brasileiras de cacau (gênero *Theobroma* L.), na botânica sistemática e geográfica. Rodriguésia 4:265–276

Duke D, Wohlgemuth E, Adams KR, Armstrong-Ingram A, Rice SK, Young DC (2021) Earliest evidence for human use of tobacco in the Pleistocene Americas. Nat Hum Behav 6:183–192. https://doi.org/10.1038/s41562-021-01202-9

Dumont JF, Lamotte S, Kahn F (1990) Wetland and upland forest ecosystems in Peruvian Amazonia: plant species diversity in the light of some geological and botanical evidence. For Ecol Manag 33–34:125–139. https://doi.org/10.1016/0378-1127

End MJ, Daymond AJ, Hadley P (2010) Technical guidelines for the safe movement of cacao germplasm (Revised from the FAO/IPGRI Technical Guidelines No. 20). CacaoNet, Biodiversity International, Montpellier

Figueira A, Janick J, Levy M, Goldsbrough P (1994) Reexamining the classification of *Theobroma cacao* L. using molecular markers. J Am Soc Hortic Sci 119:1073–1082. https://doi.org/10.21273/jashs.119.5.1078

Fouet O, Loor Solorzano RG, Rhoné B, Subía C, Calderón D, Fernández F, Sotomayor I, Rivallan R, Colonges K, Vignes H, Angamarca F, Yaguana B, Costet P, Argout X, Lanaud C (2022) Collection of native *Theobroma cacao* L. accessions from the Ecuadorian Amazon highlights a hotspot of cocoa diversity. Plants People Planet 4:605

Gama-Rodrigues AC, Müller M, Gama-Rodrigues EF, Antônio F (2021) Cacao-based agroforestry systems in the Atlantic Forest and Amazon Biomes: an ecoregional analysis of land use. Agric Syst 194:103270–103270. https://doi.org/10.1016/j.agsy.2021.103270

GBIF.org (2020) GBIF Occurrence. https://doi.org/10.15468/dl.yze9k4

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 202:18–27

Gutaker RM, Groen SC, Bellis ES, Choi JY, Pires IS, Bocinsky RK, Slayton ER, Wilkins O, Castillo CC, Negrão S, Oliveira MM, Fuller DQ, d'Alpoim Guedes JA, Lasky JR, Purugganan MD (2020) Genomic history and ecology of the geographic spread of rice. Nat Plants 6:492–502. https://doi.org/10.1038/s41477-020-0659-6

Hanada RE, Pomella AWV, Costa HS, Bezerra JL, Loguercio LL, Pereira JO (2010) Endophytic fungal diversity in *Theobroma cacao* (cacao) and *T. grandiflorum* (cupuaçu) trees and their potential for growth promotion and biocontrol of black-pod disease. Fungal Biol 114:901–910. https://doi.org/10.1016/j.funbio.2010.08.006

Harvey MG, Seeholzer GF, Cáceres AD, Winger BM, Tello JG, Camacho FH, Aponte Justiniano MA, Judy CD, Ramírez SF, Terrill RS, Brown CE, León LAA, Bravo G, Combe M, Custodio O, Zumaeta AQ, Tello AU, Bravo WAG, Savit AZ, Ruiz FWP (2014) The avian biogeography of an Amazonian headwater: the Upper Ucayali River, Peru. Wilson J Ornithol 126:179. https://doi.org/10.1676/13-135.1

Hijmans R (2023) raster: geographic data analysis and modeling. R package version 3.6-23. https://CRAN.R-project.org/package=raster

Hubert N, Duponchelle F, Nuñez J, Garcia-Davila C, Paugy D, Renno J-F (2007) Phylogeography of the piranha genera Serrasalmus and Pygocentrus: implications for the diversification of the Neotropical ichthyofauna. Mol Ecol 16:2115–2136. https://doi.org/10.1111/j.1365-294x.2007.03267.x

