The role of environment, local adaptation and past climate fluctuation on the amount and distribution of genetic diversity in the teosinte in Mexico.

Gasca-Pineda, Jaime[1*], Yocelyn T. Gutiérrez-Guerrero[1], Erika Aguirre-Planter[1], Luis E. Eguiarte[1*]

[1] Laboratorio de Evolución Molecular y Experimental, Instituto de Ecología, Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, México.

* Correspondence: Jaime Gasca-Pineda and Luis E. Eguiarte, Laboratorio de Evolución Molecular y Experimental, Instituto de Ecología, Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, AP 70-275, Coyoacán, 04510, México DF, México. fruns@unam.mx

Keywords: genetic structure, microsatellite markers, demographic inference, isolation by distance, isolation by environment

Abstract

Wild maize, commonly known as teosinte, has a wide distribution in central Mexico and inhabits a wide range of environmental conditions. According to previous studies, the environment is a determinant factor for the amount and distribution of genetic diversity. In this study, we used a set of neutral markers to explore the influence of contemporary factors and historical environmental shifts on genetic diversity, including present and three historical periods. Using a

set of 22 nuclear microsatellite loci, we genotyped 527 individuals from 29 localities. We found highly variable levels of genetic diversity (*Z. m. parviglumis* H_E= 0.3646–0.7699; *Z. m. mexicana* H_E= 0.5885–0.7671) and significant genetic structure among localities (average D_{EST}= 0.4332). Also, we recovered significant values of heterozygote deficiency (average F_{IS}= 0.1796) and variable levels of selfing (sg^2=0.0–0.3090). The Bayesian assignment analysis yielded four genetic clusters dividing the sample into subspecies, that in turn, were separated into two clusters. Environmental conditions played a strong influence in the distribution of genetic diversity, as demographic analysis and changes in species range revealed by modeling analyses were consistent. We conclude that current genetic diversity in teosinte is the result of a mixture of local adaptation and genetic isolation along with historical environmental fluctuations.

Introduction

Zea mays comprises the cultivated maize (*Zea mays* ssp. mays L.), as well as wild maize populations commonly known as teosintes (Doebley & Iltis, 1980). The most common teosinte subspecies, *Z. m. parviglumis* and *Z. m. mexicana,* diverged ca. 61,000 years ago, and both are abundant and widely distributed in central-west México (Aguirre-Liguori, Aguirre-Planter & Eguiarte, 2016; Hufford *et al.*, 2012). *Z. m. parviglumis* grows in the lowlands (<1900m above sea level) in warm and moist conditions, while *Z. m. mexicana* is found in colder and dryer highlands (>1500m above sea level) (Aguirre-Liguori, Aguirre-Planter & Eguiarte, 2016). Aside from the intrinsic interest due to their relationship with cultivated crops (Aguirre-Liguori, Aguirre-Planter & Eguiarte, 2016), the wide distribution of teosintes and the contrasting environmental conditions inhabited by each subspecies make them an attractive system to study microevolutionary processes, including genetic divergence, local adaptation, and

phylogeographic patterns. Moreover, the public resources available for *Zea mays* (for example, https://www.maizegdb.org), allow selecting genetic markers to avoid genetic-physical linkage, and in addition, a number of loci have already been validated in studies of genetic diversity in maize (Fukunaga *et al.*, 2005; Orozco-Ramírez *et al.*, 2016).

