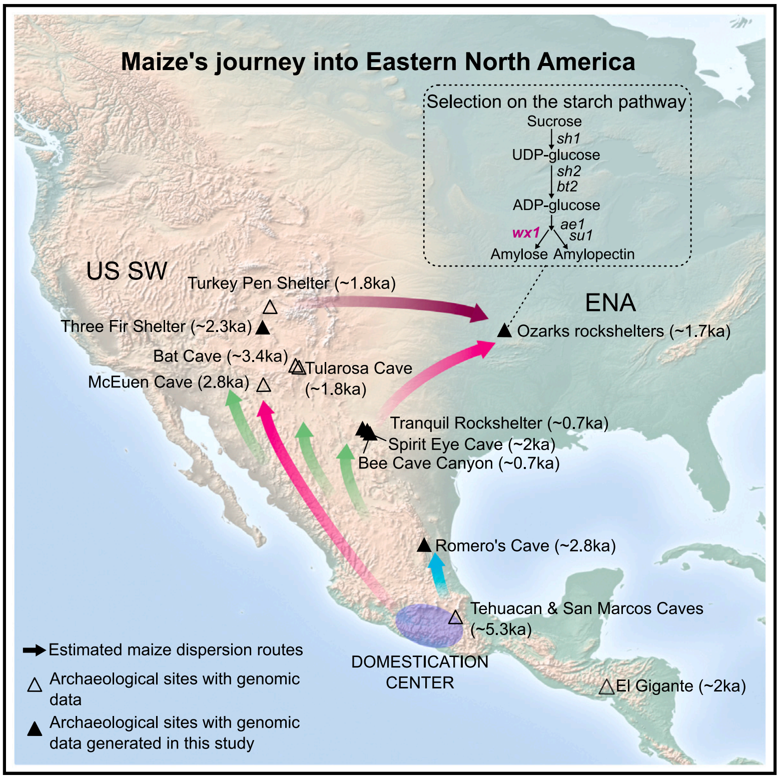
Cell

Article

The genomic origin of early maize in eastern North America

Graphical abstract


Maize's journey into Eastern North America
Selection on the starch pathway
Sucrose
sh1
UDP-glucose
sh2
bt2
ADP-glucose
wx1
ae1
su1
Amylose Amylopectin
US SW
Turkey Pen Shelter (~1.8ka)
ENA
Three Fir Shelter (~2.3ka)
Ozarks rockshelters (~1.7ka)
Bat Cave (~3.4ka)
Tularosa Cave (~1.8ka)
McEuen Cave (2.8ka)
Tranquil Rockshelter (~0.7ka)
Spirit Eye Cave (~2ka)
Bee Cave Canyon (~0.7ka)
Romero's Cave (~2.8ka)
Tehuacan & San Marcos Caves (~5.3ka)
DOMESTICATION CENTER
El Gigante (~2ka)
Estimated maize dispersion routes
Archaeological sites with genomic data
Archaeological sites with genomic data generated in this study

Highlights

- Ancient maize genomes reveal eastward dispersal into eastern North America
- Modern lineage of Northern Flint linked to 900-year-old maize from the Ozarks
- Starch metabolic pathway was repeatedly under selection during maize domestication

Authors

Jazmín Ramos-Madrigal, Gayle J. Fritz, Bryon Schroeder, ..., Shyam Gopalakrishnan, M. Thomas P. Gilbert, Nathan Wales

Correspondence

jazmin.madrigal@sund.ku.dk (J.R.-M.), nathan.wales@york.ac.uk (N.W.)

In brief

Ancient maize genomes reveal recurrent northward movements from its domestication center, culminating in two dispersals of US Southwest maize into the Ozark rockshelters in eastern North America. The 1,000-year-old maize genomes from the Ozarks provide insights into the origin of Northern Flints and the selection history of the *wx1* gene, part of the starch metabolic pathway.

Ramos-Madrigal et al., 2025, Cell *188*, 33–43
January 9, 2025 © 2024 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.
https://doi.org/10.1016/j.cell.2024.11.003

Article

The genomic origin of early maize in eastern North America

Jazmín Ramos-Madrigal,[1,14,*] Gayle J. Fritz,[2] Bryon Schroeder,[3] Bruce Smith,[4] Fátima Sánchez-Barreiro,[1] Christian Carøe,[1] Anne Kathrine Wiborg Runge,[5] Sarah Boer,[5] Krista McGrath,[5] Filipe G. Vieira,[6] Shanlin Liu,[7,8] Rute R. da Fonseca,[9] Chunxue Guo,[8] Guojie Zhang,[10] Bent Petersen,[1,11] Thomas Sicheritz-Pontén,[1,11] Shyam Gopalakrishnan,[1,12] M. Thomas P. Gilbert,[1,13] and Nathan Wales[5,*]

[1]Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, 1353 Copenhagen, Denmark
[2]Department of Anthropology, Washington University in St. Louis, St. Louis, MO 63130, USA
[3]Center for Big Bend Studies, Sul Ross State University, Alpine, TX 79832, USA
[4]Program in Human Ecology and Archaeobiology, Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA
[5]Department of Archaeology, University of York, York 10 5DD, UK
[6]Section for Geogenetics, Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark
[7]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[8]BGI-Shenzhen, Shenzhen 10059378, China
[9]Center for Macroecology, Evolution and Climate (CMEC), Center for Global Mountain Biodiversity, Globe Institute, University of Copenhagen, 2100 Copenhagen, Denmark
[10]Center of Evolutionary & Organismal Biology, Zhejiang University School of Medicine, Hangzhou 310058, China
[11]Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Bedong, Kedah 08100, Malaysia
[12]Bioinformatics, Department of Health Technology, Technical University of Denmark, 2800 Copenhagen, Denmark
[13]University Museum, Norwegian University of Science and Technology, 7012 Trondheim, Norway
[14]Lead contact
*Correspondence: jazmin.madrigal@sund.ku.dk (J.R.-M.), nathan.wales@york.ac.uk (N.W.)

https://doi.org/10.1016/j.cell.2024.11.003

SUMMARY

Indigenous maize varieties from eastern North America have played an outsized role in breeding programs, yet their early origins are not fully understood. We generated paleogenomic data to reconstruct how maize first reached this region and how it was selected during the process. Genomic ancestry analyses reveal recurrent movements northward from different parts of Mexico, likely culminating in at least two dispersals from the US Southwest across the Great Plains to the Ozarks and beyond. We find that 1,000-year-old Ozark specimens carry a highly differentiated *wx1* gene, which is involved in the synthesis of amylose, highlighting repeated selective pressures on the starch metabolic pathway throughout maize's domestication. This population shows a close affinity with the lineage that ultimately became the Northern Flints, a major contributor to modern commercial maize.

INTRODUCTION

The abundance of archaeobotanical remains coupled with isotopic evidence indicating increased human consumption of C4 plants,[1] shows that by 1,000 years before present (years BP), maize had emerged as a major crop in eastern North America (ENA). The maize cultivated in this region (Northern Flints and Southern Dent landraces) would eventually become a key contributor to modern commercial maize.[2,3] However, our understanding about the way in which maize came to dominate ENA agriculture, including the timing, dispersal routes, and history of selection, remains limited.

The earliest evidence for maize in ENA comes from phytoliths and starch grains in northeastern North America ca. 2,200 years BP.[4–6] But due to the sporadic appearance of maize in the archaeological record during this early period,[7,8] it is unclear whether maize arrived in ENA once or through multiple pulses, and the routes by which it traveled are also uncertain. Isozyme evidence and morphology of modern ENA maize show it is most closely related to landraces from the US Southwest (US SW),[9,10] suggesting transportation across the Great Plains. However, no maize macroremains or evidence of its cultivation from the time of maize arrival in ENA (~2,200 BP) has been found along potential dispersal routes between the US SW and ENA.[11] Another proposed dispersal route follows the so-called "Gilmore Corridor," which stretches from northeast Mexico across the Gulf coastal plains of Texas (Figure 1A),[12] yet definitive evidence of sustained human interaction or exchange of products

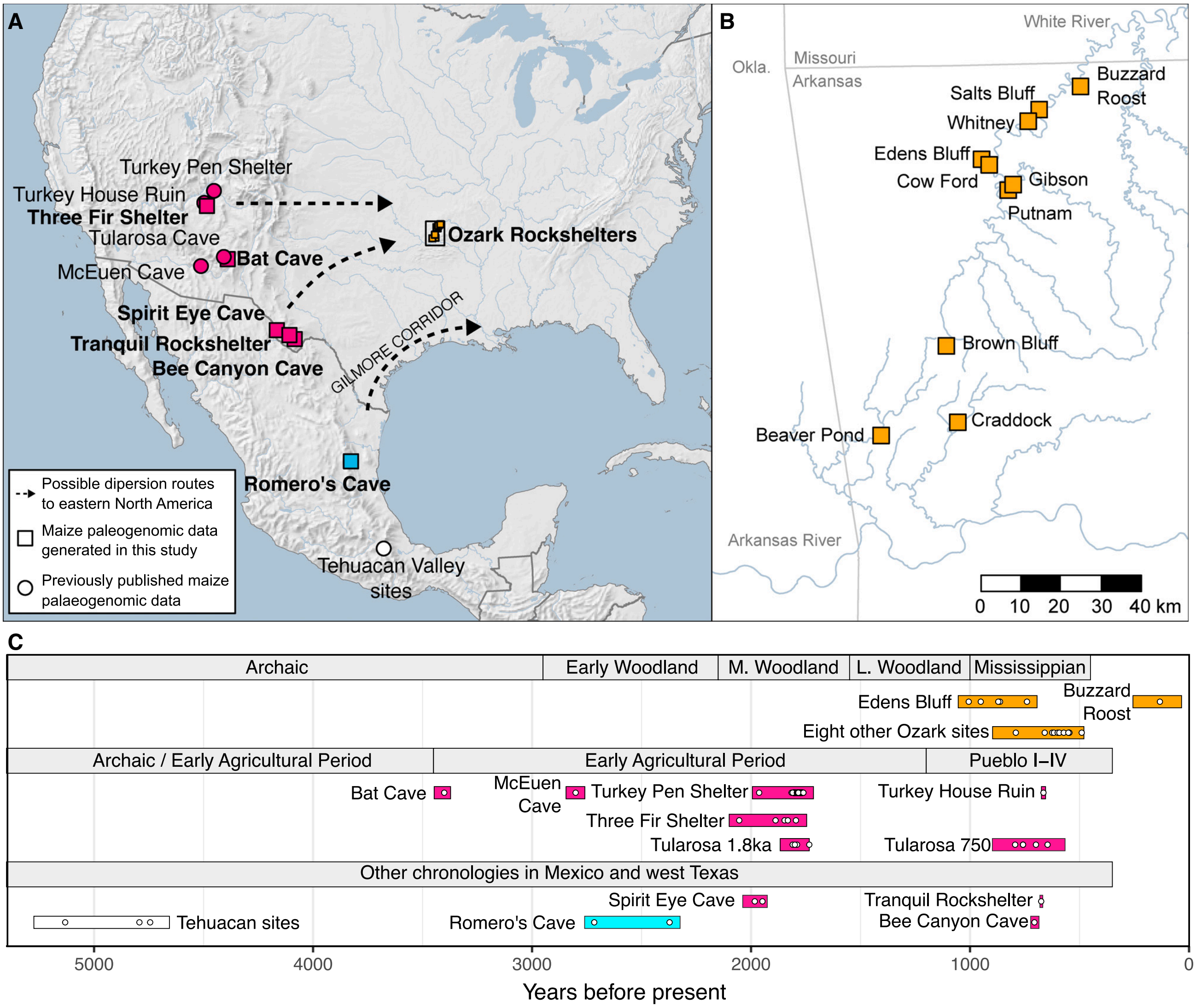


Figure 1. Geographic and temporal context of archaeological maize from North America

(A) Sites for which maize paleogenomic data is available. Genomic data were generated in this study for sites listed in bold and with square icons. Hypothesized routes for the movement of maize into ENA are indicated with arrows.

(B) Inset from (A) depicting the Ozark rockshelters, which are situated along the White River and tributaries of the Arkansas River in northwestern Arkansas.

(C) Chronology of sites and associated archaeological periods in ENA, the US SW, and Mexico and Texas (M., middle; and L., late). Site ages are based on calibrated radiocarbon dates on maize (white circles) and shown as a range for the 68% confidence interval for the oldest and youngest specimens. See Figure S1A for individually calibrated radiocarbon dates and Bayesian modeled dates of select sites.

See also Figure S1.

between northeast Mexico and ENA through this corridor is scarce.[13] Nevertheless, by the time maize arrived in ENA, it was already on its way of becoming an important part of the subsistence economy in both the US SW and northeastern Mexico,[11,14] making both dispersal routes plausible.

Two of the most intriguing questions regarding the arrival of maize in ENA are why it took so long for maize to reach ENA when it had been present 4,050 years BP in the US SW[15] and why it remained almost invisible in the archaeological record across most of the region until after 1,100 years BP.[7,8] This late introduction cannot be attributed to lack of agricultural expertise, as people in ENA had been farming an array of autochthonous crops since 4,000 years BP, including marshelder (*Iva annua* L.), chenopod (*Chenopodium berlandieri* Moq.), squash (*Cucurbita pepo* ssp. *ovifera* D.S. Decker), and sunflower (*Helianthus annuus* L.).[16] Rather, part of the delayed arrival of maize in ENA could be attributed to the time required for the crop to adapt to local conditions, although prior paleogenomic data demonstrate that some necessary adaptations were already in place 2,000 years BP in potential source regions such as the US SW.[17] Alternative hypotheses for a delayed uptake of maize include a scenario where maize primarily had ceremonial purposes in ENA until 1,200 years BP[18] or that maize farming methods may have been incompatible with cultural traditions for sowing crops of the earlier Eastern Agricultural Complex (EAC).[19]

The history of selection of maize in ENA could provide important insights on how the crop responded to the local conditions and whether certain traits were favored by farmers. Some researchers have suggested that centuries of adaptation to the short growing season and cold winters of ENA may have led to the development of the local Northern Flints landraces.[20] However, others suggest that high-yield maize was introduced at a later point in time, leading to a rapid intensification of maize agriculture, potentially with links to the development of the Mississippian cultural tradition[21] and eventually the abandonment of the EAC "lost crops" like marshelder, chenopod, maygrass (*Phalaris caroliniana* Walt.), little barley (*Hordeum pusillum* Nutt.), and erect knotweed (*Polygonum erectum* L.).[22]

To elucidate the contentious history of maize in ENA, we generated whole-genome sequencing data from 32 archaeological maize samples, ranging in age from 3,390 years BP to the present and in depth of coverage from 0.01 to 6.83× (mean ~1.36×) at the accessible regions of the maize genome (Figure 1; Tables S1 and S2). Twenty-nine of the sequenced samples have been radiocarbon dated, including six dates generated for this study (Table S1). Among the sequenced samples, eighteen maize cobs derive from ten archaeological sites in the Ozark region of northwest Arkansas (Figures 1A and 1B, orange squares). The Ozark bluff sites are renowned for their preservation of desiccated plant macrofossils, many of which are well suited to genome-wide analyses.[23,24] Radiocarbon dates for the Ozark maize samples span from ~1,000 years BP to the present (Figures 1C and S1A),[25] encompassing the period of the rapid uptake of maize agriculture in ENA. To contextualize our findings, we sequenced ancient maize genomes from other regions representing potential ancestry sources for maize in ENA. Seven of the sequenced samples come from the Tranquil Rockshelter, Bee Cave Canyon site, and Spirit Eye Cave in West Texas (Figure 1A, pink squares) and Romero's Cave in northeast Mexico (Figure 1A, turquoise square), two regions largely unexplored using paleogenomic data. Lastly, we resequenced six samples from the Three Fir Shelter (TFS),[26] located in the temperate US SW, and of one sample from Bat Cave[27] in the US SW (Figure 1A, pink square).

RESULTS AND DISCUSSION

Ancient maize dataset

We combined the 32 ancient genomes sequenced in this study with a whole-genome dataset comprising 94 domesticated maize landraces,[28–31] 23 wild maize samples,[28] and 55 ancient maize genomes[17,27,30,32,33] (Table S3). The authenticity of our ancient maize genomic data was confirmed by assessing the ancient DNA damage patterns and DNA fragment length distributions (Table S2). Additionally, we evaluated the potential correlation of substitution patterns between datasets originating from different sequencing platforms (BGI500 and Illumina) and concluded that our results are not affected by such differences (Figures S1B and S1C; Table S4).[34]

Ancestry at the potential regions of origins for ENA maize

We used multidimensional scaling (MDS) and model-based clustering analyses to explore the genetic affinities between the ancient and modern maize genomes in the dataset. The MDS analysis recovers the north-south (dim 1) and west-east (dim 2) ancestry axes that describe maize genetic diversity (Figures 2A and S2A).[27,35] Similarly, the clustering analysis assuming six ancestry components identifies previously described geographic groups: US SW, West Mexican Highland (West Mexico, from hereafter), Pan-American (comprising mainly East Mexico, Central, and northern South America),[30] Andean, South American lowland, and the wild progenitor of maize, teosinte (Figures 2A and S2B).

