DOI: 10.5281/zenodo.10208446

Past domestication of T. cacao in Latin America revealed by paleo-genomics and methylxanthines analysis

Claire Lanaud[1,2], Hélène Vignes[1,2], José Utge[3], Gilles Valette[4], Bénédicte Rhoné[1,2] , Mariella Garcia Caputi[5], Natalia Angarita[6], Olivier Fouet[1,2], Nilesh Gaikwad[7], Terry Powis[8], Ann Ciphers[9], Francisco Valdez[10], Sonia Zarillo[11], S. Quirino Olivera Nunez[12], Camilia Speller[11], Francisco Romano[6], Fred Fr. Valdez[13], Scott Raymond[14], Rey Gaston Loor Solorzano[15], Xavier Argout[1,2]

[(1)] Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR AGAP Institut, Montpellier, France
[(2)] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France
[(3)] Museum National d'Histoire Naturelle (MNHN) - UMR 7206 Eco-anthropologie, Paris, France
[(4)] Institut des Biomolécules Max mousseron – (UMR IBMM), Université de Montpellier, Montpellier, France
[(5)] Museo Antropologico y de Arte Contemporaneo (MAAC), Guayaquil, Ecuador
[(6)] Museo Nacional de Colombia (MNC), Bogota, Colombia
[(7)] University of California, Davis, USA; Gaikwad Steroidomics Lab LLC, Davis, USA
[(8)] Department of Geography and Anthropology, Kennesaw State University Kennesaw, USA.
[(9)] Universidad Nacional Autónoma de México (UNAM), Mexico, Mexico
[(10)] Institut de Recherche pour le Développement (IRD), UMR 208 PALOC, MNHN-IRD, Paris, France
[(11)] Department of Anthropology, University of British Columbia, Vancouver, Canada.
[(12)] Asociación para la Investigación Científica de la Amazonía de Perú (ASICAMPE) - Perú
[(13)] The Univ. of Texas at Austin, Austin, USA
[(14)] Department of Anthropology and Archaeology, University of Calgary, Calgary, Canada
[(15)] Instituto Nacional de Investigaciones Agropecurias, INIAP, Ecuador

ABSTRACT

Understanding the diversity of a species and its history, in order to better exploit it, is a challenge that breeders often face when creating new varieties adapted to the current environment. Many questions still remain about the first phases of cocoa domestication in Latin America. Using a multidisciplinary approach, associating archaeologists with genomics and biochemistry scientists, we tried to retrace the history of the past domestication of cocoa trees, and their migrations in South America and Central America. Plant paleo-genomics is a new field of research, that traces the evolution of cultivated plants in response to human domestication and natural selection since very ancient times. It is based on the analysis of ancient DNA (aDNA) from archaeological remains and can provide direct access to plant varieties consumed several thousand years ago. The oldest traces of cocoa consumption date back 5,500 years and have been identified in the southern Ecuadorian Amazon. New studies, reported in this paper, have been carried out from 382 archaeological samples collected mainly all along the Pacific coast of Ecuador and Colombia as well as in Central America and from diverse civilizations, the oldest of which dates back to more than 5000 years before present. Analyzes have revealed traces of ancient cocoa DNA and methylxanthines (theobromine, theophylline and caffeine) in the ceramic residues of many pre-Columbian cultures. They show a wide use of cocoa throughout the Latin America continent and at very varied times which can go back more than 5000 years. The genetic origin of varieties consumed several thousand years ago could have sometimes been identified by comparison with a large collection of modern genetic resources representing the diversity of the *T. cacao* species. These results provide new keys to unraveling the past domestication of *T. cacao* in Latin and Central America.

Keywords: *T. cacao*, ancient DNA, paleo-genomics, methylxanthines, domestication

1. Introduction

Theobroma cacao L. is originated from tropical and humid regions of South America with hotspots of diversity observed in Upper Amazon basins (Thomas *et al.* 2012) close to the frontiers between Colombia and Ecuador (Fouet *et al.*, 2022). Its diversity was classified in 10 genetic groups (Motamayor et al., 2008) and with a new genetic group recently identified by Fouet et al., (2022) as the Caqueta group.

