Supplementary Information for

Archaeological Central American Maize Genomes Suggest Ancient Gene Flow from South America

Logan Kistler[a]*, Heather B. Thakar[b], Amber M. VanDerwarker[c], Alejandra Domic[d], Anders Bergström[e], Richard J. George[c], Thomas K. Harper[d], Robin G. Allaby[f], Kenneth Hirth[d], Douglas J. Kennett[c]*

[a]Department of Anthropology, Smithsonian Institution, Washington, DC, U.S.A.
[b]Department of Anthropology, Texas A&M University, College Station, TX 77843, U.S.A
[c]Department of Anthropology, University of California, Santa Barbara, CA 93106, U.S.A
[d]Department of Anthropology, The Pennsylvania State University, State College, PA 16802, U.S.A
[e]The Francis Crick Institute, London NW1 1AT, UK.
[f]School of Life Sciences, University of Warwick, Coventry, CV4 7AL, UK.

*Logan Kistler, Douglas J. Kennett
Email: KistlerL@si.edu, kennett@anth.ucsb.edu

This PDF file includes:

Figure S1

Other supplementary materials for this manuscript include the following:

Dataset S1 (Excel file)

www.pnas.org/cgi/doi/10.1073/pnas.2015560117

Morphometric Permutation Analysis

We performed 100,000 random permutations to estimate the probability of the observed difference in each morphometric variable (Fig. S1). Each iteration of the test randomly reassigned values with replacement to maximize the number of independent simulations. The absolute difference between the simulated groups were calculated and each value was compiled to generate a sampling distribution. Permutation p-values were calculated using the proportion of the 100,000 simulations that were larger than the observed difference. All statistics and permutation test were performed using R (1).

Radiocarbon Dating

Each maize cob was subsampled for AMS ^{14}C dating at in the Human Palaeoecology and Isotope Geochemistry Laboratory at the The Pennsylvania State University. Each Samples were then pretreated with repeated baths in1MHCl and NaOH at 70 °C for 30min on a heater block. A final acid wash removed secondary carbonates formed during the base treatment. Samples were then returned to neutral pH with two 15-min baths in deionized water at 70 °C to remove chlorides, and dried on a heater block. Sample CO_2 was produced by combustion at 900 °C for 3 h in evacuated sealed quartz tubes using a CuO oxygen source and Ag wire to remove chloride compounds. Primary (OX-1) and secondary (FIRI-D/F, FIRI-H) standards and a Queets Wood background were selected to match the sample age and underwent the same chemical steps for quality assurance. Samples were graphitized at the Keck Carbon Cycle Accelerator Mass Spectrometer facility at the University of California, Irvine and AMS ^{14}C measurements were made using a modified NEC 1.5SDH-1 instrument; National Electrostatics Corporation. All ^{14}C ages were $\delta^{13}C$-corrected for mass-dependent fractionation with measured $^{13}C/^{12}C$ values (2) and calibrated using OxCal version 4.3 (3) using the IntCal13 northern hemisphere curve (4).

DNA extraction and genomic data collection

26 maize samples were prepared in the dedicated ancient DNA clean lab facilities at the Penn State Anthropology Department (n=23), and the Smithsonian Institution's Museum Support Center (n=3). Standard protocols to prevent and detect contamination were utilized (5), including strict workflow procedures, frequent cleaning with bleach and ethanol, use of complete personal protective equipment, and the preparation and sequencing of negative control reactions. Both labs are equipped with filtered air handling and are regularly decontaminated.

We extracted DNA, prepared sequencing libraries, and screened samples following established protocols for highly degraded ancient DNA (Supplemental Methods), identifying EG84, EG85, and EG90 as suitable for genomic sequencing.

DNA was extracted from archaeological maize cob (n=24), stem (n=1), and leaf (n=1) tissue exactly following the protocol described by Wales and Kistler (6). At Penn State, we prepared DNA sequencing libraries exactly as described in Kistler et al (7), and at the Smithsonian we employed the Blunt End Single Tube procedure (8) with modifications described in (9), dual indexing with primers and primer sequences described in (10), and Platinum Taq High-Fidelity (Invitrogen) for library amplification. Samples were pooled in roughly equimolar ratios, and screened on a NextSeq 550 High-output flow cell with 75bp single-end reads, and a HiSeq X10 lane with 150bp paired-end reads. Samples with sufficient endogenous DNA for genome-scale analyses were sequenced completely on a HiSeq X10. All sequencing was carried out at Admera Health, South Plainfield, NJ.

