



## Domestication of the Amazonian fruit tree *cupuaçu* may have stretched over the past 8000 years

Matheus Colli-Silva<sup>1,8</sup>, James E. Richardson<sup>2,3,4,5</sup>, Eduardo G. Neves<sup>6</sup>, Jennifer Watling<sup>6</sup>, Antonio Figueira<sup>7</sup> & José Rubens Pirani<sup>1</sup>

Amazonia, one of the largest and most biodiverse ecosystems on Earth, is a significant yet less-known arena for ancient plant domestication. Here, we traced the origins of *cupuaçu* (*Theobroma grandiflorum*), an Amazonian tree crop closely related to cacao (*T. cacao*), cherished for its flavorful seed-pulp, by employing an extensive genomic analysis based on data from four sites in Brazil. Our results indicate that *cupuaçu* is a domesticated variant of its wild relative, *cupuí* (*T. subincanum*), probably originating from the Middle-Upper Rio Negro basin. A first phase of domestication is observed through a genetic bottleneck that we estimated to have occurred 5000–8000 years before the present. Moreover, we found further reductions in genetic diversity that we estimated to have occurred during the modern era. This is consistent with a second phase of domestication that was accompanied by an increase in the geographic distribution of *cupuaçu* over the last two centuries. Unraveling *cupuaçu*'s origins adds it to the roster of plants domesticated by Amazonian indigenous people in the early to mid-Holocene. Our results suggest that Amazonia's current patterns of genetic diversity and distribution of domesticated plants were influenced by both pre-Columbian and modern histories.

<sup>1</sup>Department of Botany, Institute of Biosciences, University of São Paulo, São Paulo, Brazil. <sup>2</sup>School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland. <sup>3</sup>Environmental Research Institute, University College Cork, Cork, Ireland. <sup>4</sup>Tropical Diversity Section, Royal Botanic Garden Edinburgh, Edinburgh, UK. <sup>5</sup>Faculty of Natural Sciences, Rosario University, Bogotá, Colombia. <sup>6</sup>Museum of Archaeology and Ethnology, University of São Paulo, São Paulo, Brazil. <sup>7</sup>Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil. <sup>8</sup>Present address: Royal Botanic Gardens, Kew, Richmond, UK. ✉email: [matheus.colli.silva@alumni.usp.br](mailto:matheus.colli.silva@alumni.usp.br)

The Amazon rainforest is the largest and most biodiverse in the world<sup>1–3</sup>, home to a large number of plant species<sup>4</sup>, providing globally important ecosystem services for the stabilization of climate and carbon fixation<sup>5,6</sup>. Outstanding functional and taxonomic biodiversity of Amazonia is allied to the presence of indigenous communities who have managed plants in ways that increased agrobiodiversity over millennial timescales<sup>7</sup>. While more than 30 million people living in rural or urban settlements use or at least are familiar with many elements of the Amazonian flora, *circa* 1.5 million indigenous people from more than 400 ethnic groups rely on the forest for their livelihoods<sup>8</sup>. This number represents only a small fraction of the estimated human population before the sixteenth century, calculated to have been up to 10 million before the arrival of Europeans<sup>9</sup>.

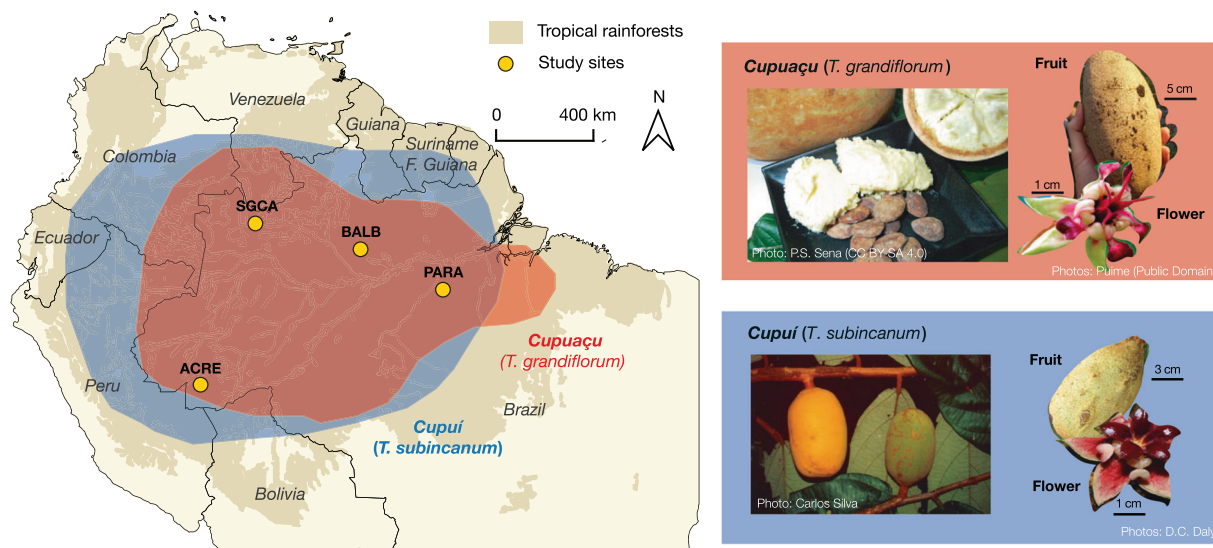
Amazonia is also the cradle of several plant species that have been used or domesticated over the past 12,000 years<sup>10</sup>. Currently, food economy of Amazonia is one of the major and most attractive activities for the region<sup>11</sup>, and dozens of plant species are believed to have gone through some degree of domestication<sup>12–14</sup>. Some species are found in the wild and have been slightly modified via artificial selection, or are eaten *in natura* and planted elsewhere, such as the *açaí* palm (*Euterpe oleracea*, Arecaceae)<sup>15</sup>, *guaraná* (*Paullinia cupana*, Sapindaceae)<sup>16</sup> and cacao (*Theobroma cacao*, Malvaceae)<sup>17,18</sup>. Other species were subjected to a strong domestication process, being selected from a wild relative with desirable traits, e.g., larger fruits with more pulp, or more seeds per pod. This is the case for pineapple (*Ananas comosus*, Bromeliaceae), with the selection of larger individual fruits<sup>19</sup>, cassava (*Manihot esculenta*, Euphorbiaceae), selected by indigenous people to be more poisonous to increase protection against herbivory and increased starch storage<sup>20</sup>, or the peach palm (*Bactris gasipaes*, Arecaceae), with the selection of varieties with varying starch and oil content used for food<sup>21</sup>.