DOI: 10.5281/zenodo.10208446

The ancient domestication of *T. cacao* has been particularly documented in Central America where it was introduced through past human-mediated dispersal, and a large number of archaeological evidence has been found showing the economic, social and cultural importance of cocoa, among the Olmecs, Maya and Aztecs civilizations (Coe and Coe, 1996; Bletter and Daly, 2006) with oldest traces of cocoa, revealed by theobromine presence, found in Olmec pottery and dating back to 3900 years BP (Powis *et al.*, 2002; Powis *et al.*, 2011). It was hypothesized that The Criollo variety, a fine flavor aromatic variety was the unique variety cultivated in Central America after differentiation of *T. cacao* in 2 sub-species (Cuatrecasas, 1964) or introduction of plants in Central America from a cradle located in Amazon region (Cheesman, 1944; Cope, 1976). Cornejo et al. (2018) proposed that Criollo was originated from Curaray groups, and introduced in Central America from few individuals.

In South America, older traces of the cocoa use and domestication, dating from 5500 years before present (BP), have been revealed in the Southern Ecuadorian Amazon (Zarillo *et al.*, 2018) where, the Nacional variety, another fine flavor variety, originated. The modern Nacional variety is an hybrid population involving three main ancestors: Criollo, Nacional and Amelonado. It was supposed to have been domesticated by a first introduction of Nacional genotypes from South Ecuador in the pacific coast, followed by a more recent introduction of Trinitario types in the pacific coast from Venezuela one century ago (Bartley 2005).

However no other information on the use of cocoa in the rest of South America during pre-Columbian times is available.

To try to clarify the possible routes of ancient cocoa domestication, we used biochemical and paleo-genomic approaches to analyze food residues found in ceramics from 20 pre-Colombian cultures widespread mainly along the pacific coast of Ecuador and Colombia, and in Central America.

2. Material and methods

2.1 Archaeological items analyzed

We analyzed 382 ceramic items from 20 pre-Colombian cultures widespread in South America (Ecuador, Colombia, Peru), mainly in the pacific coast, and in Central America (Mexico, Belize, Panama). These cultures span about 5000 years (Fig. 1).

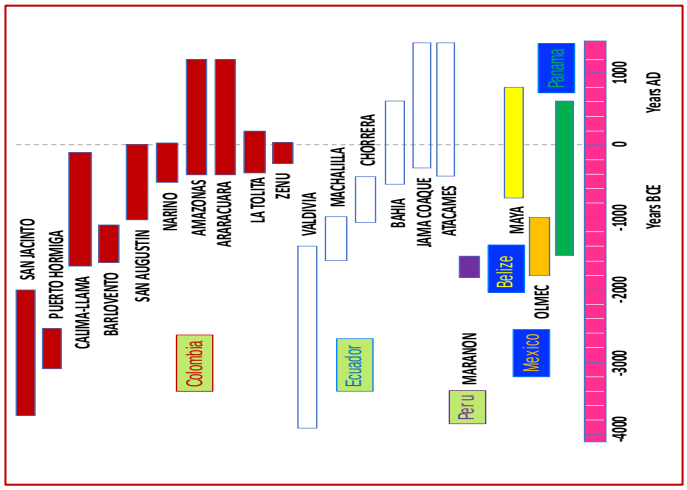

Figure 1. Associated cultural chronology for the studied archaeological items

DOI: 10.5281/zenodo.10208446

2.2 Modern T. cacao accessions taken as references

A collection of 76 *T. cacao* accessions were selected to represent the diversity of the 11 genetic groups of *T. cacao* species as identified by Fouet et al. (2022). The following groups are represented: Criollo, Caqueta (the new group identified in Colombia), Curaray, Nacional Purus, Contamana, Marañon, Iquitos, Nanay, Amelonado and Guiana. Five accessions from wild species: *Herrania nitida, T. gradiflorum, T. bicolor and T. speciosum,* were also added to this reference collection.

This reference collection was used to determine the putative origin of ancient DNA detected in the ceramic food residues.

2.3 Collection of food residues

The collection of ceramic food residues were made using rayon swabs (Copan-Italia) saturated with extraction buffer (0.1 M Tris, 0.45 M EDTA) and wiped against the inner walls of the ceramics. For Maya and Olmec samples, the interior surface of each ceramic vessel or sherd was lightly scraped using a new piece of fine-grained sandpaper, and the powder conserved for further analyses.

