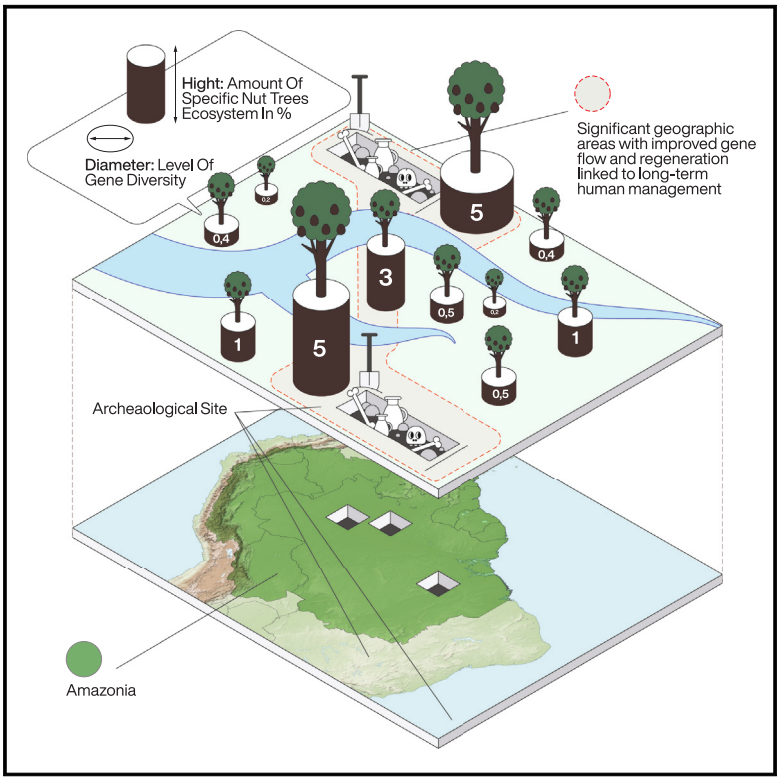
Article

Current Biology

Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin

Graphical abstract



Highlights

- New Brazil nut genome assembly allows for genetic diversity studies
- Evidence for decline in genetic diversity since the Late Pleistocene
- Significant geographic structure linked to Indigenous human management patterns
- Results highlight the value of historical insights for managing genetic diversity

Authors

Hua Wang, Victor Caetano-Andrade, Nicole Boivin, ..., Francislaide da Silva Costa, Detlef Weigel, Patrick Roberts

Correspondence

detlef.weigel@tuebingen.mpg.de (D.W.), roberts@gea.mpg.de (P.R.)

In brief

Wang and Caetano-Andrade et al. show that traditional human management is linked to increased genetic diversity in the Brazil nut, which has significantly declined since the Late Pleistocene. Their research highlights the importance of understanding historical human interactions to inform effective management strategies for vital tree species today.

Wang et al., 2025, Current Biology *35*, 1–11
February 3, 2025
https://doi.org/10.1016/j.cub.2024.12.023

CellPress

Current Biology

CellPress
OPEN ACCESS

Article

Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin

Hua Wang,[1,2,11] Victor Caetano-Andrade,[3,4,11] Nicole Boivin,[4,5,6] Charles R. Clement,[7] Wellyngton Espindola Ayala,[7] Ramiro Dario Melinski,[7] Francislaide da Silva Costa,[8] Detlef Weigel,[1,9,*] and Patrick Roberts[3,4,10,12,*]

[1]Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany
[2]College of Horticulture and Forestry sciences, Huazhong Agricultural University, Wuhan 430070, Hubei, China
[3]Department of Coevolution of Land Use and Urbanisation, Max Planck Institute of Geoanthropology, 07745 Jena, Germany
[4]Department of Archaeology, Max Planck Institute of Geoanthropology, 07745 Jena, Germany
[5]School of Social Science, University of Queensland, Brisbane, QLD 4072, Australia
[6]Griffith Sciences, University of Griffith, Nathan, QLD 4222, Australia
[7]Instituto Nacional de Pesquisas da Amazônia, Manaus, AM 69067-375, Brazil
[8]Laboratório de Evolução Aplicada, Departamento de Genética, Universidade Federal do Amazonas, Manaus, AM 69067-005, Brazil
[9]Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, 72076 Tübingen, Germany
[10]School of Archaeology, University of the Philippines, Quezon City 1101, the Philippines
[11]These authors contributed equally
[12]Lead contact
*Correspondence: detlef.weigel@tuebingen.mpg.de (D.W.), roberts@gea.mpg.de (P.R.)
https://doi.org/10.1016/j.cub.2024.12.023

SUMMARY

The Amazon rainforest is characterized by a limited number of hyperdominant trees that play an oversized role in its ecosystems, nutrient cycle, and rainfall production. Some of these, such as the Brazil nut, appear to have been intensively exploited and dispersed by Indigenous populations since their earliest arrival in this part of South America around 13,000 years ago. However, the genetic diversity—and geographic structure—of these species remains poorly understood, as does their exact relationship with past human land use. We use a new genome assembly for Brazil nut to analyze 270 individuals sampled at areas with varying intensities of archaeological evidence. We demonstrate that overall low genetic diversity, with a notable decrease since the Late Pleistocene, is accompanied by significant geographic structure, where evidence for improved gene flow and regeneration by long-term traditional human management is linked to increased genetic diversity. We argue that historical perspectives on the genetic diversity of key tree species, such as the Brazil nut, can support the development of more active management strategies today.

INTRODUCTION

The Amazon Basin is a critical carbon sink, reservoir of biodiversity, and producer of rainfall, with a central position in the functioning of the Earth system.[1] In this context, large, hyperdominant trees (defined as a set of ~220 species that account for half of the trees in the Amazon) have been considered particularly important for biogeochemical cycling.[2] Among the hyperdominant species, the ones that have documented, long-term economic and cultural relationships with human societies are disproportionately overrepresented.[3] More than four-fifths of Amazonian arboreal species, both hyperdominant and non-hyperdominant, have recorded uses by humans.[4] Alongside growing evidence that the Amazon Basin is a center of plant cultivation and domestication,[5] and has hosted various forms of ancient urbanism,[6] it is becoming clear that even in the so-called "intact" Amazonian forests, past human activities have shaped species composition, structure, and fire dynamics for millennia.[3] Nevertheless, the extent of human influence on Amazonian forests, and its interplay with the genetic pools of key tropical tree species, remains remarkably under-investigated, despite its relevance to modern conservation and management concerns.

Here, we report patterns of genetic diversity in the Brazil nut (*Bertholletia excelsa*), a monotypic genus in the Lecythidaceae pantropical family. It is one of the hyperdominant *terra firme* (upland) forest trees of the Amazon, with a critical role in carbon cycling, soil maintenance, and forest dynamics.[7] Its large, edible seeds are an important non-timber economic product coming from the Amazon Basin.[8] This tree is of particular interest, given the contributions of the agouti (*Dasyprocta* spp., Rodentia) to its dispersal, combined with outcrossing mainly by specific medium and large bees from the Apidae and Anthophoridae families, which poses certain limitations on population mixing and spatial dispersal.[9] Numerous studies have highlighted the significance of humans in the dispersal and ecology of this species from the terminal Pleistocene onward.[8] Moreover, Brazil nut is a shade-intolerant species and is most successful colonizing