To establish a framework for inferring the origins of ENA maize, we first characterized the genomic ancestry of maize from likely regions of origin, namely northeastern Mexico and the US SW. In northeastern Mexico, we found that the ~2,400-year-old maize from Romero's Cave is closely related to maize from the Pan-American group (see also Figures S2C and S3C). Today, the distribution of Pan-American maize cultivars spans from northern Mexico to lowland South America and has been identified in Central America 2,000 years BP.[36] Therefore, these results show that by ~2,500 years BP the extension of this lineage had reached Romero's Cave and suggest that maize from this same ancestry has been cultivated in northeastern Mexico for at least two millennia. In the US SW, ancient maize, including the ~2,000-year-old TFS maize sequenced in this study, clusters together with modern US landraces. In comparison with US SW maize, ancient maize from West Texas shows a different admixture pattern. Our results show that it carries ancestry from both the US SW and Mexican maize, given the placement of the Bee Cave Canyon (~700 years BP), Tranquil Rockshelter (~690 years BP), and Spirit Eye Cave (~2,000 years BP) genomes intermediate between these two groups in the MDS.

An eastward dispersal route into ENA from the US SW

We investigated the genomic ancestry of archaeological maize from the Ozark sites (ENA) in the context of its potential regions of origin. Both MDS and clustering analyses show that the ~1,000- to 440-year-old Ozark maize and ancient US SW maize have similar ancestry components and are adjacent to each other along the west-east variation axis of the MDS plot (Figure 2A), thereby supporting a US SW origin of ENA maize. Our MDS analysis also shows that modern Northern Flint accessions are placed closest to the archaeological Ozark maize, suggesting that the Ozark population was either fundamental in its creation or at least a part of the same lineage. This result is supported by the clustering analysis (Figure S2B) and outgroup-based f_3-statistics (Figure S2C). In contrast to the older Ozark maize, the ancestry profile of the youngest Ozark sample (Buzzard Roost; 275–8 years BP) reveals a mixture of not only US SW but also Pan-American maize ancestries, similar to the constitution of modern Southern Dent landraces (Figures 2A, S3B, and S3F).

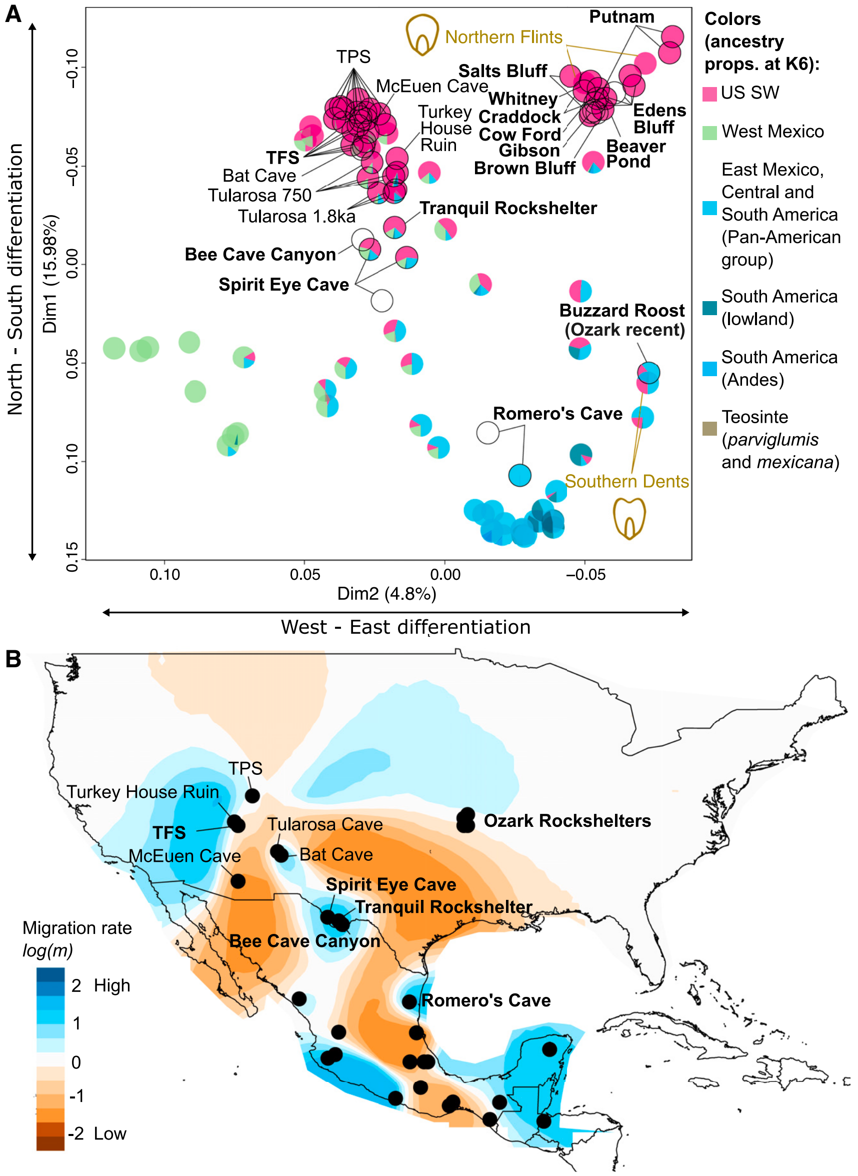


Figure 2. Spatiotemporal patterns of maize ancestry in North America uncover an eastward dispersal route into ENA

(A) MDS analysis based on whole-genome data from ancient and modern maize from North, Central, and northern South America (percentage of variance explained in parenthesis). Pie charts represent each sample's ancestry proportions estimated with ADMIXTURE assuming 6 ancestry components. Archaeological site names of ancient samples are shown in black, sites with samples sequenced in this study are indicated in bold, and names for relevant modern samples are shown in yellow. A black outline indicates ancient samples. Samples not included in the ADMIXTURE analyses are depicted as empty circles.

(B) Map showing EEMS's effective migration surface based on genetic and geographic distances for ancient and modern maize samples. Cooler and warmer colors show regions with high and low estimated migration rates, respectively. Black dots show the approximate geographic locations of the samples included in the analysis. Site names are indicated for ancient samples. Archaeological sites with samples sequenced in this study are shown in bold.

See also Figure S2.

The connections between the archaeological Ozark maize and modern maize landraces are particularly noteworthy because hybrid crosses of Northern Flints and Southern Dents created Corn Belt Dent, the principal maize cultivated in the US today.[2] Northern Flints—a group of hardy maize landraces that yield kernels with a hard, "flint-like" outer layer—were distributed throughout ENA at the time of European contact.[2] In contrast, the Southern Dents—landraces producing kernels with an indentation due to high soft starch—had a more restricted geographic range at the time of contact. Our results suggest that ancient Ozark maize originally derives from an eastward dispersion of a lineage originating in the US SW that eventually gave rise to Northern Flints in ENA. Our data are also consistent with the hypothesis that Southern Dents have a relatively recent origin, involving crosses between local ENA lineages with maize introduced from Mexico by Spanish traders in the past 500 years.[37]

To further investigate the diffusion of maize into ENA, we used EEMS[38] to estimate migration surfaces relating ancient maize samples and identify potential gene-flow barriers and routes (Figures 2B and S2D). To focus on early maize movements, we excluded modern landraces from the US that carry recent admixture (Figure S2B). The estimated migration surface identifies the region overlapping with the Central Mexican Plateau as the primary route of gene flow between Mexico and the US SW, in agreement with previous results.[27] Counter to the hypothesis that maize was transported through the Gilmore Corridor of Texas, we estimate low migration rates between East Mexican and ENA maize, leaving the central and southern Great Plains as the most likely initial migration route.

Recurrent northward movements of maize into the US SW

To fully understand maize dispersal to the US SW and later to ENA, it is essential to characterize the dynamics of maize movement north from its domestication center in Southwestern Mexico.[39,40] We used *f*-statistics-based admixture graphs[41] to reconstruct the phylogenetic relationships and potential admixture events between maize groups in North America. For this analysis, we grouped samples according to their ancestry profiles as

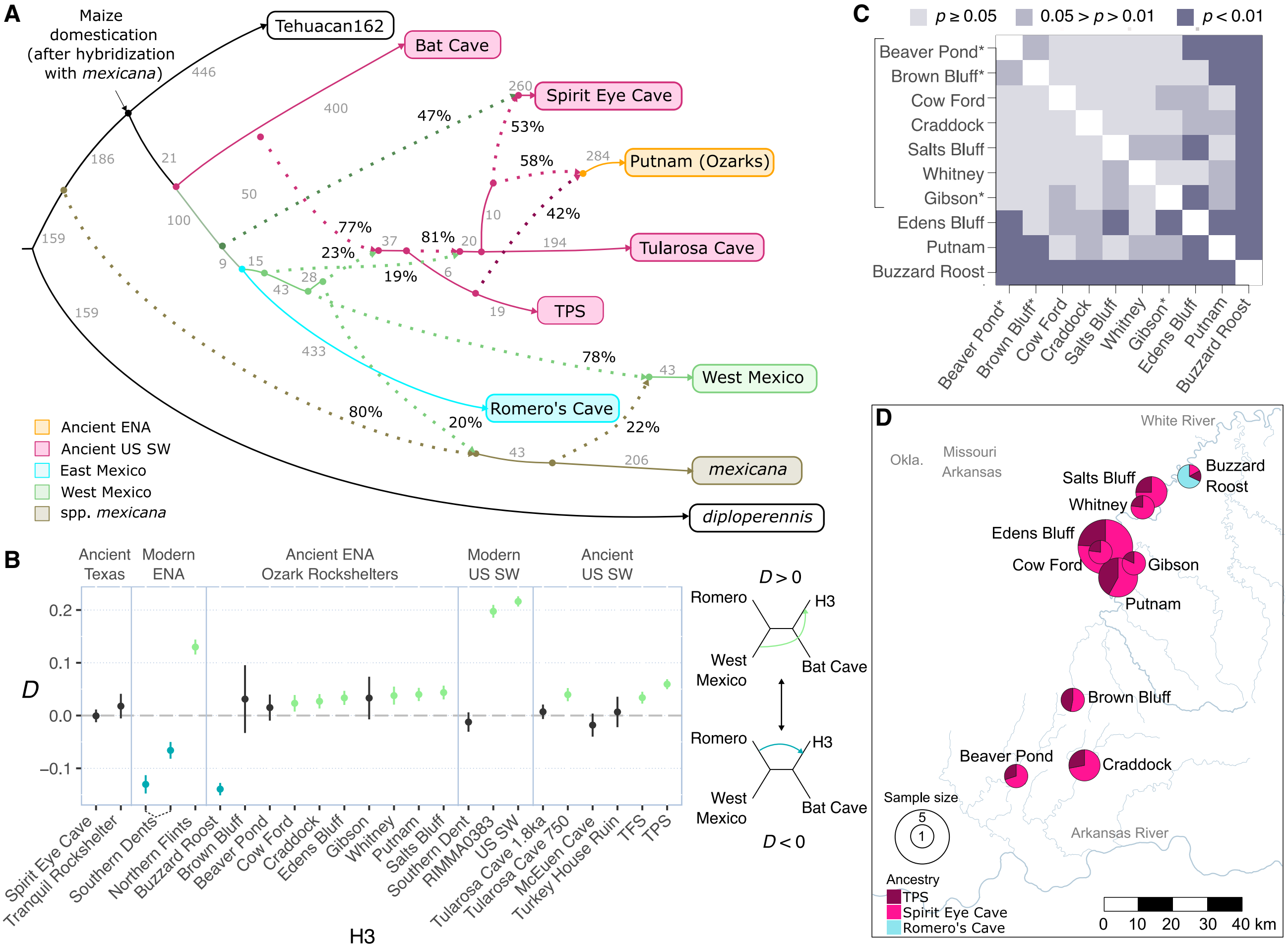


Figure 3. Phylogenetic relationships and ancestry composition of North American maize shows genetic heterogeneity in ancient maize from ENA

(A) *f*-statistic-based admixture graph showing the relationships among ancient maize in the US. Colors indicate the main ancestry groups identified with ADMIXTURE. Continuous lines indicate phylogenetic relationships between samples, with gray numbers showing the drift. Dotted lines indicate admixture events, with percentages showing the proportion derived from each lineage.

(B) Error-corrected *D*-statistic testing for gene flow between West Mexican or the ∼2,400-year-old Romero's Cave maize and the maize from the US (considering 22% *mexicana* ancestry in West Mexico maize). Individual points show the value of *D* obtained from each test. Error bars show 3.3 standard errors (SE) estimated through a block jackknife procedure.

(C) Heatmap showing the *p* values obtained for a qpWave analysis testing whether samples from pairs of Ozark sites derive from a single migration wave. Significant *p* values indicate pairs of sites for which we reject a single migration wave. Bracket shows the group of Ozark sites for which we cannot reject a single migration wave. *Samples with missing data above 90%.

(D) Pie charts showing the proportions of each of the three ancestries present in the Ozark maize (TPS = dark pink, Spirit Eye Cave = pink, and Romero's Cave = light blue), estimated using the admixture graph in (A) and the different Ozark sites. Individual pie charts show the ancestry proportions for each site and the size of the circles indicates the number of samples.

See also Figure S3.

inferred using MDS, clustering analysis, and qpWave (Figures S2B and S3A) and selected representatives of the ancestry components in US maize following our model-based clustering results. The best-fitting model recapitulated the basal relationships between the major maize lineages, showing an early split of the US SW maize, followed by the divergence of East and West Mexico maize lineages,[17,27,32] as well as gene flow between the *mexicana* subspecies and West Mexico maize[35] (Figure 3A). In agreement with previous observations,[17,27] our model also shows that ancient maize from different archaeological sites in the US SW is formed from the mixture between the population represented by the earliest maize genome from the US SW (∼3,390-year-old Bat Cave) and West Mexican maize (Figure 3A). In particular, the ∼2,000-year-old maize from the Spirit Eye Cave derives nearly half of its ancestry from the Mexican maize lineage.

That different groups of ancient maize in the US can be modeled as bearing ancestry from Southern lineages in both East and West Mexico (Figures 3A and S3D) suggests high connectivity between the US SW and northern Mexico. Although it is widely accepted that the area connecting northwest Mexico and the US SW comprises many corridors of exchange of ideas and people, the same cannot be said for northeast Mexico.[13] To explore the extent of gene flow between northern Mexico and the southern US, and to test whether East or West Mexican maize represent the most likely source of ancestry coming into the US SW at different times, we used error-corrected *D*-statistics.[42] Specifically, we tested whether different groups of maize shared more alleles with maize from East or West Mexico (Figure 3B), while accounting for the additional *mexicana* ancestry in West Mexico maize (0%–28%; Figure S3E).[35,43] We find that most modern and ancient maize from the US contains ancestry that is closely related to West Mexico landraces; however, both ENA modern Dent landraces and the recent Ozark sample (Buzzard Roost, 275–8 years BP) are exceptions to this pattern, as East Mexico (Pan-American lineage) maize represents a better source. The admixture patterns can also be observed in the MDS analysis where modern landraces from ENA are placed between the ancient Ozark samples and the East Mexico maize (Figure 2A). The ancient maize from West Texas is an interesting case, given that the Mexican ancestry in the Spirit Eye Cave and Tranquil Rockshelter is genetically equidistant to East and West Mexico maize, suggesting it could derive from a population that was either basal to both two groups or the result of symmetrical admixture from two ancestral populations.

The observation of varying proportions of West Mexican maize ancestry in the US SW maize calls for further genetic and archaeological consideration. From the genetic standpoint, the variable West Mexican ancestry could be explained by genetic heterogeneity of maize entering the US SW, continuous contact between the US SW and southern regions, or a combination of these scenarios. Archaeological records indicate that a number of different Mesoamerican crops entered the US SW at different times over the course of several millennia,[15] consistent with a continuous exchange of products between the two regions. Additionally, linguistic, paleoecological, and archaeological data suggest that maize dispersed from the domestication center in Mexico to the US SW via group-to-group diffusion,[15] which could have facilitated the continuous movement of maize in the region. Although our results provide evidence of multiple introductions of Mexican maize ancestry into ancient US SW maize, the extent of maize movement further south and the origin and distribution of West Mexico ancestry in the past remain to be investigated. Furthermore, the genetic ancestry of ancient maize from West Texas suggests that the area comprising the Central Mexican Plateau might reveal ancient maize bearing yet-undescribed genetic ancestries.

Two distinct ancestries contributed to ENA maize

We next explored the genomic diversity of ancient maize in ENA to test whether the maize genomes from the ten distinct archaeological sites can be traced back to a single or multiple ancestry sources from the US SW. For each pair of Ozark sites, we used qpWave[44] to test whether they formed a clade to the exclusion of maize from the various sites in the US SW, Romero's Cave, and modern landraces from Mexico, Central, and South America (Figure 3C). We reject the idea that maize from different sites in the Ozarks derives from a single stream of ancestry from the US SW ($p < 0.01$); instead, we find four groups with consistent admixture profiles roughly coinciding with their age: Edens Bluff, Putnam, Buzzard Roost, and a fourth group comprising all remaining sites (Figure 3C). These results are in agreement with the best-fitting admixture graphs where Ozark maize is modeled as different mixtures between two US SW lineages closely related to the Spirit Eye Cave and TPS maize (Figure 3D).