Sample reads were adapter-trimmed and paired reads were merged using AdapterRemoval 2 (11), and mapped to the maize reference genome (Zea mays B73 RefGen_v4; (12)) using the Burrowes-Wheeler Aligner *aln* function (13) with seed disabled to improve ancient DNA mapping (11) and a minimum mapping quality of 20. We used mapDamage 2.0 (14) to verify cytosine deamination profiles consistent with authentically ancient DNA, and all 5' thymine and 3' adenine residues were hard-masked within 5nt of sequence ends where deamination was most concentrated. All analyses were restricted to the strictly mappable fraction of the maize genome, as previously described (15); mappability mask previously published in (15).

Using the set of 17,672,809 SNPs described in (15), we generated pseudohaplotype SNP calls at all sites with a minimum 2x consensus, exactly as described previously (15). Using this approach we recovered 1,786,417, 3,312,860, and 2,243,175 SNPs from EG84, EG85, and EG90 respectively. In addition, we followed the approach of (16) and combined the three samples for most analyses, treating them as a single population sample. We merged the alignment files for the three samples, and re-called the SNP panel as a single set of pseudohaplotypes in this case, yielding 7,666,836 SNPs for analysis. We combined new SNP calls from El Gigante with the previously reported set of SNP calls available at (15), consisting of 109 modern genomes and 11 ancient genomes before culling for missingness during analyses. For analyses assuming SNPs in linkage equilibrium (e.g. model-based clustering), we pruned for linkage using the plink "--indep-pairwise" function. The complete SNP dataset including separate and combined El Gigante maize is available on Dryad.

Genomic analyses

Model-based clustering

We used ADMIXTURE (17) to estimate ancestry proportions under model-based clustering in maize genomes, excluding teosintes and the partially domesticated mid-Holocene genomes from Mexico (18, 19). We included all genomes with at least 25% of sites called, enforced a minimum 50% of samples called to retain a site, minimum minor allele frequency of 0.02, and using the LD-pruned SNP set. This analysis included 4,252,422 SNPs and 98 genomes, using the combined El Gigante dataset and setting k=5 as previously established (7). We ran 100 independent analyses, and compared the results by final log-likelihood (lnL). lnL values were bimodal. The majority of runs, 72%, clustered tightly together with a higher lnL range, with the remaining 28% represented a more diffuse lower tier. Among the better supported upper tier, the El Gigante genome contains an estimated 97.73%–98.04% Pan-American ancestry, and 1.95–2.26% South

American ancestry. No other ancestry cluster contributes significantly (max 0.0016%) in any run. The small proportion of South American ancestry is attributed to the Andean/Pacific group. Given genetic and historical ties between this and the Lowland lineage (7), we interpret this as a generic signature for an ancestral South American gene pool.

f-statistics

We used previously released scripts for *f4* and outgroup-*f3* calculations ((15) plink2freq.pl, f3.pl, f4.pl), and included the complete, unpruned dataset at all sites where the outgroup *Tripsacum dactyloides* was present. We used a block jackknife resampling procedure with 5Mb blocks to estimate standard error and calculate a Z-score to assess fit to the null hypothesis. Statistical significance for rejecting the null hypothesis was concluded where |Z| > 3.

Ancestry informative marker (AIM) domestication analysis discovery

We used a perl script (Dryad: AIM.pl) to calculate ancestry informativeness (I_n; (20) between pairs of populations at all SNPs called in at least half the samples from each group. Following previous research into maize domestication status (18), we designated all SNPs with $I_n \geq 0.1$ as ancestry informative markers. For domestication analysis, we compared 1) all modern domesticated maize with ≤ 20% ancestry in the highland Central American cluster with 2) all *parviglumis* and *mexicana* genomes. Highland Central American maize carries previously documented admixture from highland *mexicana* teosinte (21), and thus all sites are not reliably maize-like for I_n determination. We assessed a panel of previously identified genes associated with domestication (22) containing at least 10 AIMs in the region containing the gene, plus 10kbp upstream and downstream, yielding 199 total genes. For each genome, we calculated the proportion of teosinte-like alleles in each domestication gene region with at least 5 called alleles at AIMs, as described above. The El Gigante set of proportions could then be compared against the set of comparable values for maize and teosinte as described in the maize text.

To assess the domestication status of individual genes, we used a likelihood-based method to test whether alleles at AIMs for a given gene were more likely drawn from a population resembling modern maize or modern teosinte. We considered all AIMs in each domestication gene regions as above, and computed a gene's likelihood of originating from a reference population (maize or teosinte) on the basis of the sample allele's frequency in the reference population. Log-likelihood was therefore calculated as the sum of the natural log of the frequency of the test sample's alleles in a reference population:

$$lnL = \sum_{i}^{nAIMs} \ln f(allele)$$

Where the test sample's allele was not present in the reference population, we set the allele frequency at a nominal 0.01 to preempt the log of 0, assuming that the test allele could easily be unsampled or lost to drift. We then computed Bayes Factors (BF) following (23) as the ratio of lnL values for maize and teosinte. We concluded "substantial" evidence for one of the competing hypothesis when BF ≥ 3 or BF ≤ 1/3, and "strong" evidence when BF ≥ 10 or BF ≤ 1/10, following (23).

For visualization (Fig. 2), we normalized the log-likelihood ratios on a scale of -1 to 1 as:

$$\frac{lnL_{maize} - lnL_{teosinte}}{lnL_{maize} + lnL_{teosinte}}$$

AIM analysis for South American affinity

For South American affinity AIM analysis, we compared 1) modern domesticated maize with ≥95% combined Andean/Pacific and Lowland South American ancestry with 2) modern domesticated maize having ≤5% combined Andean/Pacific and Lowland South American ancestry, and computed I_n to identify AIMs as above. We then assessed affinity to South American ancestry by computing the proportion of these geographic AIMs carrying South American alleles in each individual sample.

Admixture graph fitting

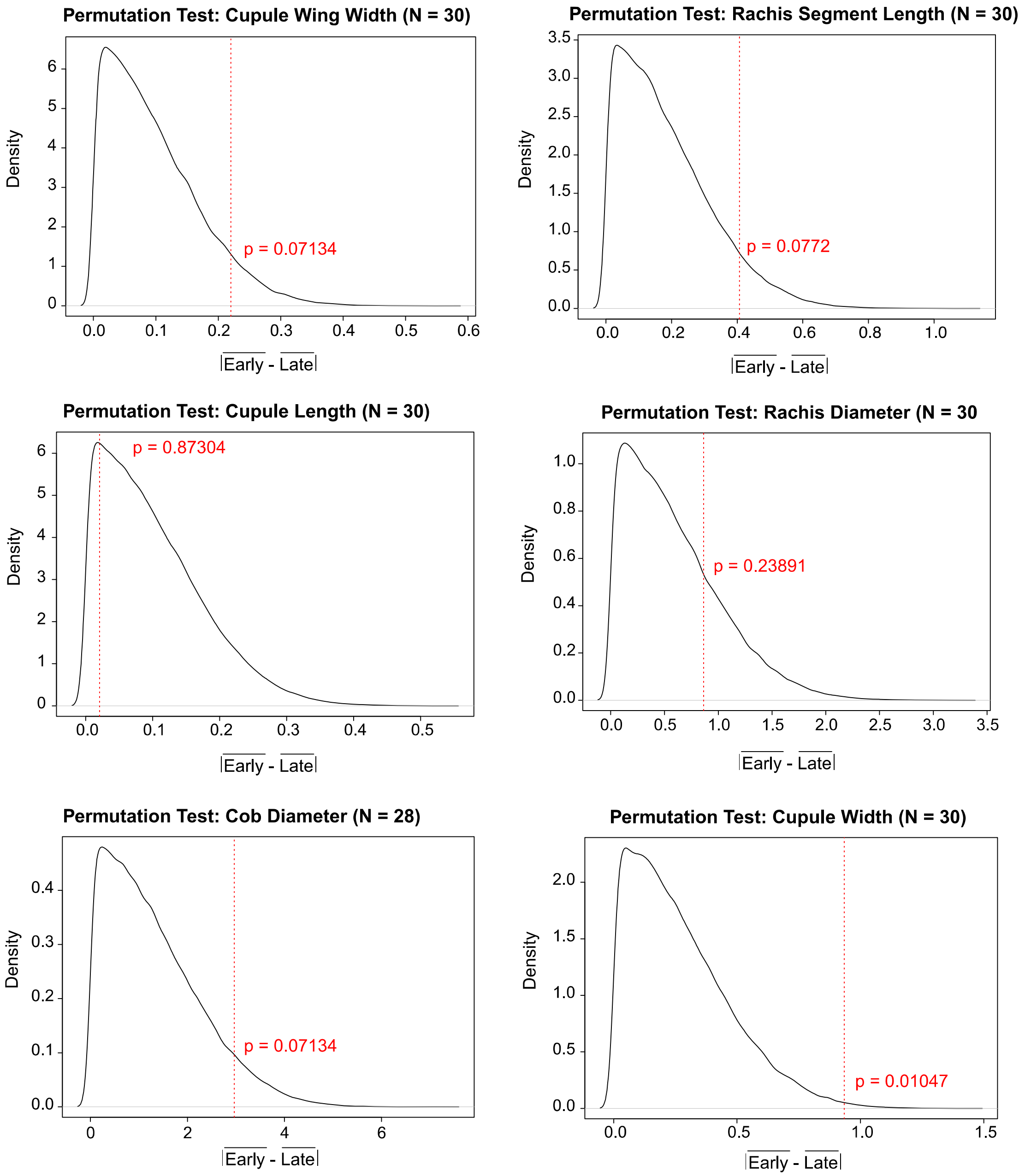
We included all samples with Pan-American or South American ancestry (≥99% on the basis of model-based clustering), plus El Gigante maize, all *parviglumis*, and *Tripsacum dactyloides* for an outgroup. We divided the Pan-American lineage into northern and southern geographic sets across the Isthmus of Panama, yielding 6 total populations. We first used AdmixTools (24) to calculate all permutations of f4-statistics as input for AdmixtureGraph (25), which we used to explore the permutation space of graphs. We first exhaustively enumerated all 3885 possible graphs relating these 6 populations with up to one admixture event, and fitted each of these to the f_4-statistics. Each graph was fit five times, retaining the best scoring fit (as evaluated using the "best_error" score). None of the graphs without any admixture events provided good fits. Among those with one admixture event, two graphs provided decent fits to the data, each with four minor outlier f_4-statistics (after these, the next best graph had nine outliers). The first of these was the topology (*Tripsacum*,(*parviglumis*,(South America,(El Gigante,(Pan-Am North,Pan-Am South))))) as shown in Figure 3a, with an admixture event from the ancestor of the South America lineage into Pan-Am South lineage. The second had the same structure, but with the admixture instead from the Pan-Am South lineage into the South American lineage. These two topologies are equivalent with respect to the f_4-statistics and thus achieved identical fits.

We then took the first of these graphs, studied the four outlier statistics that it did not perfectly predict, and hypothesized a second admixture event from the *parviglumis* lineage into the lineage ancestral to El Gigante, Pan-Am North and Pan-Am South. This improved the fit and left only one minor outlier statistic (|Z|=3.3), though the inferred admixture proportions were not stable across repeated fits. We then refitted this graph using qpGraph (24), which uses f_2 and f_3-statistics in addition to f_4-statistics, obtaining stable admixture proportions and achieving a good fit without any outlier f-statistics (largest $|Z| = 2.7$).

Genome Size Estimation

Heterochromatic knob content is the primary determinant of genome size differences in maize, and the proportion of sequence reads mappable to heterochromatic knobs has been demonstrated to reflect genome size estimates from flow cytometry (26, 27). We therefore used the proportion of reads mapping to the 180bp knob fraction of the maize genome as a proxy for genome size, following (26). We used a custom mapping strategy modeled after (28) to independently map sequence reads using a method capable of assigning reads to highly repetitive transposable element and heterochromatic knob fractions of the genome. We first generated a unique transposable element (UTE) and heterochromatic knob (knobC) reference set exactly as described in (28, 29) (reference fasta files curated on Dryad), and obtained the maize filtered gene set version ZmB73_5b_FGS_genes.fasta from (ftp.gramene.org). We created SMALT (https://www.sanger.ac.uk/tool/smalt-0/) indexes from these three reference targets with a step size of 3 and a word length of 12. We then used SMALT to first attempt mapping to the knobC set, then passed remaining unmapped reads to the UTE. Elements of the FGS also occur in the UTE, and therefore reads were treated as transposable rather than genic in origin if they could be assigned to the UTE, regardless of gene occupancy. Finally, only reads failing to map to both repetitive databases were handed down to the FGS for genic read alignment.

The proportion of all mapped reads assigned to the 180bp knob elements of the knobC fraction was summarized in terms of *RPKM*—reads per kilobase (following (29) and used for genome size analysis. In this case, $RPKM_{180\text{bp}} = R/(K \times M \times 10^{-6})$, where R is the number of reads mapped to 180bp knob elements in the knobC database, K is the combined length of the 180bp knob elements in the knobC database, and M is the total number of reads mapped to the combined knobC, UTE, and FGS genomic fractions—the complete mappable set of reads. Because of sequence-based and genomic biases in ancient DNA degradation (30), including specifically in maize (18), we did not attempt genome size estimation in archaeological maize.