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

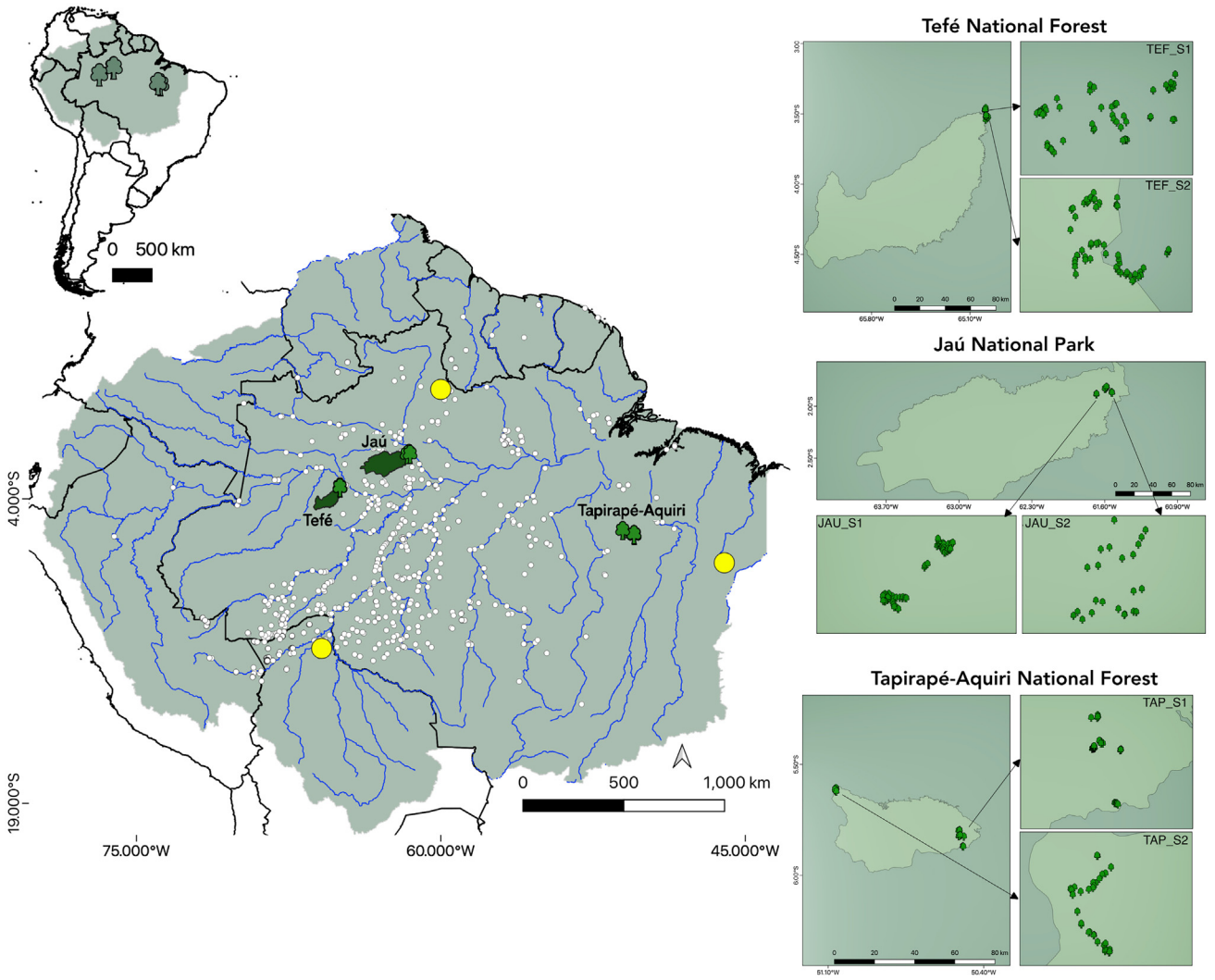


Figure 1. Geographic distribution and domestication origins of Brazil nut

Location of sampled Brazil nut (*Bertholletia excelsa*) individuals at the sites Tapirapé-Aquiri National Forest (TAP), Tefé National Forest (TEF), and Jaú National Park (JAU) within the Brazilian system of conservation units. Suffixes "_S1" and "_S2" indicate subsites with different archaeological backgrounds. Yellow dots indicate the hypothesized origins of *Bertholletia excelsa* domestication (the easternmost point is an early hypothesis by Mori and Prance, discussed in Clement et al.[16]; the northern point was hypothesized in Shepard et al.[8]; and the southwest point was proposed in Thomas et al.[14]). White dots indicate the distribution of the species in the Amazon Basin based on data from the Global Biodiversity Information Facility (GBIF) (see also Table S7).

forest areas with ample light, which are often associated with human activity.[10] Anthropogenic disturbance is, therefore, an important driver of its colonization in the landscape, as demonstrated by dendrochronological studies of Brazil nut trees, in which the patterns of growth and establishment of trees are intimately associated with Indigenous and colonial management practices that facilitate establishment of the species.[11] These human actions can be expected to have an impact on the genetic diversity of this species, with habitat disturbance to promote access to light, and dispersal of seeds, potentially affecting genetic diversity or extending particular lineages across an anthropogenic space.[8]

Brazil nut individuals are highly heterozygous, indicative of rampant outcrossing, likely associated with self-incompatibility.[12] Large-scale deforestation and fragmentation have been demonstrated to disrupt gene flow, which is harmful to the perpetuation of the species, as observed in more degraded forests where young plants have lower genetic diversity and more signs of inbreeding than adults.[12] Genetic studies, spatial modeling, linguistic and archaeological research, and field phenotypic observations have led to the proposal that the Brazil nut was incipiently domesticated and dispersed by Indigenous people,[13] with suggestions for centers of diversity and sources of dispersal in the southwestern Amazon,[12,14] eastern Amazon,[15] and/or northern Amazon[8] (Figure 1). This is supported by the fact that, at varied geographical scales, there is overall limited genetic differentiation (F_{st}) between Brazil nut populations,[12] which suggests that humans distributed Brazil nut widely during the Holocene.[5,8]

Clear relationships have also been identified between Brazil nut stands and archaeological sites along the Madeira and

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

Amazon Rivers,[14] but broader perspectives on ancient human management of these populations across the Amazon Basin have been lacking. In particular, because full genome information for this hyperdominant tree has not been available, it has been difficult to assess detailed genetic diversity from the perspective of the whole genome and, with it, to have a better understanding of the ways in which human activities, from active dispersal to promotion through opening of the canopy, have shaped Brazil nut genetics across space and time. To fill this gap in understanding of Brazil nut biology, we produced a new, high-quality reference genome, which we then used as a platform for estimating genetic diversity and structure within and between multiple locations across the Amazon Basin. Specifically, we sought to test the hypothesis that Brazil nut trees growing in areas with clear evidence of past human presence have greater genetic diversity as a result of agroforestry management practices (e.g., multiple introductions of Brazil nut, canopy opening, manuring, and disperser attraction) that favored the species. Critically, our sampled sites are located in widely spread areas, at some distance from regions where Brazil nut has been proposed to have been incipiently domesticated prior to dispersal.

RESULTS

The characteristics of the Brazil nut reference genome

The Brazil nut genome size is close to 570 Mb, as inferred with k-mers from Illumina short reads, with around 1% heterozygosity (Figure S1A). Using PacBio circular consensus sequencing and chromatin contact information, we assembled and scaffolded 17 pseudo-chromosomes with a total size of 568.1 Mb, which covered ~98.2% of the assembled sequences (578.5 Mb). The length of the 17 assembled chromosomes ranged from 28.4 to 44.4 Mb (Figures 2A, 2B, and S1B; Table S1). A long terminal repeat (LTR) assembly index (LAI) of 13.4 confirmed the high quality of this assembly. Repetitive sequences constituted about 55.4% of the entire genome, about 27% of which were LTR retrotransposons (Table S2). We predicted 38,222 protein-coding genes, with 96.1% benchmarking universal single-copy orthologs (BUSCOs) completeness (Tables 1 and S3). A total of 756 ribosomal RNA (rRNA), 5,752 transfer RNA (tRNA), 336 small nuclear RNA (snRNA), and 161 microRNA (miRNA) copies were identified.

Genome evolution of Brazil nut and synteny with other plants

Among the annotated genes, fewer than 300 genes were species-specific and did not have orthologs in the well-known model species *Arabidopsis thaliana* or any of the widely studied woody species such as grape (*Vitis vinifera*), sweet orange (*Citrus sinensis*), cottonwood (*Populus trichocarpa*), and walnut (*Juglans regia*). We found 2,525 expansions and 3,242 contractions of gene families in Brazil nut, which is within the range for the other five species. Among the 68 rapidly expanding gene families, several are associated with resistance to biotic and abiotic stresses. Wall-associated kinase-like (WAKL) proteins belong to the broader groups of receptor-like proteins (RLKs), which are well known to play a central role in pathogen defenses,[17] UDP-glycosyltransferases (UGTs) can modify both endogenous molecules relevant for defense and xenobiotics,[18] heavy-metal-associated isoprenylated plant proteins (HIPPs) are likely involved in heavy metal detoxification or tolerance,[19] and cytochrome P450s (CYP450s) are among the most central, and also the most diverse, components in the synthesis of defense-related compounds in plants[20] (Figures 2C and 2E). All of these families are expected to contribute to resilience and adaptability. Diversity is also apparent at the level of orthology, from species-specific genes to genes with single- and multiple-copy orthologs in other plants (Figure 2D). Finally, we find that the Brazil nut genome shows the highest synteny with grape (Figures S2A–S2D; Table S4).

Genome expansion of Brazil nut

Repetitive sequences account for a large fraction of the Brazil nut genome (Table S2). We assessed the age of TEs using the Kimura distance between the LTRs of Ty1 and Ty3 retrotransposons, which indicated that 90% of the insertions occurred between 2 and 40 mya, which is a long time span compared with other species, including other trees.[21] The burst of Ty3 insertions is more recent than that of Ty1, which is responsible for much of the LTR-retrotransposon content of Brazil nut (Figure 2F). There were a total of 124,479 intact LTR-retrotransposons from the Ty1 and Ty3 superfamily, and we classified these into 271 families, with 11 major Ty1 clades and 7 major Ty3 clades (Figure 2G). The most prominent Ty1 clade was Ale, and the most prominent Ty3 clade was Tekay. The distribution of substitution rates at synonymous sites (*Ks*) in duplicated, collinear regions within the Brazil nut genome indicated that the species experienced two rounds of whole-genome duplication (WGD) (Figure S3A). There was a first peak around $K_s = 1$, indicating that a more recent WGD occurred ~82 mya (estimation based on "time = $K_s/2\mu$ and $\mu = 6.1 \times 10^{-9}$ years^{-1}[22]). The second peak was around $K_s = 1.65$, corresponding to an older WGD of ~135 mya (Figure S3B).