We interpret that the observed heterogeneity in ancestry proportions among Ozark maize could be due to either genetic structure in the maize that arrived in ENA or two (or more) independent dispersals into the region: one from upland US SW (ancestry similar to TPS and TFS maize) and a second from lowland US SW (ancestry similar to Tularosa and Spirit Eye Cave maize). Swarts et al. previously found that the ∼1,800-year-old TPS maize was partially adapted for early flowering, necessary for the shorter growing season in upland US SW, and suggested the time gap in maize establishment between lowland and upland US SW was partially due to the delay in this adaptation.[17] Our results showing TPS ancestry in the Ozark maize suggest that this ancestry might have contributed to the introduction of maize to temperate regions in ENA.

Signatures of selection in the starch pathway in ENA maize

Ancient DNA research has identified temporally structured signals of selection throughout maize's domestication history,[17,27,30,32,45] with important inferences on its adaptability and roles in past diets. Considering the major role of Northern Flint landraces in the breeding of Corn Belt Dent and the finding that archaeological Ozark maize represents a close relative and possible ancestral form of Northern Flint landraces, we evaluated which genes were under selection as maize expanded into ENA. The population branch statistic (PBS)[46] was implemented to measure allele frequency differentiation in the ∼1,000- to 440-year-old Ozark maize relative to the maize from the US SW and teosinte. For every gene represented by at least 10 SNP sites in our dataset, we estimated the PBS for the following groups: ancient Ozark samples ($n = 17$), teosinte ($n = 16$), and each of the modern ancestry groups and ancient archaeological sites in the US SW independently ($n = 5–13$; Figures 4A, 4B, S4A, and S4B). We identified four genes that lie above the 99.95 quantile of the PBS distribution showing high differentiation in the Ozark maize compared with the US SW. Given that three of those genes have not been functionally characterized and their high PBS was driven by a single SNP, we focused on the *wx1* gene, where we detected two SNPs with consistently large PBS relative to Tularosa, TFS, or TPS maize from the US SW and teosinte (Figure S4C).

The *wx1* gene is involved in the conversion of ADP-glucose into amylose during starch synthesis and it is one of six key genes involved in the starch pathway[47] (Figure 4C). Most genes involved in this pathway have been previously identified as targets of selection during maize domestication[27,47] and improvement.[48,49] Notably, ancient maize genomes from the US SW showed that *su1* and *ae1* genes, which play a parallel role to *wx1* during starch synthesis, were selected upon arrival to the region.[27,45] Therefore, our results showing that *wx1* was a target of selection in ENA further highlight the importance of

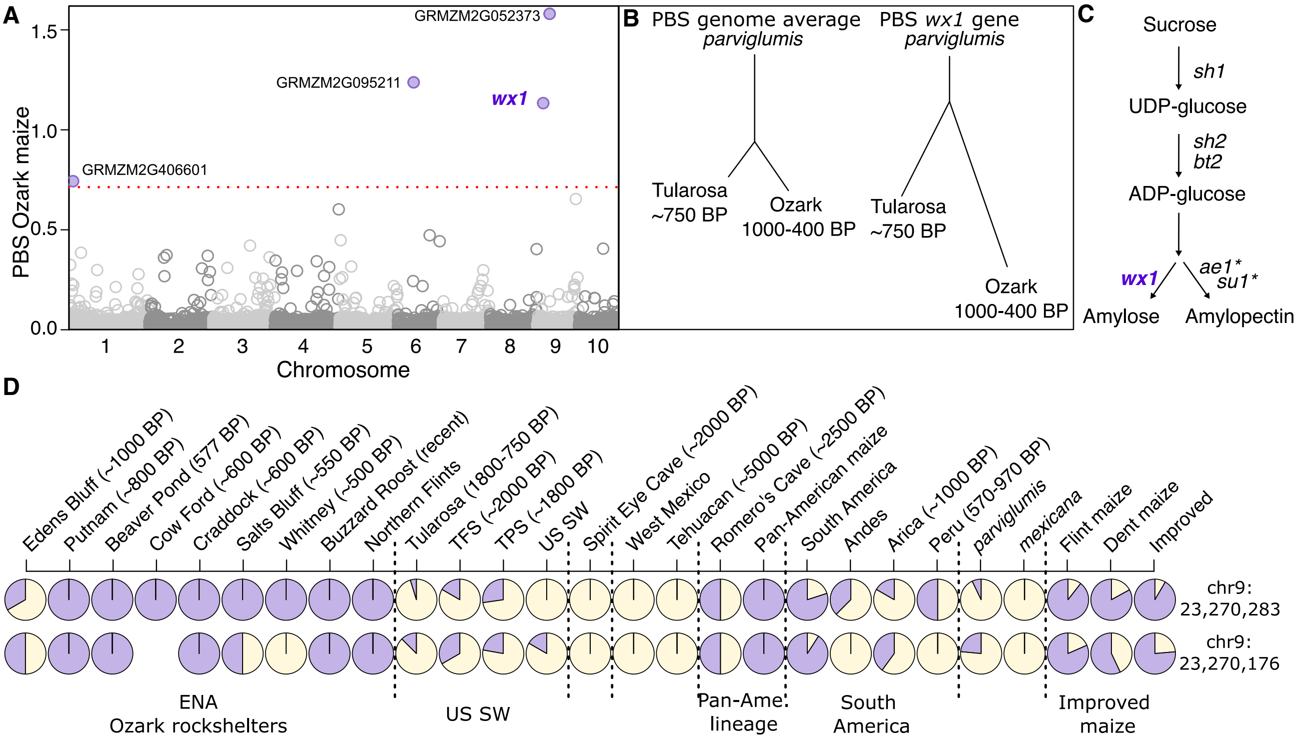


Figure 4. Signatures of selection in the starch pathway in ancient ENA maize

(A) Population branch statistic (PBS) estimated for 6,281 genes in the Ozark maize. The dotted red line shows the 99.95 quantile of the PBS distribution, and names are shown for genes above this cutoff.

(B) Trees showing the average PBS for all genes (left) and for the *wx1* gene (right).

(C) Starch pathway. *Genes previously shown to be targets of selection in ancient US SW maize.

(D) Pie charts showing the allele frequencies at the two highly differentiated SNPs in the *wx1* gene. Purple color indicates the proportion of the derived allele and white indicates the proportion of the ancestral allele.

See also Figure S4.

this metabolic pathway in the domestication history of maize in the US.

The proportions of amylose and amylopectin in maize kernels are important determinants of the kernel's structure, appearance, and texture[50]; thus, the *wx1* gene has been a target of extensive research.[47,49,51] Several mutations reducing or inactivating the function of *wx1* have been characterized that produce a type of maize best known for its low amylose content (waxy maize).[48,49] As one of the highly differentiated SNPs in the Ozark maize was a non-synonymous substitution, we performed a structural analysis[52] to investigate potential functional differences between two possible protein sequences (Figure S4D). Our structural modeling predicted different post translational modifications in the protein variants present in the Ozark and the US SW maize, suggesting a potential functional impact.

Finally, we explored the allele frequency distribution at the two highly differentiated *wx1* SNPs among different groups of maize landraces and improved maize lines from the maize hapmap2 dataset.[28] In both cases, the Ozark maize allele is fixed in most of the Ozark sites and Northern Flints and it is found in higher frequencies in the TPS, TFS, improved maize lines, maize from the Pan-American lineage, and South America (Figure 4D). Although we cannot ascertain whether the increase in frequency of the two highly differentiated SNPs occurred before or after its arrival to ENA, our results suggest that the *wx1* gene was a target of selection in the lineage leading to the Ozark maize.

Conclusions

In this study, we generated and analyzed genomic data to improve our understanding of the dispersal of maize in the US, shedding light on both its migration pathways and molecular evolution while challenging previous hypotheses. A key finding is that maize lineages were transported northward from Mexico into the US SW multiple times, bringing in new pulses of genetic diversity that ultimately shaped lineages that became invaluable to modern agronomy. Archaeology has provided evidence of crops, ideas, and people moving considerable distances between Mesoamerica and the US SW as well as northwestern Mexico,[11,13] and our results show that this movement left a mark on maize genomic diversity. We can further resolve that ancient maize from the Ozark region is descended from maize from the US SW, resulting from either multiple dispersals or the introduction of maize varieties with existing population structure. Genetic and geographic distances support a model of transportation across the central and southern Great Plains,[11] although given the limited nature of the archaeological record, the pace of this movement is unknown: potentially rapid through long-distance trade like some exotic goods[53] or potentially slow through farmer-to-farmer exchange over multiple centuries. Future work on maize microfossils from sites in the Great Plains may help resolve the pace of the dispersal, and other non-carbonized macrofossils may reveal other genetic links with modern landraces. As it stands, maize from the Putnam site in the Ozark region is the closest archaeological link to the Northern Flints, providing the best genetic evidence for the origins of this cold-adapted landrace. This knowledge could be used to guide future maize breeding programs and highlight how "peripheral" varieties of crops may become agronomically important due to advantageous traits like hardiness or temperate adaptations.

Limitations of the study

In this study, we generated and analyzed genomic sequencing data from archaeological plant remains, which are characterized by low levels of endogenous DNA and increased errors due to postmortem damage to DNA. We applied strict filtering criteria to minimize contamination (non-endogenous DNA sequences) and errors caused by postmortem damage. However, residual errors can still introduce noise into the data. Another limitation of our study is the low sequencing depth of our maize genomes, a common challenge in ancient DNA research. Low coverage reduces statistical resolution, meaning that some non-significant results may be attributed to the limited number of SNPs available for analysis.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact: Jazmín Ramos Madrigal (jazmin.madrigal@sund.ku.dk).

Materials availability

This study did not generate new reagents.

Data and code availability

Sequencing reads generated as part of this study are available at the ENA: PRJEB73480. All software used for data analyses is publicly available.

ACKNOWLEDGMENTS

Permission for destructive sampling of archaeological material was kindly granted by the Botanical Museum at Harvard University, the Center for Archaeological Investigations at Southern Illinois University Carbondale, the Center for Big Bend Studies, the University of Arkansas Collection Facility, the Osage Nation, and the University of Michigan Museum of Archaeological Anthropology. In particular, we acknowledge the invaluable assistance of Andrea A. Hunter, Tribal Historic Preservation Officer, Osage Nation; Mary Suter, Curator of Collections, University Museum, University of Arkansas; and Heather Lapham, Curator at the Center for Archaeological Investigations. We thank the Danish National High-throughput Sequencing Centre, Sarah Mak, and BGI Shenzhen for assistance in generating the sequencing data, as well as to the Computerome2 and University of York High Performance Computing service that were used to analyze the data. We would like to thank J. Víctor Moreno-Mayar, Aldo Carmona Baez, and Logan Kistler for their helpful discussion. We thank Kelly Swarts and R.G. Matson for further details on the dating of Turkey Pen Shelter. Thanks to Chris Troll, Josh Kapp, and Beth Shapiro for suggestions on the SCR library preparation method. We acknowledge DFF grants DFF-10-081390 and DFF-1325-00136, Danish National Research Foundation grant DNRF143, and Marie Skłodowska-Curie Actions grant 272927—"MaizeKey" for funding this research. R.R.d.F. is funded by the VILLUM FONDEN grant no. 25925.

AUTHOR CONTRIBUTIONS

The project was conceived by J.R.-M., M.T.P.G., and N.W. and headed by J.R.-M. and N.W. F.S.-B., C.C., A.K.W.R., N.W., S.B., K.M., S.L., G.Z., and C.G. generated data through sample preparation, laboratory work, and/or sequencing. J.R.-M. and N.W. designed the analysis and sequencing strategy. G.J.F., B. Schroeder, and B. Smith provided archaeological context. B. Schroeder excavated and curated samples. J.R.-M. performed the bioinformatic analysis with input from S.G., N.W., F.G.V., R.R.d.F., and M.T.P.G. B.P. and T.S.-P. performed protein structure modeling analyses. J.R.-M., N.W., and M.T.P.G. interpreted the results with input from G.J.F., B. Schroeder, and B. Smith. J.R.-M., N.W., and M.T.P.G. wrote the manuscript with input from all other authors. All authors revised, edited, and accepted the manuscript. Primary funding was acquired by M.T.P.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

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
  - ○ Description of the archaeological sites
- METHOD DETAILS
  - ○ aDNA laboratory work
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Data processing
  - ○ Radiocarbon dating
  - ○ Reference data and SNP calling
  - ○ Selection of outgroups
  - ○ Assessing aDNA data authenticity
  - ○ *D*-statistics to test treeness and admixture
  - ○ Annotation of the *wx1* gene

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.11.003.

Received: April 4, 2024
Revised: September 13, 2024
Accepted: November 6, 2024
Published: December 4, 2024

REFERENCES

1. Smith, B.D. (1989). Origins of Agriculture in Eastern North America. Science *246*, 1566–1571. https://doi.org/10.1126/science.246.4937.1566.
2. Hufford, M.B. (2016). Comparative genomics provides insight into maize adaptation in temperate regions. Genome Biol. *17*, 155. https://doi.org/10.1186/s13059-016-1020-2.
3. Unterseer, S., Pophaly, S.D., Peis, R., Westermeier, P., Mayer, M., Seidel, M.A., Haberer, G., Mayer, K.F.X., Ordas, B., Pausch, H., et al. (2016). A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. Genome Biol. *17*, 137. https://doi.org/10.1186/s13059-016-1009-x.
4. Hart, J.P., Brumbach, H.J., and Lusteck, R. (2007). Extending the Phytolith Evidence for Early Maize (*Zea mays* ssp. *mays*) and Squash (*Cucurbita* sp.)

in Central New York. Am. Antiq. *72*, 563–583. https://doi.org/10.2307/40035861.