INDECOPI (2016) El Indecopi entrega décima Denominación de Origem Cacao Amazonas Perú que contribuirá al desarrollo económico de más de 1,200 productores. Lima, Peru. indecopi.gob.pe

Khoshbakht K, Hammer K (2008) How many plant species are cultivated? Genet Resour Crop Evol 55:925–928. https://doi.org/10.1007/s10722-008-9368-0

Kreft H, Köster N, Küper W, Nieder J, Barthlott W (2004) Diversity and biogeography of vascular epiphytes in Western Amazonia, Yasuní, Ecuador. J Biogeogr 31:1463–1476. https://doi.org/10.1111/j.1365-2699.2004.01083.x

Lachenaud P, Motamayor JC (2017) The Criollo cacao tree (*Theobroma cacao* L.): a review. Genet Resour Crop Evol 64:1807–1820

Lachenaud P, Zhang D (2008) Genetic diversity and population structure in wild stands of cacao trees (*Theobroma cacao* L.) in French Guiana. Ann for Sci 65(3):310

Lachenaud P, Sounigo O, Gérald O (2004) Genetic structure of Guianan wild cocoa (*Theobroma cacao* L.) described using isozyme electrophoresis. Plant Genet Resour Newsl 139:24–30

Laliberté B (2012) A global strategy for the conservation and use of cacao genetic resources, as the foundation for a sustainable cocoa economy. Bioversity International, Montpellier

Laurent V, Risterucci AM, Lanaud C (1994) Genetic diversity in cocoa revealed by cDNA probes. Theor Appl Genet 88:193–198

Levis C, Costa FRC, Bongers F, Peña-Claros M, Clement CR, Junqueira AB, Neves EG, Tamanaha EK, Figueiredo FOG, Salomão RP, Castilho CV, Magnusson WE, Phillips OL, Guevara JE, Sabatier D, Molino JF, López DC, Mendoza AM, Pitman NCA, Duque A (2017) Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. Science 355:925–931. https://doi.org/10.1126/science.aal0157

Malhotra SK, Apshara SE (2014) Genetic resources of cocoa (*Theobroma cacao* L.) and their utilization—an appraisal. Indian J Genet Plant Breed 77:199–213

Morris D (1882) Cacao: how to grow and how to cure it. Government Printing Establishment, Kingston

Morrone JJ (2014) Biogeographical regionalisation of the Neotropical region. Zootaxa 3782:1

Motamayor JC, Lanaud C (2002) Molecular analysis of the origin and domestication of *Theobroma cacao* L. In: Engles JMM, Ramanatha RV, Brown AHD, Jackson MT (eds) Managing plant genetic diversity. IPGRI, Rome, pp 77–87

Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. Heredity 89:380–386. https://doi.org/10.1038/sj.hdy.6800156

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. Heredity 91:322–330. https://doi.org/10.1038/sj.hdy.6800298

Motamayor JC, Lachenaud P, DaSilvaeMota JW, Loor R, Kuhn DN, Brown JS, Schnell RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). PLoS ONE 3:e3311. https://doi.org/10.1371/journal.pone.0003311

N'Goran JAK, Laurent V, Risterucci AM, Lanaud C (1994) Comparative genetic diversity studies of *Theobroma cacao* L. using RFLP and RAPD markers. Heredity 73:589–597. https://doi.org/10.1038/hdy.1994.166

Oliveira U, Paglia AP, Brescovit AD, Carvalho CJB, Silva DP, Rezende DT, Leite FSF, Batista JAN, Barbosa JPPP, Stehmann JR, Ascher JS, Vasconcelos MF, De Marco P, Löwenberg-Neto P, Dias PG, Ferro VG, Santos AJ (2016) The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Divers Distrib 22:1232–1244

