**Gabriela Castellanos-Morales, Karen Y. Ruiz-Mondragón, Helena S. Hernández-Rosales, Guillermo Sánchez-de la Vega, Niza Gámez, Erika Aguirre-Planter, Salvador Montes-Hernández, Rafael Lira-Saade, Luis E. Eguiarte. Tracing back the origin of pumpkins (*Cucurbita pepo* ssp. *pepo* L.) in Mexico. Proceedings of the Royal Society B. doi: 10.1098/rspb.2019.1440.**

**Appendix S1. Methods**

*Chloroplast sequences*

Polymerase chain reaction (PCR) amplifications for chloroplast regions were performed in a 25 μL final reaction volume following conditions reported in [1]. The presence, size and quality of PCR products were confirmed by 1% agarose gel electrophoresis, dyed with EpiQuick DNA Stain (ET EpiGentek Group Inc.) and visualized with UV light. PCR products were sequenced at Macrogen (Macrogen USA) using forward and reverse primers. Sequences were checked for quality and assembled with CONSED 6.0 [2], taking each chloroplast region separately. We used BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to corroborate sequences and all amplified regionsshowed 99% identity score to its own region in *C. pepo*’s chloroplast complete genome (accession NC\_038229).

We downloaded the chloroplast genomes obtained by [3] from archaeological remains and modern samples (see Supplementary Table S1). Sequences for the *psbJ-petA* and *psbD-trnT* intergenic regions were extracted from these genomes. We used makeblastdb from BLAST+ [4] to organize the genomes in a database utilising the sequences from *C. pepo* ssp. *pepo* as query. We used blastn from BLAST+ to obtain coordinates within each genome (e-value < 1x10–10 for the BLAST query), and used extractseq from EMBOSS [5] to get the targeted sequences from each genome.

All sequences were aligned with ClustalW 2.0 [6] and checked manually. Sequence fragments for all regions were concatenated with DnaSP v5.10.1 [7] for a final alignment of 1,724 bp.

To perform the assignment test, we first conducted a discriminant analysis for principal components (DAPC) [8, 9] retaining 100 PC, and then we ran the *optimize.a.score* function to determine the best number of PC to explain our data. Accordingly, we retained 9 PCs for the final DAPC analysis, and plotted the group membership probability for each sampled individual.

To test for isolation by distance (IBD), we used concatenated sequences to obtain Nei’s pairwise *FST* considering the samples from Mexican *C. pepo* ssp. *pepo*, *C. pepo* ssp. *fraterna* and *C. pepo* ssp. *ovifera* with hierfstat [10]. We performed a Mantel test with 10,000 permutations to assess significance with vegan [11].

*Nuclear microsatellite loci*

To amplify 12 nuclear microsatellite loci, we used a multiplex approach in a 15 μl final volume following the conditions reported in [12], and verified PCR products in 2 % agarose gels. Amplicons were sent to the Roy J. Carver Biotechnology Center at the University of Illinois, USA for genotyping (biotech.illinois.edu). Electropherograms were analysed with PeakScanner (Applied Biosystems).

To determine the best number of K resulting from the Structure [13] analysis, we estimated an optimal *K* according to [14]’s test as implemented in Structure Harvester [15]. This test suggested a value of *K* = 2; but we felt this *K* value was not biologically informative, and it has been previously stated that [14]’s test has a tendency to estimate an optimal *K* = 2 [16]. Therefore, we decided to follow [13]’s suggestion to define a *K* value of biological meaning, and we also assessed the L(*K*) curve and chose the point where L(*K*) has the highest value and the lowest variance, which resulted in *K* = 3.

*ABC analyses*

We used data for nine nuclear microsatellite loci and two partitioned chloroplast regions (*petA*-*psb*J and *psbD-trnT*(GGU)) for three samples (*C. pepo* ssp. *fraterna*, *C. pepo* ssp. *ovifera* and *C. pepo* ssp. *pepo*) for hypothesis testing with DIYABC [17] (Supplementary Figure S1).