5. St-Pierre, C.G., and Thompson, R.G. (2015). Phytolith Evidence for the Early Presence of Maize in Southern Quebec. Am. Antiq. *80*, 408–415.
6. Albert, R.K., Kooiman, S.M., Clark, C.A., and Lovis, W.A. (2018). Earliest microbotanical evidence for maize in the Northern Lake Michigan Basin. Am. Antiq. *83*, 345–355. https://doi.org/10.1017/aaq.2018.10.
7. Simon, M.L. (2017). Reevaluating the evidence for Middle Woodland maize from the Holding site. Am. Antiq. *82*, 140–150. https://doi.org/10.1017/aaq.2016.2.
8. Simon, M.L., Hollenbach, K.D., and Redmond, B.G. (2021). New Dates and Carbon Isotope Assays of Purported Middle Woodland Maize from the Icehouse Bottom and Edwin Harness Sites. Am. Antiq. *86*, 613–624. https://doi.org/10.1017/aaq.2020.117.
9. Doebley, J., Wendel, J.D., Smith, J.S.C., Stuber, C.W., and Goodman, M.M. (1988). The origin of cornbelt maize: The isozyme evidence. Econ. Bot. *42*, 120–131. https://doi.org/10.1007/BF02859042.
10. Galinat, W.C., and Gunnerson, J.H. (1963). Spread of eight-rowed maize from the prehistoric Southwest. Botanical Museum Leaflets, Harvard University *20*, 117–160.
11. Smith, B.D. (2017). Tracing the initial diffusion of maize in North America. In Human Dispersal and Species Movement: From Prehistory to the Present (Cambridge University Press), pp. 332–348. https://doi.org/10.1017/9781316686942.014.
12. Krieger, A.D. (1948). Importance of the “Gilmore Corridor” in Culture Contacts between Middle America and the Eastern United States (Bulletin of Texas, Archaeological and Paleontological Society).
13. White, N.M., and Weinstein, R.A. (2008). The Mexican Connection and the Far West of the U.S. Southeast. Am. Antiq. *73*, 227–278. https://doi.org/10.1017/S0002731600042268.
14. Hanselka, J. (2011). Prehistoric Plant Procurement, Food Production, and Land Use in Southwestern Tamaulipas, Mexico (Washington University in St. Louis) https://doi.org/10.7936/K7Z899D6.
15. Merrill, W.L., Hard, R.J., Mabry, J.B., Fritz, G.J., Adams, K.R., Roney, J.R., and MacWilliams, A.C. (2009). The diffusion of maize to the southwestern United States and its impact. Proc. Natl. Acad. Sci. USA *106*, 21019–21026. https://doi.org/10.1073/pnas.0906075106.
16. Smith, B.D. (2006). Eastern North America as an independent center of plant domestication. Proc. Natl. Acad. Sci. USA *103*, 12223–12228. https://doi.org/10.1073/pnas.0604335103.
17. Swarts, K., Gutaker, R.M., Benz, B., Blake, M., Bukowski, R., Holland, J., Kruse-Peeples, M., Lepak, N., Prim, L., Romay, M.C., et al. (2017). Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. Science *357*, 512–515. https://doi.org/10.1126/science.aam9425.
18. Scarry, C.M. (1993). Variability in Mississippian Crop Production Strategies. In Foraging and farming in the eastern woodlands, C.M. Scarry, ed. (University Press of Florida), pp. 78–90.
19. Marston, J.M. (2021). Archaeological Approaches to Agricultural Economies. J. Archaeol. Res. *29*, 327–385. https://doi.org/10.1007/s10814-020-09150-0.
20. Hart, J.P., and Lovis, W.A. (2013). Reevaluating What We Know About the Histories of Maize in Northeastern North America: A Review of Current Evidence. J. Archaeol. Res. *21*, 175–216. https://doi.org/10.1007/s10814-012-9062-9.
21. Vanderwarker, A.M., Bardolph, D.N., and Scarry, C.M. (2018). Maize and Mississippian beginnings. In Mississippian Beginnings, G.D. Wilson, ed. (University Press of Florida), pp. 29–70. https://doi.org/10.5744/florida/9781683400103.003.0002.
22. Mueller, N.G., White, A., and Szilagyi, P. (2019). Experimental Cultivation of Eastern North America’s Lost Crops: Insights into Agricultural Practice and Yield Potential. Journal of Ethnobiology *39*, 549–566. https://doi.org/10.2993/0278-0771-39.4.549.
23. Wales, N., Akman, M., Watson, R.H.B., Sánchez Barreiro, F., Smith, B.D., Gremillion, K.J., Gilbert, M.T.P., and Blackman, B.K. (2019). Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement. Evol. Appl. *12*, 38–53. https://doi.org/10.1111/eva.12594.
24. Kistler, L., Montenegro, A., Smith, B.D., Gifford, J.A., Green, R.E., Newsom, L.A., and Shapiro, B. (2014). Transoceanic drift and the domestication of African bottle gourds in the Americas. Proc. Natl. Acad. Sci. USA *111*, 2937–2941. https://doi.org/10.1073/pnas.1318678111.
25. Fritz, G.J. (1986). Prehistoric Ozark Agriculture: The University of Arkansas Rockshelter Collections (University of North Carolina at Chapel Hill).
26. Wales, N., Carøe, C., Sandoval-Velasco, M., Gamba, C., Barnett, R., Samaniego, J.A., Madrigal, J.R., Orlando, L., and Gilbert, M.T.P. (2015). New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA. BioTechniques *59*, 368–371. https://doi.org/10.2144/000114364.
27. da Fonseca, R.R., Smith, B.D., Wales, N., Cappellini, E., Skoglund, P., Fumagalli, M., Samaniego, J.A., Carøe, C., Ávila-Arcos, M.C., Hufnagel, D.E., et al. (2015). The origin and evolution of maize in the Southwestern United States. Nat. Plants *1*, 14003. https://doi.org/10.1038/nplants.2014.3.
28. Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. *44*, 803–807. https://doi.org/10.1038/ng.2313.
29. Wang, L., Beissinger, T.M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J., and Hufford, M.B. (2017). The interplay of demography and selection during maize domestication and expansion. Genome Biol. *18*, 215. https://doi.org/10.1186/s13059-017-1346-4.
30. Kistler, L., Maezumi, S.Y., Gregorio de Souza, J., Przelomska, N.A.S., Malaquias Costa, F., Smith, O., Loiselle, H., Ramos-Madrigal, J., Wales, N., Ribeiro, E.R., et al. (2018). Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. Science *362*, 1309–1313. https://doi.org/10.1126/science.aav0207.
31. Wang, L., Josephs, E.B., Lee, K.M., Roberts, L.M., Rellán-Álvarez, R., Ross-Ibarra, J., and Hufford, M.B. (2021). Molecular parallelism underlies convergent highland adaptation of maize landraces. Mol. Biol. Evol. *38*, 3567–3580. https://doi.org/10.1093/molbev/msab119.
32. Ramos-Madrigal, J., Smith, B.D., Moreno-Mayar, J.V., Gopalakrishnan, S., Ross-Ibarra, J., Gilbert, M.T.P., and Wales, N. (2016). Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication. Curr. Biol. *26*, 3195–3201. https://doi.org/10.1016/j.cub.2016.09.036.
33. Vallebueno-Estrada, M., Rodríguez-Arévalo, I., Rougon-Cardoso, A., Martínez González, J., García Cook, A., Montiel, R., and Vielle-Calzada, J.-P. (2016). The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. Proc. Natl. Acad. Sci. USA *113*, 14151–14156. https://doi.org/10.1073/pnas.1609701113.
34. Mak, S.S.T., Gopalakrishnan, S., Carøe, C., Geng, C., Liu, S., Sinding, M.-H.S., Kuderna, L.F.K., Zhang, W., Fu, S., Vieira, F.G., et al. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. GigaScience *6*, 1–13. https://doi.org/10.1093/gigascience/gix049.
35. van Heerwaarden, J., Doebley, J., Briggs, W.H., Glaubitz, J.C., Goodman, M.M., de Jesus Sanchez Gonzalez, J., and Ross-Ibarra, J. (2011). Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. USA *108*, 1088–1092. https://doi.org/10.1073/pnas.1013011108.
36. Kistler, L., Thakar, H.B., VanDerwarker, A.M., Domic, A., Bergström, A., George, R.J., Harper, T.K., Allaby, R.G., Hirth, K., and Kennett, D.J. (2020). Archaeological Central American maize genomes suggest ancient gene flow from South America. Proc. Natl. Acad. Sci. USA *117*, 33124–33129. https://doi.org/10.1073/pnas.2015560117.

37. Galinat, W.C. (1985). Domestication and Diffusion of Maize. In Prehistoric Food Production in North America, R.I. Ford, ed. (University of Michigan Press), pp. 245–278.

38. Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet. *48*, 94–100. https://doi.org/10.1038/ng.3464.

39. Piperno, D.R., Moreno, J.E., Iriarte, J., Holst, I., Lachniet, M., Jones, J.G., Ranere, A.J., and Castanzo, R. (2007). Late Pleistocene and Holocene environmental history of the Iguala Valley, Central Balsas Watershed of Mexico. Proc. Natl. Acad. Sci. USA *104*, 11874–11881. https://doi.org/10.1073/pnas.0703442104.

40. Piperno, D.R., and Flannery, K.V. (2001). The earliest archaeological maize (*Zea mays* L.) from highland Mexico: new accelerator mass spectrometry dates and their implications. Proc. Natl. Acad. Sci. USA *98*, 2101–2103. https://doi.org/10.1073/pnas.98.4.2101.

41. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

42. Soraggi, S., Wiuf, C., and Albrechtsen, A. (2018). Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. G3 (Bethesda) *8*, 551–566. https://doi.org/10.1534/g3.117.300192.

43. Yang, N., Wang, Y., Liu, X., Jin, M., Vallebueno-Estrada, M., Calfee, E., Chen, L., Dilkes, B.P., Gui, S., Fan, X., et al. (2023). Two teosintes made modern maize. Science *382*, eadg8940. https://doi.org/10.1126/science.adg8940.

44. Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N., et al. (2012). Reconstructing Native American population history. Nature *488*, 370–374. https://doi.org/10.1038/nature11258.

45. Jaenicke-Després, V., Buckler, E.S., Smith, B.D., Gilbert, M.T.P., Cooper, A., Doebley, J., and Pääbo, S. (2003). Early allelic selection in maize as revealed by ancient DNA. Science *302*, 1206–1208. https://doi.org/10.1126/science.1089056.

46. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78. https://doi.org/10.1126/science.1190371.

47. Whitt, S.R., Wilson, L.M., Tenaillon, M.I., Gaut, B.S., and Buckler, E.S., 4th (2002). Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. USA *99*, 12959–12962. https://doi.org/10.1073/pnas.202476999.

48. Fan, L., Quan, L., Leng, X., Guo, X., Hu, W., Ruan, S., Ma, H., and Zeng, M. (2008). Molecular evidence for post-domestication selection in the Waxy gene of Chinese waxy maize. Mol. Breeding *22*, 329–338. https://doi.org/10.1007/s11032-008-9178-2.

49. Luo, M., Shi, Y., Yang, Y., Zhao, Y., Zhang, Y., Shi, Y., Kong, M., Li, C., Feng, Z., Fan, Y., et al. (2020). Sequence polymorphism of the waxy gene in waxy maize accessions and characterization of a new waxy allele. Sci. Rep. *10*, 15851. https://doi.org/10.1038/s41598-020-72764-3.

50. Li, C., Huang, Y., Huang, R., Wu, Y., and Wang, W. (2018). The genetic architecture of amylose biosynthesis in maize kernel. Plant Biotechnol. J. *16*, 688–695. https://doi.org/10.1111/pbi.12821.

51. Gu, W., Yu, D., Guan, Y., Wang, H., Qin, T., Sun, P., Hu, Y., Wei, J., and Zheng, H. (2020). The dynamic transcriptome of waxy maize (*Zea mays* L. sinensis Kulesh) during seed development. Genes Genomics *42*, 997–1010. https://doi.org/10.1007/s13258-020-00967-z.

52. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

53. Charles, D.K. (2012). Origins of the Hopewell Phenomenon (Oxford University Press) https://doi.org/10.1093/oxfordhb/9780195380118.013.0039.

54. Ramachandran, D., McKain, M.R., Kellogg, E.A., and Hawkins, J.S. (2020). Evolutionary Dynamics of Transposable Elements Following a Shared Polyploidization Event in the Tribe Andropogoneae. G3 (Bethesda) *10*, 4387–4398. https://doi.org/10.1534/g3.120.401596.

55. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes *9*, 88. https://doi.org/10.1186/s13104-016-1900-2.

56. 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.

57. 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.

58. Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356. https://doi.org/10.1186/s12859-014-0356-4.

59. Moreno-Mayar, J.V. (2022). FrAnTK: A Frequency-based Analysis ToolKit for efficient exploration of allele sharing patterns in present-day and ancient genomic datasets. G3 (Bethesda) *12*, jkab357. https://doi.org/10.1093/g3journal/jkab357.

60. Malaspinas, A.-S., Tange, O., Moreno-Mayar, J.V., Rasmussen, M., DeGiorgio, M., Wang, Y., Valdiosera, C.E., Politis, G., Willerslev, E., and Nielsen, R. (2014). bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). Bioinformatics *30*, 2962–2964. https://doi.org/10.1093/bioinformatics/btu410.

61. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

62. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

63. 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.

64. 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.

65. Cingolani, P., Platts, A., Wang, L.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 w1118; iso-2; iso-3. Fly *6*, 80–92. https://doi.org/10.4161/fly.19695.

66. Bronk Ramsey, C.B. (2009). Bayesian Analysis of Radiocarbon Dates. Radiocarbon *51*, 337–360. https://doi.org/10.1017/S0033822200033865.

67. Smiley, F.E. (1994). The Agricultural Transition in the Northern Southwest: Patterns in the Current Chronometric Data. Kiva *60*, 165–189. https://doi.org/10.1080/00231940.1994.11758264.

68. Macneish, R.S. (1964). Ancient Mesoamerican Civilization. Science *143*, 531–537. https://doi.org/10.1126/science.143.3606.531.

69. Schroeder, B., Blohm, T., and Snow, M.H. (2021). Spirit Eye Cave: Reestablishing provenience of trafficked prehistoric human remains using a composite collection-based ancient DNA approach. J. Archaeol. Sci.: Rep. *36*, 102798. https://doi.org/10.1016/j.jasrep.2021.102798.

70. Schroeder, B., and Nayapiltzin, X. (2022). A Complicated History: Collaboration with Collectors to Recover and Repatriate Indigenous Human Remains Removed from Spirit Eye Cave. Adv. Archaeol. Pract. *10*, 26–37. https://doi.org/10.1017/aap.2021.36.

71. Schroeder, B. (2022). Evidence of Late Archaic Maize Use in the Big Bend Region of West Texas. Kiva *88*, 58–83. https://doi.org/10.1080/00231940.2021.2004360.

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

73. Wales, N., Andersen, K., Cappellini, E., Avila-Arcos, M.C., and Gilbert, M.T.P. (2014). Optimization of DNA recovery and amplification from non-carbonized archaeobotanical remains. PLoS One *9*, e86827. https://doi.org/10.1371/journal.pone.0086827.

74. Gansauge, M.-T., and Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat. Protoc. *8*, 737–748. https://doi.org/10.1038/nprot.2013.038.

75. Kapp, J.D., Green, R.E., and Shapiro, B. (2021). A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. J. Hered. *112*, 241–249. https://doi.org/10.1093/jhered/esab012.

76. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115. https://doi.org/10.1126/science.1178534.

77. Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., Al-Rasheid, K.A.S., Willerslev, E., Krogh, A., and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. BMC Genomics *13*, 178. https://doi.org/10.1186/1471-2164-13-178.

78. Reimer, P.J., Austin, W.E.N., Bard, E., Bayliss, A., Blackwell, P.G., Bronk Ramsey, C., Butzin, M., Cheng, H., Edwards, R.L., Friedrich, M., et al. (2020). The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0–55 cal kBP). Radiocarbon *62*, 725–757. https://doi.org/10.1017/RDC.2020.41.

79. Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaeppler, S.M., et al. (2012). Comparative population genomics of maize domestication and improvement. Nat. Genet. *44*, 808–811. https://doi.org/10.1038/ng.2309.

80. Hufford, M.B., Bilinski, P., Pyhäjärvi, T., and Ross-Ibarra, J. (2012). Teosinte as a model system for population and ecological genomics. Trends Genet. *28*, 606–615. https://doi.org/10.1016/j.tig.2012.08.004.

81. Ross-Ibarra, J., Tenaillon, M., and Gaut, B.S. (2009). Historical divergence and gene flow in the genus *Zea*. Genetics *181*, 1399–1413. https://doi.org/10.1534/genetics.108.097238.

82. Briggs, A.W., Stenzel, U., Johnson, P.L.F., Green, R.E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M.T., Lachmann, M., et al. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. USA *104*, 14616–14621. https://doi.org/10.1073/pnas.0704665104.

83. Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., et al. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature *499*, 74–78. https://doi.org/10.1038/nature12323.

84. Moreno-Mayar, J.V., Vinner, L., de Barros Damgaard, P., de la Fuente, C., Chan, J., Spence, J.P., Allentoft, M.E., Vimala, T., Racimo, F., Pinotti, T., et al. (2018). Early human dispersals within the Americas. Science *362*, eaav2621. https://doi.org/10.1126/science.aav2621.

85. Lipson, M., and Reich, D. (2017). A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa. Mol. Biol. Evol. *34*, 889–902. https://doi.org/10.1093/molbev/msw293.

86. Leppälä, K., Nielsen, S.V., and Mailund, T. (2017). admixturegraph: an R package for admixture graph manipulation and fitting. Bioinformatics *33*, 1738–1740. https://doi.org/10.1093/bioinformatics/btx048.

87. Takuno, S., Ralph, P., Swarts, K., Elshire, R.J., Glaubitz, J.C., Buckler, E.S., Hufford, M.B., and Ross-Ibarra, J. (2015). Independent Molecular Basis of Convergent Highland Adaptation in Maize. Genetics *200*, 1297–1312. https://doi.org/10.1534/genetics.115.178327.

88. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS One *7*, e37558. https://doi.org/10.1371/journal.pone.0037558.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
6 archaeological maize samples (4 cobs, 2 kernels) from the Three Fir Shelter site in Arizona, US	This paper; Wales et al.[26]	3Fir 1255, 3Fir 1285, 3Fir 1290.5 Purple, 3Fir 1290.5 Yellow, 3Fir 1294, and 3Fir 428
2 archaeological maize samples (cobs) from the Romero's Cave in Tamaulipas, Mexico	This paper	Romero 29 and Romero 51
18 archaeological maize samples (cobs) from the Ozark rockshelters in Arkansas, US	This paper	202_Buzzard_Roost, 215_Beaver_Pond, 218_Brown_Bluff, 226_Cow_Ford, 214_Craddock, 220_Craddock, 204_Salts_Bluff, 212_Salts_Bluff, 205_Edens_Bluff, 221_Edens_bluff, 222_Edens_bluff, 223_Edens_Bluff, 224_Edens_Bluff, 216_Gibson, 203_Whitney, 207_Putnam, 209_Putnam, and 211_Putnam
5 archaeological maize samples (cobs) from the Spirit Eye Cave, Tranquil Rockshelter, and Bee Cave Canyon sites in Texas, US	This paper	BeeCaveCanyon, SpiritEyeCave_114, SpiritEyeCave_95, SpiritEyeCave_ 41P25-1012, and TranquilShelter
1 archaeological maize sample (cob) from the Bat Cave in Arizona, US	This paper; da Fonseca et al.[27]	Batcave17 (SW4Ba)
Chemicals, peptides, and recombinant proteins		
Proteinase K	Sigma-Aldrich	Cat#3115844001
Critical commercial assays		
QIAquick PCR Purification kit	QIAGEN	Cat#28104
Qubit dsDNA HS Assay Kit	Life Technologies	Cat#Q33230
NEBNext DNA Library Prep Master Mix	New England Biolabs Inc.	Cat# E6070L
MyBait target enrichment kits	MYcroarray, Ann Arbor, MI	Custom
Phusion® High-Fidelity PCR	New England Biolabs Inc.	Cat#M0531S
Deposited data		
Sequencing data for 32 ancient maize genomes	This study	Table S2; ENA: PRJEB73480
Sequencing data for 55 ancient maize genomes	Kistler et al.[30]; da Fonseca et al.[27]; Swarts et al.[17]; Ramos-Madrigal et al.[32]; Vallebueno-Estrada et al.[33]	Table S2
Sequencing data for 94 modern maize genomes (*Zea mays* subsp. *mays*)	Chia et al.[28]; Kistler et al.[30]; Wang et al.[29]; Wang et al.[31]	Table S2
Sequencing data for 23 modern wild maize genomes (*Zea mays* subsp. *parviglumis* and *Zea mays* subsp. *mexicana*)	Wang et al.[29]; Chia et al.[28]	Table S2
Sequencing data for the genomes of 2 outgroup species (*Zea diploperennis* and *Tripsacum dactylodes*)	Ramachandran et al.[54]; Chia et al.[28]	Table S2
Hapmap2	Chia et al.[28]	Table S2
Oligonucleotides		
Illumina-compatible adapters	Illumina	N/A
BGI-compatible adapters	BGI	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
AdapterRemoval 2.0	Schubert et al.[55]	https://github.com/MikkelSchubert/adapterremoval
bwa aln 0.7.12	Li and Durbin[56]	http://bio-bwa.sourceforge.net/
Picard 1.130	N/A	https://broadinstitute.github.io/picard
Genome Analysis Toolkit (GATK.3.3)	McKenna et al.[57]	https://software.broadinstitute.org/gatk/
samtools 1.2.	Li and Durbin[56]	http://samtools.sourceforge.net/
ANGSD v0.921	Korneliussen et al.[58]	https://github.com/ANGSD/angsd
FrAnTK	Moreno-Mayar J.V.[59]	https://github.com/morenomayar/FrAnTK
bamdamage	Malaspinas et al.[60]	https://bioinformaticshome.com/db/tool/bammds
R v4.1.1	R Core Team[61]	https://www.R-project.org/
plink 2.0	Chang et al.[62]	https://www.cog-genomics.org/plink/2.0/
ADMIXTURE 1.23	Alexander et al.[63]	http://dalexander.github.io/admixture/download.html
EEMS	Petkova, D. et al.[38]	https://github.com/dipetkov/eems
reemsplots2	N/A	https://github.com/dipetkov/reemsplots2
qpWave	Patterson et al.[41]	https://github.com/DReichLab/AdmixTools
qpGraph	Patterson et al.[41]	https://github.com/DReichLab/AdmixTools
TreeMix v1.13	Pickrell and Pritchard[64]	https://bitbucket.org/nygcresearch/treemix/wiki/Home
SNPeff v5	Cingolani et al.[65]	https://pcingola.github.io/SnpEff/
alphafold2.0	Jumper et al.[52]	https://github.com/google-deepmind/alphafold
OxCal 4.4.4	Bronk Ramsey[66]	https://c14.arch.ox.ac.uk/oxcal.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Description of the archaeological sites

Three Fir Shelter

The Three Fir Shelter (TFS) is located on the Black Mesa in Northern Arizona (Figure 1). This site was originally excavated in the 1980s by Francis Smiley and has yielded some of the earliest maize remains from the United States (US) Southwest.[67] We analyzed eight samples recovered from this archaeological site, which had been stored in the Center for Archaeological Investigations, Southern Illinois University Carbondale (Table S1). Two of the samples were maize kernels and the remaining six were maize cobs. Six of the samples were previously radiocarbon dated with age ranging from 2145-1947 to 1874-1719 cal. yr B.P. (95.5% CI; Table S1).[26] Two of the six cob samples did not yield sufficient endogenous DNA to be incorporated in the population genetic analyses (Table S1); these two specimens are not associated with a radiocarbon date.

Romero's Cave

The Romero's and Valenzuela's Caves are part of an archaeological assemblage near Ocampo Tamaulipas, in Northeast Mexico (Figure 1). The caves were originally excavated by Richard MacNeish in 1958 as part of his search for the origin of agriculture in Mesoamerica.[68] Maize appears in the archaeological record of these caves as early as 4000 yr B.P., however remains are sparse until approximately 2000 yr B.P. when these increase in frequency comparable with that of a crop staple. These two caves, together with the caves in the Tehuacan Valley, represent some of the few archaeological sites that have yielded non-carbonized maize remains in the region between the domestication center and the US Southwest dating to the period of maize northward expansion from the domestication center. We analyzed two samples from Romero's Cave that have been previously dated to 2667-2181 and 2839-2517 cal. yr B.P. (95.5% CI; Table S1).[45]

Sites in the Ozark region

The Ozark rockshelters are a series of archaeological sites excavated by the University Arkansas Museum between 1929 and 1934. The sites are distributed across eight counties in northwest Arkansas and one southwest Missouri county (Figure 1). We analyzed 21 samples from this archaeological assemblage. Eleven of the samples have been previously directly dated,[25] two of the samples had indirect dates from sunflower remains found in the same layer,[23] and we generated radiocarbon dates for six of the samples (Table S1). From the 21 samples sequenced 18 yielded enough endogenous DNA content and were incorporated in the population genetic analyses. Maize samples ranged in age from one recent sample 275-8 to 1063-936 cal. yr B.P. (95.5% CI).

Spirit Eye Cave (41PS25)

The Spirit Eye Cave is located in Presidio County of West Texas[69] (Figure 1). The cave has a long history of uncontrolled excavation, initial research focused on documenting and recovering material taken from the pay-to-dig history of the cave.[70] The subsequent analyses of the cultural and Indigenous ancestral remains recovered from these private collectors provide the initial understanding of the site. Fieldwork by professional archaeologists began in 2017, resulting in the recovery of numerous cultivars. For this analysis we used three maize cob remains from this site with directly generated radiocarbon dates for each sample.[71]

Bee Cave Canyon (41BS8)

The Bee Cave Canyon is located in southern Brewster County in West Texas (Figure 1). It is a large rockshelter excavated between 1928–1929 by the Museum of the American Indian, Heye Foundation. The artifacts from the 1928–1929 excavation are currently housed at the Smithsonian Institute. We analyzed one maize cob collected from the surface of the site in 2019 which was directly radiocarbon dated, yielding a calibrated age of 734–673 yr B.P.[71]

Tranquil Rockshelter (41BS1513)

The Tranquil Rockshelter is located in Brewster County of West Texas (Figure 1). The rockshelter was excavated in 2008 and 2009 by the Center for Big Bend Studies of Sul Ross State University and is unanalyzed. We analyzed one maize cob from this excavation, which was directly radiocarbon dated for this analysis, yielding a calibrated age of 718–656 yr B.P.[71]

METHOD DETAILS

aDNA laboratory work

DNA processing overview

Laboratory steps were carried out in the aDNA facilities at the University of Copenhagen and the University of York. Ancient DNA extractions and library preparations were conducted in dedicated clean rooms to minimize contamination, following the best practices, including use of full body suits and positive pressure ventilated rooms. Post-PCR steps were conducted in the facilities physically separated from the clean rooms. Unless specified below, lab work was performed at the University of Copenhagen.

Three Fir Shelter maize: deeper sequencing and target enrichment

Sequencing libraries for six maize samples (four cobs and two kernels) from the TFS were available from a previous study.[26] In that study, DNA was processed with double- and single-stranded DNA library protocols and used to compare the efficiency of the two methods. Here, we generated additional sequencing data on these existing libraries using an Illumina HiSeq 2500 in SR100 mode (Table S2).

In addition to the deeper shotgun sequencing, we performed target capture of the TFS double-stranded libraries to enrich for genomic loci of interest defined in a previous study.[27] The hybridization targets cover the exons of 348 genes, which were selected based on their potential relevance for the domestication process. To reach the necessary amount of DNA, the libraries were amplified with Phusion High-Fidelity DNA Polymerase.[72] After amplification, libraries were purified using a QIAquick PCR Purification kit and quantified using a Bioanalyzer 2100 (Agilent, Santa Clara, CA). Enrichment was performed using three custom-designed MyBait target enrichment kits (MYcroarray, Ann Arbor, MI) following the manufacturer recommendation. The custom kits targeted the same loci but used 120- 80- and 40-mer probes, with the aim of investigating capture efficiency. Libraries were pooled based on index compatibility and sample molarity and sequenced on Illumina HiSeq 2500 in SR100 mode. A description of the libraries generated and sequenced for each sample can be found in Table S2.

Romero's Cave: DNA extraction and single-stranded DNA library preparation

Two cobs from Romero's Cave were processed for DNA sequencing. A piece of each cob was pulverized using a sterile Braun Mikro Dismembrator S ball mill (B. Braun Biotech, Melsungen, Germany), a stainless steel flask and grinding ball. DNA was extracted and purified from the resulting powder following the protocol described in Wales et al.[73] DNA extracts were used to build single-stranded libraries following the preparation protocol described in Gansauge and Meyer.[74] DNA concentration in the libraries was measured using the Qubit dsDNA HS Assay Kit (Life Technologies) following the manufacturer's protocol. Libraries were sequenced on a Illumina HiSeq 2500 in SR100 mode (Table S2).

Bat Cave: DNA extraction, library preparation and deeper sequencing

The ∼3,390 yr B.P. maize sample Batcave17 (SW4Ba) from the Bat Cave in New Mexico was previously processed for DNA extraction and sequencing.[27] However, given it represents one of the oldest macrobotanical maize remains from the US Southwest, we generated additional sequencing libraries and data to increase its genome coverage. Three DNA extractions were performed, following the method described in the previous section ("Romero's Cave: DNA extraction and single-stranded DNA library preparation"). Each of the three DNA extracts were converted into double-stranded DNA libraries using the NEBNext DNA Library Prep Master Mix (E6070L, New England BioLabs) as described in Wales et al.[26] DNA concentration in the libraries was measured using the

Qubit dsDNA HS Assay Kit (Life Technologies) following the manufacturer's protocol. DNA libraries were sequenced on a Illumina HiSeq 2500 in SR80 mode. Additionally, we generated more data using the original libraries from da Fonseca et al.[27] by sequencing them on a Illumina HiSeq 2500 in SR100 mode (Table S2).

Ozark rockshelter: DNA extraction and double-stranded DNA library preparation

Six of the maize samples from the Ozark rockshelters (216_Gibson, 214_Craddock, 223_Edens_Bluff, 215_Beaver_Pond, 211_Putnam, 204_Salts_Bluff) were initially processed for DNA sequencing at the University of Copenhagen aDNA facilities. DNA was extracted and prepared into double-stranded libraries in the same manner described in the section "Romero's Cave: DNA extraction and single-stranded DNA library preparation". Sequencing libraries were pooled based on their index compatibility and sample molarity and sequenced on Illumina HiSeq 2500 in SR100 mode. We generated additional sequencing data for four of the libraries (223_Edens_Bluff, 215_Beaver_Pond, 211_Putnam, 204_Salts_Bluff) using an Illumina HiSeq 2500 in SR80 mode.

Ancient Texas: DNA extraction and library preparations

Five cob samples from sites in west Texas were processed in the aDNA facility at the University of York. DNA was extracted and prepared into double-stranded libraries in the same manner described in the section ("Romero's Cave: DNA extraction and single-stranded DNA library preparation"). Sequencing libraries were pooled based on index compatibility and molarity and sequenced on an Illumina HiSeq 2500 in SR80 mode (Table S2). To generate deep sequencing data, three samples (Spirit Eye Cave 114 and 95, and Tranquil Shelter) were extracted again and DNA was prepared using the single-stranded DNA library preparation following the Santa Cruz Reaction (SCR) protocol.[75] Each library was amplified with four indexing primers to facilitate deep sequencing. Sequencing libraries were pooled based on index compatibility and molarity and sequenced on Illumina NovaSeq 600 in PE150 mode (Table S2).

Romero's Cave and Ozark sites: BGISEQ libraries

A total of 17 samples from the Romero's Cave (n=1) and Ozark rockshelters (n=16) were sequenced using BGISEQ technology (Table S2). DNA was extracted from maize cobs following the method described in section "Romero's Cave: DNA extraction and single-stranded DNA library preparation". DNA extracts were converted into double-stranded DNA libraries using the NEBNext DNA Library Prep Master Mix (E6070L, New England BioLabs) as described in Wales et al., [26] except that BGISEQ-compatible adapters were ligated to the blunted DNA molecules. One lane per library/sample was sequenced on the BGISEQ-500 platform in SR100 mode (Table S2).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data processing

Adapter sequences, low quality stretches and leading/tailing N's were trimmed from the raw reads using AdapterRemoval 2.0.[55] Reads shorter than 30 bp after trimming were discarded and the remaining reads were mapped to the *Zea mays* ssp. *mays* reference sequence (B73-v3.25)[76] using *bwa aln* 0.7.12.[56] *Bwa* seed was disabled (-l was set to 1000) in order to prevent mapping bias due to 5' terminal substitutions caused by aDNA damage.[77] PCR-duplicates in each sequencing library were identified and removed from the resulting bam files using *Picard* 1.130 (http://picard.sourceforge.net). Reads with a mapping quality below 30, with an alternative hit, or mapping to more than one position in the reference genome (*i.e.* having the tag XT:Z and not the tag X:A:U) were discarded. Finally, reads were realigned to the reference genome using *Genome Analysis Toolkit* (GATK.3.3) and the MD-tag was recalculated using *samtools* 1.2. Sequencing results for all the ancient samples sequenced in this study are summarized in Tables S1 and S2.

To decrease the proportion of bases with C-to-T or G-to-A substitutions derived from the aDNA damage in the ancient maize samples, we trimmed 5 bases from the 5' and 3' ends of each read in all ancient samples before conducting the analyses.

Radiocarbon dating

Radiocarbon measurements were taken for the maize specimens from Salts Bluff, Edens Bluff, Spirit Eye Cave, Bee Cave, and the Tranquil Rockshelter. Published radiocarbon dates were taken from publications: Fritz[25] for the Ozark sites, Jaenicke-Després et al.[45] for Romero's Cave, da Fonseca et al.[27] for McEuen Cave, Bat Cave, and Tularosa Cave; Swarts et al.[17] for Turkey Pen Shelter; Wales et al.[26] for Three Fir Shelter; and Schroeder et al.[71] for the Spirit Eye Cave, Bee Cave, and the Tranquil Rockshelter. Dates were calibrated using OxCal 4.4.4[66] with the IntCal20 calibration curve.[78] Published data for Turkey Pen Shelter and Tularosa were modeled in a Bayesian approach according to the depositional times. For Turkey Pen Shelter, one dated sample (JK1699), was excluded from the Bayesian model as it did not meet the test for homogeneity. Likewise, one sample from the more recent phase at Tularosa Cave (SW105) was excluded from the Bayesian analysis because it did not meet the test for homogeneity. Radiocarbon calibrations are shown in Figure S1A.

Reference data and SNP calling

We compiled a dataset consisting of whole-genome data for the 32 ancient maize samples sequenced in this study, 94 maize landraces,[28–31] 23 wild maize relatives (21 subsp. *parviglumis* and 2 subsp. *mexicana* samples),[28,29] one *Tripsacum dactyloides*,[28] one *Zea diploperennis*,[54] and 55 published ancient maize samples[17,27,30,32,33] (Table S3). We obtained FASTQ files for all reference samples from the NCBI Sequence Read Archive or the European Nucleotide Archive. Sequencing reads were mapped to the B73-v2.25 reference genome using the same procedure and parameters described in the "data processing" section.

To identify single nucleotide polymorphic (SNP) sites in our dataset, we performed SNP calling using the genotype-likelihood-based method implemented in ANGSD v0.921[58] and all 153 modern samples and a subset of 30 high depth of coverage (>1×) ancient samples. For the SNP calling, we used the GATK genotype likelihood method implemented in ANGSD (-GL 2) and applied to following filters: minimum base quality of 20 (-minQ 20), minimum mapping quality of 30 (-minMapQ 30), minimum SNP *p*-value of 1e6 (-SNP_pval 1e-6), minimum number of samples without missing data per site of 50 (-minind 50), minimum per sample depth of coverage of 3 (-setMinDepthInd 3), minor allele frequency of 0.05 and excluded transitions. Additionally, to avoid incorporating highly repetitive genomic regions which are difficult to map using short reads, we applied a mappability mask to restrict the analyses to sites that can be mapped uniquely in the genome as described in Ramos-Madrigal et al.[32] Once we identified SNPs, we randomly sampled one read for every SNP and for every sample using FrAnTK.[59] Reads with mapping quality lower than 30 and bases with quality lower than 20 were discarded. This approach allowed us to co-analyse ancient and modern maize samples with varying in depth of coverages, as it is common practice in aDNA studies. The final dataset consisted of 1,826,117 transversion sites across 206 maize samples. When only a subset of the samples was used in a particular analysis or additional filters were applied we specify the number of SNPs that remained after filtering in the corresponding sections.

Selection of outgroups

We used three different outgroups: *Tripsacum* (*Tripsacum dactyloides*), *diploperennis* (*Zea diploperennins*) and *parviglumis* (*Zea mays* subsp. *parviglumis*). Each outgroup provides different levels of resolution due to their phylogenetic distance from domesticated maize, genome coverage, mappability to the maize genome, and the number of available individuals.

Tripsacum is the most distant outgroup and, because there is no evidence of admixture with *Zea* species, it is commonly used as an outgroup in maize studies (e.g. [28,30,32,79]). However, because it is evolutionarily distant, sequencing data from *Tripsacum* maps only to highly conserved regions of the maize genome, limiting the number of SNPs for analysis. We used *Tripsacum* to estimate error rates (where the choice of outgroup has minimal effect) and as an ancestral genome for polarizing the site frequency spectrum in population branch statistic analyses (restricted to conserved regions).

Diploperennis, a closer relative of maize, is more suitable for certain analyses. However, extensive gene flow within the *Zea* genus[80,81] makes it less ideal when studying the relationships of maize with *mexicana* or *parviglumis* (two wild maize subspecies). A *D*-statistic test *D*(*Tripsacum*, *diploperennis*; *mexicana*, *parviglumis*) yielded positive significant results (*z*-score ~3.916), indicating that the available *diploperennis* genome likely carries *mexicana* admixture. We used this genome as an outgroup when using *Tripsacum* significantly reduced the number of available SNPs, such as in treemix analyses where we incorporate several low-coverage ancient genomes. It was also used in the admixture graphs to root the topology of the tree and in most *D*-statistic tests, where the potential admixture with *mexicana* does not affect the results.

Parviglumis, is one of the three wild maize subspecies for which several genomes are available. However, gene flow between *parviglumis* and domesticated maize in regions where they overlap geographically is common,[80] making it a less ideal outgroup. We used *parviglumis* as an outgroup in f_3-statistic tests, which need allele frequencies for each population and benefit from having multiple individuals. Since we focused on ancient and modern maize from the US, outside *parviglumis* distribution, we do not expect gene flow to affect the results.

Assessing aDNA data authenticity

aDNA damage patterns

Unless treated to specifically remove deaminated bases, ancient DNA sequencing reads are characterized by an increase of C-to-T and G-to-A substitutions towards the 5' and 3' ends respectively.[82] These damage patterns are often used to assess the authenticity of sequencing data derived from ancient samples. We estimated the proportion of the different substitutions with respect to the reference genome in all ancient samples sequenced in this study using *bamdamage*.[60] Quality thresholds were set to –mapquality 30 and –basequality 20. Substitution patterns were consistent with those observed in other ancient maize specimens confirming the authenticity of the data (Table S2).

Estimating type-specific error rates

To further evaluate the quality and authenticity of the data we estimated relative error rates using ANGSD v0.921[58] as described in Orlando et al.[83] This method estimates the excess of derived substitutions in a given sample compared to a high quality genome using a maximum likelihood approach. Maize landrace RIMMA1010 and HapMap2 sample TDD39103 (*Tripsacum dactyloides*) were used as high quality and outgroup genomes, respectively. In both cases we used a majority count consensus sequence using ANGSD v0.921 built with reads with a minimum mapping quality of 30 and base quality 20. Error rates in the ancient samples are comparable to those obtained in similar studies,[17,27,30,32] and can be mostly attributed to C-to-T and G-to-A transitions derived from the aDNA damage (Table S2). To decrease the biases that this extra error might cause, transitions were excluded from the subsequent analyses except when specified otherwise.

Comparison between BGI and Illumina platforms

It has been demonstrated that DNA sequencing data from BGISEQ and Illumina platforms display similar characteristics.[34] Here, we further explored the potential biases derived from using these two different platforms by generating paired data for six maize samples (Romero_29, 204_Salts_Bluff, 211_Putnam, 214_Craddock, 215_Beaver_Pond and 223_Edens_Bluff) (Table S2). Our results replicate previous observations showing no substantial differences in error profiles, aDNA damage patterns, average fragment length, GC content and endogenous content (Table S4).[34] Additionally, we explore potential correlations in the data substitution patterns using *D*-statistics (Figures S1B and S1C). We find no bias in the results associated with differences in sequencing in these two different platforms.

aDNA damage patterns, GC content and fragment length in BGISEQ and Illumina data

For each of the six paired samples we estimated the average fragment length, 5' and 3' aDNA terminal aDNA damage and GC content in the reads mapped to the maize (B73-v3.25) genome after removing PCR duplicates and performing quality filtering as described in the "data processing" section. Fragment length and aDNA terminal damage was estimated using *bamdamage*.[60] We observed similar values when comparing the data obtained from the two sequencing platforms. Consistent with previous findings, we observe no statistically significant difference in GC content or terminal damage between sequencing platforms (paired *t*-test, *p*-value > 0.4).[34] Although we observe a slight increase in average fragment length in data derived from the BGI sequencing platform, we do not find a statistically significant difference (paired *t*-test, *p*-value = 0.03301).

Relative error rates in BGISEQ and Illumina data

For each of the six paired samples we estimated type-specific error rates as described in the "assessing aDNA data authenticity" section. We do not find statistically significant differences in the error rates in data from the two sequencing platforms (paired *t*-test, *p*-value = 0.38) (Table S2).

D-statistics assessing differences between BGISEQ and Illumina data

We used *D*-statistics to investigate potential spurious correlations between samples sequenced using the same sequencing chemistry. *D*-statistics were computed using FrAnTK[59] as described in the "D-statistics using FrAnTK" section below. We estimated *D*-statistics of the form D(X^{BGISEQ}, $X^{ILLUMINA}$; 202_BR, RIMMA0409), where X^{BGISEQ} and $X^{ILLUMINA}$ are the same sample sequenced in BGISEQ-500 or Illumina, 202_BR is the youngest maize sample sequenced with BGISEQ-500 and RIMMA0409 is a modern maize landrace. If no biases inherent to the sequencing exist, we expect $D \sim 0$, alternatively significant deviation from $D \sim 0$ towards positive values would indicate the $H2^{BGISEQ}$ and 202_BR are artificially closer. We found no significant deviation from $D \sim 0$ (Figure S1B). Additionally, we evaluated the *Z*-scores obtained from the tests *D*($H1^{ILLUMINA}$, H2; 202_BR, TIL15) and *D*($H1^{BGISEQ}$, H2; 202_BR, TIL15) for paired $H1^{ILLUMINA}$ and $H1^{BGISEQ}$ samples, where H2 represents all the samples in the reference panel and TIL15 corresponds to one of the *parviglumis* samples. For each paired comparison we restricted the test to sites that were non-missing in both the Illumina and BGISEQ-500 data. The distribution of *Z*-scores from tests involving the same H1 sample are expected to be identical in the absence of any sequencing bias. We find no statistically significant difference in the distributions suggesting there is not any bias derived from both platforms that can affect this type of analyses (Figure S1C).

Multidimensional scaling analysis

We performed a multidimensional scaling (MDS) analysis to explore the genetic relationships of the ancient and modern maize samples. Starting from the SNP dataset described in the "reference dataset" section, we discarded samples with more than 90% missing data, with the exception of the Spirit Eye Cave, the Tranquil Rockshelter, and samples with black outline and white filling in Figure S2A. The final dataset consisted of 184 samples. We performed an MDS analysis on the entire dataset, after excluding wild maize samples (Figure S2A), and excluding wild maize and landraces from South America (Figure 2A). In each case, identity-by-state pairwise distances between samples were estimated using plink2.0[62] and the *cmdscale* function from R[61] was used to perform the MDS analysis.

ADMIXTURE analysis

To investigate the genetic structure in ancient and modern maize samples we used ADMIXTURE 1.23.[63] We included ancient and modern maize landraces as well as wild *parviglumis* and *mexicana* samples in the dataset. Starting from the SNP dataset described in "reference dataset and SNP calling" section, we discarded samples with more than 90% missing data, with the exception of the Spirit Eye Cave and the Tranquil Rochshelter samples that were included with 95% missing data. The final dataset consisted of 175 samples. ADMIXTURE was run assuming 2 to 7 admixture clusters (K={2..7}). For each value of K, we ran 100 replicates starting on different seed values and kept the replicate with the best likelihood (Figures 2A and S2B). The results from ADMIXTURE estimating seven ancestry components were used to define groups among modern maize landraces that represent the main geographic ancestry components. Samples with at least 99% ancestry for each of the main components were grouped in the f_3- and *D*-statistics tests.

EEMS analysis

EEMS[38] was used to estimate and visualize potential migration routes and barriers. EEMS estimates effective migration rates on a geographic space based on the genetic distance and geographic coordinates of a set of samples. Note that EEMS does not consider geographic features such as mountains and valleys, which could affect the geographic distance between samples. Instead, the migration corridors and barriers inferred by EEMS can sometimes be attributable to geographic barriers.

To avoid that recent maize movements could interfere with our inference of past migration routes, we excluded modern maize landraces from the US. We ran EEMS on three different datasets: (1) all ancient genomes from the US, Romero's Cave and modern genomes from Mexico, (2) 1000-3000 year-old ancient genomes from Mexico and the US, and (3) ≤2000 year-old ancient genomes from the US and modern genomes from Mexico. We used the SNP dataset described in the "reference data and SNP calling" section. Geographical coordinates for each of the samples is indicated in Table S1. For each dataset, we set the number of demes (nDemes) to 300, and ran 2,000,000 iterations of the MCMC algorithm (numMCMCIter), with a burn-in (numBurnIter) of 1,000,000 iterations. In each case, we assessed the convergence of the run based on the MCMC chain. Results were plotted using reemsplots2 package (https://github.com/dipetkov/reemsplots2). We obtained similar results for the three dataset, where EEMS inferred a migration barrier coinciding with the Gilmore corridor and a potential gene flow corridor across the Great Plains. Results are presented in Figures 2B (dataset 1) and S2D (datasets 2 and 3).

Outgroup f3-statistics

We used outgroup *f3*-statistics as implemented in FrAnTK[59] to measure the amount of shared drift between the ancient maize and modern samples (Figure S2C). We used the SNP dataset described in the "reference dataset and SNP calling" section. Samples were grouped following the groups identified in the ADMIXTURE analysis for modern maize and qpWave (supplemental information section ''Identifying homogeneous ancestry clusters using qpWave'') analysis for ancient maize. Twenty-one *parviglumis* samples were used as the outgroup.

Identifying homogeneous ancestry groups

We used qpWave in order to identify groups of ancient maize samples that derived from a single migration wave. In brief, qpWave uses f_4-statistics to estimate the minimum number of migrations or source populations required to explain a group of test samples or populations. It does so, by estimating all the possible f_4-statistics of the form:

$$f_4(\text{Left}_{\text{FIXED}}, \text{Left}_2;\ \text{Right}_1, \text{Right}_{\text{FIXED}})$$

Where the 'Left' corresponds to the test populations and the 'Right' corresponds to the source populations. We used qpWave in two ways: (1) to test whether samples from the different sites in the Ozarks were consistent with a single migration wave, and (2) to identify groups of samples among the ancient and modern US maize that were differentially related to maize from outside the US ancestry cluster.

Migration waves into the Ozarks

We used qpWave to test if the samples from different sites in the Ozark rockshelters derived from the same or different migration waves. For the 'Right' populations in the test we used *diploperennis* as the fixed outgroup and selected groups of modern and ancient maize representing the main geographic groups: Andean maize, Mexican Highlands, Pan-American maize, South America lowland, Romero's Cave, Bat Cave, Tularosa Cave, Spirit Eye Cave and Turkey Pen Shelter. Modern samples were grouped according to the ancestry clusters identified in the ADMIXTURE analysis (Table S3) and ancient samples were grouped according to the site and approximate age (Tables S1 and S3). For the 'Left' populations, we tested all possible pairs of Ozark sites grouping the samples according to their site (Table S1). We used the SNP dataset described in the MDS analysis section and set the *allsnps* option in qpWave to 'YES' in order to maximize the number of sites available for each test. We identify four groups among Ozark sites, each one consistent with a single migration wave (Figure 3C).

US maize differentially related to Mexican and South American landraces

D-statistics and qpGraph admixture modeling show that maize in the US carries varying proportions of ancestry derived from Mexican landraces. We used qpWave in order to identify groups of ancient and modern maize in the US that carry different proportions or sources of Mexican maize ancestry. For the 'Right' populations we used *diploperennis* as the fixed outgroup and selected modern landraces representatives of the main geographic groups outside the US: Andean maize, the Mexican Highlands, Pan-American maize and South American lowlands. For the 'Left' we populations we tested all possible pairs of ancient and modern maize samples individually. We used the SNP dataset described in the MDS analysis section and set the *allsnps* option in qpWave to 'YES' to maximize the number of sites available for each test. We expect that pairs of samples that carry similar proportions and ancestry sources of Mexican maize will be consistent with a single migration wave with this set of 'Right' populations. Our results show that at least four migration waves of ancestry from Mexican maize are necessary to explain the ancestry in the ancient maize from the US (Figure S3A). Overall, we identify five groups that are consistent with a single migration wave in this setup: (1) the two most ancient samples from the US SW (Batcave_17 and McEuen_43), (2) upland US SW (Three Fir Shelter and Turkey Pen Shelter), (3) ancient Texas maize, (4) a third group formed by the Tularosa Cave and the Ozark sites, and (5) the recent sample from the Ozark Buzzard Roost site.

Treemix graphs

We used TreeMix v. 1.13[64] to model the phylogenetic relationships of the ancient maize from the US and the Romero's Cave maize. We ran TreeMix in two datasets: one including ancient and modern maize from the US (Figure S3B), a second one including ancient and modern maize from Central and South America (Figure S3C). In each case we used FrAnTK[59] to estimate per population allele frequencies starting from the dataset described in the MDS analysis section and grouped the sample according to the ancestry clusters identified in ADMIXTURE (for the modern samples) and qpWave analysis (for the ancient samples; Tables S1 and S2). TreeMix was run assuming 0 to 10 migration edges and for each number of migrations a total of 10 replicates starting at different seed values were run and the replicate with the best likelihood was kept. The US maize dataset consists of 23 groups samples and 51,652 transversion sites. The Central and South American dataset consisted of 17 groups of samples and 194,930 transversion sites.

Admixture graphs modeling

We evaluated the evolutionary relationships of different groups of ancient maize using admixture graphs as implemented in qpGraph[41] and the dataset described in the "reference data and SNP calling" section. In brief, qpGraph estimates branch length and admixture proportions of a predefined admixture graph and evaluates its fit based on the estimated and expected f_4-statistics among a set of samples. To obtain the best fitting admixture graph(s) we followed a procedure similar to the one described in Moreno-Mayar et al.[84] First, we built a base graph with representatives of the main genetic groups contributing to the ancestry of maize in the

US as shown by the MDS and clustering analyses (Figures 2A and S2B), and then incorporated each of the ancient US maize groups one by one. Admixture graphs were evaluated based on the *z*-score of the f_4-statistic with the worst fit and the score. We considered a graph had a good fit if the absolute value of the worst f_4-statistic's z-score was $\leq$ 3.33. Additionally, where more than one graph fitted the data we used the qpGraph score to select those with best fit; if two or more graphs had a difference in their overall score of $\leq$3 (p=0.05) we considered they had equally good fit.[85]

The following groups were included in the base graph: the 3,390 year-old maize genome from Bat Cave (representative of the initial migration of maize into the US Southwest), the 2,424 year-old Romero's Cave maize (representative of the Pan-American maize given its basal position in this lineage), the modern West Mexico highland maize, wild *mexicana* maize (contributes to highland maize), the 5,310 year-old maize genome from the Tehuacan Valley (represents an early lineage equidistant to all domesticated maize as the root of domesticated maize[32]) and *Zea diploperennis* as outgroup. To build the base graph, we started by using *admixturegraph* R package[86] to list all the possible tree topologies including all six groups. For each tree topology we used qpGraph to estimate the branch lengths and evaluated the obtained score and worst fitting *z*-score. Since none of the trees had a good fit (|z|>3.33), we selected the topology with the best score and added a migration edge to all possible branches using *admixturegraph* R package and fitted the graphs using qpGraph. From the resulting graphs we selected the ones with the best fit and repeated the process of adding a migration edge. After incorporating two migration edges, we obtained eight admixture graphs that fitted the data in all cases recovering the bi-directional admixture between wild *mexicana* admixture and West Mexico maize.[35,87] Starting from these eight admixture graphs, we added the remaining ancient maize groups sequentially in the following order: Turkey Pen Shelter, Tularosa Cave, Spirit Eye Cave and Ozark's Putnam. Each group was first added as a non-admixed branch, and then as a mixture of two branches (admixed) considering all possible combinations of branches. We evaluated the resulting graphs and selected the ones with the best fit to move to the next ancient maize group.

Estimating admixture proportions

The best admixture graph models show that the maize from the Putnam site is a mixture of two lineages: one that is most closely related to maize in the Spirit Eye Cave in Texas (58%) and a second one most closely related to maize in the Turkey Pen Shelter in upland US SW (42%). To estimate the admixture proportions of these ancestries for the remaining sites in the Ozarks rockshelter, we incorporated each of them independently to the best model before incorporating the Putnam maize. For the recent sample from the Buzzard Roost site, which carries additional ancestry from the Pan-American maize lineage, we incorporated this sample to the graph in Figure 3A in order to model its ancestry. For each of the Ozark sites we selected the graph(s) with the best fit as described in the previous section. Admixture graphs for each of the Ozark sites are available in the figshare repository under the DOI: https://doi.org/10.6084/m9.figshare.27287871. Ancestry proportions estimated for each site are shown in Figure 3D.

D-statistics to test treeness and admixture

D-statistics using FrAnTK

We used *D*-statistics as implemented in FrAnTK[59] in order to evaluate the phylogenetic placement and potential gene flow between the ancient and modern samples. In particular, we tested key features obtained in the admixture graph model in the f_4-statistics admixture (Figure 3A) and Treemix graphs (Figures S3B and S3C). Similar to the f_3-statistics, we used the SNP dataset described in the "reference data and SNP calling" section. We assessed the significance of the tests through a weighted block jackknife procedure over 5.5 kb blocks which account for the linkage disequilibrium observed in the maize genome.[28] Deviations from D=0 were presumed significant if the observed Z-score was above or below 3.33 (|Z|>3.33). Each test performed is described below.

Romero's Cave samples are part of the Pan-American group: Admixture and MDS results showed the Romero's Cave samples shared most of their ancestry with the Pan-American maize (Figures 2A and S2B). Furthermore, TreeMix admixture graphs suggested the Romero's Cave maize split from the common ancestor of Pan-American and South American maize. We used *D*-statistics to test if the ancient Romero maize was equidistant to every pair of Pan-American and South American maize landraces, as suggested by the admixture graph, and using *diploperennis* as outgroup. Results were consistent with Romero maize splitting from the common ancestor of the lowland South American and Pan-American maize lineages (|Z|$\leq$ 3.3). The oldest maize remains in the Romero's and Valenzuela's Caves date back to 4,000-4,500 yr B.P., but it was not until ~2,400 yr B.P. that human populations in the area started cultivating maize at an abundance comparable to that of a food staple. Our results show that modern maize in the region derives from the same lineage that was present since 2,700 yr B.P.. However, despite being the only other archaeological site midway from the domestication center and towards the US southwest, samples in the Romero's Cave are most likely a different migration wave northward from the domestication center than the one that gave rise to landraces in the US.

Present-day maize in the US carries varying proportions of Mexican Highland and Pan-American maize: Admixture graph modeling showed ancient maize from the US SW derives from a mixture of the initial introduction of maize and Mexican maize (Figure 3A). We used *D*-statistics to test the extent and sources of Mexican maize ancestry in the ancient maize from the US. We tested whether maize from the different archaeological sites in the US and modern US landraces share more alleles with maize from Mexico than the 3390-year-old Bat Cave (as representative of the initial maize that was introduced into the US SW). In particular, we computed a *D*-statistic of the form *D*(Bat Cave, H3; Mexican maize, Tripsacum), where H3 corresponds to all ancient and modern maize from the US and Mexican maize corresponds to the two genetic groups of maize in Mexico (Mexican Highlands in the West and the pan-American lineage in the East). Our results show that all ancient and modern maize from the US, except for the 2700-year-old McEuen Cave maize, shares more alleles with Mexican maize compared to the Bat Cave sample (Figure S3D). The McEuen Cave maize sample represents the second oldest sample from the US SW (after Bat Cave sample), for which genomic data has been generated.[27] The fact that the McEuen Cave sample does not carry additional Mexican ancestry could indicate that the first wave of Mexican maize ancestry occurred only after ∼2700 yr B.P., or that, if it arrived earlier, it did not reach all maize cultivated in the region.

Identifying the best admixture source for Ozark Buzzard Roost sample: Clustering analysis (Figure 2A), D-statistics (Figure 3C) and admixture graphs (Figure S3B) show that the recent sample from the Ozark Buzzard Roost site is a mixture of Ozark's maize ancestry and ancestry most closely related to maize in East Mexico, Central and South America, similar to Southern Dent landraces. To identify the best source for the non-Ozark ancestry we estimated a *D*-statistic of the form *D*(Putnam site, Buzzard Roost; H3, *diploperennis*), where H3 represents all potential sources of ancestry (Figure S3F). Our results show that for modern Southern Dent landraces the Pan-American lineage is the best source of admixture and in the case of the recent sample from the Buzzard Roost site Romero's Cave maize and Pan-American lineage are the best sources of admixture.

Error-corrected D-statistics using ANGSD

D-statistics and admixture graph modeling showed ancient and modern maize from the US carries varying proportions of ancestry from Mexican maize. We next tested whether this ancestry came from the Highland Mexican maize in the West, the Pan-American maize in the East or the ancestor of both using *D*-statistics. We computed a test of the form *D*(West Highland Mexico, Romero's Cave; Bat Cave, H3), where H3 represents ancient and modern maize from the US. Since in most cases three of the samples are ancient genomes, *D*-statistics were computed using ANGSD doabbababa2 funcion, which accounts for differential error in the samples potentially derived from ancient DNA damage.[42] Additionally, given that Highland Mexican maize carries additional ancestry from the wild subsp. *mexicana*,[35,43] which would decrease the shared alleles between West Mexican maize and H3, we considered varying proportions of subsp. *mexicana* (0-28%) in West Mexico maize. To do so, we computed a test of the form *D*(subsp. *mexicana*, Romero's Cave; Bat Cave, H3) and subtracted varying proportions of it from the corresponding test with West Highland Mexico instead of subsp. *mexicana*.

ANGSD -doAncError was used to estimate error rates for each of the samples included in the tests as described in the section "estimating type-specific error rates" and using *Z. diploperennis* as the ancestral genome and landrace RIMMA1010 as the perfect genome. The *Z. diploperennis* FASTA sequence was masked using the mappability mask described in Ramos-Madrigal et al.[32] *D*-statistics were estimated using ANGSD doabbababa2 restricting to reads with mapping quality ≥30 and bases with quality ≥ 20. Additionally, a mappability mask was applied to the Bat Cave sample in order to restrict to regions that can be unambiguously mapped.[32] A block jackknife procedure over 5500 bp blocks was used to obtain confidence intervals for each test. Samples were pooled per group, according to the groups defined by the qpWave and model-based clustering analyses as indicated in Table S3.

Our results show that West Mexico maize is the best source for the Mexican ancestry in maize in the US SW and Ozark sites (Figures 3C and S3E), consistent with the admixture graph (Figure 3A). Contrastingly, the best source for the Southern Dents, sweet corn and the recent sample from the Ozark Buzzard Roost site is the Romero's Cave maize. Finally, the admixture in maize from the Tranquil Rockshelter and Spirit Eye Cave in Texas is from maize that is equidistant to the West Mexico and Romero's Cave maize, also consistent with the admixture graph (Figure 3A).

Population Branch Statistic analysis

We used the Population Branch Statistic to measure changes in allele frequencies in the Ozark maize since its divergence from maize in the US Southwest. The PBS identifies SNPs that show strong changes in allele frequencies in a focal population compared to a contrast population and an outgroup. We used 16 *parviglumis* samples as an outgroup and tested the following groups from the US Southwest as contrast population: Three Fir Shelter (n=5), Tularosa Cave 1.8ka (n=9), Tularosa Cave 750 (n=10), Turkey Pen Shelter (n=13), present-day maize from Eastern US (n=9), present-day maize from the US Southwest (n=6) and all ancient maize from the US Southwest combined (n=27). For the Ozark maize group we included all samples with the exception of the recent sample from the Buzzard Roost site (n=17).

To estimate the PBS, we used the genotype-likelihoods (GL) approach implemented in ANGSD v0.931 to account for the low coverage in the data. This approach has been previously demonstrated to work with medium to low coverage data from ancient samples.[27] First we estimated GL for each of the populations in the test at sites with a minimum depth of coverage of 3 and maximum missingness of 50% using the GATK model (-GL 2) implemented in ANGSD. Reads with mapping quality below 30, bases with quality below 20 and transitions were discarded. The GL were used to obtain maximum-likelihood estimates of the 2-dimensional site frequency spectrum for all possible pairs of maize populations using realSFS.[88] Then, we calculated per-site weighted F_{ST} between pairs of populations using realSFS. The F_{ST} estimates were used to compute PBS for each gene and for the different arrangements as described in Yi et al.[46] We only considered genes with a minimum of 10 SNP sites. Results are shown in Figures 4A, S4A, and S4B.

Annotation of the *wx1* gene

SNPeff analysis

The PBS analysis identified two SNPs overlapping with the *wx1* gene with high PBS in the Ozark maize (G/A substitution at chr9:23,270,176 and T/A substitution at chr9:23,270,283). In both cases the derived allele (polarized using *Tripsacum*) is found at high frequency in the Ozark maize (0.73 and 0.85) compared to the US Southwest (0.16-0.22 and 0.04-0.27) and both alleles are segregating in *parviglumis*.

To annotate these nucleotide substitutions and to evaluate their functional impact used SNPeff v5.[65] GATK *HaplotypeCaller*[57] was used to call genotypes for sample 204_Salts_Bluff, which carried the alleles with high frequency among Ozark maize samples and had the highest depth of coverage (5.02$\times$ at the mappable regions of the genome). We did not perform any further filtering of the genotype calls, given we were only interested in estimating the functional impact on the differentiated SNPs. SNPeff was run using the genotype calls and the SNPeff pre-built Zea_maysv3_29 database. The first SNP (G/A, chr9:23270176) is located in a *wx1* intron, while the second SNP (T/A, chr9:23270283) is located in the fifth exon of the *wx1* gene and leads to an amino acid substitution (aspartate to valine) in the protein.

To visualize the variation around the two differentiated SNPs, we used ANGSD to obtain allele counts for the region around the SNPs (Figure S4C). Read with mapping quality $\leq$30 and bases with quality $\leq$20 were discarded. Figure S4C shows that the alleles found in at these two SNPs in the Ozark maize co-occur in most samples where they are present. Since these two SNPs are located only 107 bp apart, their co-occurrence may suggest they are linked. Additionally, the fact linkage-disequilibrium in maize breaks quickly could explain why we find only two SNPs with high PBS in the *wx1* gene.[28]

Alphafold2.0 protein structure modeling

We used alphafold2.0 to investigate the potential impact of the Ozark maize amino acid substitution in the 3D structure of the WAXY1 protein.[52] We obtained the WAXY reference protein sequence (P04713) from the UniProt database and created two versions of the protein sequence: one with the aspartate at position 180 (Ozark maize version) and a second one with a valine at position 180 (US SW version). We then used alphafold2.0 to reconstruct and compare the two 3D structures. Alphafold predicts that the A180D amino acid substitution is located at a surface accessible site but it does not lead to a change in the protein structure (Figure S4D). A further post-translational modification prediction analysis suggested that a tyrosine (Y) three positions upstream from the A180D amino acid substitution is phosphorylated in the US SW version of the protein but not in the Ozark maize version.

Supplemental figures

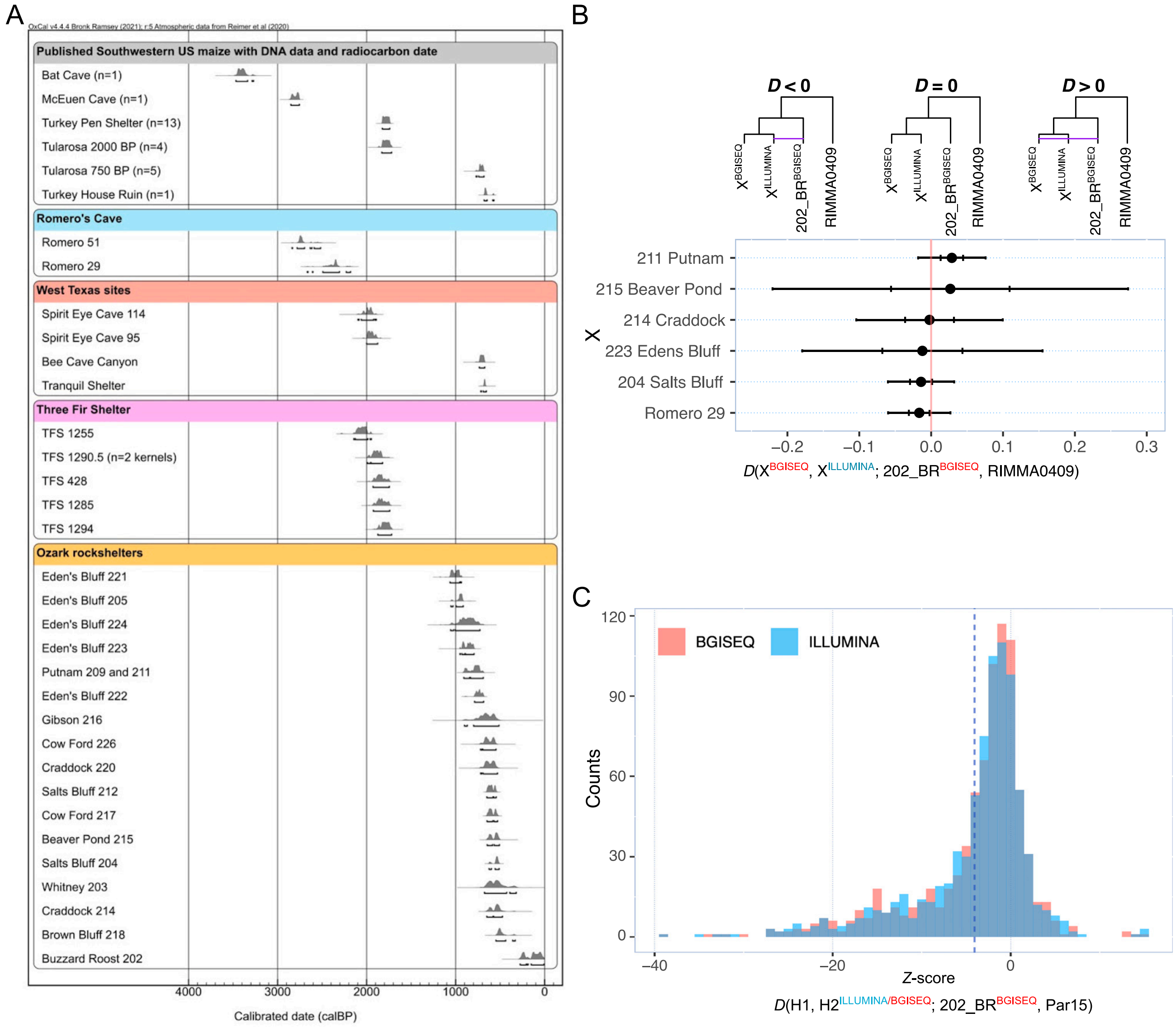


Figure S1. Radiocarbon calibrations and examination of potential biases generated by different sequencing platforms (Illumina HiSeq 2500 and BGISEQ-500), related to Figure 1

(A) Radiocarbon calibrations for individual samples and published assemblages.

(B) *D*-statistic of the form $D(X^{BGISEQ}, X^{ILLUMINA}; 202\_BR^{BGISEQ}, RIMMA0409)$, where X^{BGISEQ} and $X^{ILLUMINA}$ represent the paired samples sequenced in BGISeq-500 and Illumina HiSeq 2500, and 202_BR represents the youngest Ozark sample sequenced with BGISeq-500. Individual points show the value of *D* obtained for each test and error bars show 3.3 SE estimated through a block jackknife procedure. Significant deviation from $D \sim 0$ toward positive values would indicate that X^{BGISEQ} and 202_BR^{BGISEQ} are artificially closer. We do not find any significant deviation from $D \sim 0$.

(C) Distribution of *Z* scores obtained from a *D*-statistic of the form D(H1, H2, 202_BR, Par15), where H1 represents all samples in the whole-genome dataset (n = 239) and H2 represents one of the paired sequenced samples sequenced with Illumina (blue) or BGISeq-500 (red). We find no statistically significant difference in the *Z* score distributions (*ks*.test, p value > 0.05).

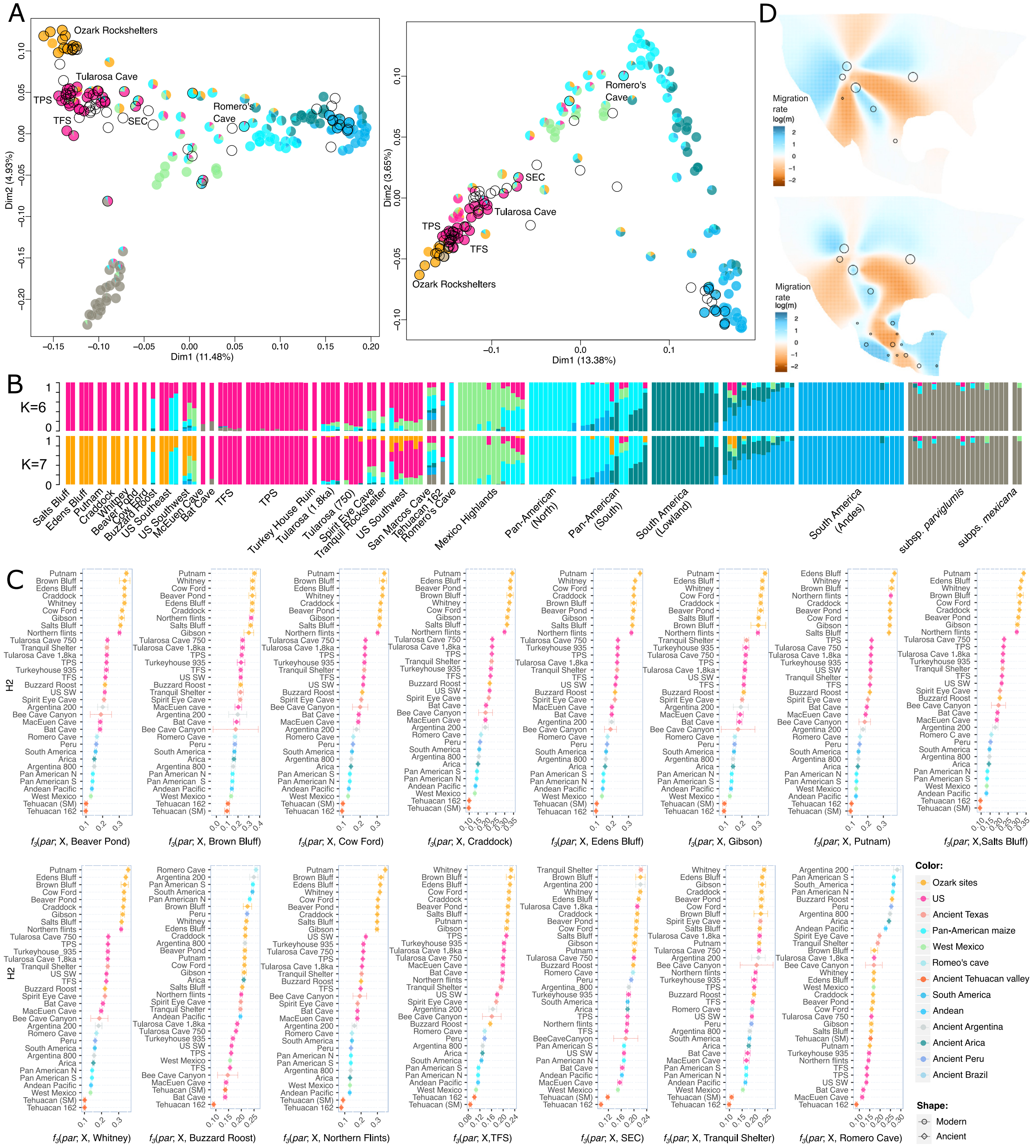


Figure S2. Characterization of maize ancestry geographical patterns and migration routes, related to Figure 2

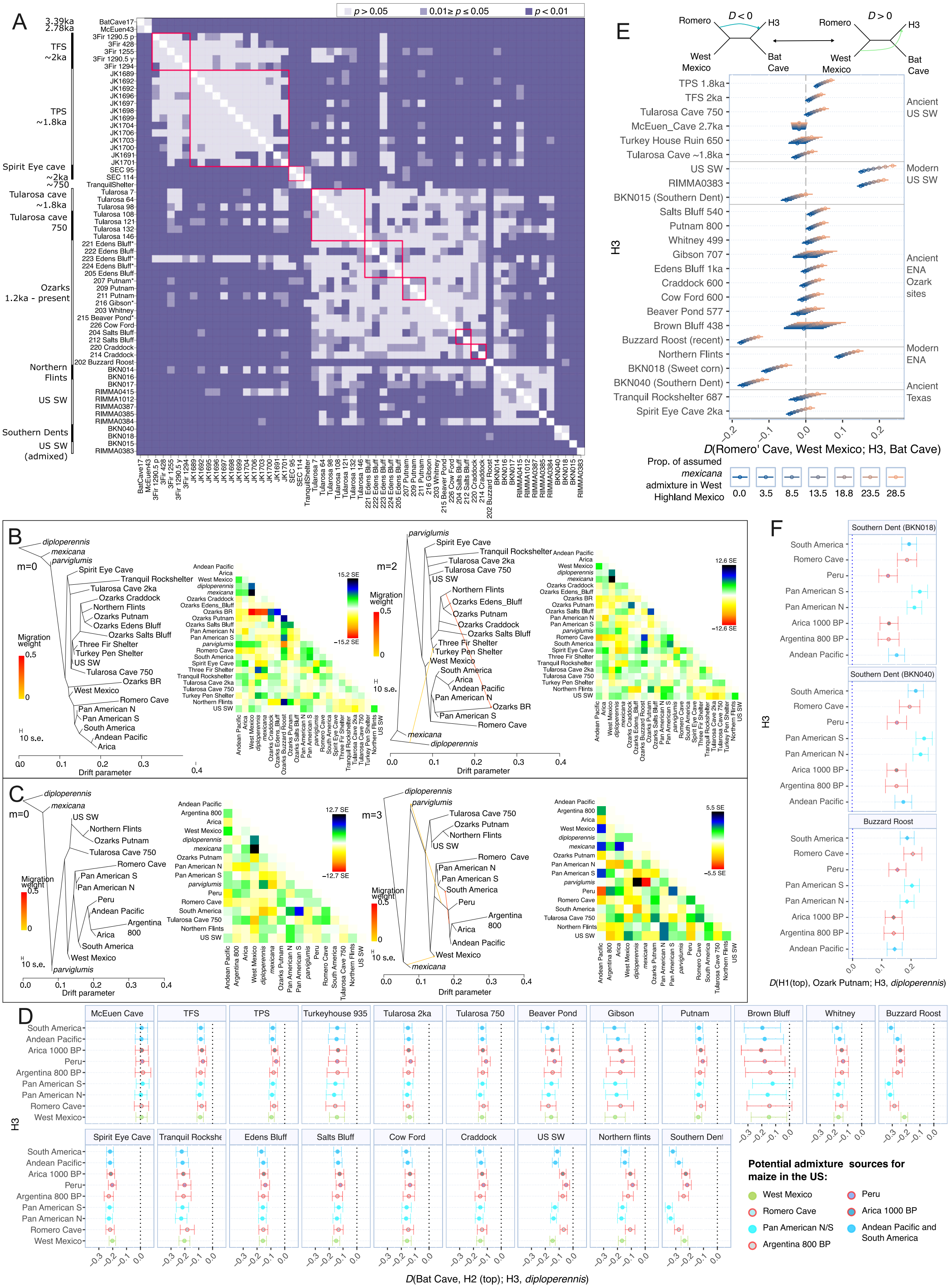
(A) MDS analysis including ancient, wild (teosinte), and domesticated maize (left) and excluding wild maize (right). Pie charts represent individual maize genomes, colors show the admixture proportions obtained from an ADMIXTURE analysis assuming 7 ancestry components (B), and empty circles represent samples not in the ADMIXTURE analysis. Ancient samples are indicated with a black outline. Names for relevant ancient samples are shown (TPS, Turkey Pen Shelter; TFS, Three Fir Shelter; and SEC, Spirit Eye Cave).

(B) Unsupervised clustering analysis using ADMIXTURE and assuming 6 and 7 ancestry components. Vertical bars represent different maize genomes, different colors show the ancestry components, and the proportion of each color represents the ancestry proportions.

(legend continued on next page)

(C) Outgroup f_3-statistics for the maize genomes from the Ozark rockshelters, modern Northern Flints, Three Fir Shelter, Spirit Eye Cave, Tranquil Rockshelter, and Romero's Cave. Each point indicates the f_3-statistic estimate. Error bars show 3.3 SE calculated using a block jackknife procedure. Colors indicate the different archaeological sites and modern maize groups in the y axis. Shapes indicate whether the samples are modern or ancient.

(D) EEMS results showing the estimated effective migration surfaces based on genomic and geographic data for datasets 2 (top; 1,000- to 3,000-year-old ancient genomes from Mexico and the US) and 3 (bottom; $\leq$2,000-year-old ancient genomes from the US and modern genomes from Mexico). Cooler and warmer colors indicate regions with high and low migration rates, respectively. Circles show the demes used by EEMS, which broadly correspond to the location of the samples included in the analysis.



(legend on next page)

Figure S3. Phylogenetic relationships and admixture patterns in ancient maize, related to Figure 3

(A) *qpWave* results for pairs of ancient maize samples and using *diploperennis* and representatives of maize ancestry groups outside the US as outgroups. A pink outline shows clusters of ancient maize genomes from the same archaeological sites. *Samples with missing data above 90%.

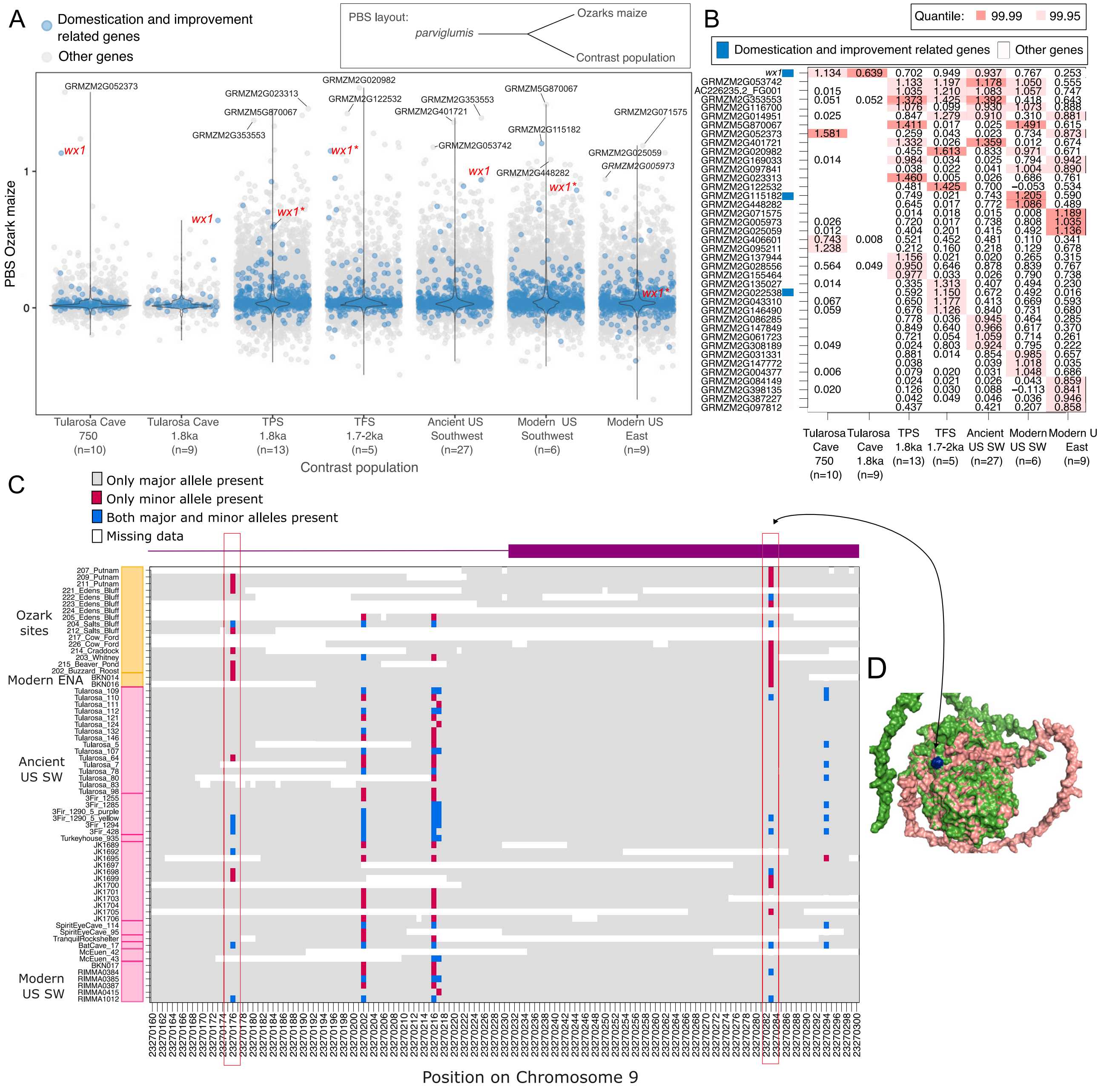
(B) Treemix admixture graphs focusing on maize from North America and including the new genomes from the Ozark Rockshelters, TFS, Spirit Eye Cave, and Tranquil Rockshelter (assuming 0 and 2 admixture events). Trees show the relationships between samples, and arrows show admixture events and their color shows the admixture proportion as indicated in the legend. Heatmaps show the residual values for each of the trees. Ozarks BR corresponds to the Buzzard Roost Ozark sample.

(C) Treemix admixture graphs focusing on maize from South America and including the ancient Romero's Cave maize (assuming 0 and 3 admixture events). Notation similar to that of (B).

(D) *D*-statistic tests of the form *D*(Bat Cave, H2; H3, *diploperennis*) testing potential ancestry sources for the admixture found in the ancient and modern US maize. For each ancestry group from the US (H2; top labels in individual panels), we tested whether they shared significantly more alleles with different maize groups outside the US (H3) compared with the Bat Cave maize. Individual points show the value of *D* obtained for each test and error bars show 3.3 SE estimated through a block jackknife procedure. All of the US maize groups (except for McEuen Cave) show significant negative results indicating admixture with the maize group in H3 (Z score < -3.3), with more negative values of *D* indicating better admixture sources (e.g., the Pan-American maize group is a better admixture source for Southern Dents).

(E) Error-corrected *D*-statistic of the form *D*(Romero's Cave, West Mexico; H3, Bat Cave) to test whether the Romero's Cave or West Mexico maize is the best source for the admixture in the different groups of maize in the US (H3). Individual points show the value of *D* obtained for each test and error bars show 3.3 SE estimated through a block jackknife procedure. Colors indicate different proportions of additional wild *mexicana* ancestry considered for the West Mexico maize. Tests significantly deviating from 0 indicate US maize (H3), with admixture that is most similar to West Mexico ($D > 0$) or Romero's cave maize ($D < 0$).

(F) *D*-statistic of the form *D*(H1, Putnam; H3, *diploperennis*) testing for admixture from potential sources in East Mexico and Central and South America in the two Southern Dent genomes from the maize HapMap2 (BKN018 and BKN040) and the Ozark maize genome from the Buzzard Roost site. Individual points show the value of *D* obtained for each test and error bars show 3.3 SE estimated through a block jackknife procedure. The recent maize sample from the Buzzard Roost site displays similar patterns to the Southern Dents, with the Pan-American and Romero's Cave maize showing the larges*t D* values.



	Tularosa Cave 750 (n=10)	Tularosa Cave 1.8ka (n=9)	TPS 1.8ka (n=13)	TFS 1.7-2ka (n=5)	Ancient US SW (n=27)	Modern US SW (n=6)	Modern US East (n=9)
wx1 🟦	1.134	0.639	0.702	0.949	0.937	0.767	0.253
GRMZM2G053742			1.133	1.197	1.178	1.050	0.555
AC226235.2_FG001	0.015		1.035	1.210	1.083	1.057	0.747
GRMZM2G353553	0.051	0.052	1.373	1.425	1.392	0.418	0.643
GRMZM2G116700			1.076	0.099	0.930	1.073	0.888
GRMZM2G014951	0.025		0.847	1.279	0.910	0.310	0.881
GRMZM5G870067			1.411	0.017	0.025	1.491	0.615
GRMZM2G052373	1.581		0.259	0.043	0.023	0.734	0.873
GRMZM2G401721			1.332	0.026	1.359	0.012	0.674
GRMZM2G020982			0.455	1.613	0.833	0.971	0.671
GRMZM2G169033	0.014		0.984	0.034	0.025	0.794	0.942
GRMZM2G097841			0.038	0.022	0.041	1.004	0.890
GRMZM2G023313			1.460	0.005	0.026	0.686	0.761
GRMZM2G122532			0.481	1.425	0.700	−0.053	0.534
GRMZM2G115182 🟦			0.749	0.021	0.743	1.205	0.590
GRMZM2G448282			0.645	0.017	0.772	1.086	0.489
GRMZM2G071575			0.014	0.018	0.015	0.008	1.189
GRMZM2G005973	0.026		0.720	0.017	0.738	0.808	1.035
GRMZM2G025059	0.012		0.404	0.201	0.415	0.492	1.136
GRMZM2G406601	0.743	0.008	0.521	0.452	0.481	0.110	0.341
GRMZM2G095211	1.238		0.212	0.160	0.218	0.129	0.678
GRMZM2G137944			1.156	0.021	0.020	0.265	0.315
GRMZM2G028556	0.564	0.049	0.950	0.646	0.878	0.839	0.767
GRMZM2G155464			0.977	0.033	0.026	0.790	0.738
GRMZM2G135027	0.014		0.335	1.313	0.407	0.494	0.230
GRMZM2G022538 🟦			0.592	1.150	0.672	0.492	0.016
GRMZM2G043310	0.067		0.650	1.177	0.413	0.669	0.593
GRMZM2G146490	0.059		0.676	1.126	0.840	0.731	0.680
GRMZM2G086285			0.778	0.036	0.945	0.464	0.285
GRMZM2G147849			0.849	0.640	0.966	0.617	0.370
GRMZM2G061723			0.721	0.054	1.059	0.714	0.261
GRMZM2G308189	0.049		0.024	0.803	0.924	0.795	0.222
GRMZM2G031331			0.881	0.014	0.854	0.985	0.657
GRMZM2G147772			0.038		0.039	1.018	0.035
GRMZM2G004377	0.006		0.079	0.020	0.031	1.048	0.686
GRMZM2G084149			0.024	0.021	0.026	0.043	0.859
GRMZM2G398135	0.020		0.126	0.030	0.088	−0.113	0.841
GRMZM2G387227			0.042	0.049	0.046	0.036	0.946
GRMZM2G097812			0.437		0.421	0.207	0.858

Figure S4. Signatures of selection in the ancient Ozark maize, related to Figure 4

(A) Violin plots showing Ozark maize PBS estimated using different contrast populations. PBS was estimated for genes with a minimum of 10 SNP sites. The names for genes above the 99.99 quantile are shown. *Cases where the *wx1* gene is shown but did not pass the 99.95 quantile threshold. Genes are colored in blue if they correspond to previously identified genes associated with maize domestication (n = 420) and improvement traits (n = 486).

(B) Heatmap showing the genes above the 99.95 (light pink) and 99.99 (dark pink) quantile of the PBS distribution for different contrast populations.

(C) Alleles present in the genomic region surrounding the two SNPs with high PBS for the Ozark maize. Each row represents a different sample from Eastern North America (orange) or the US Southwest (pink). Each column represents a position in the genome. Colors indicate whether the sample carries only the major allele (gray), only the minor allele (red), both alleles (blue), or lack coverage (white). The two high PBS sites are marked with red squares.

(D) Reconstruction of the WAXY1 protein structure using alphafold2.0. Region highlighted in blue shows the location of the amino acid substitution in the Ozark maize (corresponding to the position chr9:23,270,283).