Despite the great biodiversity of Amazonia, many species remain overlooked. An interesting but relatively unexplored example is *cupuaçu* (*Theobroma grandiflorum*, Malvaceae), a fruit tree crop related to cacao found growing in the Amazon Basin,

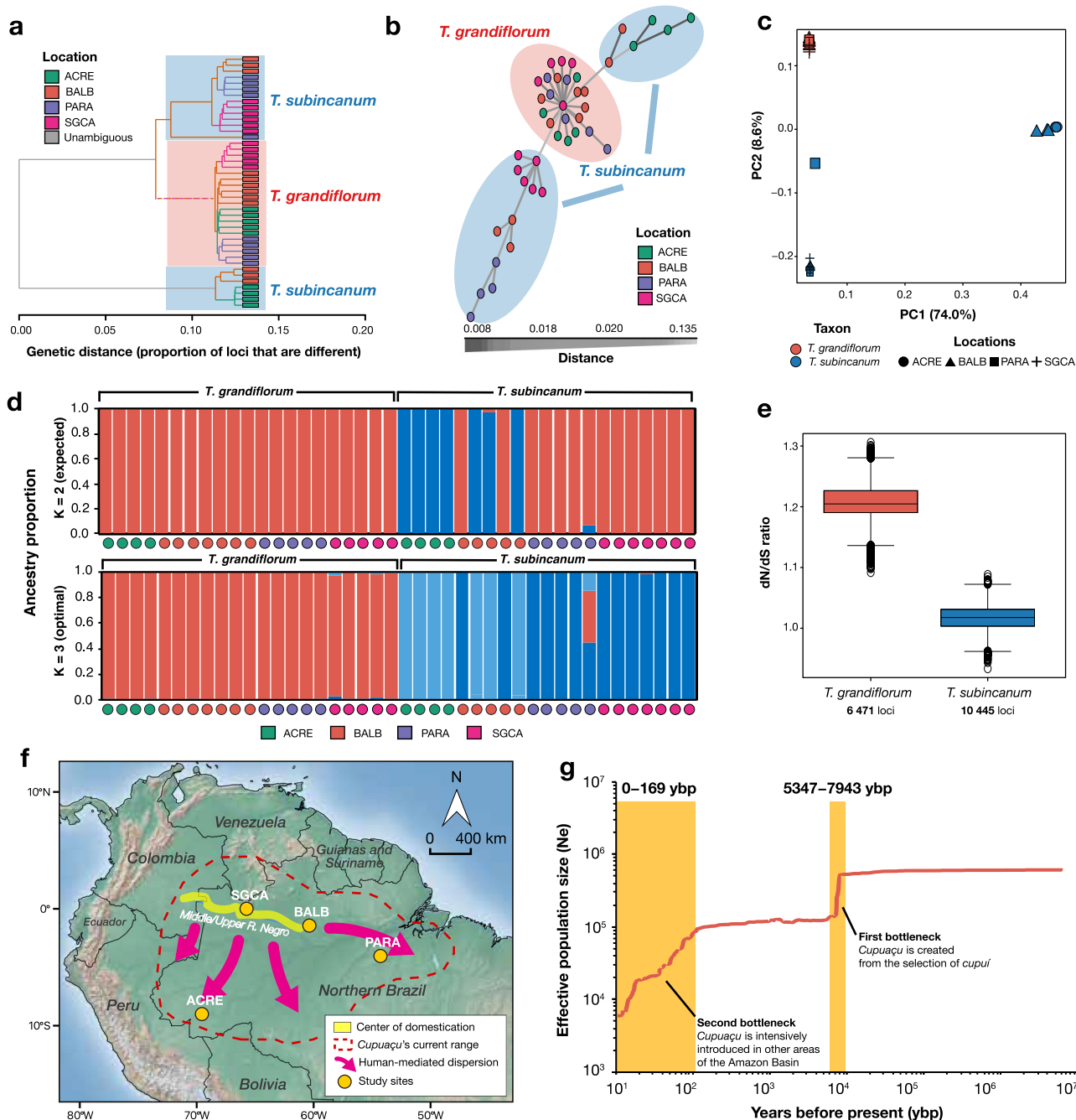
mostly in Brazil (Fig. 1)<sup>22,23</sup>. *Cupuaçu* seed-pulp is used to produce a range of desserts<sup>24</sup>, and the seeds can also be used to produce *cupulate*, a candy equivalent to chocolate, but with a different taste. Although trade and cultivation of *cupuaçu* is currently associated with local community-based production, its economic importance has grown in recent decades. *Cupuaçu* trees are mostly found in Brazil, and estimates by Brazilian authorities show that its production reached over 21,000 tons in 2017, generating an income of almost 55 million Brazilian reais (over 11 million US dollars) for that year<sup>25</sup>. In fact, the issue of control of *cupuaçu*-derived products once resulted in an international dispute in the early 2000s, involving the improper use of the name “*cupuaçu*” as a trademark by a Japanese company<sup>26</sup>.

*Cupuaçu*'s closest relative is *cupuí* (*T. subincanum*), and in recent phylogenies the two species have consistently emerged as sister-groups<sup>27–30</sup>. Morphologically, *cupuí* is very similar to *cupuaçu*, but with smaller fruits and seeds and more pubescent leaves. They both share almost the same geographic distribution, except that *cupuí* extends further into Western Amazonia towards Colombia, Ecuador, and Venezuela<sup>22,23</sup> (Fig. 1; Supplementary Data 1; Supplementary Methods). The geographic distribution of *cupuaçu* tends to be associated with human settlements. The seeds that are enveloped by its edible pulp are discarded and germinate, forming seedlings that grow near anthropized areas. Etymologically, *cupuaçu* and *cupuí* also show a close relationship, as the two words have a common Tupi-Guarani root (*kupu*-, “like cacao” and *-uasú*, “large”). Repeated cases of natural and artificial fertile hybrids between *cupuaçu* and *cupuí* have also been reported<sup>31–34</sup>.

Traditionally, *cupuaçu* has been defined as “incipiently domesticated”, i.e., a taxon whose populations have been subject to some degree of domestication, with an expected reduction in genetic variability, but whose average phenotype would still be within the range of its wild form<sup>35</sup>. A second study demonstrated *cupuaçu*'s limited genetic diversity<sup>36</sup>, which would be an unexpected pattern for a species traditionally considered as wild. Studies of crop domestication increasingly use genomic data<sup>37,38</sup>, and recent studies have shown that *T. cacao*, *cupuaçu*'s close



**Fig. 1 Presenting *cupuaçu* and *cupuí*.** Overview of the major collections of *T. grandiflorum* (*cupuaçu*) and its closest wild relative, *T. subincanum* (*cupuí*), in the Amazon Basin (Supplementary Methods, Supplementary Fig. 1, and Supplementary Data 1). *Cupuí* is characterized by smaller fruits, typically measuring 7–15 cm in length and 5–8 cm in width, with reddish-purple structures in the flowers. In contrast, *cupuaçu* produces larger fruits, ranging from 10–25 cm (or even larger) in size, accompanied by yellowish-green flower structures. The pulp of *cupuaçu* fruits is highly valued and widely used in the preparation of a variety of desserts. The collections were conducted at multiple location sites, including Xapuri-Acre (ACRE), Balbina-Amazonas (BALB), Tapajós-Pará (PARA), and São Gabriel da Cachoeira-Amazonas (SGCA).



relative, underwent strong domestication ~3600 years before the present (yBP), with the introduction of some individuals from Western Amazonia into Mesoamerica<sup>17</sup>. There is also archaeogenomic evidence of cacao consumption in Western Amazonia by 5200 yBP<sup>18</sup>. Here, we employed population RAD-seq data to determine the domestication history of *cupuaçu*. We demonstrate that *cupuaçu* is a domesticated form derived from *cupuí*, shedding light on its origins and geographic history. Our findings uncover intriguing patterns of genetic diversity and distribution among domesticated plants in Amazonia, providing valuable insights into the influence of both pre-Columbian and modern histories in the region.

**Results and discussion**

**Cupuaçu as a domesticated form.** We used multiple approaches to determine genomic relationships and compare genetic diversity

between *cupuaçu* and *cupuí*, each of which indicate that *cupuaçu* samples are all nested as a genetic subset within *cupuí* specimens (Fig. 2a–c; Supplementary Fig. 4). *Cupuaçu* is therefore not a sister species of *cupuí*, as suggested in phylogenetic studies, but is instead nested within it. One of the signatures of domestication in the genome is a reduction in genetic diversity when compared to a potential wild form of the crop<sup>13,39,40</sup>. Our findings reveal that the genetic diversity of *cupuaçu* is lower than that of *cupuí* ( $\pi_{cupuaçu} = 1.44 \times 10^{-4} < \pi_{cupuí} = 4.56 \times 10^{-4}$ ,  $p$ -value < 0.001 for a pairwise Mann–Whitney  $U$  test; Supplementary Table 3; see Methods). Additionally, *cupuaçu* individuals were found to be less genetically diverse within their populations than *cupuí* individuals (Fig. 2a, b; *cupuaçu* weighted  $F_{ST} = 0.118 \pm 0.078$  versus  $F_{ST} = 0.491 \pm 0.330$  in *cupuí*,  $p$ -value < 0.001 for a pairwise Mann–Whitney  $U$  test; Supplementary Tables 4, 5). This was further supported by the strong genetic homogeneity observed in *cupuaçu* individuals from all locations, with  $K = 3$  as the optimal

**Fig. 2 Populational genomics and demographic history of cupuaçu and cupuí.** This figure presents the findings from our study on the populational genomics and demographic history of *cupuaçu* (*Theobroma grandiflorum*) and *cupuí* (*T. subincanum*). The figure consists of seven panels representing different analyses conducted in the study. **a** An UPGMA tree is shown, depicting the genetic relationships between samples from the four study locations. The branches on the tree represent the inferred ancestral ranges of the samples. The study locations are color-coded for clarity (see Supplementary Methods and Supplementary Fig. 6 for details). **b** A haplotype network illustrates the genetic distances among samples of *T. grandiflorum* (*cupuaçu*) and *T. subincanum* (*cupuí*). This network provides insights into the genetic diversity and relationships between the two species. **c** PCA plot reveals the genetic relationships between *T. grandiflorum* and *T. subincanum* samples collected from the study locations. This analysis aids in understanding the population structure and differentiation of the two species. **d** Results for STRUCTURE analysis to infer the relationships among the species by grouping them into  $K = 2$  and  $K = 3$  optimal clusters.  $K = 2$  represents the expected clustering for the two species (*T. grandiflorum* and *T. subincanum*). Different ancestry groups are colored by different colors. Supplementary Methods and Supplementary Data 3 provide further details on this analysis. **e** Mutation loads and positive selection *loci* candidates in *T. grandiflorum* and *T. subincanum* samples were examined by calculating the ratio of nonsynonymous (dN) to synonymous (dS) changes across 1000 bootstrap replicates. This analysis focused on regions of the genome potentially under positive selection, with the number of candidate *loci* indicated on the *x*-axis. See Supplementary Methods and Supplementary Data 4 for more information. **f** Based on genomic data, the biogeographic history of *cupuaçu* was reconstructed, indicating its origin in the Middle/Upper Rio Negro, Northwestern Brazil. Additionally, the figure shows the extensive introduction of *cupuaçu* to various sites within the Amazon Basin and beyond. The study sites are marked as Xapuri-Acre (ACRE), Balbina-Amazonas (BALB), Tapajós-Pará (PARA), and São Gabriel da Cachoeira-Amazonas (SGCA). **g** Demographic history of *cupuaçu*, inferred using Stairway Plot 2. The *y*-axis represents the effective population size ( $N_e$ ), which is proportional to nucleotide diversity. Decreases in  $N_e$  indicate bottleneck events, as highlighted in the plot. Additional details can be found in the Methods and Supplementary Data 5.

number of ancestor groups as inferred by STRUCTURE analysis (Fig. 2d; see also Methods, Supplementary Fig. 5 and Supplementary Data 3). Expected heterozygosity levels of the identified ancestry group corresponding to *cupuaçu* were also low ( $H_E = 0.06$ ; see Supplementary Data 3). All of these patterns are consistent with a reduction in genetic diversity in *cupuaçu* as a consequence of inbreeding and a lack of introgression from *cupuí* to *cupuaçu*.

Additionally, our results suggest that *cupuaçu* conforms to the “cost-of-domestication” hypothesis<sup>41</sup>, as evidenced by the excessive mutation loads, particularly an abundance of non-synonymous changes in *cupuaçu* compared to *cupuí* in candidate *loci* for selection (Fig. 2e; Supplementary Data 3; Supplementary Methods). Furthermore, in theory, domestication typically leads to fewer regions under selection, controlling most of the desired features with larger effects<sup>42,43</sup>. Our findings align with this expectation (6471 candidate *loci* have increased in frequency in *cupuaçu* vs. 10,445 candidate *loci* in *cupuí*; Fig. 2e; Supplementary Data 3; Supplementary Methods), although only a subset of these genes would have been positively selected for, with the majority of them increasing in frequency due to hitchhiking.

**Cupuaçu in time and space.** A stairway plot (Fig. 2g), along with the analyses outlined above, indicates a loss of genetic diversity in *cupuaçu*. This plot also suggests that *cupuaçu* was the product of two phases of domestication, the first of which occurred long before European colonization, in the mid-Holocene, somewhere in Northwestern Amazonia (Fig. 2f). The precise locality of domestication cannot be determined based on the limitations in the extent of our geographic sampling. However, earliest diverging *cupuaçu* specimens are found to be most closely related to *cupuí* samples from São Gabriel da Cachoeira and Balbina, in the Middle-Upper Rio Negro Basin (Fig. 2b; Supplementary Fig. 6; Supplementary Data 5), suggesting that initial domestication occurred somewhere around these localities.

Dating analyses indicate two major bottlenecks: one between 5347–7943 yBP, which would mark the initial domestication of *cupuaçu*, and a second one, which would represent a second phase of domestication, from *c.* 169 years ago (Fig. 2g; Supplementary Data 4). Between these two bottleneck events, we interpret that *cupuaçu* may have been dispersed only locally and over limited distances, which is consistent with the observed stabilization of the effective population size seen in Fig. 2g, as also theoretically suggested<sup>40</sup>. Additionally, samples of *cupuaçu* from all areas are not very genetically distinct from each other (Fig. 2b, d), suggesting a single initial domestication process in

the Middle/Upper Rio Negro region. However, Allaby et al.<sup>44</sup> showed how apparent monophyly and geographic origin of a crop might result from protracted domestication with individuals selected from multiple localities. Therefore, the initial domestication process would have been followed by the introduction of *cupuaçu* to all other areas of its current geographic distribution during the second phase of domestication (Fig. 2f).

#### Cupuaçu and the history of human occupation in Amazonia.

Our dating of the domestication of *cupuaçu* matches archaeological evidence of the early occupation of the Amazon by indigenous societies from the early Holocene<sup>13,14</sup>. It is also consistent with evidence that revealed systems of plant cultivation and management from the same period<sup>45–48</sup>, which included fructiferous trees<sup>47,48</sup>. In fact, the Rio Negro Basin has been occupied since at least 8500 yBP, while Northwestern Amazonia is considered among the very earliest hubs of human settlement and plant cultivation in the Americas<sup>46–50</sup>. Our dating somewhat precedes earliest archeological evidence for the use of *cupuaçu* that dates to the late Holocene (4200–500 yBP) and all known sites with archaeobotanical remains in the Middle-Upper Rio Negro Basin<sup>9</sup>.

People from the Northwestern Amazon Basin belong to at least twenty ethnic groups that speak languages from three linguistic families: Arawak, Tukano and Makú<sup>51</sup>. The forest in this region is composed of *terra-firme* upland vegetations which do not flood, *campinarana* forests that display stunted vegetation that grow on well-drained, nutrient poor sandy soils, and of *igapó* forests flooded throughout the year. *Theobroma* species are mostly associated with the *terra-firme* forests, growing over clayed, non-flooded soils up to 1000 meters elevation<sup>22,23</sup>, and *cupuaçu* would have been mostly planted in areas that allowed its adequate growth. Thus, in the Middle-Upper Rio Negro, the initial cultivation of *cupuaçu* would have been associated with other plants that are better adapted to non-flooded areas, such as cassava<sup>52</sup>. Archaeological and anthropological evidence indicate a gradual increase in the practice of creating agrobiodiversity by indigenous people since the early Holocene<sup>48,53</sup>. Evidence of the first “dark earths” (“*terras pretas*”)—fertile and productive lands resulting from human management—also date to the mid Holocene<sup>53,54</sup>. We argue our results add to the evidence that suggests that during the Holocene indigenous groups in Amazonia tended to have a broad-spectrum diet, including the use and management of palms, legumes, and other useful fruit trees<sup>9</sup>, such as *cupuaçu*.

Our results provide evidence against the idea that *cupuaçu* originated as a wild entity in the Eastern Amazon Basin<sup>13,14,20,22</sup>, as well as the idea that the species would be the product of extinct megafauna selection<sup>55</sup>, and point to a much faster, more recent, and intense human influence. Our findings also indicate that the introduction of *cupuaçu* to other regions of the Amazon Basin intensified only over the last two centuries (Fig. 2g), which coincides with historical evidence<sup>43</sup>. The popularization of *cupuaçu* from late 1880, but especially after 1970<sup>56</sup>, reflects major historical events in the Brazilian Amazon such as, respectively, the late nineteenth century rubber boom<sup>57</sup> and the large influx of migrants from other areas of Brazil promoted by the military governments of the 1970s and 1980s. During this time, there was an intensive influx of people from outside the Amazon Basin followed by a major decline in indigenous populations, with many locals being pushed away towards the inner forests to avoid attacks from rubber tappers. From this period on, the use and promotion of *cupuaçu* in popular culture would have increased, along with its commercial value<sup>52</sup>.

Our findings indicate that *cupuaçu* was domesticated prior to cacao<sup>17</sup>. This is perhaps not surprising given that cacao was likely first used for its pulp by indigenous societies<sup>18</sup>, and it was only later domesticated to produce a chocolate-like beverage from the seeds. The early use of pulp in both species would have made *cupuaçu* more attractive than cacao if a greater quantity and/or quality of pulp were found in the former or in representatives of its progenitor *cupuí*. But what is extraordinary about the origin of *cupuaçu* is that the artificial selection pressure imposed by humans on populations of its progenitor, *cupuí*, was intense enough to develop morphological traits that were sufficiently distinct for taxonomists to recognize it as a different species<sup>22</sup>.

It is important to acknowledge some limitations of our study, particularly with respect to the limited geographic sampling of populations. Results based on a sample size, consisting of only four locations scattered across the vast Amazon Basin (see Methods and Supplementary Methods) should be treated with a degree of caution. Consequently, the area we propose as the center of domestication is comparatively broad, encompassing a region between two locations that are over 400 km apart (Fig. 2f). Future collections in additional locations may provide more refined information about the plant's center of origin and a better understanding of changes in genetic diversity over time. It is plausible that such research may lead to a more precise timeframe for *cupuaçu*'s domestication in comparison to the over two millennia period (5347–7943 yBP) we outlined here. This presents an exciting opportunity for further investigations, focusing not only on cacao's crop wild relatives but also on other valuable plants in Amazonia from other seed-bearing plant families that may have undergone similar domestication processes, akin to that of *cupuaçu*.

## Methods

**Sampling, DNA extractions and sequencing.** A total of 47 specimens, comprising 24 individuals of *T. grandiflorum* (*cupuaçu*) and 23 individuals of *T. subincanum* (*cupuí*), were collected from four different locations: Acre, Amazonas-Balbina, Pará-Tapajós, and Amazonas-São Gabriel da Cachoeira (Fig. 1). The sampling aimed to cover a significant portion of the known geographic range of the species, as defined in previous taxonomic studies<sup>22–24</sup> (refer to Supplementary Methods for more details). Following the recommended sampling framework and routine for population genomic studies<sup>58</sup>, six individuals per species were collected from each of the four locations, except for Pará-Tapajós, where five *cupuí* individuals were collected. Field expeditions for sampling took place between October and November 2021, and

during collection we avoided collecting intentionally planted specimens (see Supplementary Methods).

DNA extractions were performed on freshly collected leaf samples using QIAGEN's DNeasy Plant Mini Kit according to the manufacturer's instructions. For genomic analysis, restriction-site associated DNA sequencing (RAD-seq) data were generated by Floragenex (Oregon, USA). The DNA was digested with the *Pst*I restriction enzyme, followed by sonication, ligation of sequencing adapters, and selection of 300–500 bp fragments for PCR amplification and library preparation. The resulting libraries were sequenced on an Illumina NovaSeq6000 platform, producing around 500 million 100 bp single-end reads.

**Post-sequencing genomic manipulation and SNP calling.** Reads were demultiplexed and assembled to each sample using the “stacks” v. 2.62 pipeline in R Environment<sup>59,60</sup>. Quality control of the sequencing data was assessed using “FastQC” v. 0.1.3 in R<sup>61</sup>. After demultiplexing, quality trimming (with a quality threshold of 30), alignment, and data normalization, an average of 3.9 million reads per sample were obtained (Supplementary Table 2 and Supplementary Data 2). The “QuasR” v. 1.38.0 package in R<sup>62</sup> was used for data normalization. Five samples were excluded from further analysis due to generating relatively fewer reads compared to the other samples (see Supplementary Table 1).

Resulting data were mapped against the reference genome of *T. cacao* (Criollo cultivar B97-61/B2)<sup>63</sup> using the “ShortRead” v. 4.2 package in R<sup>64</sup>. Variant calling was performed on the aligned files using “bcftools” v. 1.16, and SNP filtering was conducted using “vcftools” v. 0.1.16 in Python v. 3.10.8 environment<sup>65,66</sup>. SNP filtering criteria retained only SNPs with a genotype call rate above 95% across all individuals and a minor allele count of three, with genotypes containing less than three reads being recoded. The mean read depth of the samples was  $2462.8 \pm 803.9$  (average  $\pm$  standard deviation; see also Supplementary Figs. 1, 2), and the mean proportion of mapped reads (including SNPs and indels) was  $206,614 \pm 49,733$  per sample (Supplementary Table 3; Supplementary Fig. 3; Supplementary Methods), which represented an average of  $93.9\% \pm 2.7\%$  of mapped variants per sample. This resulted in a final filtered and aligned variant calling file containing 42 samples from eight populations (21 samples per species, 4–6 samples per population) and 339,034 variant sites.

**Genomic structure analysis and ancestral range reconstructions.** We conducted a series of complementary analyses to examine the genomic structure of *cupuaçu* and *cupuí* populations. Firstly, we constructed a genetic distance tree using “poppr” v. 2.9.3 package in R<sup>67</sup> to illustrate the relatedness between each sample. Secondly, we generated a minimum spanning haplotype network and performed a principal component analysis (PCA) using the R packages “adegenet” v. 2.1.10<sup>68</sup> and “igraph” v. 1.3.5 in R, respectively. Additionally, we examined overall nucleotide diversity by comparing the average paired-F<sub>ST</sub> and average  $\pi$  statistics across all populations for the 10 chromosomes. This analysis, performed using “bcftools” v. 1.16<sup>65</sup>, involved defining a window frame of 10 kbp. Statistical differences between each pair of populations were evaluated using a Mann-Whitney-Wilcoxon test in Phyton, with a significance threshold of  $p$ -value  $< 0.001$ .

We utilized STRUCTURE v. 2.3.4<sup>69</sup> to determine ancestral groups. This allowed us to assign individuals to potential ancestry populations and calculate overall membership proportions among them. For the K populations, we ran eight analyses with K varying from 1 to 8, representing each group sampled here (two species, each collected from four sites). Each set of analysis

consisted of three independent runs, and we obtained average scores from these runs. The parameters were set with 10,000 reps and a 1000 burn-in period after verifying that the likelihood values stabilized within these values. To visualize and decide the optimal K populations based on our data, we used the “pophelper” v. 2.3.1 package in R<sup>70</sup>.

Furthermore, we conducted Bayesian phylogenetic inference based on whole-genomic SNPs<sup>71</sup> (see Supplementary Methods). The resulting phylogenetic tree served as a basis for ancestral area reconstruction analysis to estimate the historical geographic distribution of the studied groups. For this analysis, we used the “BioGeoBEARS” v. 1.1.2 package in R<sup>72,73</sup>, considering the study locations adopted in this research (see Supplementary Methods, Supplementary Tables 6 and 7 and Supplementary Data 6 for more details).

**Selection estimates and testing for the “cost-of-domestication” hypothesis.** To investigate the “cost-of-domestication” hypothesis<sup>41</sup>, we conducted gene annotation for the coding regions and calculated dN/dS ratios for each chromosome, to identify *loci* from coding regions. We employed a linkage-disequilibrium approach to identify partial and incomplete selective sweeps, using two LD-based scores: iHS and XP-EHH, with the “rehh” v. 2.0.2. package<sup>74</sup>. After remapping genes against a reference genome, we identified regions of candidate genes under selection, most of these hitchhiked with much fewer genes truly being positively selected. From this set of candidate *loci*, we measured the ratio of synonymous to nonsynonymous mutations for each site and chromosome, comparing the changes between *cupuaçu* and *cupuí* samples. For detailed methodology and outputs, please refer to the Supplementary Methods and Supplementary Data 3.

**Demographic history of *cupuaçu*.** Demographic history of *cupuaçu* was investigated to determine the time of origin of *cupuaçu* lineages. We employed the “easySFS” and “*ada*” inference tools in Java to estimate a folded site-frequency spectrum (SFS), which describes the distribution of allele frequencies across variant sites in the populations<sup>75</sup>. A 4-population model was used, with each location site (Acre, Balbina, Pará, and São Gabriel da Cachoeira in Brazil) represented. The resulting SFS file was then incorporated into the Stairway Plot 2 program in Python<sup>76,77</sup>.

Stairway Plot 2<sup>76</sup> utilizes folded SNP frequency spectra data and has demonstrated good performance, particularly for estimating effective population size in non-model organisms with unknown ancestral allele frequencies. We followed a similar approach to the one used for *T. cacao*<sup>17</sup>, conducting a demographic inference with mutation rates  $\mu = 3.1 \times 10^{-9}$  mutations  $\times$  bp<sup>-1</sup>  $\times$  generations and assuming a generation time of three years, as typically reported for *T. grandiflorum*<sup>78,79</sup>. Based on these parameters, we created an input model file for the demographic analysis, simulating 100 independent samples and incorporating four breaking-points, as recommended by the package developers<sup>78</sup>. A comprehensive list and description of the outputs from this analysis can be found in Supplementary Data 4, along with additional details in the Supplementary Methods section.

**Reporting summary.** Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

Variant call files are deposited at the European Variation Archive (<https://www.ebi.ac.uk/eva/?eva-study=PRJEB61195>). Raw FASTQ sequencing data were deposited at the SRA from NCBI via BioProject (<https://www.ncbi.nlm.nih.gov/sra/PRJNA940113>).

Received: 5 June 2023; Accepted: 19 October 2023;

Published online: 01 November 2023

### References

- Hubbell, S. P. et al. How many tree species are there in the Amazon and how many of them will go extinct? *Proc. Natl Acad. Sci.* **105**, 11498–11504 (2008).
- ter Steege, H. et al. Hyperdominance in the Amazonian tree flora. *Science* **342**, 1243092 (2013).
- Cardoso, D. et al. Amazon plant diversity revealed by a taxonomically verified species list. *Proc. Natl Acad. Sci.* **114**, 10695–10700 (2017).
- Antonelli, A. et al. Amazonia is the primary source of Neotropical biodiversity. *Proc. Natl Acad. Sci.* **115**, 6034–6039 (2018).
- Doughty, C. E. et al. Drought impact on forest carbon dynamics and fluxes in Amazonia. *Nature* **519**, 78–82 (2015).
- Esquivel-Muelbert, A. et al. Tree mode of death and mortality risk factors across Amazon forests. *Nat. Commun.* **11**, 5515 (2020).
- Arroyo-Kalin, M. & Riris, P. Did pre-Columbian populations of the Amazonian biome reach carrying capacity during the Late Holocene? *Philos. Trans. R. Soc. B: Biol. Sci.* **376**, 20190715 (2020).
- Quijano-Vallejos, P., Veit, P., Tipula, P. & Reyter, K. Undermining rights: indigenous lands and mining in the Amazon. *World Resources Institute* **1**, 1–136 (2020).
- Neves, E. G. et al. Peoples of the Amazon before European Colonization. *Amazon Assess. Rep.* **2021** **8**, 1–8.39 (2021).
- Levis et al. Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. *Science* **355**, 925–931 (2017).
- Dufour, D. L., Piperata, B. A., Murrieta, R. S. S., Wilson, W. M. & Williams, D. D. Amazonian foods and implications for human biology. *Ann. Hum. Biol.* **43**, 330–348 (2016).
- Miller, R. P. & Nair, P. K. R. Indigenous agroforestry systems in Amazonia: from prehistory to today. *Agrofor. Syst.* **66**, 151–164 (2006).
- Clement, C., De Cristo-Araújo, M., Coppens D'Eeckenbrugge, G., Alves Pereira, A. & Picanço-Rodrigues, D. Origin and domestication of native Amazonian crops. *Diversity* **2**, 72–106 (2010).
- Clement, C. R. et al. The domestication of Amazonia before European conquest. *Proc. R. Soc. B: Biol. Sci.* **282**, 20150813 (2015).
- de Oliveira, M. do S. P. & Schwartz, G. Açai—*Euterpe oleracea*. *Exotic Fruits* **1**, 1–5 (2018).
- Luiz Atroch, A. & do Nascimento Filho, F. J. Guarana—*Paullinia cupana* Kunth var. *sorbilis* (Mart.) Ducke. *Exotic Fruits* **1**, 225–236 (2018).
- Cornejo, O. E. et al. Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Commun. Biol.* **1**, 1–12 (2018).
- Zarrillo, S. et al. The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. *Nat. Ecol. Evol.* **2**, 1879–1888 (2018).
- Smith, M. K. and Botella, J. R. (2020) ‘*Ananas comosus* pineapple’, in CABI Books, CABI International.
- Alves-Pereira, A. et al. A population genomics appraisal suggests independent dispersals for bitter and sweet manioc in Brazilian Amazonia. *Evol. Appl.* **13**, 342–361 (2019).
- Clement, C. R. et al. Origin and dispersal of domesticated peach palm. *Front. Ecol. Evol.* **5**, 1–19 (2017).
- Cuatrecasas, J. Cacao and its allies, a taxonomic revision of the genus *Theobroma*. *Contr. U. S. Natl Herb.* **35**, 379–614 (1964).
- Colli-Silva, M. & Pirani, J. R. Estimating bioregions and undercollected areas in South America by revisiting Byttnerioideae, Helicteroideae and Sterculioideae (Malvaceae) occurrence data. *Flora* **271**, 151688 (2020).
- Bletter, N. & Daly, D. C. Cacao and its relatives in South America. In *Chocolate in Mesoamerica: A Cultural History of Cacao*. (ed. McNeil, C. L.) 31–68 (University Press of Florida, 2006).
- IBGE [Instituto Brasileiro de Geografia e Estatística]. Produção de cupuaçu. <https://www.ibge.gov.br/explica/producao-agropecuaria/cupuacu/br> (2023).
- Vale, M. M., Alves, M. A. & Pimm, S. L. Biopiracy: conservationists have to rebuild lost trust. *Nature* **453**, 26–26 (2008).
- Whitlock, B. A. & Baum, D. A. Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene vicilin. *Syst. Botany* **24**, 128–138 (1999).
- Sousa Silva, C. R. & Figueira, A. Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. *Plant Syst. Evol.* **250**, 93–104 (2004).
- Borrone, J. W., Meerow, A. W., Kuhn, D. N., Whitlock, B. A. & Schnell, R. J. The potential of the WRKY gene family for phylogenetic reconstruction: an example from the Malvaceae. *Mol. Phylogenet. Evol.* **44**, 1141–1154 (2007).