Osorio-Guarín JA, Berdugo-Cely J, Coronado RA, Zapata YP, Quintero C, Gallego-Sánchez G, Yockteng R (2017) Colombia a source of cacao genetic diversity as revealed by the population structure analysis of germplasm bank of *Theobroma cacao* L. Front Plant Sci. https://doi.org/10.3389/fpls.2017.01994

Pebesma E, Bivand R (2005) Classes and methods for spatial data in R. R News 5(2):9–13

Pebesma EJ, Bivand R (2023) Spatial data science: with applications in R, 1st edn. CRC Press, Boca Raton

Pittier H (1924) Apuntes sobre cacao: las tres especies fundamentales de las cuales provienen todas las variedades cultivadas. Hacienda 19:364–365

Pittier H, Chevalier A (1925) L'origine hybride des cacaoyers cultivés. J Agr Trad Bot Appl 6:344–349

Pittier H, Ducke A, Chevalier A (1926) L'Origine géographique et botanique des Cacaoyers et l'utilité de leur greffage. J Agr Trad Bot Appl 6:344–349

Pound FJ (1938) Cacao and witches' broom disease (Marasmius perniciosus) of South America, with notes on other species of *Theobroma*: report on a visit to Ecuador, the Amazon Valley and Colombia. American Cocoa Research Institute, Port of Spain

Pound FJ (1945) A note on the cocoa population of South America. In: Report and Proceedings of the 1945 cocoa conference, London, pp 131–133

Preuss P (1901) Der Kakao, seine Kultur und Aufbereitung [Cocoa, its cultivation and preparation]. In: Expedition Nach Central- Und Südamerika (Chap. 13, pp 167–277). Colonial Economic Committee, Berlin. Reprinted 1985: Archives of Cocoa Research 3, 15–115. International Office of Cocoa and Chocolate, Brussels

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

Rezende EA, Ribeiro MTF (2009) O cupuaçu é nosso? Aspectos atuais da biopirataria no contexto brasileiro. Revista De Gestão Social e Ambiental 3:53–74

Richardson JE, Whitlock BA, Meerow AW, Madriñán S (2015) The age of chocolate: a diversification history of *Theobroma* and Malvaceae. Front Ecol Evol. https://doi.org/10.3389/fevo.2015.00120

Rouse JW, Haas RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the great plains with ERTS. NASA Spec Publ 351:309

Sambuichi RHR, Vidal DB, Piasentin FB (2012) Cabruca agroforests in southern Bahia, Brazil: tree component, management practices and tree species conservation. Biodivers Conserv 21:1055–1077. https://doi.org/10.1007/s10531-012-0240-3

Santos P, Negri AJ (1997) A comparison of the normalized difference vegetation index and rainfall for the Amazon and Northeastern Brazil. J Appl Meteorol Climatol 36:958–965. https://doi.org/10.1175/1520-0450(1997)036%3c0958:ACOTND%3e2.0.CO;2

Schultes RE (1958) A synopsis of the genus Herrania. J Arnold Arbor 39:217–295

Schultes RE (1984) Amazonian cultigens and their northward and westward migrations in pre-Columbian times. In: Stone D (ed) Pre-Columbian plant migration. Harvard University Press, Cambridge

Schumann K (1886) Sterculiaceae. In: Martius CFP, Eichler AG (eds) Flora brasiliensis. Leipzig, 12(3), 1–114, tab. 1–24

Sereno ML, Albuquerque PSB, Vencovsky R, Figueira A (2006) Genetic diversity and natural population structure of Cacao (*Theobroma cacao* L.) from the Brazilian Amazon evaluated by microsatellite markers. Conserv Genet 7:13–24. https://doi.org/10.1007/s10592-005-7568-0

Silva CRS, Albuquerque PSB, Ervedosa FR, Mota JWS, Figueira A, Sebbenn AM (2010) Understanding the genetic diversity, spatial genetic structure and mating system at the hierarchical levels of fruits and individuals of a continuous *Theobroma cacao* population from the Brazilian Amazon. Heredity 106:973–985. https://doi.org/10.1038/hdy.2010.145