2.4 Methylxanthine analyses

Three methylxanthine components present in *T. cacao* were analysed: theobromine, theophylline and caffeine.

Biochemical analyses were carried as described in Zarrillo et al (2018) from one swab wiped against the inner walls of the ceramics. The presence of any methylxanthine was considered as positive when the methylxanthine amount was > 700 pg/swab.

2.5 Laboratory environment for ancient DNA (aDNA) analyses

Ancient DNA is very degraded. We took adequate precautions to prevent contamination by modern, undamaged DNA, which would have been preferentially amplified if present in the DNA extracts. All pre-PCR experiments were conducted under sterile conditions in a laboratory dedicated to ancient DNA analyses (MNHN laboratory located in the Musée de l'Homme in Paris) as shown in Fig. 2.

All post-PCR experiments were carried out at the CIRAD genomics laboratory (UMR AGAP Institute, Montpellier-France).

Figure 2. First steps of the experiments made in the sterile environment of the white room of the MNHN laboratory (Paris) to avoid modern DNA contaminations.

DOI: 10.5281/zenodo.10208446

2.6 Extraction of ancient DNA

Four swabs, wiped against the inner walls of the ceramics, were used for each aDNA extraction made. Extraction was made with the Quiagen DNeasy PowerLyzer PowerSoil Kit and according to the kit described procedure except for the binding step where the concentration of the C4 saline solution was increased.

Several negative samples were added to the analyses.

2.7 Library construction

Libraries construction were made according Meyer and Kircher (2010), modified by the P2GM, MNHN lab.

Most of the libraries were directly sequenced in pair ends on a Novaseq 6000 Illumina sequencer directly after library construction except for a part of them that were subjected to a DNA capture procedure at the beginning of the experimentations, according to Zarillo *et al.*, 2018.

2.8 Bioinformatic sequences treatment

Illumina paired-end reads were first cleaned for adapter removal using Cutadapt v3.4 and Trim Galore v 0.6.6 (Martin, 2011) and mapped on the *Theobroma cacao* whole genome sequence (V2) (Argout et al., 2017), after discarding reads shorter than 30 nucleotides. The pre-selected aDNA sequences, were then BLASTed against the NCBI NT database, and only sequences for which the first hit was *T. cacao* or *Herrania* were respectively considered for further analyses;

The presence of *T. cacao* (or *Herrania*) sequence was considered as positive in a sample when at least 5 *T. cacao* (or *Herrania*) first hit sequences were identified in a sample.

2.9 SNP identification and analyses

Sequences selected as *T. cacao* or *Herrania* ancient DNA sequences were blasted on the V2 cocoa genome (Argout et *al.*, 2017) and only sequences mapping in a unique location were retained for SNP analyses.

SNP from homologous modern sequences of the reference collection were extracted.

Structure analyses (Prichard *et al.,* 2000) and Nei genetic distances (Nei, 1978) were then carried out individually for each archaeological sample and the reference collection, using the common corresponding set of SNP markers.

3. Results

3.1 Methylxanthine analyses

Among the 334 archaeological samples analyzed from South America, 125 are positive for theobromine presence, 27 for theophylline presence and 134 for caffeine presence (Fig.3). 26,9% contain at least two methylxanthine among the three present in *T. cacao* seeds, generally theobromine and caffein, so probably reflecting a large use of *T. cacao*.The presence of theobromine as well as of caffeine, can be observed at least in one archaeological item of all culture studied.

DOI: 10.5281/zenodo.10208446

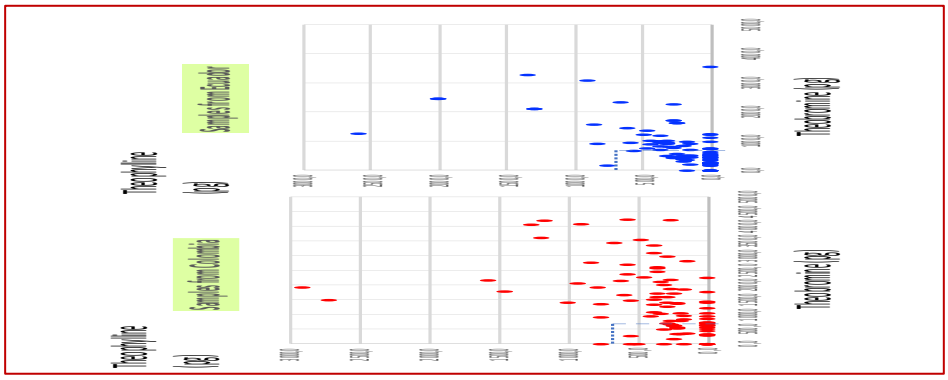

Figure 3. Distribution of theophylline and theobromine in Ecuadorian and Colombian archaeological samples

3.2 Ancient DNA analyses

Selection of ancient DNA sequences specific to *T. cacao/Herrania*

We constructed libraries from 171 selected ceramic items positive for methylxanthine presence for most of them.

The food residues contain DNA from a mix of several species of bacteria, fungus, plant, and often with many similarities between sequences from different species. So filters were necessary to identify specifically the *T. cacao* sequences.

After aDNA sequences mapping against the cocoa genome, a second filter was necessary to identify the sequences specific to *T. cacao*. A blast against data sequences from the NCBI NT international database was then carried out and a selection of sequences for which the similarities with *T. cacao* sequences are the highest ("first hit cacao") among all species sequences from the database was made.

A total of 617654 sequences, from all samples, could be mapped on the cocoa genome and verified for their ancient DNA characteristics (small size fragments, typical DNA miscoding at the ends of the aDNA fragments…).

After blast against the NCBI NT database, 19836 (3,21%) were identified as "first hit" *T. cacao* sequences, representing specific *T. cacao* sequences, and 1135 (0,18%) were identified as first hit *Herrania* sequences.

The negative controls were not positive for the presence of *T. cacao* or *Herrania* sequences.

As for methylxanthine presence *T. cacao* specific sequences could be observed in nearly all ceramic associated cultures from South America:

Among 123 archaeological items analyzed from 15 South American cultures (60 from Ecuador, 61 from Colombia, and 2 from Peru), the majority of them being positive for the presence of theobromine, 36 samples from Ecuador, 36 samples from Colombia, and 2 samples from Peru were positive for ancient *T. cacao* DNA.

The presence of *Herrania* sequences was also observed in 10 cultures from South America: with 13 samples from Ecuador and 20 samples from Colombia positive for the presence of *Herrania* DNA.

These results show a wide use of *T. cacao* and wild species in South America

Genetic origin of aDNA sequences

Only archaeological samples with a minimum of 20 SNP markers were considered for these analyses, which resulted in the possible analysis of 67 samples.

DOI: 10.5281/zenodo.10208446

Some examples of structure analyses are represented on Fig. 4.

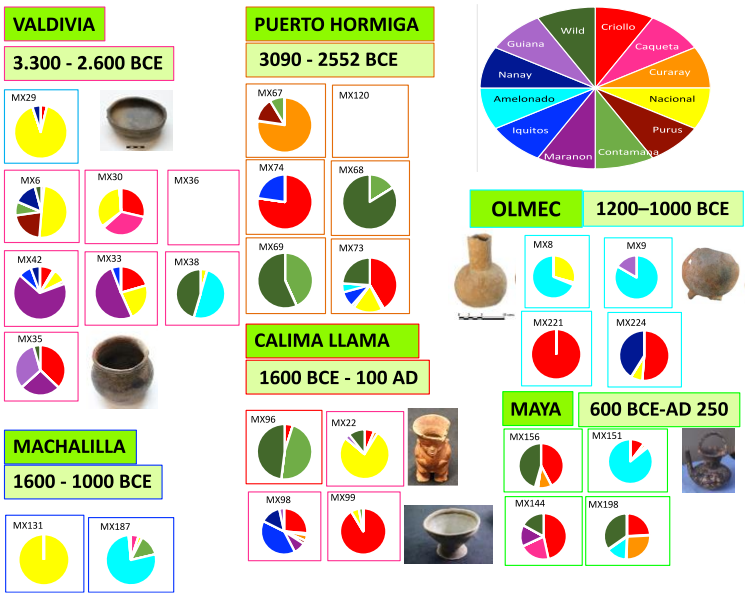

Figure 4. Examples of genotype membership proportion of *T. cacao* genetic groups revealed individually for each archaeological items in several studied cultures, and illustrating the diversity of origins observed for a same culture.

A great diversity of origins was observed among the samples from most of cultures. In the oldest ones, the Valdivia, we can found genotypes related to Marañon and Contamana genetic groups, suggesting interactions with this region, but also genotypes related to Nacional, Criollo, and Amelonado groups. We can find also the contribution of wild species in the cacao consumed by these populations. A similar diversity of origins is found in more recent Ecuadorian cultures as Machalilla and Chorrera, with items completely Nacional or hybrid forms involving Nacional with other origins including Amelonado.

In the cultures from Colombia, a similar diversity can be also observed, with some genotypes highly related to Criollo observed in Puerto Hormiga and Calima-Llama cultures.

These results indicate that the Criollo forms already existed for 5000 years in Colombia.

This diversity was also observed among the Olmec and Maya samples. Criollo was not the unique variety consumed, but contributions of Amelonado, Nacional or Iquitos genetic groups can be observed as early as in the Olmec period.

All these results obtained by structure analyses were also confirmed by Nei genetic distances analyses.

4. Discussion-Conclusion

Our results, based on methylxanthines and ancient *T. cacao* DNA presence in the ceramic food residues, have shown that cocoa has been widely used in the pacific coast of South America for over 5000 years. Their presence were found inside ceramic items used for rituals as well as for domestic activities. We found positive samples in nearly all cultures from Central and South America, except Narino culture for which only one sample was analysed, confirming the wide use of *T. cacao* in Central and South America.

Our analyses have also shown that the wild relative species were also largely used in a majority of South American cultures, and in Maya culture in Central America.

Structure analyses revealed a large diversity of origins, from Amazonia, generally present in all cultures of the South American coast, and in the oldest cultures as Valdivia and Puerto Hormiga cultures. This diversity probably reflects multiple interactions between cultures from the Pacific coast and Amazonian regions as already reported (Valdez, 2008; Lathrap, 2013; Clement et al., 2015, Calagan et al., 2022). A genetic mixing was observed in most of the archaeological samples, corresponding to diverse origins successively introduced and used, or to hybrid occurring naturally between cocoa trees from several origins cultivated at the same place. This diversity was favorable to the production of varieties adapted to their new environment.

In Central America, during the Olmec and Mayan occupation, Criollo was not the only cultivated variety found, contrary to the assumptions made previously, but genotypes related to Amelonado and Nacional were also observed, suggesting possible interactions with Ecuador from where all these genotypes were present and could have been introduced, and with Colombia where existed genotypes very close to the Criollo.

Thus, paleo-genomic approaches have strongly challenged our first hypotheses on the domestication of cocoa and help us to better understand and manage the genetic resources available. They show the complex history of the domestication of cocoa varieties that have relied on different human migratory waves since the beginning of the Holocene and on many trade exchanges within the Amazon and with the Pacific coast that have probably accompanied the multiple movements of cocoa trees. It is this human history that, along with natural selection due to the environment, has shaped the cocoa varieties currently cultivated.

Acknowledgements

We gratefully acknowledge Yannick Marie and Agnes RASTETTER who managed the sequencing activities of this work, which were mostly carried out on the iGenSeq core facility of ICM (Institut du Cerveau et de la Moelle épinière, Paris).

We thank the I-Site MUSE and Valrhona for their financial support of this project. This work, part of the MUSE Amazcacao project, was publicly funded through ANR (the French National Research Agency) under the "Investissement d'avenir" programme with the reference ANR-16-IDEX-0006.

We thank INPC (Instituto National de Patrimonio Cultural – Ecuador), MAAC (Museo Antropologico y de Arte Contemporaneo-Ecuador), ICANH (Instituto Colombiano de Antropologia e historia – Colombia), Museo Nacional de Colombia, for giving us access to archaeological samples conserved in museum reserves.

We thank CRC (Cocoa Research Center – Trinidad and Tobago) and CATIE (Centro Agronómico Tropical de Investigación y Enseñanza- Costa Rica) for providing a part of the modern *T. cacao* accessions studied in this work.

References

Argout, X. *et al.* (2017) The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. BMC Genomics 18, 730. doi.org/10.1186/s12864-017-4120-9

Bartley, B. G. D. (2005) The genetic diversity of cacao and its utilization. (CABI Publishing, Oxfordshire, UK.

Bletter N, Daly DC (2006) Chocolate in Mesoamerica: A Cultural History of Cacao, ed McNeil CL (Univ Press of Florida, Gainesville, FL), pp 31–68.

Callaghan, R., Montenegro, A., Fitzpatrick , S.M. (2022). The Effects of Intra- and Interannual Wind and Current Variation on Sailing Raft Travel along the Pacific Coast of the Americas. *In WAVES OF INFLUENCE, Pacific Maritime Networks Connecting Mexico, Central America, and Northwestern South America.* Dumbarton oaks Pre-Columbian symposia and colloquia. Colin McEwan and Frauke Sachse Eds.

Cheesman, E. E. (1944) Notes on the nomenclature, classification and possible relationships of cocoa populations. Trop. Agric. 21, 144–159, 8.

Clement, C.R., Denevan, W.M., Heckenberger, M.J., Junqueira, A.B., Neves, E.G.,Teixeira, W.G., Woods, W.I. (2015) The domestication of Amazonia before European conquest.Proc. R. Soc. B282: 20150813.http://dx.doi.org/10.1098/rspb.2015.0813

Coe, S. D. & Coe, M. D. (2013).The True History of Chocolate, 3rd Ed. (Thames and Hudson, New York)

Cornejo, O.E., Yee, MC., Dominguez, V. *et al*. (2018) Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Commun Biol* **1**, 167. https://doi.org/10.1038/s42003-018-0168-6

Cuatrecasas, J. (1964). Cacao and Its Allies, a Taxonomic Revision of the Genus Theobroma." in Systematic Plant Studies, 379–614. Contributions from the United States National Herbarium. Washington, DC: Smithsonian Institution Press.

Fouet, O., Loor Solorzano *et al.* (2022). Collection of native Theobroma cacao L. accessions from the Ecuadorian Amazon highlights a hotspot of cocoa diversity. Plants, People, Planet, 4(6), 605–617. https://doi.org/10.1002/ppp3.10282

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29, 1682–1684 doi: 10.1093/bioinformatics/btt193.

Lathrap, D.W. (2013). La Antigüedad e Importancia de las Relaciones de Intercambio a Larga Distancia en los Trópicos Húmedos de Sudamérica Precolombina. In Aldo Bolaños. Amazonas: Ruta milenaria-Tome 1- p161-179- Ediciones Copé, Peru.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal**,** [S.l.], v. 17, n. 1, p. pp. 10-12,. ISSN 2226-6089. doi:https://doi.org/10.14806/ej.17.1.200.

Meyer M., Kircher M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb Protoc. (6):pdb.prot5448. doi: 10.1101/pdb.prot5448. PMID: 20516186.

Motamayor, J. C. *et al.* (2008) Geographic and genetic population differentiation of he Amazonian chocolate tree (*Theobroma cacao L*). PLoS One 3, e3311.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. Jul;89(3):583-90. doi: 10.1093/genetics/89.3.583. PMID: 17248844; PMCID: PMC1213855.

Ondov BD, Bergman NH, Phillippy AM. (2011) Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. Sep 30;12:385. doi: 10.1186/1471-2105-12-385. PMID: 21961884; PMCID: PMC3190407.

Powis TG, *et al.* (2002) The origins of cacao use in Mesoamerica. Mexicon 30:35–38.

Powis, T. G., Cyphers, A., Gaikwad, N. W., Grivetti, L. & Cheong, K. (2011) Cacao use and the San Lorenzo Olmec. *Proc Natl Acad Sci USA* 108, 8595. https://doi.org/10.1073/pnas.1100620108.

Pritchard, J. K., Stephens M. and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155 (2): 945–59.

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

Zarrillo, S., Gaikwad, N., Lanaud, *et al.* (2018). The use and domestication of *Theobroma cacao* during the mid-Holocene in the upper Amazon. Nature Ecology & Evolution, 2(12), 1879–1888. https://doi.org/10.1038/s41559-018-0697-x.