We first ensured that parameters yielded datasets similar to the empirical dataset by conducting initial simulations to refine the values for each prior and to select the summary statistics that were more informative for each molecular marker [18 – 20]. Initial priors were based on previously reported data, and broad priors were chosen to account for the absence of information to constrain priors for model parameters [21]. Estimated date of divergence between *C. pepo* ssp. *fraterna* and *C. pepo* ssp. *ovifera* was 1.33 Ma (0.47 – 2.4 HDP 95; [1]). Moreover, the reported dates for the oldest archaeological remains from *C. pepo* ssp. *pepo* from Mexico are 9,000 years before present in Guilá Naquitz [22, 23] and 7,600-9,000 years before present from Infiernillo cave in Ocampo, Tamaulipas, as well 5,200-6,000 years before present in Ocampo’s cave in the same area [24]. As pumpkins are annual plants, we considered generation time to correspond to one year. Accordingly, we set time of coalescence priors as t0 = current time; t1 = domestication period (normal prior from 5,000 to 100,000 years ago with a mean of 52,500 ± 47,500 s.d. by considering that lineage divergence may predate domestication) and t2 = lineage divergence period (normal prior from 100,000 to 2,500,000 years ago with a mean of 1,300,000 ± 1,200,000 s.d.) and restrained the analysis so that t2 > t1.

We also considered that *C. pepo* ssp. *pepo* and *C. pepo* ssp. *ovifera* might have undergone changes in their effective population sizes (*Ne*) after domestication as is suggested by their levels of genetic variation. Therefore, we included in the four scenarios changes in the effective population sizes for *C. pepo* ssp. *pepo*’s ancestral lineage(*NeaP*) in t1 and for *C. pepo* ssp. *ovifera* (*NeaO*) in t1-*a* (considering *a* with a normal distribution from 1,000 to 9,000 with a mean of 5,000 ± 4000, because archaeological remains for *C. pepo* ssp. *ovifera* are younger than those for *C. pepo* ssp. *pepo*). We set all effective population sizes with a uniform distribution, with past effective population size for *C. pepo* ssp. *pepo*’s ancestral lineageset as *NeaP* = 1,000 to 80,000 and for *C. pepo* ssp. *ovifera* set as *NeaO* = 100,000 to 1,000,000; and current effective population sizes for *C. pepo* ssp. *pepo* set as *Nepe* = 100 to 200,000, for *C. pepo* ssp. *ovifera* set as *Neovi* = 100 to 80,000 and for *C. pepo* ssp. *fraterna* set as *Nefra* = 100 to 10,000. For scenario 4 we set migration rate among *C. pepo* ssp. *ovifera* and *C. pepo* ssp. *fraterna* as *ra* and 1-*ra* for *C. pepo* ssp. *pepo*’s origin,with *ra* values from 0.001 to 0.999.

Mutation model for nuclear microsatellite loci was set as stepwise mutation model (SMM) with a minimum mean mutation rate (per locus per generation) of 1x10–7 and a maximum of 1x10–4, and an individual mutation rate of 1x10–8 and a maximum of 1x10–3. For chloroplast region *petA-psbJ* we set the mutation model to HKY with a mean mutation rate (per site per generation) from 1x10–11 to 1x10–8, and for *psbD*-*trnT* the mean mutation rate was set to 1x10–9 to 1x10–7.

A total of 51 summary statistics were calculated considering all markers (nine microsatellite loci and two partitioned chloroplast regions) and three genetic lineages or samples (*C. pepo* ssp. *ovifera*, *C. pepo* ssp. *fraterna* and Mexican *C. pepo* ssp. *pepo*) –in accordance to results from clustering analyses described in methods. One sample and two sample summary statistics are provided. One sample summary statistics for nine microsatellite loci for each sample were mean number of alleles, mean genetic diversity (6 summary statistics), while two sample (pairwise) summary statistics were mean number of alleles, mean genetic diversity and *FST* (9 summary statistics). For each chloroplast region, one sample summary statistics were mean number of haplotypes, number of segregating sites, mean of pairwise differences and Tajima’s D (12 summary statistics for each region), and two sample (pairwise) summary statistics were number of segregating sites and *FST* (6 summary statistics for each region) (Table A1).

We ran 4,000,000 simulations to evaluate ABC scenarios and parameters. Accordingly, prior fit to the observed data recovered that all microsatellite summary statistics and most summary statistics for chloroplast regions from simulated data were similar to those from observed data. Exceptions were eight summary statistics for scenario 1, seven summary statistics for scenario 2, 11 summary statistics for scenario 3, and five summary statistics for scenario 4 (Table A1).

To define which scenario has the best adjustment to our observed data, we obtained the posterior probability for each scenario by a direct approach and by the logistic regression approach implemented in DIYABC. For the direct approach, the posterior probabilities of each scenario is approximated by the fraction of simulations produced by each scenario closer to the real data; whereas for the logistic approach, a linear regression between a parameter and the vector of summary statistics of 1% of simulated datasets closest to the observed dataset is obtained, and the intercept condition where observed and simulated summary statistics coincide is used to estimate the posterior probability of each scenario [19, 25 – 27]. DIYABC also provides additional insight into the accuracy and power of model selection by estimating confidence intervals (C.I.) for the posterior probabilities of each scenario, where for adequate resolution in scenario choice we would expect that the C.I. of the model with the highest posterior probability does not overlap with the C.I. of the next supported scenario [19].

For the best-supported scenario, we report the median of each temporal and demographic parameters drawn from the posterior distribution [28] estimated with DIYABC 2.1.0. For this, a local regression on 1% simulations closest to the observed data is obtained, and the bias of the mean parameter estimates is assessed by generating 1,000 simulated datasets after a logit transformation of parameters. The logit transformation is implemented to reduce the inequality of variances among parameters [28]. The average relative bias, estimated by DIYABC, considers the differences between simulated points estimates and the true value, divided by the true value, and averaged over the 1,000 simulated datasets [18].

We evaluated the ability of the ABC analysis to discriminate between scenarios and to estimate type I and type II error rates for the direct and logistic approaches. In order to do so, 1,000 new datasets are simulated with each scenario, using each scenario as the null or alternative hypothesis, and the number of times the correct scenario has the highest posterior probability is summarized [17, 19, 26]. Therefore, type I error is the proportion of pseudo-observed datasets generated under a particular scenario that support a different scenario (false negative), and type II error is the proportion of pseudo-observed datasets generated under alternative scenarios that support the selected scenario (false positive) [19]. Then, to evaluate if the best-supported scenario could successfully reproduce the observed data, we simulated 10,000 pseudo-observed datasets from the posterior, and performed model checking implemented in DIYABC using the summary statistics not used during the inference step (Supplementary Figure S2) [18].

**Table A1.** Prior checking of summary statistics used to test scenarios in ABC analyses. One sample summary statistics for nine microsatellite loci for each sample were mean number of alleles, mean genetic diversity (6 summary statistics), and two sample (pairwise) summary statistics were mean number of alleles, mean genetic diversity and *FST* (9 summary statistics). For each chloroplast region, one sample summary statistics were mean number of haplotypes, number of segregating sites, mean of pairwise differences and Tajima’s D (12 summary statistics for each region), and two sample (pairwise) summary statistics were number of segregating sites and *FST* (6 summary statistics for each region). Asterisks denote those summary statistics in each scenario that showed statistical differences from the values estimated from observed data. Exceptions were eight summary statistics for scenario 1, seven summary statistics for scenario 2, 11 summary statistics for scenario 3, and five summary statistics for scenario 4.

| Summary statistic | Observed value | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 |
| --- | --- | --- | --- | --- | --- |
| NAL\_1\_1 | 10.778 | 0.3633 | 0.8523 | 0.6872 | 0.4405 |
| NAL\_1\_2 | 3.889 | 0.1884 | 0.1888 | 0.1842 | 0.189 |
| NAL\_1\_3 | 2.556 | 0.4459 | 0.4468 | 0.4501 | 0.447 |
| HET\_1\_1 | 0.579 | 0.0969 | 0.2206 | 0.1594 | 0.1254 |
| HET\_1\_2 | 0.566 | 0.1862 | 0.1866 | 0.1829 | 0.1866 |
| HET\_1\_3 | 0.268 | 0.3478 | 0.3476 | 0.3495 | 0.3484 |
| N2P\_1\_1&2 | 11.222 | 0.266 | 0.3782 | 0.29 | 0.2722 |
| N2P\_1\_1&3 | 11.000 | 0.3062 | 0.8537 | 0.5033 | 0.4235 |
| N2P\_1\_2&3 | 5.222 | 0.1522 | 0.1522 | 0.1402 | 0.1523 |
| H2P\_1\_1&2 | 0.601 | 0.0929 | 0.2022 | 0.1449 | 0.1171 |
| H2P\_1\_1&3 | 0.623 | 0.0928 | 0.2502 | 0.1484 | 0.1383 |
| H2P\_1\_2&3 | 0.585 | 0.1426 | 0.1425 | 0.1262 | 0.1427 |
| FST\_1\_1&2 | 0.258 | 0.8126 | 0.5863 | 0.643 | 0.7429 |
| FST\_1\_1&3 | 0.439 | 0.8578 | 0.9247 | 0.751 | 0.914 |
| FST\_1\_2&3 | 0.521 | 0.6576 | 0.6579 | 0.6461 | 0.6572 |
| NHA\_2\_1 | 1.000 | 0.1161 | 0.1547 | 0.1472 | 0.1142 |
| NHA\_2\_2 | 2.000 | 0.668 | 0.6676 | 0.6654 | 0.6685 |
| NHA\_2\_3 | 1.000 | 0.4622 | 0.4621 | 0.4621 | 0.462 |
| NSS\_2\_1 | 0.000 | 0.1161 | 0.1547 | 0.1472 | 0.1142 |
| NSS\_2\_2 | 12.000 | 0.9783\* | 0.9782\* | 0.973\* | 0.9784\* |
| NSS\_2\_3 | 0.000 | 0.4622 | 0.4621 | 0.4621 | 0.462 |
| MPD\_2\_1 | 0.000 | 0.1161 | 0.1547 | 0.1472 | 0.1142 |
| MPD\_2\_2 | 1.500 | 0.8904 | 0.8909 | 0.8848 | 0.8912 |
| MPD\_2\_3 | 0.000 | 0.4622 | 0.4621 | 0.4621 | 0.462 |
| DTA\_2\_1 | 0.000 | 0.432 | 0.6594 | 0.6223 | 0.4324 |
| DTA\_2\_2 | -2.236 | 0.0007\*\*\* | 0.0007\*\*\* | 0.001\*\*\* | 0.0007\*\*\* |
| DTA\_2\_3 | 0.000 | 0.5035 | 0.5034 | 0.5033 | 0.5033 |
| NS2\_2\_1&2 | 12.000 | 0.8333 | 0.5035 | 0.4319 | 0.6037 |
| NS2\_2\_1&3 | 0.000 | 0.0213\* | 0.124 | 0.019\* | 0.0583 |
| NS2\_2\_2&3 | 12.000 | 0.5683 | 0.568 | 0.4908 | 0.5683 |
| HST\_2\_1&2 | 0.997 | 0.8904 | 0.6476 | 0.6519 | 0.819 |
| HST\_2\_1&3 | 0.000 | 0.0502 | 0.5809 | 0.0303\* | 0.3609 |
| HST\_2\_2&3 | 0.684 | 0.0705 | 0.0701 | 0.053 | 0.0702 |
| NHA\_3\_1 | 5.000 | 0.1797 | 0.2026 | 0.1963 | 0.1842 |
| NHA\_3\_2 | 1.000 | 0.0759 | 0.0757 | 0.0752 | 0.0756 |
| NHA\_3\_3 | 2.000 | 0.6973 | 0.6969 | 0.6971 | 0.6963 |
| NSS\_3\_1 | 9.000 | 0.2606 | 0.3481 | 0.3293 | 0.2582 |
| NSS\_3\_2 | 0.000 | 0.0759 | 0.0757 | 0.0752 | 0.0756 |
| NSS\_3\_3 | 1.000 | 0.6849 | 0.6847 | 0.6849 | 0.684 |
| MPD\_3\_1 | 1.477 | 0.2738 | 0.4567 | 0.3954 | 0.2876 |
| MPD\_3\_2 | 0.000 | 0.0759 | 0.0757 | 0.0752 | 0.0756 |
| MPD\_3\_3 | 0.520 | 0.8571 | 0.8569 | 0.8579 | 0.8564 |
| DTA\_3\_1 | 0.006 | 0.4192 | 0.7925 | 0.7225 | 0.4673 |
| DTA\_3\_2 | 0.000 | 0.4661 | 0.465 | 0.4647 | 0.4648 |
| DTA\_3\_3 | 1.554 | 0.9729\* | 0.973\* | 0.973\* | 0.9728\* |
| NS2\_3\_1&2 | 11.000 | 0.1204 | 0.0451\* | 0.0346\* | 0.0731 |
| NS2\_3\_1&3 | 9.000 | 0.035\* | 0.2694 | 0.0298\* | 0.1214 |
| NS2\_3\_2&3 | 3.000 | 0.0126\* | 0.0126\* | 0.008\*\* | 0.0124\* |
| HST\_3\_1&2 | 0.590 | 0.5042 | 0.0048\*\* | 0.0027\*\* | 0.3859 |
| HST\_3\_1&3 | -0.180 | 0.0021\*\* | 0.1803 | 0.0008\*\*\* | 0.1625 |
| HST\_3\_2&3 | 0.861 | 0.04\* | 0.0401\* | 0.0284\* | 0.0399\* |

**References**

1. Castellanos-Morales G, Paredes-Torres LM, Gámez N, Hernández-Rosales HS, Sánchez-de la Vega G, Barrera-Redondo J, Aguirre-Planter E, Vázquez-Lobo A, Montes-Hernández S, Lira-Saade R, et al. 2018. Historical biogeography and phylogeny of *Cucurbita*: insights from ancestral area reconstruction and niche evolution. *Molecular Phylogenetics and Evolution* 128: 38-54. doi: 10.1016/j.ympev.2018.07.016.
2. Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Research* 8: 195-202. doi: 10.1101/gr.8.3.195.
3. Kistler L, Newsom LA, Ryan TM, Clarke AC, Smith BD, Perry GH. 2015. Gourds and squashes (*Cucurbita* spp.) adapted to megafaunal extinction and ecological anachronism through domestication. *Proceedings of the National Academy of Science of the United States of America* 112(49): 15107-15112. doi: 10.1073/pnas.1516109112.
4. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421. doi: 10.1186/1471-2105-10-421.
5. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6): 276-277. doi: 10.1016/S0168-9525(00)02024-2.
6. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettingan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948. doi: 10.1093/bioinformatics/btm404.
7. Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452. doi: 10.1093/bioinformatics/btp187.
8. Jombart T, Pontier D, Dufour A-B. 2009. Genetic markers in the playground of multivariate analysis. *Heredity* 102: 330-341. doi: 10.1038/hdy.2008.130.
9. Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11: 94. doi: 10.1186/1471-2156-11-94.
10. Goudet J. 2004. Hierfstat, a package for R to compute and test hierarchical *F*-statistics. *Molecular Ecology Notes* 5(1): 184-186. doi: 10.1111/j.1471-8286.2004.00828.x.
11. Dixon P. 2009. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* 14(6): 927-930. doi: 10.1111/j.1654-1103.2003.tb02228.x.
12. Sánchez-de la Vega G, Castellanos-Morales G, Gámez N, Hernández-Rosales HS, Vázquez-Lobo A, Aguirre-Planter E, Jaramillo-Correa JP, Montes-Hernández S, Lira-Saade R, Eguiarte LE. 2018. Genetic resources in the “calabaza pipiana” squash (*Cucurbita argyrosperma*) in Mexico: genetic diversity, genetic differentiation and distribution models. *Frontiers in Plant Science* 9: 400. doi: 10.3389/fpls.2018.00400.
13. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multiloci genotype data. *Genetics* 155(2): 945-959.
14. Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14(8): 2611-2620.
15. Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4(2): 359-361.
16. Janes JK, Miller JM, Dupuis JR, Malefant RM, Gorrell JC, Cullingham CI, Andrew RL. (2017) The K2 conundrum. *Molecular Ecology* 26: 3594-3602.
17. Cornuet JM, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin JM, Estoup A. 2014. DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphisms, DNA sequences and microsatellite data. *Bioinformatics* 30: 1187-1189. doi: 10.1093/bioinformatics/btt763.
18. Cornuet JM, Ravigne V, Estoup A. 2010. Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics* 11: 401. doi: 10.1186/1471-2105-11-401.
19. Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Molecular Ecology* 19: 2609-2625.
20. Chan LM, Brown JL, Yoder AD. 2011. Integrating statistical genetic and geospatial methods to bring new power to phylogeography. *Molecular Phylogenetics and Evolution* 59: 523-537. doi: 10.1016/j.ympev.2011.01.020.
21. Knowles LL, Alvarado-Serrano DF. 2010. Exploring the population genetic consequences of the colonization process with spatio-temporally explicit models: insights from coupled ecological, demographic and genetic models in montane grasshoppers. *Molecular Ecology* 19: 3727-3745.
22. Smith BD. 1997. The initial domestication of *Cucurbita pepo* in the Americas 10,000 years ago. *Science* 276: 932-934. doi: 10.1126/science.276.5314.932.
23. Flannery KV. 1986. *Guila Naquitz: archaic foraging and early agriculture in Oaxaca, Mexico*. Orlando, FL, USA: Academic Press.
24. Hanselka JK. 2017. Revisiting the archaeobotanical record of Romero’s cave in the Ocampo region of Tamaulipas, Mexico. *Journal of Ethnobotany* 37(1): 38-59. https://doi.org/10.2993/0278-0771-37.1.37.
25. Beaumont MA. 2008. Joint determination of topology, divergence time and immigration in population trees. In: Matsamura S, Forster P, Rrenfrew C, eds. *Simulation, Genetics and Human Prehistory* (pp. 135-154). Cambridge, UK: McDonald Institute for Archaeological Research.
26. Guillemaud T, Beaumont MA, Ciosi M, Cornuet J-M, Estoup A. 2010. Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity* 104: 88-99.
27. Robert CP, Cornuet J-M, Marin J-M, Pillai NS. 2011. Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Science of the United States of America* 108: 15112-15117. doi: 10.1073/pnas.110200108.
28. Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet J-M. 2004. Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *Bufo marinus*. *Evolution* 58: 2021-2036.