Previous studies dealing with teosintes genetic diversity were focused on genetic resources and domestication processes (Fukunaga *et al.*, 2005; van Heerwaarden *et al.*, 2010), evolutionary analyses based on particular genes (Moeller, Tenaillon & Tiffin, 2007; Ross-Ibarra, Tenaillon & Gaut, 2009), and genomic adaptation (Fustier *et al.*, 2017; Aguirre-Liguori *et al.*, 2017; 2019). However, due to their specific objectives, the majority of these studies used broad geographic scales or a large number of localities with few individuals, while other studies involved a higher number of individuals, yet using few molecular markers, either isozyme or microsatellites (Buckler, Gaut & McMullen, 2006; Fukunga *et al.*, 2005). More recently, some studies implemented SNPs (Pyhäjärvi *et al.*, 2013; Aguirre-Liguori *et al.*, 2017). In general, these studies suggested that genetic diversity and the distribution of teosintes are strongly affected by environmental factors (Aguirre-Liguori *et al.*, 2017). Thus, it is expected that historical events like the Last Glacial Maximum (LGM) and the Younger dryas (Berger, 1990) influenced the distribution of teosinte populations, with a concurrent influence on their genetic diversity. Moreover, aside from the environmental factors shaping teosinte genetic diversity, previous studies left open questions about the fine details of microevolution, such as genetic structure, past demographic events, selfing rates, and the geographical distribution of genetic diversity. Hence, relevant questions have remained: Are there high levels of genetic structure among teosinte populations, or they comprise a genetic continuum? Is the amount and distribution of current genetic diversity influenced by the past climate changes? Are local factors,

like selfing and local adaptation shaping genetic diversity? And, using neutral molecular data, is there evidence of gene flow between subspecies?

Recent advances in coalescent theory (Wu & Drummond, 2011; Nikolic & Chevalet, 2014) used jointly with distribution modeling analyses for present and past environmental conditions (Hijmans *et al.*, 2005; Thuiller *et al.*, 2019), have proved to be powerful tools to understand the processes underlying the observed levels and patterns of genetic variation in plants and animals (Aguirre-Liguori *et al.*, 2019; Castellanos-Morales *et al.*, 2019). In this sense, neutral molecular markers, such as microsatellites allow for a wide variety of robust analyses (Selkoe & Toonen, 2006; Guichoux *et al.*, 2011; Kalia *et al.*, 2011) including the inference of demographics and migration process. Even though the many appealing analyses that can be performed using microsatellites, recent studies highlighted the need of a high number of loci (>20) along with a comprehensive geographic representation of the species (including large number of populations and individuals), in order to achieve reliable conclusions (Aguirre-Liguori *et al.*, submitted).

In this study, we used a set of 22 microsatellite loci to conduct thorough analyses of the amount and geographic distribution of genetic diversity for a total of 29 teosinte populations in Mexico, 14 of *Z. m. parviglumis* and 15 of *Z. m. mexicana*. The studied localities encompassed a wide range of environmental conditions and most of the geographic distribution of both subspecies (Aguirre-Liguori *et al.*, 2017). We estimated the genetic diversity, genetic structure, inbreeding and selfing levels, and identified genetic barriers among 29 localities of teosinte. Additionally, we used present and past environmental data to assess the influence of past environmental conditions in the demographic history of this species.

Materials and methods

Sampling, DNA extraction and genotyping

We sampled 29 sites including 14 localities of *Zea mays parviglumis* and 15 localities of *Zea mays mexicana* subspecies (Fig. 1, Table S1, Supplementary Materials) encompassing a broad altitudinal and environmental gradient of teosinte populations. Seeds were collected from 10-26 plants per site, and one seed per plant was grown in a greenhouse obtaining a total of 527 plants. DNA extraction was carried out from fresh leaf tissue using a modified version of the CTAB extraction method (Doyle & Doyle, 1987).

To evaluate the genetic diversity of teosintes, we amplified a set of 22 microsatellite loci previously used for maize genetic diversity analyses. To avoid physical linkage, loci were selected considering their chromosomal position and distance within chromosomes (>150cM). The microsatellites names and primer sequences are reported in Supplementary Materials (Table S2). The PCR products were visualized on 2% agarose gels stained with ethidium bromide. The fluorescent-labeled PCR products were genotyped in the UIUC Core Sequencing Facility (https://unicorn.biotech.illinois.edu/) using GeneScan LIZ 600 size standard (Applied Biosystems). Allele sizes were determined using Genemapper v 3.7 (Applied Biosystems). Presence of genotyping errors and null alleles were assessed using MICRO-CHECKER v 2.2.3 (Van Oosterhout *et al.*, 2004), applying the Bonferroni correction for multiple comparisons. To evaluate potential bias due to null alleles in estimates of genetic structure, we used the FreeNa software (Chapuis & Estoup, 2007) to implement a statistical test for differences between the standard F_{ST} versus the F_{ST} with the ENA (Excluding Null Alleles) correction.