Genetic variation and demographic history of Brazil nut populations shaped by human activity

To understand the population histories of Brazil nut growing in different parts of the Amazon Basin, and their connection to human activity, we sampled 270 individuals from three protected areas of the Brazilian Amazon, Tapirapé-Aquiri National Forest (TAP), Tefé National Forest (TEF), and Jaú National Park (JAU). For each of the sites, we sampled two subpopulations with different archaeological backgrounds. The subsites at TAP are separated by ~60 km, and archaeological surveys indicate human occupation in the region as far back as 5,700 years calibrated years before present (cal BP).[23] The Brazil nut trees at TAP_S1 are growing on a *terra preta* (Amazonian dark earth) archaeological site that was occupied from 1,200 to 330 cal BP,[23] while there is no clear archaeological evidence for human activity at the TAP_S2 subsite. At present, Brazil nut populations in TAP are not intensively managed because a large mining enterprise was granted rights to this area by the Brazilian government and local communities are prohibited from accessing these tree populations.

The TEF_S1 and TEF_S2 subsites are on opposite banks of one of the tributaries of the Tefé River, separated by ~5 km. The trees at TEF_S2 are growing on an archaeological complex with extensive evidence of past human activity that includes *terra*

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

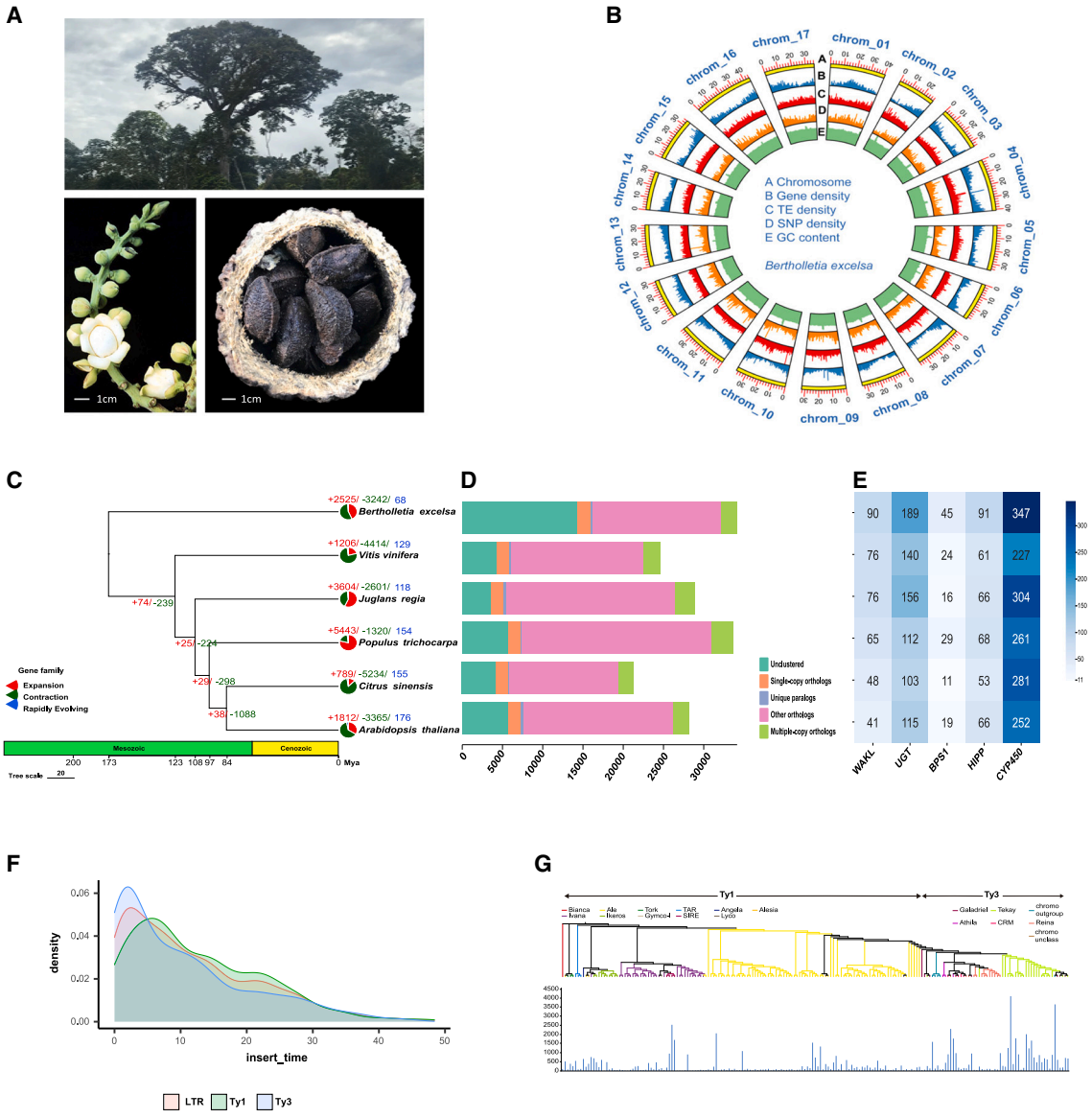


Figure 2. Morphology and genomic characteristics of Brazil nut

(A) A Brazil nut tree in the Amazon rainforest (top), a flower (lower left), and a seed pod (lower right).

(B) Circos plot of the genome assembly, with TE density, SNP density, and GC content along the 17 pseudo-chromosomes. Lengths in Mb.

(C) Fraction of expanding, contracting, and rapidly evolving gene families in Brazil nut and five other exemplary dicotyledonous angiosperms. "+" (red) and"−" (green) denote gene expansions and contractions, and the numbers of rapidly evolving families are in blue. The scale on the x axis shows the estimated divergence times.

(D) Distribution of different types of genes in the species shown in (C).

(E) The sizes of five gene families that are rapidly evolving in Brazil nut and their comparison with the five other species shown in (C).

(F) The insertion times of major TE classes.

(G) Phylogenetic analysis of intact LTR-transposons (top) and their clade-level classification and copy number (bottom) (see also Figures S1–S3 and Tables S2 and S4).

preta and ceramics, in addition to the occurrence of many other useful plants, such as rubber tree (*Hevea brasiliensis*, Euphorbiaceae), caiaué (*Elaeis oleifera*, Arecaceae), and açaí (*Euterpe precatoria*, Arecaceae).[24] At TEF_S1, although Brazil nut individuals are not growing on a *terra preta* archaeological site, historical sources have revealed that the Brazil nut stands at both TEF_S1 and TEF_S2 were likely constantly managed throughout European colonization until the present,[8] which makes these sites a special case of long-term pre- and post-colonial human environmental manipulation. The subsites at JAU are separated by ~15 km, with *terra preta* present in the vicinity of JAU_S2. At JAU_S1 there are no evident signs of past human presence and the riverine community at the site was established over the last century.

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

Table 1. Genome assembly and annotation statistics for Brazil nut

Property	Measurement
Assembly size	606,485,445 bp
Combined length of pseudo-chromosomes	595,900,818 bp
N50 scaffold length	35.3 Mb
Size of retrotransposons	159.1 Mb
Size of DNA transposons	112.5 Mb
Size of total repeat sequences	314.7 Mb
GC content	34.5%
Protein-coding genes	38,222
Transcript isoforms	45,049
Mean mRNA length	4,811 bp
Mean coding sequence length	989 bp
Mean intron length	1,295 bp
BUSCO completeness	96.1%

See also Tables S2 and S3.

Table 2. Nucleotide diversity across different Brazil nut sites and subsites

Site	Subsite	Evidence of human presence	Accessions	π ($\times 10^{-3}$)
Tapirapé-Aquiri National Forest (TAP)	all	N/A	69	1.18
	TAP_S1	none	32	1.22
	TAP_S2	*terra preta*, ceramics, lithics, carbonized seeds	37	1.24
Tefé National Forest (TEF)	all	N/A	112	1.25
	TEF_S1	near *terra preta*	51	1.29
	TEF_S2	*terra preta*, ceramics	61	1.22
Jaú National Park (JAU)	all	N/A	89	1.27
	JAU_S1	near *terra preta*	52	1.34
	JAU_S2	none	37	1.28

Nucleotide diversity (π) in different sites and subsites; each subsite has been marked with its archaeological situation (see also Figure S4).

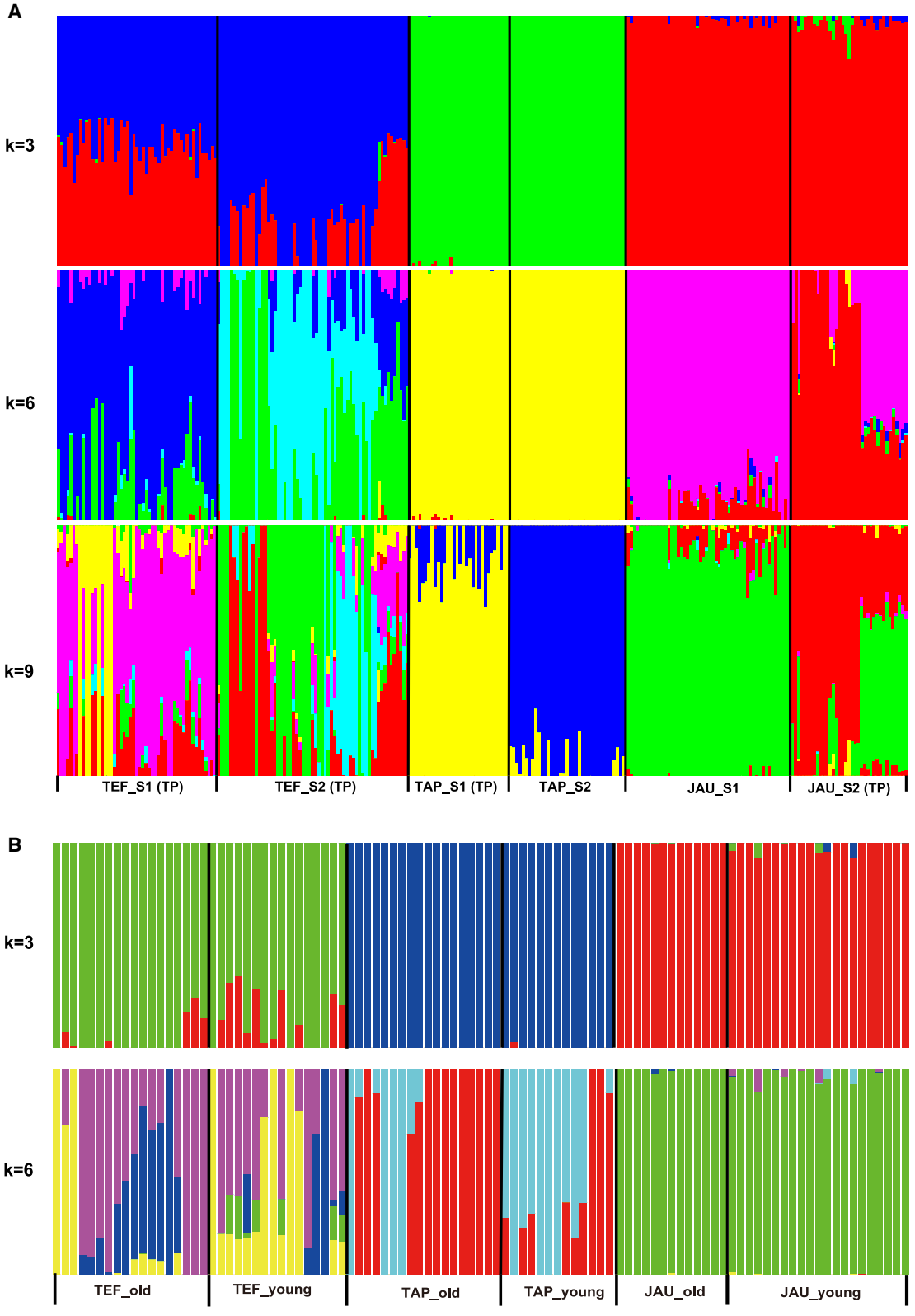
We genotyped our collection of 270 Brazil nut individuals using double-digest restriction site associated DNA sequencing (ddRAD-seq) to cover approximately 5% of the reference genome (Table S5). We identified 1,16,306 single-nucleotide polymorphisms (SNPs), with a ratio of 1.33 for nonsynonymous to synonymous SNPs, and 10,125 small indels (<10 bp) (Table S6). Nucleotide diversity (π) was similar for the three collection sites (averages from 1.62 to 1.81 $\times 10^{-4}$ bp^{-1}; Table 2; Figure S4A), as was linkage equilibrium, with decay being slower than that in other wild tree species[25] (Figure S4B). There was clear isolation by distance, with F_{st} = 0.26 and 0.24 between TAP and the other two sites, while F_{st} between TEF and JAU was only 0.07. Between subsites, the F_{st} was low, from 0.02 to 0.07. ADMIXTURE revealed an optimal value of $k = 9$ subpopulations (Figure 3A). In a principal-component analysis (PCA), the first three principal components were consistent with the ADMIXTURE inferences, which also showed that only the subpopulations from Tefé diverged notably from each other (Figure S5A). TreeMix analysis identified significant gene flow between two of the western sites (JAU_S1 to TEF_S2) (Figure S6). When we classified the populations by age according to tree diameter at breast height (DBH) (young group = DBH $\leq$ 80 cm; old group = DBH $\geq$ 150 cm), we found the young group from site TEF to have a stronger gene flow from the JAU population than the old group from TEF (Figure 3B). Estimates of historical effective population size (N_e) using the stairway plot method[26] indicated a decline of overall Brazil nut genetic diversity since the last glacial maximum (LGM). A prior demographic bottleneck, consistent with a known period of environmental change, the Gelasian epoch (2.5–1.8 mya), was more moderate (Figure 4A). Using SMC++ gave similar results. After a short phase of apparent population stability, the genetic diversity of Brazil nut has apparently been decreasing until today (Figure 4B).

DISCUSSION

Compared with other large tree species exploited by humans, the Brazil nut populations sampled in this study have a relatively low level of genetic diversity,[25] displaying a notable decline since the Late Pleistocene. Today the main, and perhaps only, significant natural disperser of the Brazil nut apart from humans is the red-rumped agouti (*Dasyprocta leporina*), along with related species, all of which have small ranges and, therefore, are not conducive to long-distance dispersal, imposing further limits on genetic diversity. As humidity increased during the transition to the Holocene, the more open forests of the glacial period would have been replaced by denser forests where Brazil nut has difficulty recruiting new trees,[10] perhaps leading to even more reduced opportunities for genetic exchange until the arrival of humans. Despite limitations to its dispersal, Brazil nut has been ecologically very successful, being among the 227 hyperdominant species of Amazon flora[2] and among the most dominant species for carbon storage and productivity.[7] A GO enrichment analysis of species-specific genes hints at pathways that may be of relevance for environmental adaptation (Figure S5B). For example, phosphate ion transport function may be of particular significance, given that soil phosphate availability in soil is low across much of the Amazon Basin.[27]

It is highly likely that humans played a major role in the success of Brazil nut, as previous studies suggested that patterns of extant genetic diversity have been shaped by human influence.[8,14] By local burning and forest disturbance, humans may have "released" Brazil nut growth and increased the chances of this long-lived pioneer to establish itself in the canopy.[10,11] In addition, humans have planted Brazil nut trees around their settlements for a long time.[8] These practices have been seen as a key part of the process of domestication of Amazonian landscapes, a process that has transformed species abundance and floristic composition.[28] The continued decline of the effective population size of Brazil nut may be associated with the well-documented movements of peoples who had new approaches to land management, such as the intensification of manioc

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023



(legend on next page)

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

cultivation and interethnic trade networks.[8] Despite Brazil nut's relatively low genetic diversity, or perhaps in part because of it, our study has been able to pick up geographic differences in genetic variation. Reinforcement of gene flow between populations of Brazil nut, associated with human settlements, has been proposed before.[14] We observe the same situation in our study, with the only significant gene flow between two subpopulations from different sites, separated by 380 km, likely due to recent human intervention (Figure S6).

Our data show that the distribution of genetic variation, when seen through the perspective of population division and population structure, is closely related to archaeological evidence for past human presence as well as the continuous forest management to the present day. At the inter-site level, the greater diversity at the TEF sites provides support for Indigenous human communities contributing to an increase in Brazil nut genetic diversity, given the well-attested archaeological evidence for pre-Columbian human presence and present plant manipulation in this region.[8,24] At the intra-site level, it is also possible to discern a richer genetic background in TAP and JAU subpopulations on *terra preta* than in subpopulations without evidence of anthropogenic soils (Figure 3A). Indigenous forest management has, therefore, likely played an important role in stemming overall declines in Brazil nut genetic diversity, with incipient domestication and human management of the landscape even reversing trends of decline in genetic diversity and encouraging regional variability and genetic richness.

Our data provide evidence that, throughout the Holocene, Indigenous populations across the Amazon have significantly inserted themselves into the ecology and seed dispersal mechanisms of not only field crops but also trees.[16] As a result of limited dispersal by natural means, Brazil nut has low, and overall declining, genetic diversity. Industrial or large-scale deforestation is worsening this trend, with Brazil nut in logged areas having drastically reduced genetic diversity,[12] which poses a major problem for this species and its role as a critical ecological component of Amazonian ecosystems.[2,7] Our analysis confirms the very different influences that contemporary local communities and past human societies have had on the genetic reservoir of this species, with the introduction of different lineages and management of dynamic forest landscapes increasing genetic diversity in different regions, even if certain lineages are being selected for and promoted across the forest at large scale. We argue that the promotion of active Indigenous approaches to Brazil nut management through agroforestry management for nut harvesting and mobile maintenance of dynamic succession in certain forest areas offers the potential to increase regional and basin-wide genetic diversity, balancing challenges of declining genetic diversity with its status as a key economic and ecological species for the Amazonian biome at large. Moreover, our study highlights the potential and importance of combining archaeological and genomic approaches to exploring the genetic and anthropogenic histories of key tree species in the tropics through time. We posit that studies of this type on Brazil nut in different regions, and on different related tree species, have the potential to yield insights into long-term human effects on tree genetics and their significance for contemporary management strategies.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Patrick Roberts (roberts@gea.mpg.de).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

Data for the whole assembled genomes, genome resequencing, and ddRAD-seq have been deposited in the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). Accession numbers are listed in the key resources table. This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research was carried out in accordance with Brazilian legislation and was registered in the Biodiversity Authorization and Information System (SisBio) (registration no. 64134-2) and in the National System for the Management of Genetic Heritage and Accessed Traditional Knowledge (SisGen) (registration no. A34C6A8). This work was supported by the Max Planck Society and the Novozymes Prize of the Novo Nordisk Foundation. We thank Susan Trumbore and Johannes Krause for encouragement and early discussions.

AUTHOR CONTRIBUTIONS

H.W., V.C.-A., D.W., and P.R. designed the research. H.W. produced the genome assemblies, performed population genetic analyses, and assisted with sampling. V.C.-A. performed sampling and assisted in data interpretation. N.B. and C.R.C. assisted in the design of the sampling strategy. D.W. assisted in data interpretation. P.R. organized the sampling work and analyzed data. H.W. and P.R. wrote the manuscript with assistance from the other authors.

DECLARATION OF INTERESTS

D.W. holds equity in Computomics, which advises plant breeders. D.W. also consults for KWS SE, a globally active plant breeder and seed producer.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Plant materials and sampling sites description
- METHOD DETAILS
  - Genome size estimation, sequencing, and assembly
  - Pseudo-chromosomes assignment by using Hi-C
  - Genome annotation

Figure 3. Population structure analysis of Brazil nut

Analysis of population structure using ADMIXTURE.

(A) Samples ordered by site and subsite. "TP" denotes locations with *terra preta*.

(B) Samples ordered by site and age. "Old" suffix denotes individuals with tree diameter at breast height (DBH) of at least 150 cm, "young" individuals with DBH of no more than 80 cm. Individuals with intermediate DBH are excluded, leaving 99 out of 270 samples for the present analysis (see also Figures S4–S6).

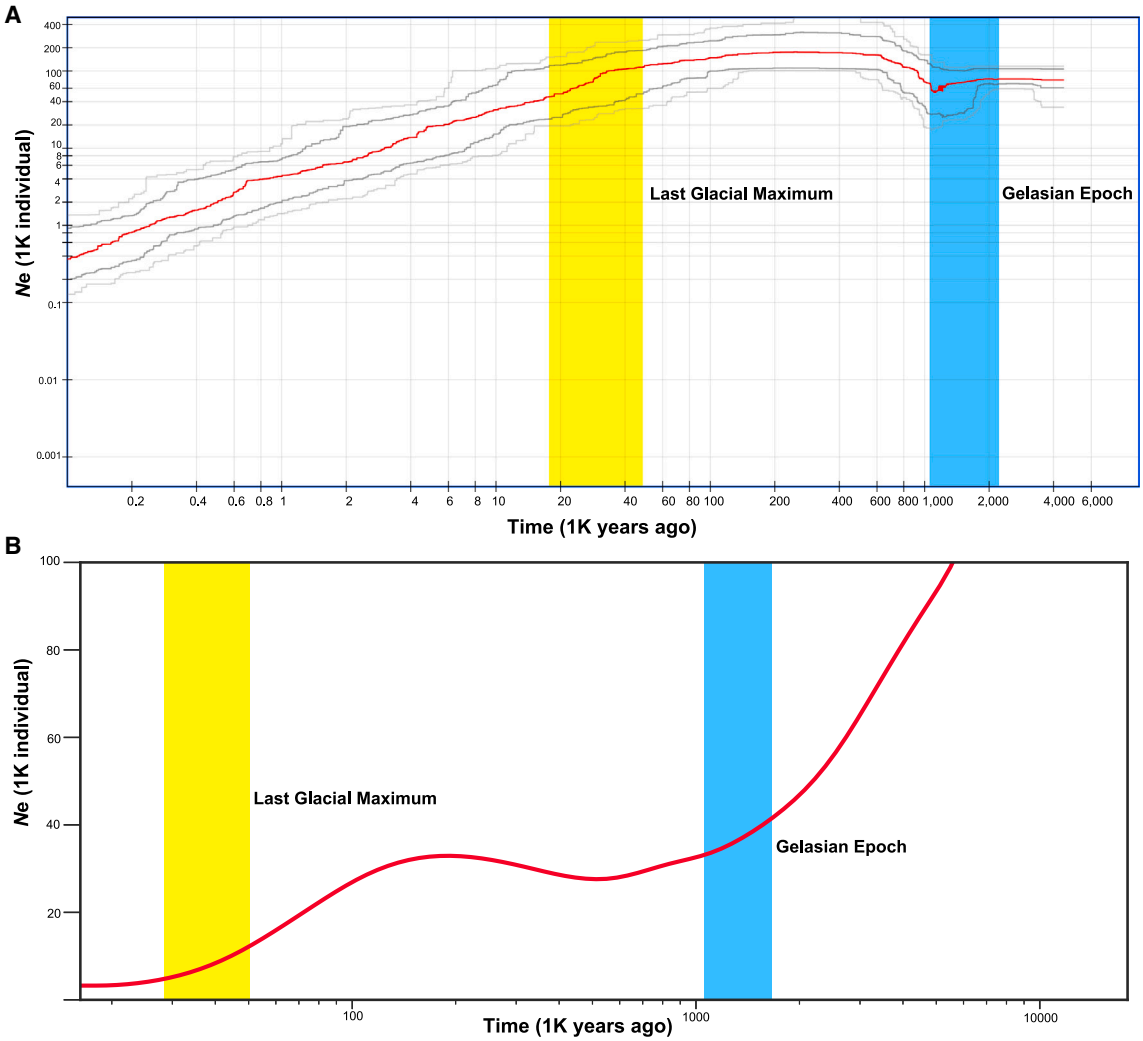


Figure 4. Demographic history and population bottleneck of Brazil nut

(A) The stairway plot method indicates a bottleneck during the Gelasian epoch (blue shading), which is notorious for climate upheaval, including many glacial cycles. After a period of stability, beginning just before the last glacial maximum (LGM, yellow shading), the effective population size N_e continuously declined. The light gray line indicates the 0.01 confidence intervals, and the dark gray line indicates the 0.05 confidence intervals.

(B) A broadly similar picture is seen when using SMC++ to infer N_e.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.12.023.

Received: May 14, 2024
Revised: November 4, 2024
Accepted: December 9, 2024
Published: January 24, 2025

REFERENCES

1. Flores, B.M., Montoya, E., Sakschewski, B., Nascimento, N., Staal, A., Betts, R.A., Levis, C., Lapola, D.M., Esquível-Muelbert, A., Jakovac, C., et al. (2024). Critical transitions in the Amazon forest system. Nature *626*, 555–564. https://doi.org/10.1038/s41586-023-06970-0.

2. Ter Steege, H., Prado, P.I., Lima, R.A.F., Pos, E., de Souza Coelho, L., de Andrade Lima Filho, D., Salomão, R.P., Amaral, I.L., de Almeida Matos, F.D., Castilho, C.V., et al. (2020). Biased-corrected richness estimates for the Amazonian tree flora. Sci. Rep. *10*, 10130. https://doi.org/10.1038/s41598-020-66686-3.

3. Levis, C., Costa, F.R., Bongers, F., Peña-Claros, M., Clement, C.R., Junqueira, A.B., Neves, E.G., Tamanaha, E.K., Figueiredo, F.O., Salomão, R.P., et al. (2017). Persistent effects of pre-Columbian plant domestication on Amazonian forest composition. Science *355*, 925–931. https://doi.org/10.1126/science.aal0157.

4. Coelho, S.D., Levis, C., Baccaro, F.B., Figueiredo, F.O.G., Pinassi Antunes, A., Ter Steege, H., Peña-Claros, M., Clement, C.R., and Schietti, J. (2021). Eighty-four per cent of all Amazonian arboreal plant individuals are useful to humans. PLoS One *16*, e0257875. https://doi.org/10.1371/journal.pone.0257875.

5. Clement, C.R., Casas, A., Parra-Rondinel, F.A., Levis, C., Peroni, N., Hanazaki, N., Cortés-Zárraga, L., Rangel-Landa, S., Alves, R.P., Ferreira, M.J., et al. (2021). Disentangling domestication from food production systems in the Neotropics. Quaternary *4*, 4. https://doi.org/10.3390/quat4010004.

6. Prümers, H., Betancourt, C.J., Iriarte, J., Robinson, M., and Schaich, M. (2022). Lidar reveals pre-Hispanic low-density urbanism in the Bolivian Amazon. Nature *606*, 325–328. https://doi.org/10.1038/s41586-022-04780-4.

7. Fauset, S., Johnson, M.O., Gloor, M., Baker, T.R., Monteagudo M, A., Brienen, R.J.W., Feldpausch, T.R., Lopez-Gonzalez, G., Malhi, Y., ter Steege, H., et al. (2015). Hyperdominance in Amazonian forest carbon cycling. Nat. Commun. *6*, 6857. https://doi.org/10.1038/ncomms7857.

8. Shepard, G.H., and Ramirez, H. (2011). "Made in Brazil": Human Dispersal of the Brazil Nut (*Bertholletia excelsa*, Lecythidaceae) in Ancient Amazonia. Econ. Bot *65*, 44–65. https://doi.org/10.1007/s12231-011-9151-6.

9. Martins, K., Santos, R.d.S.O.d., Campos, T.d., and Wadt, L.H.d.O. (2018). Pollen and seed dispersal of Brazil nut trees in the southwestern Brazilian Amazon. Acta Amaz. *48*, 217–223. https://doi.org/10.1590/1809-4392201800021.