Silva FB, Shimabukuro YE, Aragão LEOC, Anderson LO, Pereira G, Cardozo F, Arai E (2013) Large-scale heterogeneity of Amazonian phenology revealed from 26-year long AVHRR/NDVI time-series. Environ Res Lett 8:024011–024011. https://doi.org/10.1088/1748-9326/8/2/024011

Soria JV (1970) Principal varieties of cocoa cultivated in tropical America. Cocoa Growers' Bull 19:12–21

Sweeney M, McCouch S (2007) The complex history of the domestication of rice. Ann Bot 100:951–957. https://doi.org/10.1093/aob/mcm128

Thomas E, van Zonneveld M, Loo J, Hodgkin T, Galluzzi G, van Etten J (2012) Present spatial diversity patterns of *Theobroma cacao* L. in the neotropics reflect genetic differentiation in Pleistocene refugia followed by human-influenced dispersal. PLoS ONE 7:e47676. https://doi.org/10.1371/journal.pone.0047676

Tuomisto H, Ruokolainen K (1997) Biodivers Conserv 6:347–357. https://doi.org/10.1023/a:1018308623229

Ulian T, Diazgranados M, Pironon S, Padulosi S et al (2020) Unlocking plant resources to support food security and promote sustainable agriculture. Plants People Planet 2(5), 421–445. https://doi.org/10.1002/ppp3.10145

van Hall CJJ (1914) Cocoa. Macmillan and Co, London, pp 1–515

Van Rossum G, Drake FL Jr (1995) Python reference manual. Centrum voor Wiskunde en Informatica, Amsterdam

Wang L, Beissinger TM, Lorant A, Ross-Ibarra C, Ross-Ibarra J, Hufford MB (2017) The interplay of demography and selection during maize domestication and expansion. Genome Biol. https://doi.org/10.1186/s13059-017-1346-4

Watling J, Shock MP, Mongeló GZ, Almeida FO, Kater T, De Oliveira PE, Neves EG (2018) Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. PLoS ONE 13:e0199868–e0199868. https://doi.org/10.1371/journal.pone.0199868

Zarrillo S, Gaikwad N, Lanaud C, Powis T, Viot C, Lesur I, Fouet O, Argout X, Guichoux E, Salin F, Solorzano RL, Bouchez O, Vignes H, Severts P, Hurtado J, Yepez A, Grivetti L, Blake M, Valdez F (2018) The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. Nat Ecol Evol 2:1879–1888. https://doi.org/10.1038/s41559-018-0697-x

Zequeira-Larios C, Santiago-Alarcon D, MacGregor-Fors I, Castillo-Acosta O (2021) Tree diversity and composition in Mexican traditional smallholder cocoa agroforestry systems. Agrofor Syst 95:1589–1602. https://doi.org/10.1007/s10457-021-00673-z

Zhang D, Motilal L (2016) Origin, dispersal, and current global distribution of cacao genetic diversity. In: Bailey BA, Meinhardt LW (eds) Cacao diseases. Springer, Cham, pp 3–31. https://doi.org/10.1007/978-3-319-24789-2_1

Authors and Affiliations

Matheus Colli-Silva[1,2] · James Edward Richardson[3,4,5,6] · Antonio Figueira[7] · José Rubens Pirani[1]

✉ Matheus Colli-Silva
matheus.colli.silva@alumni.usp.br

1 Department of Botany, Institute of Biosciences, University of São Paulo, São Paulo, Brazil

2 Present Address: Royal Botanic Gardens, Kew, Richmond, Surrey, UK

3 School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland

4 Environmental Research Institute, University College Cork, Cork, Ireland

5 Tropical Diversity Section, Royal Botanic Garden Edinburgh, Edinburgh, UK

6 Faculty of Natural Sciences, Rosario University, Bogotá, Colombia

7 Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil