Title: Multi-proxy evidence highlights a complex evolutionary legacy of maize in South America

Authors: – Logan Kistler[1,2]*, S. Yoshi Maezumi[3,4], Jonas Gregorio de Souza[3], Natalia A.S. Przelomska[1,5], Flaviane Malaquias Costa[6], Oliver Smith[7], Hope Loiselle[1,8], Jazmín Ramos-Madrigal[7], Nathan Wales[9], Eduardo R. Ribeiro[1], Ryan R. Morrison[2], Claudia Grimaldo[10], Andre P. Prous[11], Bernardo Arriaza[12], M. Thomas P. Gilbert[7,13], Fabio de Oliveira Freitas[14]*, Robin G. Allaby[2]*

Affiliations:

[1]Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA.

[2]Department of Life Science, University of Warwick, Coventry CV4 7AL, UK.

[3]Department of Archaeology, College of Humanities, University of Exeter, Laver Building, North Park Road, Exeter EX4 4QE, UK.

[4]Department of Geography and Geology, The University of the West Indies, Mona Campus, Jamaica, JM.

[5]Center for Conservation Genomics, Smithsonian Conservation Biology Institute, National Zoo, Washington, DC, 20008, USA.

[6]University of São Paulo, Escola Superior de Agricultura Luis de Queiroz, Piracicaba, SP 13418-900, Brazil.

[7]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark.

[8]Department of Anthropology, University of Washington, Denny Hall 314, Seattle, WA 98195 USA.

[9]Department of Archaeology, University of York, King's Manor, York, YO1 7EP, UK.

[10]Department of Oncology, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford, OX3 7DQ, UK.

[11]Museu de Historia Natural/Jardim Botânico da Universidade Federal de Minas Gerais, Belo Horizonte, MG 31270-901, Brazil.

[12]Instituto de Alta Investigación, Universidad de Tarapacá, 15101 Arica, Chile.

[13]Norwegian University of Science and Technology, University Museum, 7491 Trondheim, Norway.

[14]Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF, CEP 70770-901 Brazil.

*Correspondence to: KistlerL@si.edu; Fabio.Freitas@embrapa.br; R.G.Allaby@warwick.ac.uk

Abstract: Domesticated maize evolved from wild teosinte under human influences in Mexico beginning around 9,000 BP, traversed Central America by ~7,500 BP, and spread into South America by ~6,500 BP. Landrace and archaeological maize genomes from South America suggest that the ancestral population to South American maize was brought out of the domestication center in Mexico and became isolated from the wild teosinte gene pool before traits of domesticated maize were fixed. Deeply structured lineages then evolved within South America out of this partially domesticated progenitor population. Genomic, linguistic, archaeological, and paleoecological data suggest that the southwestern Amazon was a secondary improvement center for partially domesticated maize. Multiple waves of human-mediated dispersal are responsible for the diversity and biogeography of modern South American maize.

One Sentence Summary: Maize lineages diversified while the domestication process was still underway, and the Southwest Amazon was a secondary improvement center for partially domesticated maize in South America.

Main Text:

Maize (*Zea mays* ssp. *mays*) evolved from wild Balsas teosinte (*Zea mays* ssp. *parviglumis*, hereafter *parviglumis*) in modern-day lowland Mexico beginning around 9,000 years ago (*1*), and spread to dominate food production systems throughout much of the Americas by the beginning of European colonization in the fifteenth century. Archaeological and genetic data from ancient DNA studies have highlighted aspects of maize natural history, including the evolution and fixation of agricultural traits, and adaptation of maize to diverse new environments (*2*–*6*). Archaeological remains establish that maize was brought to the southwestern US and the Colorado Plateau by ~4,000 BP (*7*), traversing Panama by ~7,500 BP (*8*), and arriving in Coastal Peru (*9*), the Andes (*10*), and lowland Bolivian Amazon (*11*) between ~6,500 and 6,300 BP (Fig. 1; Table S1). Today, maize is a staple food species, yielding over 6% of all food calories for humans, plus more in livestock feed and processed foods (*12*).

Maize domestication is thought to have occurred once, with little subsequent gene flow from *parviglumis* (*13*, *14*). However, archaeogenomic evidence reveals maize was only partially domesticated in Mexico by ~5,300 BP (*2*, *3*), carrying a mixture of wild-type and maize-like alleles at loci involved in the domestication syndrome. For example, the domestic-type *TGA1* gene variant responsible for eliminating the tough teosinte fruitcase was already present by this time period (*2*), whereas other loci associated with changes to seed dispersal and starch production during domestication still carried wild-type variants (*2*, *3*). The state of partial domestication sets these archaeogenomes apart from modern fully domesticated maize, which carries a complete, stable set of domestication alleles conferring the domesticated phenotype. This partially domesticated maize was grown in Mexico well after maize had become established in South America, which raises the question of how South American maize came to possess the full complement of fixed domestication traits. To reconcile archaeobotanical and genomic data concerning the domestication and dispersal history of maize in South America, we sequenced maize genomes from forty indigenous landraces and nine archaeological samples from South America (Fig. 1; Tables S2, S3), and analyzed them alongside published modern (n=68) and ancient (n=2) maize and teosinte genomes (*15*).

Model-based clustering highlights extensive admixture and population overlap between maize populations, but we observe several robust lineages (*15*) (Fig. 1): i) The Andes and the Pacific coast of South America, ii) lowland South America, including the Amazon and Brazilian Savanna, iii) North America north of the domestication center, and iv) highland Mexico and

Central America, previously observed to contain introgression from wild *Z. mays* ssp. *mexicana* (*14*, *16*). We also observe a widespread 'pan-American' lineage spanning from northern Mexico into lowland South America. In a previous analysis based on multiple nuclear microsatellites, maize formed a monophyletic subset of teosinte, with South American lineages as the most derived elements in a phylogenetic tree (*13*). This pattern has been interpreted as evidence for a single episode of domestication followed by dispersal culminating in the Andes after maize became established throughout the rest of the range of cultivation (*13*). However, archaeological evidence for persistent maize cultivation indicates it was established in numerous locations throughout South America by ~6500-4000 BP regionally. On the basis of this information, we propose that South American maize was carried away from the Mesoamerican domestication center soon after initial stages of domestication, and may have been one of several partially domesticated maize lineages that independently fissioned from the primary gene pool following the onset of domestication in Mexico (Fig. 2).

Using f_4 statistics (*17*), we observe asymmetry in *parviglumis* ancestry among modern maize populations (Fig. 2). This reveals that maize-*parviglumis* gene flow was ongoing in some lineages after others became reproductively isolated. Whereas later gene flow from *Z. mays* ssp. *mexicana*—a highland subspecies of teosinte—is well documented in some maize (*6*, *14*, *16*), this finding contradicts the assumption that dispersal and diversification throughout the Americas happened only after the severance of gene flow from *parviglumis* (*13*, *14*). Thus, while South American maize became reproductively isolated from the wild progenitor when it was carried away from the domestication center, maize lineages remaining in Mexico underwent continued crop-wild gene flow before diversifying into extant landraces over subsequent millennia. The pan-American lineage shows excess shared ancestry with *parviglumis* relative to all other major groups (Fig. 2B), suggesting that this group emerged from the domestication center and dispersed after other maize lineages became regionally established. Because the pan-American lineage carries excess *parviglumis* ancestry relative to the strictly South American lineages, it appears to represent a second episode of maize dispersal from Mesoamerica, reinforcing two major waves of maize movement into South America as previously suggested (*5*).

The genomes of two ancient maize cobs from the Tehuacan Valley of Mexico at ~5300 BP recently revealed a state of partial domestication—a mixture of maize- and *parviglumis*-like alleles at loci involved in domestication (*2*, *3*). This is puzzling, given the sustained use of domesticated maize from ~6500 BP onward in South America (Fig. 1; Table S1) (*11*, *18*). However, principal components analysis and f_3 statistics reveal considerable genomic distance between these two Mesoamerican archaeogenomes (Figs. 1; S2), and f_3 statistics confirm that the SM10 genome (*3*) is more maize-like while the Tehuacan162 genome (*2*) is more *parviglumis*-like (Fig. S2). In total, the two genomes are from the same region and time period, and both are partially domesticated, but otherwise they appear to represent independent samples out of a diverse semi-domesticated population containing an array of domestic and wild-type alleles.

Given the state of partial domestication observed in the Tehuacan and San Marcos genomes (*2*, *3*), early South American maize emerging from their common ancestral population would likely also have been a partially domesticated form of maize containing an assortment of wild and domestic alleles. This ancestral population likely harbored the building blocks for fully domesticated maize, but lacked the allelic fixation and linkage of the modern domesticated crop.

We expect that in this ancestral semi-domesticated population, domestication loci under ongoing selection would have been continually de-coupled from their chromosomal neighborhood through recombination (*19*, *20*), resulting in an enrichment of the original parviglumis genomic background near domestication genes relative to its genome-wide retention. If the domestication syndrome was fully established in the common ancestor of all extant maize, no modern *parviglumis* genome should carry this enriched affinity to domestication loci to differing degrees in different maize lineages, since the same background would have become fixed in their common ancestor. However, if South American maize became isolated while fundamental domestication was still ongoing, as we hypothesize, then components of the *parviglumis* genomic background are expected to differ between early stratified maize lineages. Therefore in this case, modern *parviglumis* genomes would carry a specifically South American or non-South American affinity for the enriched wild-type background near domestication loci.

We compared *D*-statistics (*21*) across the whole genome (D_{WG}) and within 10kb of 186 known domestication loci (D_{dom}) to test for these asymmetrical *parviglumis* contributions between pairs of extant South American and non-South American maize around domestication genes (*15*). We found that *parviglumis* enrichment associated with domestication is highly patterned among major ancestry groups, with several *parviglumis* genomes associated exclusively with either South American or non-South American D_{dom} enrichment, and a significant association with ancestry overall (Fig. 2C; χ^2 test p=2.74 x 10^{-6}). That is, we observe that *parviglumis* ancestry is enriched near domestication genes in a pattern demonstrating that domestication-associated selection was still ongoing after the stratification of the major extant lineages from their semi-domesticated ancestral population. This pattern validates a model where the ancestral population in South America was itself only partially domesticated during its dispersal away from the domestication center.

In total, we find support for a model of stratified domestication in maize (Fig. 2). The initial stages of maize domestication likely occurred only once within a diverse wild Balsas River Basin gene pool, as previously suggested (*13*). However, before the domestication syndrome was fixed and stable, multiple lineages separated and selection pressures on domestication loci continued independently outside of the primary domestication center. Some of these divergent semi-domesticated populations likely led to terminal lineages lacking sufficient diversity and ecological context to continue the domestication process. Others, like ancestral South American maize, evolved into fully domesticated lineages under continuing anthropogenic pressures.

The earliest evidence places maize in the southwestern Amazon by ~6,500 BP (*11*), a region serving as a geographic interface of the lowland and Andean/Pacific genetic lineages (Fig. 1). We hypothesize that the southwestern Amazon may have been a secondary improvement center for the partially domesticated crop before the divergence of the two South American groups. When maize arrived, southwestern Amazonia was a plant domestication hotspot (*22*). Additionally, microfossil assemblages (*11*, *22*) reveal the presence of polyculture (mixed cropping) from ~6,500 cal BP onward, such that a new crop species could be integrated into existing food production systems supporting domestication activities.

Pollen and phytolith data demonstrate a west-to-east pattern of maize expansion across the Amazon, and show that maize was consistently present from ~4,300 BP onward in the eastern Amazon (*18*). Initially, maize in the eastern Amazon was part of a polyculture agroforestry system combining annual crop cultivation with wild resource use and low-level management through burning (*18*). Maize cultivation proceeded alongside the progressive enrichment of edible forest species and subsequent waves of new crop arrivals, including sweet potato (~3,200 BP), manioc (~2,250 BP) and squash (~600 BP). The development of anthropogenically enriched Amazonian Dark Earth soils ~2,000 BP (*23*) enabled the expansion and intensification of maize cultivation, likely increasing carrying capacity to sustain growing populations in the eastern Amazon (*18*). The extant endemic maize lineage in lowland South America likely originated with this long-term process involving millennia of evolving land use practices.

Several landraces and two archaeogenomes (~700 BP) in eastern Brazil also show strong genetic links to Andean maize near the southwestern Amazon (Fig. 3). This pattern closely mirrors linguistic patterns linking Andean, Amazonian, and eastern Brazilian maize cultivation, and suggests a second major west-to-east cultural expansion of maize traditions. A loanword for maize with possible Andean origins was transmitted from Amazonian Arawak languages—most likely originating in southwest Amazonia (*28*)—into Macro-Jê stock languages in the Brazilian savanna and Atlantic coast (*24*) (Fig. S3). Archaeological evidence suggests this expansion occurred ~1200-1000 BP with the spread of a cultural horizon of geometric enclosures and mound ring villages throughout southern Amazonia, and ring villages in the central Brazilian savannas and the Atlantic coast (Figs. 3; S4) (*25–27*). This process is roughly contemporaneous with archaeological Andean-admixed genomes in the area. Thus, Arawak speakers likely brought non-local Andean/Pacific maize lineages into a landscape where maize was an established component of long-term land management and food production strategies.

Finally, we quantified the mutation load in maize genomes—the accumulation of potentially deleterious alleles due to drift and selection (*16*)—using a phylogenetic framework to estimate evolutionary constraint (*15*). We observe that South American lineages carry a higher mutation load than other maize lineages. Mutation load increases linearly with distance from the domestication center and is linked with ancestry, and the Andean/Pacific group carries the highest burden of potentially deleterious variants (Fig. 4; (*15*)). The mutation load in the Andes has been attributed to selection for high-altitude adaptations (*16*), but the elevated mutation load in lowland maize also suggests a history of shared selection and drift effects prior to highland adaptation. These processes would likely have included a founder episode as maize was carried into South America, persistent selection pressures for regional adaptation, and the latter stages of domestication after isolation from the founding gene pool. We also find that Andean and Pacific maize from ~1000 BP into the early colonial period has a low mutation load compared with its modern Andean/Pacific counterparts (Wilcoxon p=0.002477; (*15*); Fig. 4); although still elevated compared with non-South American lineages. It is possible that Andean maize experienced a wave of deleterious allele accumulation as human and crop populations were disrupted by changes caused by the arrival of Europeans (*28*). Alternatively, the increasing mutation load in modern crops could represent the ongoing effects of burdensome allele accumulation over nine millennia of human intervention.

References and Notes:

1. D. R. Piperno, A. J. Ranere, I. Holst, J. Iriarte, R. Dickau, Starch grain and phytolith

evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proc. Natl. Acad. Sci.* **106**, 5019–5024 (2009).

2. J. Ramos-Madrigal *et al.*, Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. *Curr. Biol.* **26**, 3195–3201 (2016).

3. M. Vallebueno-Estrada *et al.*, The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. *Proc. Natl. Acad. Sci.* **113**, 14151–14156 (2016).

4. V. Jaenicke-Despres *et al.*, Early allelic selection in maize as revealed by ancient DNA. *Science.* **302**, 1206–1208 (2003).

5. F. O. Freitas, G. Bandel, R. G. Allaby, T. A. Brown, DNA from primitive maize landraces and archaeological remains: implications for the domestication of maize and its expansion into South America. *J. Archaeol. Sci.* **30**, 901–908 (2003).

6. R. R. da Fonseca *et al.*, The origin and evolution of maize in the American Southwest. *Nat. Plants*. **1**, 1–5 (2015).

7. W. L. Merrill *et al.*, The diffusion of maize to the southwestern United States and its impact. *Proc. Natl. Acad. Sci.* **106**, 21019–21026 (2009).

8. D. R. Piperno, K. H. Clary, R. G. Cooke, A. J. Ranere, D. Weiland, Preceramic Maize in Central Panama: Phytolith and Pollen Evidence. *Am. Anthropol.* **87**, 871–878 (1985).

9. A. Grobman *et al.*, Preceramic maize from Paredones and Huaca Prieta, Peru. *Proc. Natl. Acad. Sci.* **109**, 1755–9 (2012).

10. M. B. Bush *et al.*, A 6900-year history of landscape modification by humans in lowland Amazonia. *Quat. Sci. Rev.* (2016), doi:10.1016/j.quascirev.2016.03.022.

11. S. O. Brugger *et al.*, Long-term man-environment interactions in the Bolivian Amazon: 8000 years of vegetation dynamics. *Quat. Sci. Rev.* (2016), doi:10.1016/j.quascirev.2015.11.001.

12. F. A. O. of the United Nations, FAOSTAT statistics database (2018), doi:http://www.fao.org/faostat/.

13. Y. Matsuoka *et al.*, A single domestication for maize shown by multilocus. *Proc. Natl. Acad. Sci*. **99**, 6080–6084 (2002).

14. J. van Heerwaarden *et al.*, Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc. Natl. Acad. Sci.* **108**, 1088–1092 (2011).

15. Supplemental materials online.

16. L. Wang *et al.*, The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).

17. N. Patterson *et al.*, Ancient Admixture in Human History. *Genetics*. **192**, 1065–1093 (2012).

18. S. Y. Maezumi *et al.*, The legacy of 4,500 years of polyculture agroforestry in the eastern Amazon. *Nat. Plants* (2018), doi:10.1038/s41477-018-0205-y.

19. W. G. Hill, A. Robertson, The effect of linkage on limits to artificial selection. *Genet. Res.*

8, 269–94 (1966).

20. M. W. Feldman, S. P. Otto, F. B. Christiansen, Population genetic perspectives on the evolution of recombination. *Annu. Rev. Genet.* **30**, 261–295 (1996).
21. R. E. Green *et al.*, A draft sequence of the Neandertal genome. *Science*. **328**, 710–22 (2010).
22. J. Watling *et al.*, Direct archaeological evidence for Southwestern Amazonia as an early plant domestication and food production centre. *PLoS One*. **13**, e0199868 (2018).
23. B. Glaser, Prehistorically modified soils of central Amazonia: a model for sustainable agriculture in the twenty-first century. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* (2007), doi:10.1098/rstb.2006.1978.
24. E. R. Ribeiro, Tapuya connections: language contact in eastern Brazil. *Liames*. **9**, 61–76 (2009).
25. J. G. De Souza *et al.*, Pre-Columbian earth-builders settled along the entire southern rim of the Amazon. *Nat. Commun.* (2018), doi:10.1038/s41467-018-03510-7.
26. M. J. Heckenberger *et al.*, Pre-Columbian urbanism, anthropogenic landscapes, and the future of the Amazon. *Science.* (2008), doi:10.1126/science.1159769.
27. I. Wüst, C. Barreto, The Ring Villages of Central Brazil: A Challenge for Amazonian Archaeology. *Lat. Am. Antiq.* **10**, 3–23 (1999).
28. C. R. Clement, 1492 and the loss of amazonian crop genetic resources. II. Crop biogeography at contact. *Econ. Bot.* (1999), doi:10.1007/BF02866499.
29. R. S. Walker, L. A. Ribeiro, Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings. Biol. Sci.* **278**, 2562–7 (2011).
30. A. Cooper, H. N. Poinar, Ancient DNA: Do it right or not at all. *Science.* **289**, 1139 (2000).
31. N. Wales, K. Andersen, E. Cappellini, M. C. Ávila-Arcos, M. T. P. Gilbert, Optimization of DNA recovery and amplification from non-carbonized archaeobotanical remains. *PLoS One*. **9** (2014), doi:10.1371/journal.pone.0086827.
32. M. Dodt, J. T. Roehr, R. Ahmed, C. Dieterich, FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel).* **1**, 895–905 (2012).
33. M. Kircher, in *Ancient DNA: Methods and Protocols*, B. Shapiro, M. Hofreiter, Eds. (Springer, 2012).
34. Y. Jiao *et al.*, Improved maize reference genome with single-molecule technologies. *Nature* (2017), doi:10.1038/nature22971.
35. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013) (available at http://arxiv.org/abs/1303.3997).
36. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009), doi:10.1093/bioinformatics/btp324.
37. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**,

2078–9 (2009).

38. A. McKenna *et al.*, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010), doi:10.1101/gr.107524.110.

39. J. M. Chia *et al.*, Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* (2012), doi:10.1038/ng.2313.

40. P. S. Schnable *et al.*, The B73 maize genome: complexity, diversity, and dynamics. *Science*. **326**, 1112–5 (2009).

41. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* (2011), doi:10.1093/bioinformatics/btr011.

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

43. A. W. Briggs *et al.*, Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38** (2010).

44. N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, D. Reich, Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20130624–20130624 (2014).

45. R. Bukowski *et al.*, Construction of the third-generation Zea mays haplotype map. *Gigascience*. **7** (2018), doi:10.1093/gigascience/gix134.

46. T. S. Korneliussen, A. Albrechtsen, R. Nielsen, ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*. **15**, 356 (2014).

47. C. C. Chang *et al.*, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* (2015), doi:10.1186/s13742-015-0047-8.

48. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

49. P. Skoglund *et al.*, Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science.* (2012), doi:10.1126/science.1216304.

50. A. D. Martin, K. M. Quinn, J. H. Park, MCMCpack : Markov Chain Monte Carlo in R. *J. Stat. Softw.* (2011), doi:10.18637/jss.v042.i09.

51. S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* (2011), doi:10.1101/gr.113985.110.

52. R. S. Harris, thesis, The Pennsylvania State University (2007).

53. A. Siepel, D. Haussler, Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. *Mol. Biol. Evol.* (2004), doi:10.1093/molbev/msh039.

54. E. V. Davydov *et al.*, Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).

55. G. M. Cooper *et al.*, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–13 (2005).

56. L. Kistler, R. Ware, O. Smith, M. Collins, R. G. Allaby, A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res.* **45** (2017), doi:10.1093/nar/gkx361.

57. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. *R Found. Stat. Comput. Vienna, Austria.* (2016).

58. M. B. Hufford *et al.*, Comparative population genomics of maize domestication and improvement. *Nat. Genet.* (2012), doi:10.1038/ng.2309.

59. D. L. Payne, in *Handbook of Amazonian Languages*, D. C. Derbyshire, G. K. Pullum, Eds. (Mouton de Gruyter, Berlin, 1991), vol. 3, pp. 355–499.

60. B. N. I. of H. and A. H. (IPHAN), National Register of Archaeological Sites (2018), (available at http://portal.iphan.gov.br/).

61. J. Stephen Athens *et al.*, Early Prehistoric Maize in Northern Highland Ecuador. *Lat. Am. Antiq.* (2016), doi:10.7183/1045-6635.27.1.3.

62. M. B. Bush, D. R. Piperno, P. A. Colinvaux, A 6,000 year history of Amazonian maize cultivation. *Nature* (1989), doi:10.1038/340303a0.

63. S. Mora Camargo, *Cultivars, anthropic soils, and stability: a preliminary report of archaeological research in Araracuara, Colombian Amazonia* (University of Pittsburgh Latin American archaeology reports ; no. 2, Pittsburgh, 1991).

64. N. A. Sublette Mosblech, A. Chepstow-Lusty, B. G. Valencia, M. B. Bush, Anthropogenic control of late-Holocene landscapes in the Cuzco region, Peru. *The Holocene*. **22**, 1361–1372 (2012).

65. M. B. Bush *et al.*, Holocene fire and occupation in Amazonia: records from two lake districts. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **362**, 209–18 (2007).

66. A. C. Roosevelt, *Parmana: Prehistoric Maize and Manioc Subsistence Along the Amazon and Orinoco* (1980).

67. L. Hilbert *et al.*, Evidence for mid-Holocene rice domestication in the Americas. *Nat. Ecol. Evol.* (2017), doi:10.1038/s41559-017-0322-4.

68. M. B. Bush, M. C. Miller, P. E. De Oliveira, P. A. Colinvaux, Two histories of environmental change and human disturbance in eastern lowland Amazonia. *The Holocene*. **10**, 543–553 (2000).

69. R. M. Bird, O. F. D. Jr., E. T. de Carvalho, *Rev. Arqueol. São Paulor*, in press.

70. J. Iriarte *et al.*, Evidence for cultivar adoption and emerging complexity during the mid-Holocene in the La Plata basin. *Nature*. **432**, 614–7 (2004).

71. L. Perry *et al.*, Early maize agriculture and interzonal interaction in southern Peru. *Nature*. **440**, 76–9 (2006).

72. R. L. Burger, N. J. Van Der Merwe, Maize and the Origin of Highland Chavin Civilization: An Isotopic Perspective. *Am. Anthropol.* **92**, 85–95 (1990).

73. D. M. Pearsall, K. Chandler-Ezell, J. A. Zeidler, Maize in ancient Ecuador: results of

residue analysis of stone tools from the Real Alto site. *J. Archaeol. Sci.* **31**, 423–442 (2004).

74. R. M. Rosenswig, D. M. Pearsall, M. A. Masson, B. J. Culleton, D. J. Kennett, Archaic period settlement and subsistence in the Maya lowlands: new starch grain and lithic data from Freshwater Creek, Belize. *J. Archaeol. Sci.* **41**, 308–321 (2014).

75. M. R. Arford, S. P. Horn, Pollen evidence of the earliest maize agriculture in Costa Rica. *J. Lat. Am. Geogr.* **3**, 108–115 (2004).

76. D. R. Piperno, Identifying crop plants with phytoliths (and starch grains) in Central and South America: A review and an update of the evidence. *Quat. Int.* **193**, 146–159 (2009).

77. D. J. Kennett *et al.*, Pre-pottery farmers on the Pacific coast of southern Mexico. *J. Archaeol. Sci.* **37**, 3401–3411 (2010).

78. D. J. Rue, Archaic Middle American Agriculture and Settlement: Recent Pollen Data from Honduras. *J. F. Archaeol.* (1989), doi:10.1179/jfa.1989.16.2.177.

79. V. Jaenicke-Despres, B. D. Smith, in *Histories of Maize*, J. Staller, R. Tykot, B. Benz, Eds. (Academic Press, Burlington, MA, 2006), pp. 83–96.

80. S. A. Hall, Early maize pollen from Chaco Canyon, New Mexico, USA. *Palynology*. **34**, 125–137 (2010).

81. C. Bronk Ramsey, Bayesian analysis of radiocarbon dates. *Radiocarbon*. **51**, 337–360 (2009).

82. A. G. Hogg *et al.*, SHCal13 Southern Hemisphere Calibration, 0–50,000 Years cal BP. *Radiocarbon* (2013), doi:10.2458/azu_js_rc.55.16783.

83. D. Schaan *et al.*, New radiometric dates for precolumbian (2000-700 B.P.) earthworks in western Amazonia, Brazil. *J. F. Archaeol.* **37**, 132–142 (2012).

84. S. Saunaluoma, Pre-Columbian earthworks in the Riberalta region of the Bolivian Amazon. *Amaz. Antropol.* **2** (2010).

85. H. Prümers, in *Amazonía: Memorias de las Conferencias Magistrales del 3er Encuentro Internacional de Arqueología Amazónica*, S. Rostain, Ed. (Ekseption Publicidad, Quito, 2014), pp. 73–89.

86. E. G. Neves *et al.*, "Pesquisa e Formação nos Sítios Arqueológicos Espinhara e Sol de Campinas do Acre - PESC" (University of São Paulo, São Paulo, 2016).

87. I. Wüst, thesis, University of São Paulo (1990).

88. P. I. Schmitz, Arqueologia de Goiás: sequencia cultural e datações de C14. *Anuário Divulg. Científica*. **3/4**, 1–19 (1976).

89. V. Calderón, in *Programa Nacional de Pesquisas Arqueológicas: resultados preliminares do terceiro ano*, PRONAPA, Ed. (Museu Paraense Emílio Goeldi, Belém, 1969), pp. 161–168.

90. C. Perota, thesis, University of São Paulo (1979).

91. M. D. Andreatta, Projeto Anhangüera de Arqueologia de Goiás (1975-1985). *Rev. do Mus. Paul.* **33**, 143–156 (1985).

92. P. I. Schmitz, M. B. Ribeiro, A. S. Barbosa, M. O. Barbosa, A. F. Miranda, *Caiapônia - Arqueologia nos cerrados do Brasil Central* (UNISOS, São Leopoldo, 1986).

93. E. R. Oliveira, thesis, University of São Paulo (2005).

Acknowledgements: We thank Admera Health for assistance with sequence data collection, and we thank Dolores Piperno for comments on the manuscript. **Funding:** Work was supported by Natural Environment Research Council Independent Research Fellowship NE/L012030/1 to LK, and a sub-award from Science and Technology Facilities Council grant ST/K001760/1 [PI Thomas Meagher, co-I Peter Kille] to LK and RGA. **Author contributions:** Study conceptualization and design: LK, FdOF, RGA; Sample acquisition: FdOF, APP, CG, BA, MTPG; Genomic data collection: LK, FdOF, OS, NW, RM; Genomic data analysis: LK, NASP; Archaeology and linguistic background and interpretation: JGdS, SYM, FdOF, FMC, ER; Interpretation and integration of results: LK, SYM, JGdS, FMC, JR-M, NW, FdOF, RGA; Visualization: LK, SYM, JGdS, HL, NASP; Manuscript drafting: LK, SYM, JGdS, with input from NASP, FdOF, and RGA. All authors reviewed and contributed to the final manuscript. **Competing interests:** We declare no competing interests. **Data and materials availability:** Raw sequence data, NCBI Sequence Read Archive accession SRP152500. In-house scripts for data handling and analysis (allele frequency estimation, *f* and *D* statistic calculation, genome alignment conformation for mutation load analysis, and exclusion amplification duplicate removal), genome-wide GERP scoring details, genomic mappability bed file, SNP calls, and mapDamage results are available on the Dryad Digital Repository doi:TBD. Germplasm for newly sequenced maize landraces is curated at the Embrapa gene bank in Brasilia, Brazil and Programa Cooperativo de Investigaciones en Maíz in Peru, which provided sample material for this study to FdOF and CG. Archaeological samples from Santa, Chorrillos, Ica, and Jujuy were originally obtained via the PSUM Archaeological Project, Paurarku Archaeological Project, and Samaca Archaeological Project, facilitated by archaeologists Victor Pimentel, Kevin Lane, David Beresford-Jones, and Hugo Yacobaccio.

Figures and legends

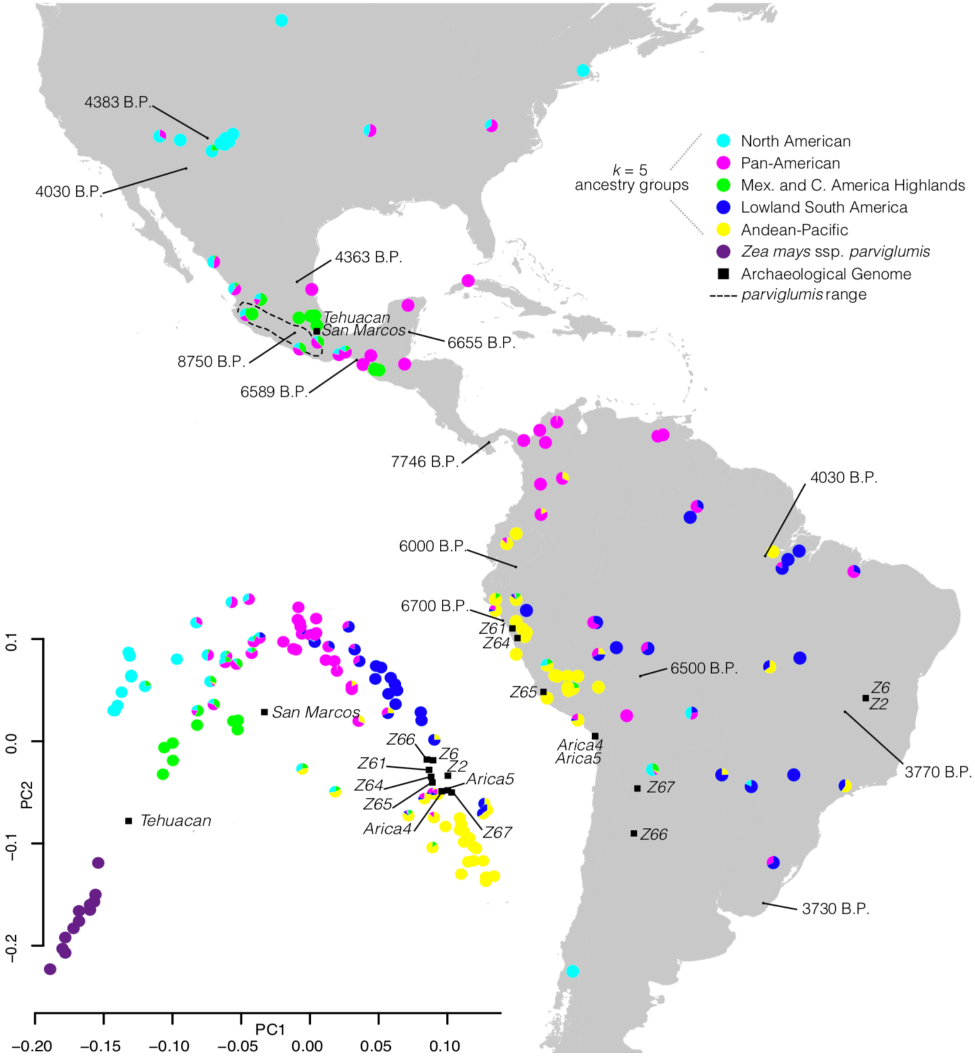


Figure 1 – Distribution and ancestry proportions of maize genomes, and principal components analysis of maize and *parviglumis* genomes. Pie colors reflect ancestral proportions estimated via model-based clustering (*k*=5) of modern maize genomes (*15*). Archaeological genomes were projected onto the PCA to mitigate degradation biases (*15*). Dates reflect early regional maize archaeobotanical remains (Table S1 and Fig. S1).

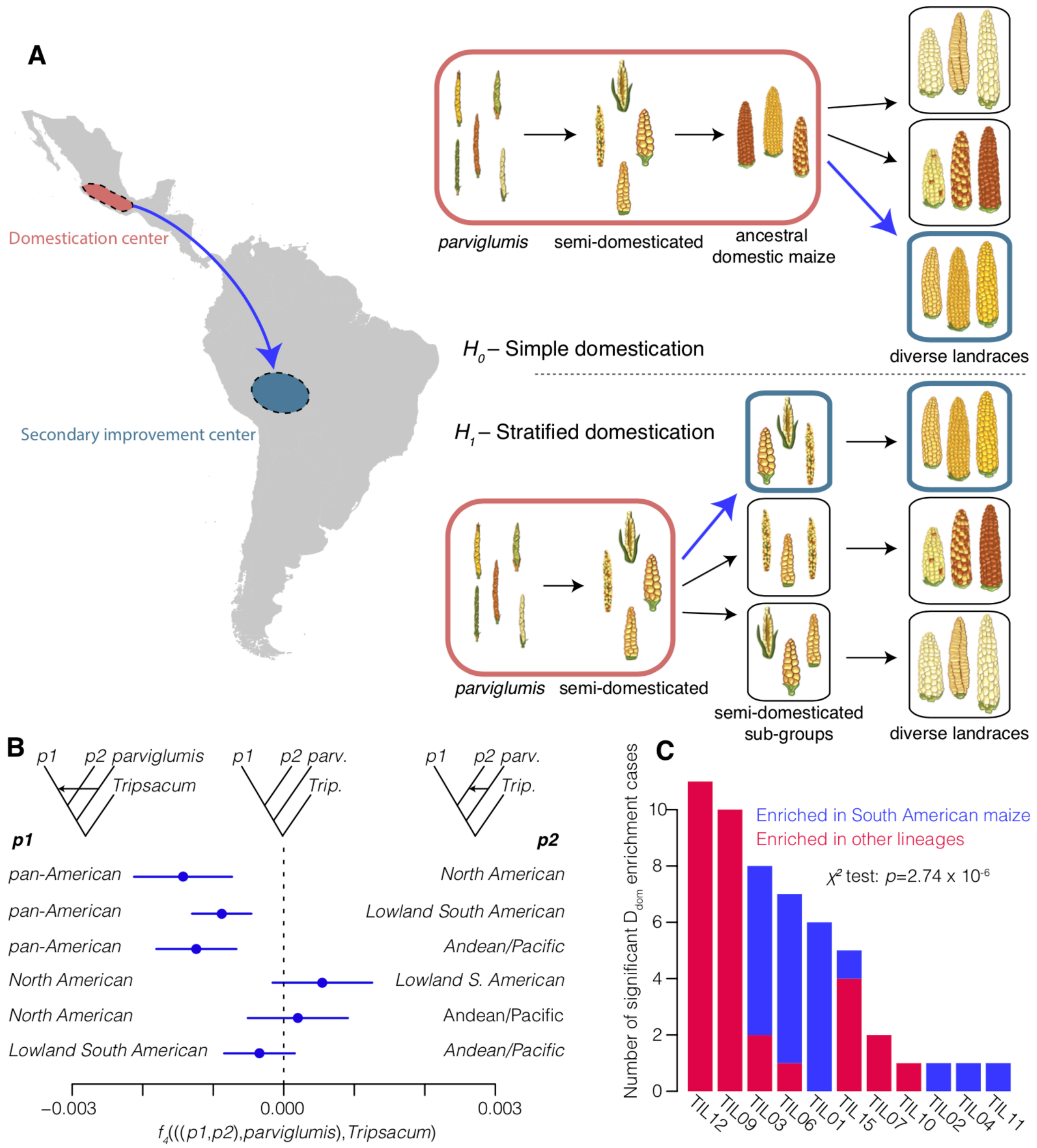


Figure 2 – A stratified domestication model for maize. **A)** Schematic comparing the conventional domestication model under which maize became fully domesticated and then dispersed throughout the Americas, vs. a stratified domestication model in which partially domesticated sub-populations became reproductively isolated before the fixation of the domestication syndrome. **B)** f_4 statistics demonstrating excess allele sharing between the pan-American lineage and wild *parviglumis* compared with other maize, revealing non-uniform crop-wild gene flow after initial domestication. Bars are three standard errors under a block jackknife (*15*). **C)** Barplot of enriched *parviglumis* contributions to ancestry near domestication genes, where each bar is a *parviglumis* genome contributing to South American maize (blue) or other maize (red) D_{dom} enrichment. Geographic segregation in D_{dom} enrichment among *parviglumis* genomes suggests that the domestication syndrome was not yet fixed in a common domesticated ancestor of modern maize.

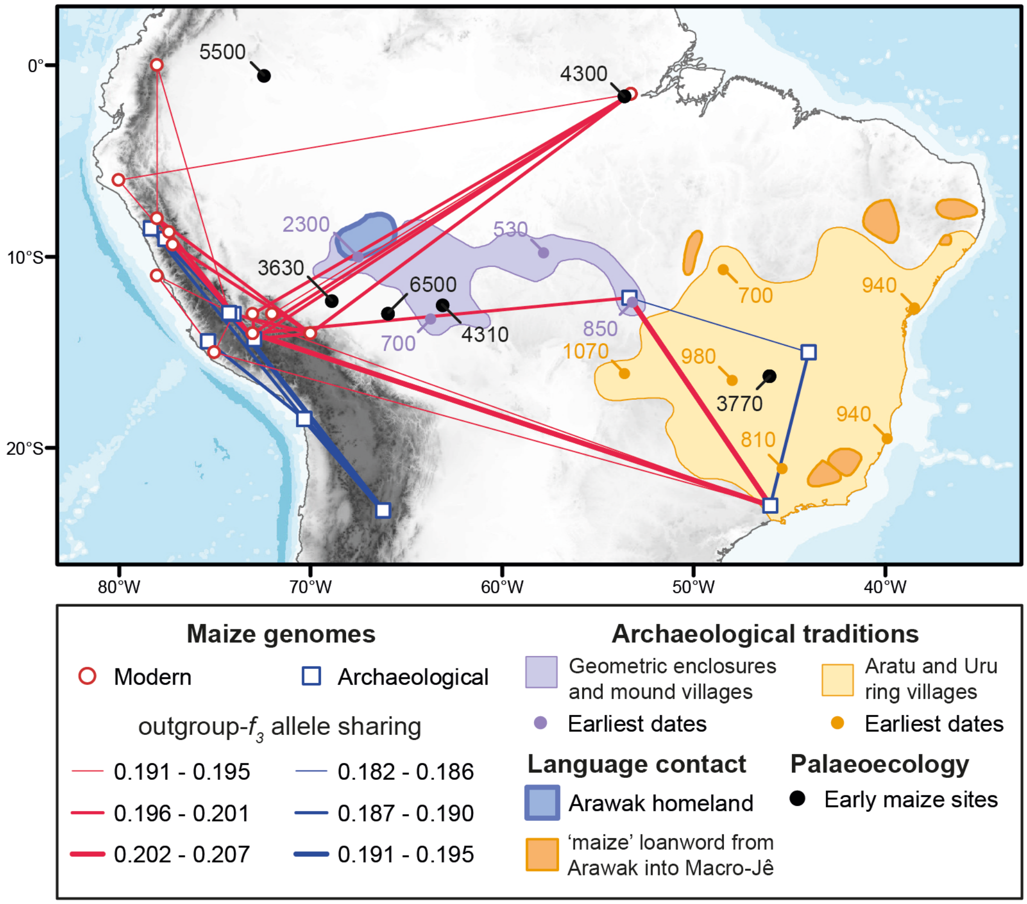


Figure 3 – Genomic relatedness overlapping linguistic and archaeological patterns in lowland South America. Maize genomes with ≥50% Andean/Pacific ancestry and ≥99% South American ancestry are connected by lines with the two other genomes with which they share the highest outgroup-f_3 value. Geometric enclosures and mound ring villages of southern Amazonia broadly coincide with the expansion of Arawak languages, whereas the Uru and Aratu ring villages coincide with the distribution of Macro-Jê languages (*15*) (Figs. S3, S4). Only the earliest regional dates for each archaeological tradition are shown (see Table S4). Macro-Jê languages borrowing an Arawak loanword for 'maize' are based on (*24*). Arawak homeland is shown approximately in the modern location of Apurinã following (*29*).

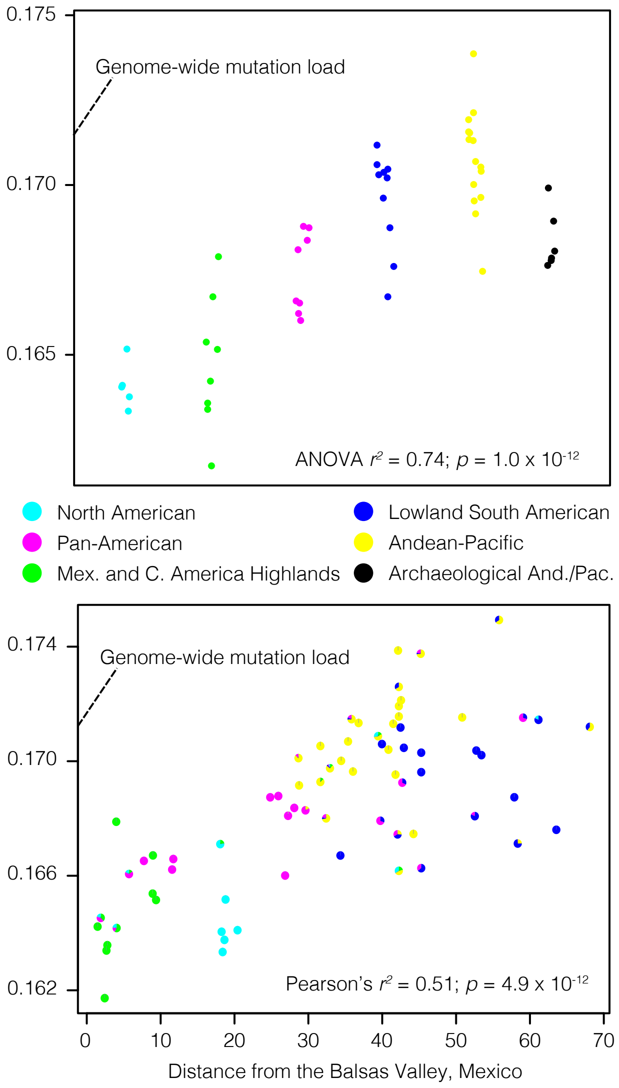


Figure 4 – Genome-wide mutation load across ancestry groups (non-admixed samples only in top panel), and load compared with distance to the domestication center. Mutation load is calculated as a proportion of the theoretical maximum load over observed SNPs, and ancient load scores are re-scaled for missingness using a Procrustes transformation (*15*). Euclidean distance in degrees to the Balsas River Valley is shown.

Supplemental Materials:

Materials and Methods

Figures S1-S4

Tables S1-S4

References (*30-92*)

Supplementary Materials for

Multi-proxy evidence highlights a complex evolutionary legacy of maize in South America

Logan Kistler[1,2]*, S. Yoshi Maezumi[3,4], Jonas Gregorio de Souza[3], Natalia A.S. Przelomska[1,5], Flaviane Malaquias Costa[6], Oliver Smith[7], Hope Loiselle[1,8], Jazmín Ramos-Madrigal[7], Nathan Wales[9], Eduardo R. Ribeiro[1], Ryan R. Morrison[2], Claudia Grimaldo[10], Andre P. Prous[11], Bernardo Arriaza[12], M. Thomas P. Gilbert[7,13], Fabio de Oliveira Freitas[14]*, Robin G. Allaby[2]*

*Correspondence to: KistlerL@si.edu; Fabio.Freitas@embrapa.br; R.G.Allaby@warwick.ac.uk

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S4
Tables S1 to S4

Materials and Methods

Materials

Modern maize landrace accessions originally collected in traditional farming contexts were sampled from the Embrapa germplasm collection in Brasilia and Programa Cooperativo de Investigaciones en Maíz in Peru. Landrace details are provided in Dataset S1. Archaeological sample details are provided in Table S2.

```
/mnt/data/logan/maize/S_Am/ancient/bwa/fullBams/mapDamage
```

Methods

Modern DNA isolation and library preparation

We attempted seed germination from two seeds of each modern landraces and harvested ~100mg of first-leaf tissue from successful seedlings for DNA extraction (n=30). For 10 landraces where germination failed, we isolated DNA directly from seed tissue. DNA isolation followed a conventional CTAB protocol: Tissue was ground in liquid nitrogen with a sterile mortar and pestle and incubated for 24 hours with gentle agitation in lysis buffer (2% w/v CTAB; 100mM Tris pH 8; 20mM EDTA; 1.4M NaCl; 2% w/v PVP; 0.5% v/v β-Mercaptoethanol for seeds only), then subjected to two rounds of chloroform purification. The recovered aqueous fraction was mixed with 1.5 volumes Qiagen Buffer AW1 and bound to a silica spin column. The column was washed twice with Qiagen Buffer AW2 and once with acetone, air-dried for 5 minutes, and eluted in 100uL TE buffer. DNA was quantified by Qubit, and 1µg DNA was sheared in a 100uL volume using a bioruptor ultrasonicator to a target size of 350bp. We used a 2-stage SPRI bead size selection on sheared DNA to target 250bp-450bp fragments following the KAPA Hyper Prep Library protocol size selection guidelines (KR0961, v. 4.15) with AMPure XP beads (Beckman Coulter A63880). We then used NEBNext Ultra II kits (E7645) to prepare PCR-free Illumina-compatible libraries following the manufacturer's protocol, using Illumina TruSeq Nano DNA LT single indexed adapters diluted to a 1:4 ratio (from Illumina FC-121-4001 and FC-121-4002). Final libraries were purified using 1.2 volumes of SPRI beads, quantified by qPCR using the NEB Quant kit (E7630), pooled in equimolar ratios, and sequenced across six lanes of a HiSeq X10 instrument.

Ancient DNA isolation and library preparation

All ancient DNA handling up to the point of library pooling (no PCR was used) was carried out in dedicated ancient DNA clean lab facilities at the University of Warwick and University of Copenhagen, with strict observation of established protocols to prevent and detect contamination (*30*), including sequencing and analysis of control negative libraries.

We isolated DNA from 16 ancient maize samples at Warwick using the above protocol with the following modifications: Lysis incubation was extended to 72 hours, 5 volumes of binding buffer were used, and elution was in a final volume of 60µL. Additionally, DNA was freshly extracted from two ancient maize samples from the Arica site at the University of Copenhagen following the protocol described in ref (*31*), and shipped to Warwick for library preparation. We prepared PCR-free sequencing libraries from ancient DNA isolates as follows: We treated 53.5 µL DNA with 2 µL NEB FFPE DNA repair mix and 6.5 µL FFPE DNA Repair Buffer (M6630). We purified the enzyme reaction using Qiagen MinElute kits (28004) with an elution volume of 50 µL, and carried out NEB Ultra ii PCR-free library preparation as above, with modifications: Ligation time was extended from 15 to 30 minutes, and 1.5 volumes of SPRI beads were used for final purification. 18 ancient maize libraries and a negative control library prepared in parallel were pooled in equimolar ratios and sequenced on one MiSeq v3 150 cycle flowcell to assess endogenous DNA content. On the basis of a preliminary screen for endogenous DNA content, we selected nine libraries suitable for whole genome sequencing, and pooled them to sequence across six lanes of a HiSeq X10 instrument. Additional tissue from these nine samples was also sent for AMS radiocarbon dating at Beta Analytic (Table S2).

Read processing and alignment

Paired modern reads underwent adapter trimming and 3' quality trimming following the first base quality score below 20, if present, using Flexbar (*32*). For ancient samples, forward and reverse reads were merged with a minimum base quality of 20 following (*33*), and only merged reads were carried forward into analysis.

We used BWA to map all reads directly to the soft-masked maize reference genome (*Zea mays* B73 RefGen_v4 (*34*)), requiring a minimum mapping quality of 20. For modern reads, we used bwa-mem (*35*) with default settings, and for ancient datasets consisting of shorter reads, we used bwa-backtrack (bwa *aln* (*36*)) with seed disabled for improved mapping in the presence of terminal mismatches introduced through base misincorporation (). Although we used PCR-free libraries, we observed some redundancy from exclusion amplification duplicates on the HiSeq-X patterned flowcells. Migration of molecules during exclusion amplification creates localized duplicates, so that 96 separate flowcell tiles can be expected to yield largely independent sets of starting molecules. Therefore, we removed possible flowcell duplicates by separating read alignments by tile, independently removing duplicates with the samtools (*37*) *rmdup* command, and re-combining the resulting duplicate-removed files into a single read alignment, thereby avoiding removal of independent starting molecules with identical genomic coordinates by chance (Dryad: /scripts/rmdup_by_tile.pl). We used GATK (*38*) to locally realign reads around short indels.

We acquired maize HapMap2 data for maize landraces and wild *Zea mays* ssp. *parviglumis* (*39*) and additional published landrace data (*16*) from the NCBI sequence read archive, and processed reads exactly as above for modern paired reads, except we used the samtools *rmdup* function as normal for duplicate removal.

Given the extreme repetitive content of the maize genome (*40*), we assessed sitewise mappability of short reads based on the method described in (*2*): We used Jellyfish (*41*) to summarize all 35mers in the soft-masked maize reference genome (Zea mays B73 RefGen_v4 (*34*)) and mapped unique 35mers back to the genome using bwa *aln* (*36*). We filtered the resulting bam file to remove reads with one or more suboptimal mapping locations differing from the source location by only one mismatch. We used samtools *mpileup* (*37*) to summarize positional depth of coverage, so that any position covered by 18 reads satisfied the requirement that the majority of position-containing 35mers are uniquely mappable at the 1-mismatch level. This method estimated 21% of the maize reference assembly to be uniquely mappable at the majority (18 read) level, consistent with previous estimates (*2*). However, we opted for a more conservative approach requiring all 35mers covering a position to be uniquely mappable at the 1-mismatch level with a 20nt buffer masked around non-unique positions, retaining 10.04% of the reference assembly (mappability bed file available on Dryad: /misc/ZeaV4.map35.plus20.bed).

For all ancient DNA datasets, we used mapDamage 2.0 (*42*) to assess DNA degradation characteristics in read alignments. Treatment with FFPE enzyme mix includes uracil removal, so the recovered damage profiles are thought to primarily derive from 5-methyl-cytosine deamination to thymine rather than cytosine-to-uracil deamination recovered through an adenine template (*43*). As such, damage is less prominent than may be expected for libraries without pre-treatment (*44*), but is unambiguously prominent in all ancient samples (all mapDamage outputs available in Dryad: /mapDamage/).

SNP selection

We analyzed the LLD set of high-quality linkage-validated SNPs from hapmap3 (*45*). We also carried out *de novo* SNP discovery using ANGSD (*46*) over our modern South American samples and the hapmap2 (*39*) panel using the command “angsd -bam [bamList.txt] -GL 1 -out [outfileStem] -doMaf 2 -doMajorMinor 1 -SNP_pval 1e-2 -sites [mappableRegions.angsd] -doCounts 1 -doGeno 12 -doPost 1”, and we filtered the output to require coverage by a minimum of 20 samples to consider a position. Because of the high repetitive element content of the maize genome during to transposable elements and recent whole genome duplication (*40*), we discovered SNPs only in the uniquely mappable fraction as above, and strictly filtered the highest-coverage SNPs likely to originate in non single-copy regions as follows: We excluded all sites in the top .5% of coverage from any single sample, the top 1% of coverage shared between any 2 independent samples, and sites in the top 5% of coverage in 5 or more samples. Before culling on this basis, coverage at some sites was extreme even in the hapmap3 LLD set, whereas this pruning strategy substantially curtailed SNP coverage toward the poissonian expectation. We worked only with bi-allelic SNPs, and for all analysis except mutation load (see below), we only analyzed SNPs with a minor allele frequency ≥0.02 and at least half of all samples called. The complete unpruned SNP set (n=17,672,809 sites) is included on Dryad in PLINK format: “/snpCalls”.

SNP calling

We analyzed SNPs using a pseudohaplotype approach so as to include the inbred lines from hapmap2 without biases with regards to heterozygosity. For each sample, we used samtools *mpileup* (*37*) to summarize positional support at SNP sites from curated bam files, and selected a base at random supporting an allele with at least two independent reads. We used PLINK 1.9 (*47*) to prune SNPs for linkage disequilibrium using the option ‘--indep 5kb 5 2’ for analyses assuming linkage independence (PCA and model-based clustering). SNP pseudohaplotype calls used for analysis are available on dryad in plink format: “/snpCalls”.

Ancestry estimation using model-based clustering

We ran Admixture (*48*) over the SNP set of modern maize pruned for LD with *k* range from 2-10 for each of 10 replicates with independent starting seeds, and visualized the results from each replicate with the highest log-likelihood. Given extensive overlap between substructured populations, we selected *k=5* as a level for further analysis based on observed geographic structure and consistency among replicate runs.

Principal components analysis

To perform a principal components analysis of genomic variation including variably degraded archaeological genomes (144,757 to 7,365,319 SNPs called), we first carried out a PCA using only modern genomes, and used Procrustes-based projection to add ancient genomes in turn, after ref (*49*). Using the LD-pruned and maf-filtered SNP set (above) in modern genomes, we performed a reference PCA using PLINK 1.9 (*47*). We then independently repeated the procedure adding one ancient genome in each case, and used the MCMCpack (*50*) R package to carry out a Procrustes transformation to estimate the best fit to the reference panel based on the first three principal components. The dilation, translation, and rotation values from the Procrustes transformation were then applied to the complete ancient sample-containing eigenvector matrix, and the ancient sample was projected accordingly onto the reference PCA.

Mutation load estimation

We used last (*51*) to align 27 repeat-masked plant genomes independently to the unmasked maize v4 reference genome, then used the sequence of *maf-convert*, *axtChain*, *chainMergeSort*, *chainPreNet*, *chainNet*, *netToAxt*, *axtToMaf*, and *MultiZ* to make multi-alignment with 20nt minimum aligned blocks (http://genome.ucsc.edu/) (*52*). We used *splitMaf* to make single-chromosome mafs and *maf2fasta* to make fasta multi-alignments by chromosome. We used a perl script (Dryad: /scripts/collapseFA.pl) to collapse indels to match the frame of the maize reference genome, and a perl script (Dryad: /scripts/matchMasking.pl) to lift masking from the repeat-masked maize genome onto all aligned genomes. We then excluded the maize genome from all model fitting and sitewise calculations. We extracted fourfold-degenerate sites from the chromosome 10 alignment using a perl script (Dryad: /scripts/quickDegen.pl) to fit a neutral model for calculating evolutionary constraint, and fit model the neutral model using *phyloFit* (*53*). The resulting neutral model tree is on Dryad: /GERP/chr10.4degen.tre. We then used GERP++ *gerpcol* (*54*) to calculate rejected substitution (RS) scores at all genomic sites with ≥3 aligned non-maize genomes.

For estimating genome-wide mutation load, we considered all SNP sites with ≥ 4 expected substitutions and an RS score ≥ 2, signifying a significant level of constraint (*55*). We polarized derived and ancestral alleles according to *Sorghum bicolor* and, if *Sorghum* was not represented, *Setaria italica.* Because the expected substitutions value, and therefore the RS score, are variable according to number of genomes present at the position, we summarized the mutation load as:

(sum of RS scores at all sites with a derived allele) / (sum of RS scores at all sites with a base called)

This value gives the proportion of the theoretical maximum number of rejected substitutions for a panel of potentially informative SNPs. Genome-wide expected substitutions and RS scores are on Dryad: /GERP/Zea_mays.allChr.rates.gz. We did not estimate mutation load in the inbred accessions from HapMap2, given that recessive deleterious alleles may be purged during inbreeding.

Because of sequence-based and genomic biases in ancient DNA degradation (*56*), including specifically in maize (*2*), we re-scaled ancient mutation loads to correct for potential biases introduced by non-random missing data. For each ancient genome overlapping at least 10,000 scorable SNPs, we recalculated genome-wide mutation loads for all modern samples using only SNPs present in the ancient sample, as well as the load score for the ancient sample, and culled the top and bottom 20% of modern estimates to conservatively remove outliers. We then performed a procrustes transform using the R package MCMCpack (*50*, *57*) to refit the mutation load panel from the SNP-restricted modern datasets to the reference set of complete-genome mutation load estimates. We then added the ancient sample to the SNP-restricted dataset, applied the dilation and translation values from the procrustes transform to the complete set, and used the resulting re-scaled value as the corrected mutation load score for ancient samples.

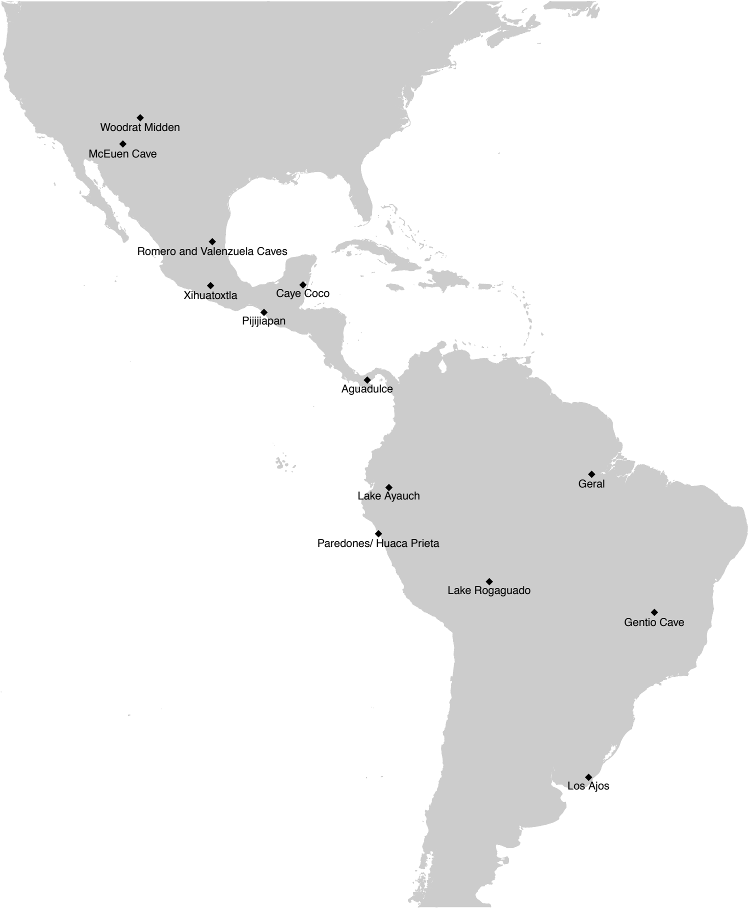
Domestication gene ancestry enrichment

We used a perl script (Dryad: /scripts/dStat.pl) for estimating *D*-statistics and standard error using an unweighted jackknife procedure in 5Mb blocks from SNP data in PLINK format. We first estimated genome-wide D-statistics in the form $D(((X, Y), \textit{parviglumis}), \textit{Tripsacum})$ for all pairs of South American (X; Andean/Pacific plus Lowland ancestry ≥99%) vs. non-South American (Y; Andean/Pacific plus Lowland ancestry <1%) maize genomes compared with 11 *parviglumis* genomes (D_{WG}). We then estimated D-statistics on a subset of the genome within 10kb of 186 genes previously documented as having been involved during the evolution of domestication (*58*) (D_{dom}), and identified sets of individuals whose genome-wide and domestication-gene *D*-statistics differed significantly—non-overlapping at the level of 2 standard errors in each test estimated using an unweighted block jacknife. Incongruent D_{dom} and D_{WG} statistics signify an enrichment of teosinte ancestry associated with domestication genes in either genome X ($D_{dom} < D_{WG}$) or genome Y ($D_{dom} > D_{WG}$).

f_3 and f_4 statistic estimation

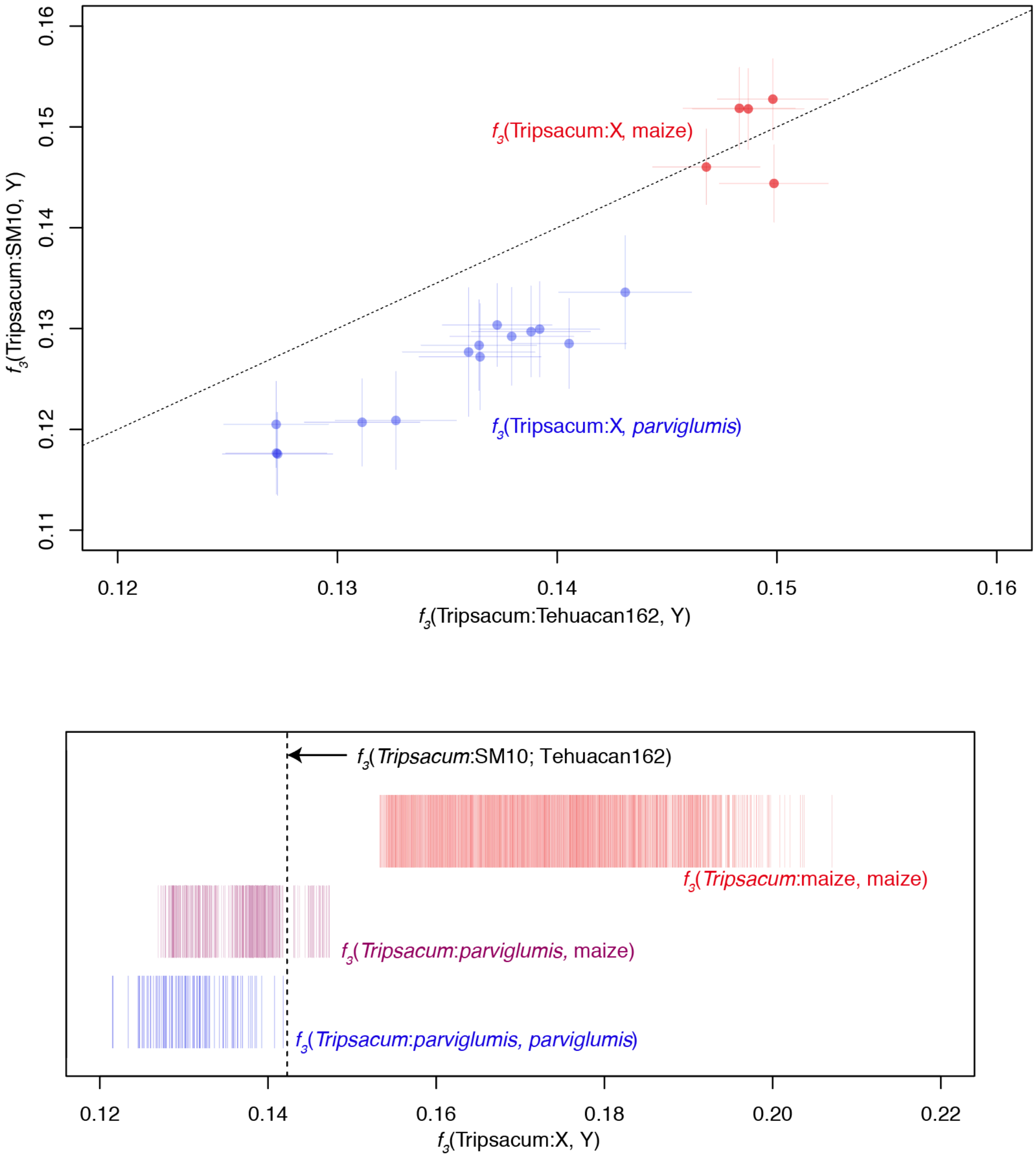
We used an in-house perl script (Dryad: /scripts/plink2freq.pl) to estimate sitewise allele frequency for input populations, and two other in-house perl scripts (Dryad: /scripts/f3.pl ; /scripts/f4.pl) to estimate f_3 and f_4 statistics and standard error using an unweighted jackknife procedure in 5Mb blocks from SNP data in plink format, following the equations in (*17*).

Fig. S1.



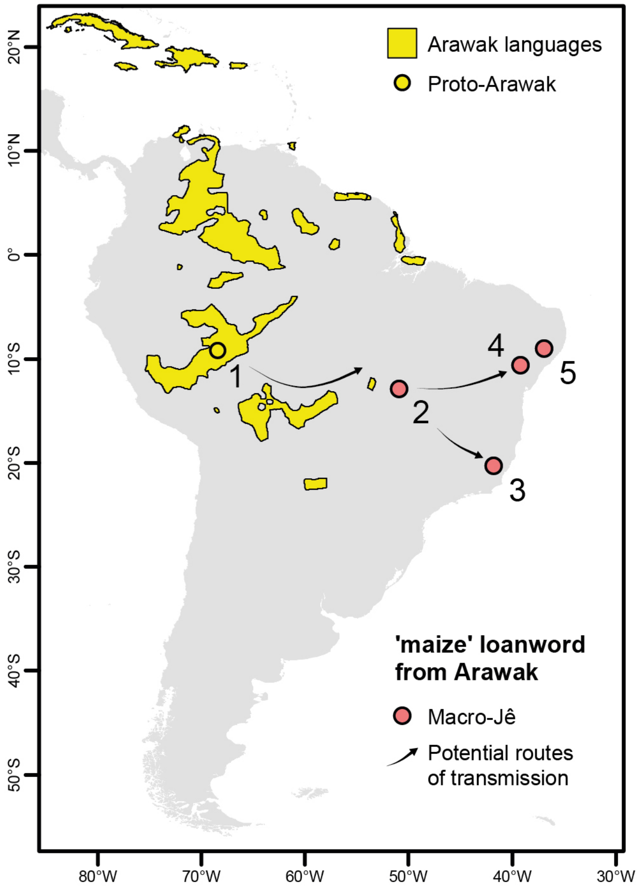
Early regional archaeological maize remains. Named sites correspond with dates on Figure 1, and additional information including type of remains and full references on Table S1.

Fig. S2.



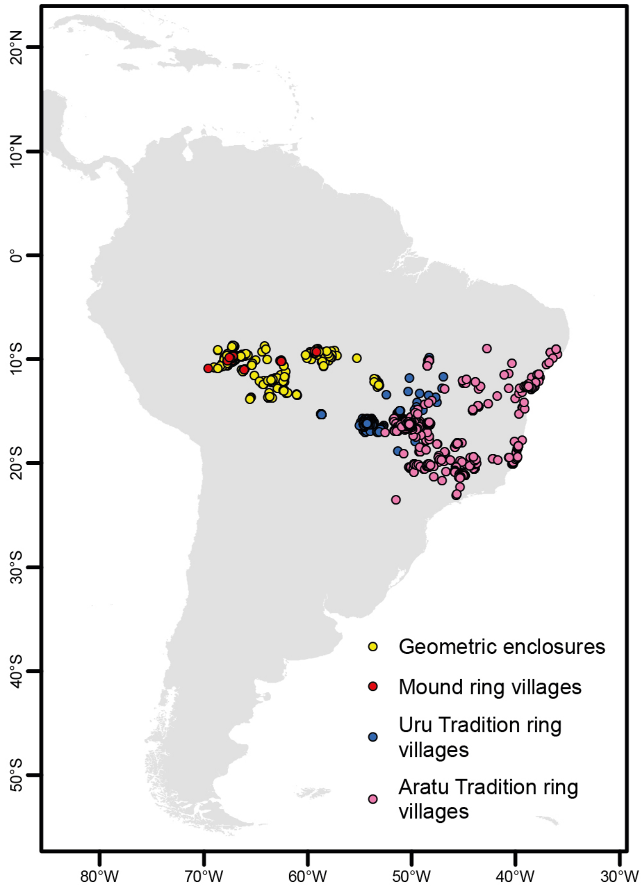
Outgroup-*f_3* comparison of two ancient partially domesticated maize samples from the Tehuacan Valley—SM10 (*3*) and Tehuacan162 (*2*). Above, Tehuacan162 is significantly more *parviglumis*-like than SM10. Error bars are three standard errors estimated using an unweighted block jackknife. Below, vertical bars are all vs. all maize and teosinte comparisons organized as pairs of maize, maize and *parviglumis*, and pairs of *parviglumis*. SM10 and Tehuacan162 show allele sharing typical of a maize-*parviglumis* comparison.

Fig. S3.



Distribution of languages of the Arawak family and ‘maize’ loanwords into other families. 1) Proto-Arawak ***mariki** (*59*); 2) Karajá **maki**; 3) Puri ***maky*** and Coroado ***maheky***; 4) Kipeá **masiki** and Dzubukuá **madiki**; 5) Yatê **máltʃi**. Arawak homeland as per (*29*).

Fig. S4.



Distribution of archaeological traditions discussed in the main text, data from the Brazilian National Register of Archaeological Sites (*60*).

Table S1.

Site Name	Latitude	Longitude	cal B.P.	Reference	Evidence
Lake San Pablo	0.22	-78.22	4900	(*61*)	pollen
Lake Ayauch	-3.04	-78.03	6000	(*62*)	pollen
Lake Sauce	-6.7	-76.21	6320	(*10*)	pollen
Abeja	-0.57	-72.4	5500	(*63*)	pollen
Huaypo	-13.4	-72.13	2800	(*64*)	pollen
Lake Gentry	-12.33	-68.87	3630	(*65*)	pollen
Lake Rogaguado	-13	-65.93	6500	(*11*)	pollen
Parmana	7.86	-65.77	1600	(*66*)	macro
Monte Castelo	-12.55	-63.09	4310	(*67*)	phytoliths
Geral	-1.64	-53.59	4030	(*68*)	pollen
Gentio Cave	-16.25	-46.03	3770	(*69*)	macro
Los Ajos	-33.7	-53.96	3730	(*70*)	phytoliths
Waynuna	-15.27	-72.75	4000	(*71*)	phytoliths and starch
Paredones/ Huaca Prieta	-7.93	-79.29	6700	(*9*)	macro, phyotoliths and starch
Chavin	-9.59	-77.18	2800	(*72*)	macro, stable isotopes
Real Alto	-2.37	-80.72	4750	(*73*)	phytoliths and starch
Xihuatoxtla	18.32	-99.53	8750	(*1*)	phytoliths and starch
Caye Coco	18.4	-88.39	6655	(*74*)	starch
Laguna Martínez	10.6	-85.35	5512	(*75*)	pollen
Aguadulce	8.33	-80.64	7746	(*76*)	phytoliths
Pijijiapan	15.49	-93.09	6589	(*77*)	phytoliths and pollen
Lake Yojoa	14.94	-88.02	5464	(*78*)	pollen
Romero and Valenzuela Caves	22.97	-99.32	4363	(*79*)	macro
Chaco Canyon	36.08	-108.01	4383	(*80*)	pollen
McEuen Cave	33.31	-110.1	4030	(*7*)	macro

Regionally early maize remains shown in Figure 1, Figure 3, and Figure S1.

Table S2.

Reference number	Altitude (m above sea level)	Collection site	AGE (C14 BP)	Calibrated Date BP (95.4%)	Latitude	Longitude
Z2	700	Peruaçu Valley - Januaria - Boquete Cave	570 +- 60	650-490	-15.00	-44.00
Z6	700	Peruaçu Valley - Januaria - Lapa da Hora Cave	630+- 60	660-515	-15.00	-44.00
Z61	200	Site EC-11, lower Santa valley, Ancash, Peru	800 +/- 30	730-660	-8.53	-78.34
Z64	3990	Site Cho9 Machay D, Chorrillos, Ancash, Peru	630 +/- 30	646-587 (58.7%) 573-535 (36.7%)	-9.06	-77.56
Z65	220	Site H13, lower Ica valley, Ica, Peru	970 +/- 30	920-770	-14.43	-75.34
Z66	2465	Argentina	1010 +/- 30	929-798	-27.34	-66.55
Z67	3700	Site Chayal, Susques, Jujuy, Argentina	100 +/- 30	253-225 (11.8%) 143- present (83.6%)	-23.24	-66.21
Arica4	N/A coastal	Arica, Chile	990 +/- 30	925-790	-18.47	-70.32
Arica5	N/A coastal	Arica, Chile	780 +/- 30	725-655	-18.47	-70.32

Archaeological maize genomes sequenced and analyzed in this study. Date calibration was done in OxCal (*81*) using the SHCal13 calibration curve (*82*).

Table S3.

Lab ID	Country/ Region	Province/ Municipality	Landrace/ race	Latitude	Longitude
C1[b]	Peru/ La libertad	Pacasmayo	Alazan	-6.00	-80.00
C10[b]	Peru/ Lima	Chancay	Huachano	-11.00	-78.00
C11[b]	Peru/ Apurimac	Abancay	Huancavelicano	-14.00	-73.00
C12[b]	Peru/ Ica	Nazca	Iqueño	-15.00	-75.00
C13[b]	Peru/ Arequipa	Arequipa	Kculli	-17.00	-72.00
C14[b]	Peru	Lambayeque	Mochero	-6.00	-80.00
C15[b]	Peru/ Ayacucho	Huanta	Morocho	-13.00	-74.00
C17[b]	Peru/ Amazonas	Luya	Morocho Cajabambino	-6.00	-78.00
C18[b]	Peru/ La libertad	Pacasmayo	Pagaladroga	-7.00	-80.00
C19[b]	Peru/ Ancash	A. Raymondi	Pagaladroga	-9.00	-77.00
C20[b]	Peru/ Apurimac	Abancay	Paro	-13.00	-73.00
C21[b]	Peru/ San Martin	Huallaga	Piricinco	-7.00	-77.00
C22[b]	Peru/ Apurimac		Piscorunto	-14.14	-72.49
C23[b]	Peru/ Ancash	pallasca	Rabo de Zorro	-8.00	-78.00
C26[b]	Peru/ Puno	Azangaro	Uchuquilla	-14.00	-70.00
C27[a]	Brazil/ Acre	Feijo	Alho	-8.10	-70.18
C28[a]	Brazil/ Rondonia	Guajara minirm	Atiti-Nhae-Kowo - (roxo)	-10.47	-65.20
C29[a]	Brazil/ Parana	Francisco Beltrao	Avati Moroti	-26.04	-53.04
C3[b]	Peru/ Apurimac	Antabamba	Chullpi	-14.00	-73.00
C30[a]	Paraguay		Avati Moroti guapi	-22.00	-58.00
C31[a]	Brazil/ Para	Portel	Batin	-1.58	-50.48
C32[a]	Brazil/ Sao Paulo		Caigang	-22.00	-51.00
C33[a]	Brazil/ Para	Prainha	Cateto	-1.50	-53.27
C34[a]	Brazil/ Para	Altamira	Cristal	-3.15	-52.13
C35[a]	Brazil/ Maranhao	Bela Vista	Dente de Burro	-3.45	-45.17
C36[a]	Brazil/ Acre	Tarauacá	Entrelaçado	-8.09	-70.46
C38[a]	Brazil/ Sao Paulo		Complexo Guanani	-23.00	-46.00
C39[a]	Brazil/ Rondonia	Guajara minirm	Mole	-10.47	-65.20
C40[a]	Brazil/ Acre	Xapuri	Palha Roxa Acreana	-10.40	-68.25
C42[a]	Brazil/ Mato Grosso Sul	Amambai	Pipoca Roxa	-23.06	-55.13
C43[a]	Brazil/ Para	Sen Jose Porfilio	Pontinha	-2.35	-51.56
C44[a]	Brazil/ Rondonia	Guajara minirm	Semi-Dent	-10.47	-65.20
C45[a]	Brazil/ Mato Grosso	Parque Indigena Xingu, Waura village	Xingu- Waura	-12.14	-53.34
C46[a]	Brazil/ Goias	Ilha bananal, Caraja Village	Caraja	-11.34	-50.40
C47[a]	Brazil/ Roraima	Boa Vista		2.48	-60.40
C48[a]	Brazil/ Roraima	Caracarai		1.49	-61.07
C5[b]	Peru/ Junin	Huancayo	Confite puntiagudo	-12.00	-75.00
C6[b]	Peru/ Apurimac	aymaraes	Cuzco	-13.00	-72.00
C8[b]	Peru/ Madre de Dios	Tahuamanu	Enano	-11.00	-70.00

C9[b]	Peru/ Apurimac	Andahuaylas	Granada	-14.00	-73.00

Maize landraces analyzed in this study. Superscript “a” denotes samples from the Embrapa germplasm collection, “b” denotes samples obtained from PCIM (El Programa del Maíz, Peru).

Table S4.

Site	cal BP	Lat	Long	Type	Location	Reference
Severino Calazans	2350-2160	-10.03	-67.51	Geometric enclosure (*geoglyph*)	SW Amazon (Acre, Brazil)	(*83*)
Tumichucua	2295-1870	-11.15	-66.16	Geometric enclosure	SW Amazon (Riberalta, Bolivia)	(*84*)
MT-FX-13	930-740	-12.38	-53.19	Geometric enclosure (*Xinguano fortification*)	S Amazon (Mato Grosso, Brazil)	(*26*)
Bella Vista	735-675	-13.26	-63.71	Geometric enclosure (*ring ditch*)	SW Amazon (Baures, Bolivia)	(*85*)
Mt04	625-505	-9.81	-57.82	Geometric enclosure	S Amazon (Mato Grosso, Brazil)	(*25*)
Sol de Campinas	930-795	-10.06	-67.31	Mound ring village	SW Amazon (Acre, Brazil)	(*86*)
MT-SL-29	1240-930	-16.12	-53.60	Ring village (*Uru*)	Cerrado (Mato Grosso, Brazil)	(*87*)
GO-CP-02	1275-920	-16.96	-51.58	Ring village (*Aratu*)	Cerrado (Goiás, Brazil)	(*88*)
Guipe	1175-760	-12.70	-38.45	Ring village (*Aratu*)	Atlantic coast (Bahia, Brazil)	(*89*)
Monsarás	1050-810	-19.52	-39.88	Ring village (*Aratu*)	Atlantic coast (Espírito Santo, Brazil)	(*90*)
GO-NI-06	1225-785	-16.47	-47.98	Ring village (*Aratu*)	Cerrado (Goiás, Brazil)	(*91*)
Sapucaí Phase	955-670	-21.08	-45.36	Ring village (*Aratu*)	Cerrado (Minas Gerais, Brazil)	(*92*)
Estiva 2	890-560	-10.68	-48.44	Ring village (*Aratu*)	Cerrado (Tocantins, Brazil)	(*93*)

Early regional dates associated with archaeological traditions discussed in the text, and mapped on Figure 3.