10. Scoles, R., and Gribel, R. (2021). Growth and survival over ten years of Brazil-nut trees planted in three anthropogenic habitats in northern Amazonia. Acta Amaz. *51*, 20–29. https://doi.org/10.1590/1809-4392202001462.

11. Andrade, V.L.C., Flores, B.M., Levis, C., Clement, C.R., Roberts, P., and Schongart, J. (2019). Growth rings of Brazil nut trees (*Bertholletia excelsa*) as a living record of historical human disturbance in Central Amazonia. PLoS One *14*, e0214128. https://doi.org/10.1371/journal.pone.0214128.

12. Sujii, P.S., Martins, K., Wadt, L.H.d.O., Azevedo, V.C.R., and Solferini, V.N. (2015). Genetic structure of Bertholletia excelsa populations from the Amazon at different spatial scales. Conserv. Genet. *16*, 955–964. https://doi.org/10.1007/s10592-015-0714-4.

13. Gribel, R., Lemes, M.R., Bernardes, L.G., Pinto, A.E., and Sheppard, G. (2007). Phylogeography of Brazil-nut tree (*Bertholletia excelsa*, Lecythidaceae): evidence of human influence on the species distribution. Anais do Encontro Anual da" Association for Tropical Biology and Evolution"(ATBC).

14. Thomas, E., Alcázar Caicedo, C., McMichael, C.H., Corvera, R., and Loo, J. (2015). Uncovering spatial patterns in the natural and human history of Brazil nut (*Bertholletia excelsa*) across the Amazon Basin. J. Biogeogr. *42*, 1367–1382. https://doi.org/10.1111/jbi.12540.

15. Mori, S.A., and Prance, G.T. (1990). Taxonomy, ecology, and economic botany of the Brazil nut (Bertholletia excelsa Humb. & Bonpl.: Lecythidaceae). Adv. Econ. Bot. *8*, 130–150. https://www.jstor.org/stable/43927571.

16. Clement, C.R., de Cristo-Araújo, M., Coppens D'Eeckenbrugge, G., Alves Pereira, A., and Picanço-Rodrigues, D. (2010). Origin and domestication of native Amazonian crops. Diversity *2*, 72–106. https://doi.org/10.3390/d2010072.

17. Verica, J.A., and He, Z.-H. (2002). The cell wall-associated kinase (WAK) and WAK-like kinase gene family. Plant Physiol. *129*, 455–459. https://doi.org/10.1104/pp.011028.

18. Gharabli, H., Della Gala, V., and Welner, D.H. (2023). The function of UDP-glycosyltransferases in plants and their possible use in crop protection. Biotechnol. Adv. *67*, 108182. https://doi.org/10.1016/j.biotechadv.2023.108182.

19. de Abreu-Neto, J.B., Turchetto-Zolet, A.C., de Oliveira, L.F.V., Zanettini, M.H.B., and Margis-Pinheiro, M. (2013). Heavy metal-associated isoprenylated plant protein (HIPP): characterization of a family of proteins exclusive to plants. FEBS J. *280*, 1604–1616. https://doi.org/10.1111/febs.12159.

20. Hansen, C.C., Nelson, D.R., Møller, B.L., and Werck-Reichhart, D. (2021). Plant cytochrome P450 plasticity and evolution. Mol. Plant *14*, 1244–1265. https://doi.org/10.1016/j.molp.2021.06.028.

21. Zhang, X.T., Chen, S., Shi, L.Q., Gong, D.P., Zhang, S.C., Zhao, Q., Zhan, D.L., Vasseur, L., Wang, Y.B., Yu, J.X., et al. (2021). Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. Nat. Genet. *53*, 1250–1259. https://doi.org/10.1038/s41588-021-00895-y.

22. Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155. https://doi.org/10.1126/science.290.5494.1151.

23. Silveira, M.I.d., Rodrigues, M.C.L., Oliveira, E.R.d., and Losier, L.-M. (2008). Seqüência Cronológica de Ocupação na Área do Salobo (Pará). Revista de Arqueologia *21*, 61–84. http://repositorio.museu-goeldi.br/handle/mgoeldi/287.

24. Lopes, R.C.d.A. (2018). A Tradição Polícroma da Amazônia no contexto do médio rio Solimões (AM). PhD thesis (Universidade Federal de Sergipe).

25. Hu, G., Feng, J., Xiang, X., Wang, J., Salojärvi, J., Liu, C., Wu, Z., Zhang, J., Liang, X., Jiang, Z., et al. (2022). Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. Nat. Genet. *54*, 73–83. https://doi.org/10.1038/s41588-021-00971-3.

26. Liu, X., and Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. Nat. Genet. *47*, 555–559. https://doi.org/10.1038/ng.3254.

27. Huntingford, C., Zelazowski, P., Galbraith, D., Mercado, L.M., Sitch, S., Fisher, R., Lomas, M., Walker, A.P., Jones, C.D., Booth, B.B.B., et al. (2013). Simulated resilience of tropical rainforests to CO2-induced climate change. Nature Geosci. *6*, 268–273. https://doi.org/10.1038/ngeo1741.

28. Levis, C., Flores, B.M., Moreira, P.A., Luize, B.G., Alves, R.P., Franco-Moraes, J., Lins, J., Konings, E., Peña-Claros, M., Bongers, F., et al. (2018). How People Domesticated Amazonian Forests. Front. Ecol. Evol. *5*, 171. https://doi.org/10.3389/fevo.2017.00171.

29. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics *33*, 2202–2204. https://doi.org/10.1093/bioinformatics/btx153.

30. Workman, R., Timp, W., Fedak, R., Kilburn, D., Hao, S., and Liu, K. (2018). High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. protocols.io. 10.17504/protocols.io.4vbgw2n.

31. Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054. https://doi.org/10.1038/nmeth.4035.

32. Cheng, H.Y., Concepcion, G.T., Feng, X.W., Zhang, H.W., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

33. Chakraborty, M., Baldwin-Brown, J.G., Long, A.D., and Emerson, J.J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. *44*, e147. https://doi.org/10.1093/nar/gkw654.

34. Guan, D., McCarthy, S.A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics *36*, 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025.

35. Lam, K.K., LaButti, K., Khalak, A., and Tse, D. (2015). FinisherSC: a repeat-aware tool for upgrading *de novo* assembly using long reads. Bioinformatics *31*, 3207–3209. https://doi.org/10.1093/bioinformatics/btv280.

36. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. *3*, 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

37. Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science *356*, 92–95. https://doi.org/10.1126/science.aal3327.

38. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. USA *117*, 9451–9457. https://doi.org/10.1073/pnas.1921046117.

39. Tempel, S. (2012). Using and understanding RepeatMasker. In Mobile Genetic Elements: Protocols And Genomic Applications (Humana Press), pp. 29–51. https://doi.org/10.1007/978-1-61779-603-6_2.

40. Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. *35*, W265–W268. https://doi.org/10.1093/nar/gkm286.

41. Ou, S., and Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol. *176*, 1410–1422. https://doi.org/10.1104/pp.17.01310.

42. Yan, H., Bombarely, A., and Li, S. (2020). DeepTE: a computational method for *de novo* classification of transposons with convolutional neural network. Bioinformatics *36*, 4269–4275. https://doi.org/10.1093/bioinformatics/btaa519.

43. Lomsadze, A., Burns, P.D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. *42*, e119. https://doi.org/10.1093/nar/gku557.

44. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics *24*, 637–644. https://doi.org/10.1093/bioinformatics/btn013.

45. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512. https://doi.org/10.1038/nprot.2013.084.

46. Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666. https://doi.org/10.1093/nar/gkg770.

47. Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom. Bioinform. *2*, lqaa026. https://doi.org/10.1093/nargab/lqaa026.

48. Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7.

49. Dunn, N.A., Unni, D.R., Diesh, C., Munoz-Torres, M., Harris, N.L., Yao, E., Rasche, H., Holmes, I.H., Elsik, C.G., and Lewis, S.E. (2019). Apollo: democratizing genome annotation. PLoS Comput. Biol. *15*, e1006790. https://doi.org/10.1371/journal.pcbi.1006790.

50. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

51. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

52. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W.Z., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

53. Wang, J., Li, L., Qi, H., Du, X., and Zhang, G. (2016). RestrictionDigest: A powerful Perl module for simulating genomic restriction digests. Electron. J. Biotechnol. *21*, 36–42. https://doi.org/10.1016/j.ejbt.2016.02.003.

54. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

55. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

56. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

57. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

58. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; *iso-2*; *iso-3*. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

59. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.

60. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

61. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. *8*, e1002967. https://doi.org/10.1371/journal.pgen.1002967.

62. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M., and Yang, T.-L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics *35*, 1786–1788. https://doi.org/10.1093/bioinformatics/bty875.

63. Tian, Y., Lu, G., and Zhai, G. (2024). SMC++: Masked Learning of Unsupervised Video Semantic Compression. Preprint at arXiv. https://doi.org/10.48550/arXiv.2406.04765.

64. Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., and Liu, J. (2022). WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol. Plant *15*, 1841–1851. https://doi.org/10.1016/j.molp.2022.10.018.

65. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 157. https://doi.org/10.1186/s13059-015-0721-2.

66. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment, Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

67. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

68. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 2461. https://doi.org/10.1093/molbev/msaa131.

69. Yang, Z.H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. *13*, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555.

70. De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269–1271. https://doi.org/10.1093/bioinformatics/btl097.

71. MMA (2006). Plano de manejo para uso múltiplo da Floresta Nacional do Tapirapé-Aquiri (Ministério Do Meio Ambiente).

72. Silveira, M.I., Kern, D.C., Berredo, J.F., Costa, J.A., and da Costa, M.L. (2016). Um milênio de ocupações arqueológicas com manchas de terra preta em floresta na região de Carajás, Pará, Brasil. Boletim do Museu Paraense Emílio Goeldi-Ciências Naturais *11*, 11–31. https://doi.org/10.46357/bcnaturais.v11i1.457.

73. Borges, S.H. (2004). Species poor but distinct: bird assemblages in white sand vegetation in Jau National Park, Brazilian Amazon. Ibis *146*, 114–124. https://doi.org/10.1111/j.1474-919X.2004.00230.x.

74. Leonardi, V. (2013). Os Historiadores E Os Rios: Natureza E Ruína Na Amazônia Brasileira (Editora UnB).

75. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011.

76. Lang, P.L.M., Weiß, C.L., Kersten, S., Latorre, S.M., Nagel, S., Nickel, B., Meyer, M., and Burbano, H.A. (2020). Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. Mol. Ecol. Resour. *20*, 1228–1247. https://doi.org/10.1111/1755-0998.13168.

77. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. *34*, 1812–1819. https://doi.org/10.1093/molbev/msx116.

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Bertholletia excelsa leaves	A individual preserved in 'Parque das Castanheiras', Manaus, Brazil (3°04'48.2"S 59°56'05.0"W)	N/A
270 Brazil Nut wild individuals	samples from three protected areas of the Brazilian Amazon during 2018-2019	N/A
Chemicals, peptides, and recombinant proteins		
SPRIselect™ beads	Beckman Coulter	Cat# B23317
37% Formaldehyde solution	Sigma-Aldrich	Cat# F8775
1X PBS, pH 7.4	Thermo Fisher	Cat# 10010023
DNase / RNase- free distilled water	Invitrogen	Cat# 10977015
ApeKI	New England Biolabs	Cat# R0643S
BfaI	New England Biolabs	Cat# R0568S
Critical commercial assays		
SMRTbell Template Prep Kit 1.0	Pacific Biosciences	Cat# 100-259-100
Sequel SMRT cells 1M v2	Pacific Biosciences	Cat# 101-008-000
NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina®	New England Biolabs	Cat# E7490
AMPure PB Beads	Pacific Biosciences	Cat# 100-256-900
Qubit 3.0 Fluorometer	Life Technologies	Cat# Q33216
Dovetail™Omni-C® Kit	Dovetail, USA	N/A
Deposited data		
Bertholletia excelsa genome assembly	National Center for Biotechnology Information (NCBI)	NCBI: RJNA1099931
Bertholletia excelsa genome CCS sequence	National Center for Biotechnology Information (NCBI)	NCBI: RJNA1099931
270 *Bertholletia excelsa* dd RAD-seq sequence	National Center for Biotechnology Information (NCBI)	NCBI: RJNA1099931
RNA-seq data of *Bertholletia excelsa* leaf	National Center for Biotechnology Information (NCBI)	NCBI: RJNA1099931
Software and algorithms		
Jellyfish	Vurture et al.[29]	https://www.cbcb.umd.edu/software/jellyfish/
GenomeScope	Workman et al.[30]	https://github.com/schatzlab/genomescope
FALCON-Unzip	Chin et al.[31]	https://github.com/PacificBiosciences/FALCON_unzip
Hifiasm	Cheng et al.[32]	https://i5k.nal.usda.gov/bio_data/1394330
Quickmerge	Chakraborty et al.[33]	https://github.com/mahulchak/quickmerge
Purge_dups	Guan et al.[34]	https://github.com/dfguan/purge_dups
FinisherSC	Lam et al.[35]	http://kakitone.github.io/finishingTool/
Juicebox	Durand et al.[36]	https://hpc.ilri.cgiar.org/juicebox-software
3D-DNA pipeline	Dudchenko et al.[37]	https://github.com/aidenlab/3d-dna
RepeatModeler2	Flynn et al.[38]	https://github.com/Dfam-consortium/RepeatModeler/blob/master/RELEASE-NOTES
RepeatMasker	Tempel et al.[39]	http://www.repeatmasker.org/
LTR_FINDER	Xu et al.[40]	https://github.com/oushujun/LTR_FINDER_parallel
LTR_retriever	Ou et al.[41]	https://github.com/oushujun/LTR_retriever

(*Continued on next page*)

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deep TE	Yan et al.[42]	https://github.com/LiLabAtVT/DeepTE
GeneMark-ES Suite	Lomsadze et al.[43]	https://genemark.bme.gatech.edu/
Augustus	Stanke et al.[44]	https://github.com/nextgenusfs/augustus
Trinity	Haas et al.[45]	https://github.com/trinityrnaseq/trinityrnaseq
PASA	Haas et al.[46]	https://github.com/PASApipeline/PASApipeline
ProtHint	Bruna et al.[47]	https://github.com/gatech-genemark/ProtHint
EVidenceModeler	Haas et al.[48]	https://github.com/EVidenceModeler/EVidenceModeler
Apollo	Dunn et al.[49]	https://github.com/ApolloAuto/apollo
BUSCO	Simão et al.[50]	https://busco.ezlab.org/
Diamond	Buchfink et al.[51]	https://github.com/drostlab/rdiamond
InterProScan	Jones et al.[52]	https://github.com/ebi-pf-team/interproscan
RestrictionDigest	Wang et al.[53]	https://github.com/JINPENG-WANG/RestrictionDigest
Fastp	Chen et al.[54]	https://github.com/OpenGene/fastp
BWA	Li et al.[55]	http://bio-bwa.sourceforge.net/
GATK	McKenna et al.[56]	https://gatk.broadinstitute.org/hc
VCFtools	Danecek et al.[57]	https://github.com/vcftools/vcftools
SnpEff	Cingolani et al.[58]	https://pcingola.github.io/SnpEff/
ADMIXTURE	Alexander et al.[59]	https://dalexander.github.io/admixture/
Plink	Purcell et al.[60]	https://github.com/insilico/plink
TreeMix	Pickrell et al.[61]	https://github.com/carolindahms/TreeMix
PopLDdecay	Zhang et al.[62]	https://github.com/BGI-shenzhen/PopLDdecay
easySFS	N/A	https://github.com/isaacovercast/easySFS
SMC++	Tian et al.[63]	https://github.com/popgenmethods/smcpp
WGDI	Sun et al.[64]	https://github.com/SunPengChuan/wgdi
OrthoFinder	Emms et al.[65]	https://github.com/davidemms/OrthoFinder
MAFFT	Katoh et al.[66]	https://github.com/GSLBiotech/mafft
trimal	Capella-Gutierrez et al.[67]	https://github.com/inab/trimal
IQ-tree	Minh et al.[68]	https://github.com/iqtree/iqtree2
PAML	Yang et al.[69]	https://github.com/abacus-gene/paml
CAFÉ	De Bie et al.[70]	https://github.com/hahnlab/CAFE
Other		
Tair	N/A	https://www.arabidopsis.org/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Plant materials and sampling sites description

The Brazil Nut individual used for PacBio *de novo* genome assembly was preserved in the 'Parque das Castanheiras', Manaus, Brazil (3°04'48.2"S 59°56'05.0"W). For diversity analyses, a total of 270 Brazil Nut wild individuals were sampled from three protected areas of the Brazilian Amazon during 2018-2019. (Table S7).

The first site was the Tapirapé-Aquiri National Forest ("TAP") in eastern Amazonia. It is dominated by submontane rainforest and alluvial rainforest.[71] Most archaeological sites in the region occur on terraces in the meanders of the local creeks.[72] The trees from the stand TAP_S2 are growing on patches of Amazonian Dark Earth, *terra preta*, that extend over the region in patches between 600 x 400 m and 200 x 150 m, concomitantly or not, with depths varying between 60 cm and 1 m, with archaeological dates going back to the eighth century CE and material culture discoveries including ceramic fragments, lithic material, and carbonised seeds.[72] The trees growing at TAP_S1 are at the edge of the park area, and are visited sporadically by local people from the surrounding community, but there is no archaeological evidence in the area.

The Tefé National Forest ("TEF") site is composed of dense ombrophilous forest, mainly *terra firme* habitats but also igapó flooded forests. The Brazil Nut trees at both the TEF_S1 and TEF_S2 subsites are part of the agroforestry system of contemporary communities, and forest management is connected to the harvesting of Brazil Nut seeds, as well as practices such as the cutting of lianas to

Please cite this article in press as: Wang et al., Long-term human influence on the demography and genetic diversity of the hyperdominant *Bertholletia excelsa* in the Amazon Basin, Current Biology (2024), https://doi.org/10.1016/j.cub.2024.12.023

release tree growth, the planting of trees,[8] the cultivation of cassava (*Manihot esculenta*), and the collection of other non-timber forest resources such as Açaí (*Euterpe precatória*) and Pupunha (*Bactris gasipaes*). Trees at the Ponta da Castanha archaeological site (TEF_S2) are growing on *terra preta*, and ceramic fragments have been found.[24] Although the trees in TEF_S1 apparently do not directly overlap an archaeological site, the region has many indications that it was extensively occupied in the past, such as ceramics, funerary urns, and *terra preta* soils. Finally, the Jaú National Park (JAU), which is the third largest tropical forest park in the world, with vast areas of *terra firme* and campinarana and igapó forests,[73] is located on the west bank of the lower Rio Negro. No detailed archaeological surveys have been carried out in the areas where the trees were sampled. JAU_S1 is an old Brazil Nut stand, located in an area that was documented as Indigenous territory in the 18th century and later became a site for rubber exploration.[74] The site was gradually abandoned over the decades since the second rubber boom ended. We observed *terra preta* near JAU_S1. The Brazil Nut trees in JAU_S2 occur among agricultural fields and home gardens in the Cachoeira community, which settled there during the first and second rubber boom. No archaeological evidence was found at JAU_S2.

METHOD DETAILS

Genome size estimation, sequencing, and assembly

Jellyfish was used to calculate k-mers,[75] and GenomeScope was used to estimate genome size.[29] Young leaves were flash frozen on dry ice and transferred to the Instituto Nacional de Pesquisas da Amazonia (INPA) in Manaus. For PacBio (Menlo Park, CA, USA) circular consensus sequencing (CCS), high molecular weight DNA was extracted according to a published protocol.[30] The BluePippin (Sage Science, Beverly, MA, USA) system was used for size selection. SMRTbell libraries (16–20 kb) were constructed according to manufacturer's instructions. Libraries were sequenced on a PacBio Sequel II instrument, generating 29.1 Gb (estimated 50x genome coverage) of high fidelity (HiFi) reads. PacBio HiFi reads were assembled with FALCON/FALCON-Unzip[31] and Hifiasm[32] with parameters optimised for diploid species. We used the FALCON-phased version to combine the two draft genome assemblies with the quickmerge methodology[33] using the FALCON assembly as the query and the Hifiasm assembly as the reference (options: hco 5.0 c 1.5 l 19000000 ml 10000). We purged the Haplotigs of the merged assembly with Purge_dups[34] then used FinisherSC[35] to further increase the contiguity of the assembly (Table S1).

Pseudo-chromosomes assignment by using Hi-C

A Hi-C library for chromatin contact information and scaffolding was constructed with the Omni-C Library Preparation Kit (Dovetail Genomics, Scotts Valley, CA, USA) according to the manufacturer's instructions (https://dovetailgenomics.com/products/omni-c-product-page/), and sequenced on a HiSeq 3000 instrument (Illumina, San Diego, CA, USA) to generate 55 Gb of 2x150 bp paired-end reads. The Hi-C reads were transformed into chromatin contact maps with Juicer[36] (v.1.6.2), and the information was used to scaffold the non-redundant PacBio contigs with the 3D-DNA pipeline[37] with a haploid model. This procedure yielded 17 pseudo-chromosomes.

Genome annotation

TE annotation: We used RepeatModeler2[38] (https://github.com/Dfam-consortium/RepeatModeler) to find de novo repeats in the genome. The repeat library was imported to RepeatMasker[39] (version 4.1.1) to identify and cluster repetitive elements that were masked in the assembly with the sensitive mode. LTR-TEs were predicted using LTR_FINDER[40] (v.1.07) with parameters '-D 20000 -d 1000 -L 700 -l 100 -p 20 -C -M 0.9'. Results from the above annotation steps were integrated using LTR_retriever (v.2.6).[41] The results from each software package were integrated to generate a comprehensive, non-redundant TE library, which was used as input to produce a non-redundant TE annotation. DeepTE was used to further classify unclassified TEs.[42] Gene annotation: We first used BRAKER2(v 2.1.5), which combines GeneMark-ET[43] and AUGUSTUS[44] to integrate evidence from orthologous proteins, transcriptomes and *ab initio* gene prediction. Leaf RNA-seq reads was imported into the Trinity *de novo* assembly and genome-guided assembly pipelines with default parameters,[45] and transcripts were imported into PASA.[46] ProtHint[47] was also used for predicting and scoring hints based on the UniProt plant protein database (accessed on 1 February 2021). EVidenceModeler-1.1.1[48] was used to combine the different predictions into weighted consensus gene structures. With the combined result, PASA was used to update the EVidenceModeler consensus predictions, adding UTR annotations and models for alternatively spliced isoforms. Limited manual refinement was done with Apollo.[49] BUSCO[50] (version 5.1.0) was used to evaluate annotation completeness (Table S3). Out of 2,326 eudicot conserved genes, 2,235 (96.1%) were annotated in the Brazil Nut assembly, of which 2114 (90.9%) were complete single-copy BUSCO genes. Gene functions were assigned by searching the predicted proteins against public databases with DIAMOND[51] with e-value < 1e^{-5}, including the UniProt and the KEGG (Kyoto Encyclopedia of Genes and Genomes) databases. We aligned the proteins against the InterProdatabase using InterProScan[52] to identify protein domains and transmembrane helices and to assign gene ontology (GO) terms.

dd RAD-seq sequencing and SNP calling

To generate diversity data, we simulated ddRAD-seq fragments *in silico* using RestrictionDigest,[53] then chose the restriction enzymes ApeKI and BfaI, because a digest of genomic DNA with this combination was predicted to produce the most fragments in the 300-500 bp range. The ddRAD-seq library was constructed as described,[76] with libraries sequenced on a HiSeq 3000 instrument with 2x150 bp paired end reads. For variant calling, adaptors and low-quality bases (Q < 30) were trimmed from raw reads using

fastp,[54] and the filtered clean reads were aligned against the genome assembly BWA[55] with default parameters (Table S5). For diversity analysis, we used only SNPs and small indels up to 10 bp identified using GATK4[56] following the best practices workflow (https://gatk.broadinstitute.org/). HaplotypeCaller and GenotypeGVCFs were used to call variants from all samples. A hard filter was used with criteria 'QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0'. The filtered VCF file was subjected to quality control by VCFtools[57] with the following criteria ' –max-missing 0.9 –maf 0.05 –mac 3 –minDP 3 –minQ 30 '. The SnpEff[58] program was used to annotate SNPs and large-effect SNPs with modification of the start or stop codons, and alternative splice sites were extracted for further analysis (Table S6).

Population structure and demographic history

ADMIXTURE[59] was used to infer ancestral populations among Brazil Nut accessions with different k values (from 1 to 10), using 2,000 bootstraps. k = 9 had the smallest cross-validation error. For the subset of young and old individuals, k = 3 had the smallest cross-validation error. PLINK1.9[60] and VCFtools 0.1.16[57] were used to calculate population divergence statistics, including nucleotide diversity (π) and genetic differentiation (F_{st}). TreeMix[61] was used to identify gene flow among populations. Linkage disequilibrium decay was calculated using PopLDdecay[62] (version 3.31) with default parameters. The site-frequency spectrum (SFS) was calculated using easySFS (https://github.com/isaacovercast/easySFS) with parameters '-f -a –proj 520'. We took the maximum-likelihood estimate of the folded SFS as input to estimate population demography history with the stairway plot algorithm[26] using 200 bootstraps, a generation time of 18 years, and a mutation rate per generation per site of 6.5×10^{-9}. SMC++ was also used to estimate population demography history with the same mutation rate and generation time and parameter -c 2000.[63] GO enrichment analyses of species-specific genes were carried out with the OmicShare platform (www.omicshare.com/tools). Significance of enrichment was determined using Fisher's exact test, with *P* values adjusted using Benjamini–Hochberg multiple-hypothesis-testing correction.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome evolution analyses

We collected the longest open reading frame for each gene and identified self-to-self DIAMOND[51] hits of proteins at least 30 aa long and e-value < 1e-5. The results were imported to WGDI[64] to identify syntenic blocks and substitutions per synonymous site (Ks). OrthoFinder[65] (version 2.3.3) was used to classify orthogroups of proteins from Brazil Nut and five other representative species, *Vitis vinifera* (V2.0), *Citrus sinensis* (Csi_valencia_1.0), *Populus trichocarpa* (V3.0), *Juglans regia* (V2.0), and *Arabidopsis thaliana* (TAIR10). A tree was constructed based on 1,331 single-copy genes of the six species. We used MAFFT[66] to construct protein alignments for each single-copy gene family and removed gaps from the alignments using trimal.[67] IQ-tree 2[68] was used to construct a maximum likelihood tree with a Best-fit model of JTT + F + R10 and 1000 bootstrap replicates. The divergence time of each node in the phylogenetic tree was estimated based on the JC69 model in the MCMCTree program from PAML.[69] The MCMCTree was run for 1,000,000 generations, with a burn-in 10,000 iterations to a stable state. The divergence time between *J. regia* and *A. thaliana* (108 Mya)[77] was used for calibration. For gene family expansion and contraction analysis, the ancestral gene content of each cluster at each node was investigated with CAFE 3.1.[70] Based on the phylogeny and gene numbers per orthogroup in each species, the gene family expansions/contractions at each branch were determined with P-value < 0.01. Expansion and contraction events were estimated by comparison to the RCAs (recent common ancestors) of the species.