30. Richardson, J. E., Whitlock, B. A., Meerow, A. W. & Madriñán, S. The age of chocolate: a diversification history of *Theobroma* and Malvaceae. *Front. Ecol. Evol.* **3**, 1–14 (2015).
31. Addison, G. & Tavares, R. Hybridization and grafting in species of *Theobroma* which occur in Amazonia. *Evolution* **6**, 380–386 (1952).
32. Silva, C. R. S., Venturieri, G. A. & Figueira, A. Description of Amazonian *Theobroma* L. collections, species identification, and characterization of interspecific hybrids. *Acta Bot. Bras.* **18**, 333–341 (2004).
33. Rivas, L. H. et al. Genetic diversity in natural populations of *Theobroma subincanum* Mart. in the Brazilian Amazon. *Genet. Mol. Res.* **12**, 4998–5006 (2013).
34. Arenas-de-Souza, M. D., Rossi, A. A. B., Varella, T. L., Silveira, G. F. & Souza, S. A. M. Stigmatic receptivity and pollen viability of *Theobroma subincanum* Mart.: fruit species from the Amazon region. *Rev. Bras. Frutic.* **38**, e757 (2016).
35. Clement, C. R. 1492 and the loss of Amazonian crop genetic resources. I. The relation between domestication and human population decline. *Econ. Bot.* **53**, 188–202 (1999).
36. Alves, R. M., Sebbenn, A. M., Artero, A. S., Clement, C. & Figueira, A. High levels of genetic divergence and inbreeding in populations of *cupuassu* (*Theobroma grandiflorum*). *Tree Genet. Genomes* **3**, 289–298 (2006).
37. Smith, O. et al. A domestication history of dynamic adaptation and genomic deterioration in Sorghum. *Nat. Plants* **5**, 369–379 (2019).
38. McAlvay, A. C. et al. *Brassica rapa* domestication: untangling wild and feral forms and convergence of crop morphotypes. *Mol. Biol. Evol.* **38**, 3358–3372 (2021).
39. Gepts, P. The contribution of genetic and genomic approaches to plant domestication studies. *Curr. Opin. Plant Biol.* **18**, 51–59 (2014).
40. Gaut, B. S., Seymour, D. K., Liu, Q. & Zhou, Y. Demography and its effects on genomic variation in crop domestication. *Nat. Plants* **4**, 512–520 (2018).
41. Lu, J. et al. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* **22**, 126–131 (2006).
42. Paterson, A. H. What has QTL mapping taught us about plant domestication? *New Phytol.* **154**, 591–608 (2002).
43. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
44. Allaby, R. G., Fuller, D. Q. & Brown, T. A. The genetic expectations of a protracted model for the origins of domesticated crops. *Proc. Natl Acad. Sci. USA* **105**, 13982–13986 (2008).
45. Lombardo, U. et al. Early Holocene crop cultivation and landscape modification in Amazonia. *Nature* **581**, 190–193 (2020).
46. Mora, S. Early inhabitants of the Amazonian tropical rain forest: a study of humans and environmental dynamics (Unpublished doctoral thesis). University of Calgary, Calgary, AB. (2001)
47. Watling, J. et al. Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. *PLOS ONE* **13**, e0199868 (2018).
48. Shock, M. P., Moraes, C. & de, P. A floresta é o domus: a importância das evidências arqueobotânicas e arqueológicas das ocupações humanas amazônicas na transição Pleistoceno/Holoceno. Boletim do Museu Paraense Emílio Goeldi. *Ciências Humanas* **14**, 263–289 (2019).
49. Piperno, D. R. The origins of plant cultivation and domestication in the New World Tropics: patterns, process and new developments. *Curr. Anthropol.* **52**, S453–S470 (2011).
50. Morcote-Ríos, G. & Bernal, R. Remains of palms (Palmae) at archaeological sites in the New World: a review. *Bot. Rev.* **67**, 309–350 (2001).
51. Neves, E. G., Petersen, J. B., Bartone, R. N. & Augusto Da Silva, C. Historical and socio-cultural origins of Amazonian dark earth. In *Amazonian Dark Earths*. (eds Lehmann, J. et al.) 29–50 (Kluwer Academic Publishers, 2003).
52. Furquim, L. P. et al. Facing change through diversity: resilience and diversification of plant management strategies during the mid to late holocene transition at the Monte Castelo Shellmound. *SW Amazonia. Quat.* **4**, 8 (2021).
53. Neves, E. G. Biodiversidade e agrobiodiversidade como legados de povos indígenas. in *Povos Tradicionais e Biodiversidade no Brasil: contribuições dos povos indígenas, quilombolas e comunidades tradicionais para a biodiversidade, políticas e ameaças* (eds Cunha, M. C., Magalhães, S. B. & Adams, C.) 10–85 (Sociedade Brasileira para o Progresso da Ciência).
54. Lacerda, F. G. & Vieira, E. R. C. 'O celeiro da Amazônia': agricultura e natureza no Pará na virada do século XIX para o XX. *Topoi (Rio de Janeiro)* **16**, 157–181 (2015).
55. Guimarães, P. R., Galetti, M. & Jordano, P. Seed dispersal anachronisms: rethinking the fruits extinct Megafauna ate. *PLoS ONE* **3**, e1745 (2008).
56. Homma, A. K. O., Carvalho, R. A. & Menezes, A. J. Extrativismo e plantio racional de cupuaçuzeiros no Sudeste Paulista: a transição inevitável. in *Extrativismo Vegetal na Amazônia: histórica, ecologia, economia e domesticação* (ed. Homma, A. K. O.) 297–306 (Empresa Brasileira de Pesquisa Agropecuária, 2014).
57. Hecht, S. B. Environment, development and politics: capital accumulation and the livestock sector in Eastern Amazonia. *World Dev.* **13**, 663–684 (1985).
58. Nazareno, A. G., Bemmels, J. B., Dick, C. W. & Lohmann, L. G. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Mol. Ecol. Resour.* **17**, 1136–1147 (2017).
59. Rochette, N. C., Rivera-Colón, A. G. & Catchen, J. M. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.* **28**, 4737–4754 (2019).
60. R. Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2021).
61. Kassambara, A. fastqcr: Quality Control of Sequencing Data. R package version 0.1.3 (2023).
62. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinform.* **31**, 1130–1132 (2014).
63. Argout, X. et al. The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2010).
64. Morgan, M. et al. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
65. Li, R. et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
66. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
67. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. *Poppr*: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281 (2014).
68. Jombart, T. *ade4*: an R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405 (2008).
69. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
70. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2016).
71. Ronquist, F., van der Mark, P. & Huelsenbeck, J. P. Bayesian phylogenetic analysis using MrBayes. In *The Phylogenetic Handbook*. (ed. Lemey, P.) 210–266 (Cambridge University Press, 2009).
72. Matzke, N. J. Probabilistic historical biogeography: new models for founder-event speciation, imperfect detection, and fossils allow improved accuracy and model-testing. *Front. Biogeogr.* **5**, 242–248 (2013).
73. Matzke, N. J. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in Island Clades. *Syst. Biol.* **63**, 951–970 (2014).
74. Gautier, M. & Vitalis, R. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**, 1176–1177 (2012).
75. Sousa, V. & Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* **14**, 404–414 (2013).
76. Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559 (2015).
77. Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol.* **21**, 1–9 (2020).
78. Alves, R. M. et al. Canopy replacement used in the evaluation of cupuassu tree genotypes in the state of Pará. *Rev. Bras. Frutic.* **42**, 1–11 (2020).
79. Falcão, M., de, A. & Lleras, E. Aspectos fenológicos, ecológicos e de produtividade do cupuaçu – *Theobroma grandiflorum* (Willd. ex Spreng.) Schum. *Acta Amazonica* **13**, 725–735 (1983).

## Acknowledgements

This study was funded by multiple sources, including the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) under Finance Code 001, and the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) through Grant IDs 2020/01375-1 and 2020/10206-9. Additionally, AF acknowledges support from a CNPq fellowship (Grant ID 310645/2021-2). M.C.-S. extends special gratitude to Renato Ramos and Paulo Sano for facilitating access to a cluster funded by the “PAT (Plano de Ação Territorial) Espinhaço Mineiro” initiative, which received funding from the Brazilian Environment Ministry (MMA) through the ProEspécies Initiative ([www.proespecies.eco.br](http://www.proespecies.eco.br)). Furthermore, we thank Douglas C. Daly for granting permission to use photos of *cupuí*, which were utilized in creating certain elements of Fig. 1 of this work.

## Author contributions

M.C.-S.: conceptualization, methodology, formal analysis, writing the original draft. A.F.: reviewing and editing the manuscript, data validation. J.E.R.: supervision, reviewing and editing the manuscript, data validation. E.G.N. and J.W.: reviewing and editing the manuscript, data validation. J.R.P.: supervision, project administration, funding acquisition, reviewing and editing the manuscript, data validation.

## Competing interests

The authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43247-023-01066-z>.

**Correspondence** and requests for materials should be addressed to Matheus Colli-Silva.

**Peer review information** *Communications Earth & Environment* thanks Robin Allaby, Hanno Schaefer, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Erika Buscardo, Aliénor Lavergne, Joseph Aslin. A peer review file is available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023