Supplemental Fig. S1. Permutation analysis using morphometric variables of Early (4,340-4020 cal. BP) and Late (2,300-1,900 cal. BP). Permutation tests were performed using 100,000 random iteration to estimate the probability of the observed difference in each morphometric variable. Each iteration of the test randomly reassigned values with replacement to maximize the number of independent simulations. The absolute difference between the simulated groups were calculated and each value was compiled to generate a sampling distribution. Permutation p-values were calculated using the proportion of the 100,000 simulations that were larger than the observed difference. Three maize samples with dates after 1900 BP were exclude in the tests.

References

1. R CORE TEAM, R: a language and environment for statistical computing. 2020 (2020).

2. M. Stuiver, H. A. Polach, Discussion reporting of 14C data. *Radiocarbon* **19**, 355–363 (1977).

3. C. B. Ramsey, Methods for summarizing radiocarbon datasets. *Radiocarbon* **59**, 1809–1833 (2017).

4. P. J. Reimer, *et al.*, IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).

5. T. L. Fulton, B. Shapiro, "Setting up an ancient DNA laboratory" in *Ancient DNA*, Methods in Molecular Biology., B. Shapiro, *et al.*, Eds. (Springer New York, 2019), pp. 1–13.

6. N. Wales, L. Kistler, "Extraction of ancient DNA from plant remains" in *Ancient DNA*, (Humana Press, 2019).

7. L. Kistler, *et al.*, Multiproxy evidence highlights a complex evolutionary legacy of maize in South America. *Science* **362**, 1309–1313 (2018).

8. C. Carøe, *et al.*, Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **9**, 410–419 (2018).

9. S. S. T. Mak, *et al.*, Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *GigaScience* **6** (2017).

10. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* **2010**, pdb.prot5448-pdb.prot5448 (2010).

11. M. Schubert, S. Lindgreen, L. Orlando, AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).

12. Y. Jiao, *et al.*, Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).

13. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

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

15. L. Kistler, *et al.*, Data from: Multi-proxy evidence highlights a complex evolutionary legacy of maize in South America. *Dryad*, Dataset 10.5061/dryad.70t85k2.

16. R. R. da Fonseca, *et al.*, The origin and evolution of maize in the Southwestern United States. *Nature Plants* **1**, 14003 (2015).

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

18. J. Ramos-Madrigal, *et al.*, Genome sequence of a 5,310-year-old maize cob provides insights into the early stages of maize domestication. *Curr. Biol.* **26**, 3195–3201 (2016).

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

20. N. A. Rosenberg, L. M. Li, R. Ward, J. K. Pritchard, Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422 (2003).

21. J. van Heerwaarden, M. B. Hufford, J. Ross-Ibarra, Historical genomics of North American maize. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12420–12425 (2012).

22. M. B. Hufford, *et al.*, Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).

23. A. F. Jarosz, J. Wiley, What are the odds? A practical guide to computing and reporting Bayes factors. *J. Problem Solving* **7** (2014).

24. N. Patterson, *et al.*, Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).

25. K. Leppälä, S. V. Nielsen, T. Mailund, admixturegraph: an R package for admixture graph manipulation and fitting. *Bioinformatics* **33**, 1738–1740 (2017).

26. Y. Jian, *et al.*, Maize (*Zea mays* L.) genome size indicated by 180-bp knob abundance is associated with flowering time. *Sci. Rep.* **7**, 5954 (2017).

27. J.-M. Chia, *et al.*, Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).

28. C. M. Diez, E. Meca, M. I. Tenaillon, B. S. Gaut, Three groups of transposable elements with contrasting copy number dynamics and host responses in the maize (*Zea mays* ssp. *mays*) genome. *PLoS Genet.* **10**, e1004298 (2014).

29. M. I. Tenaillon, M. B. Hufford, B. S. Gaut, J. Ross-Ibarra, Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* **3**, 219–229 (2011).

30. L. Kistler, R. Ware, O. Smith, M. Collins, R. G. Allaby, A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res.* **45**, 6310–6320 (2017).

Dataset S1 Legend

Sample details for previously published modern maize and teosinte genomes included in analyses, including data source, SRA accession numbers, 180bp knob RPKM, location details, and inferred ancestry proportions.