Population Genetics Analyses

We estimated basic population genetics summary statistics: observed number of alleles (*Na*), the rarefied allelic richness (*Ar*, rarefied to the minimum sample size), the observed (H_O) and expected heterozygosity (H_E), using the package "adegenet" version 2.1.1 (Jombart, 2008) implemented in R (R Core Team, 2015).

To detect deviations from Hardy-Weinberg proportions, we performed a test of excess or deficiency of heterozygotes by means the fixation index (F_{IS}) (Wright, 1965; Weir & Cockerham, 1984) using Genepop version 4.6 (Rousset, 2008). Its significance was calculated using the Fisher exact test with 10,000 dememorisation steps, 1,000 batches, and 5,000 iterations per batch. To approximate the degree of selfing in teosinte localities, we calculated sg^2 based on multilocus comparison of linkage disequilibria using the software RMES (David *et al.*, 2007).

To evaluate the genetic structure among localities, we used the R package DEMEtics (Gerlach *et al.*, 2010) to estimate the Jost's D_{EST}. We selected this measure because standard F_{ST} may underestimate population differentiation due to the high mutation rates of microsatellites (Balloux & Lugon-Moulin, 2002), and large population sizes of teosintes. We calculated the 95% confidence intervals (C.I.) as a measure of significance; if D_{EST} C.I. values did not approximate zero, we took the value as significant. Additionally, we performed a Principal Component Analysis (PCA) to evaluate the allelic differences at the individual level using the sampling localities as point classes with "adegenet" R package.

To detect genetic grouping among individuals and localities, we ran Structure v2.3 (Pritchard, Stephens & Donnelly, 2000) using the admixture model with correlated allele frequencies (Falush, Stephens & Pritchard, 2003). We performed 25 independent runs for each value of *K* from 1 to 30 using 1,000,000 MCMC steps followed by 500,000 steps as burn-in. To find the most likely number of clusters (*K*) we used the method proposed by Evanno, Regnaut &

Goudet (2005) implemented in Structure Harvester (Earl & vonHoldt, 2012). Also, all lnP(D) values were plotted to choose the *K* value with the highest likelihood and lowest variance. We performed the analyses for the whole sample, as well for each subspecies alone.

Gene flow was approximated by means of the recent migration rates (*m*) estimated by BayesASS 3.0.3 (Wilson & Rannala, 2003). As BayesASS does not perform well with multiple gene pools within populations (Wilson & Rannala, 2003), we used the genetic clusters obtained by Structure. To ensure good chain-mixing, we performed several test runs varying the values of parameters to achieve acceptance rations of 0.25-60 as recommended by BayesASS authors. The analysis was performed using $10x10^9$ Markov chain Monte Carlo iterations, sampling every 100,000 steps and discarding the first 10% iterations as burn-in period.

In order to detect genetic barriers among localities, we implemented the Monmonier algorithm (Monmonier, 1973) using adegenet. To perform this analysis, we used a Gabriel graph as a connection network among localities and used the localities' scores obtained from the previously estimated PCA as distance matrix. The distance threshold between immediate neighbors was chosen as an abrupt decrease between connected points. We used the function optimize.monmonier to obtain the best boundaries.

Isolation by Distance and Isolation by Environment analyses

To approximate Isolation by Distance (IBD) patterns for each subspecies, we performed standard Mantel tests (Mantel, 1967) using pairwise Nei genetic distances between localities versus the log10 of geographic distances in meters. The Mantel tests and their significance (10,000 random replications) were estimated by means of the R package vegan (Oksanen *et al.*, 2009), while geographic distances were obtained using the raster package (Hijmans, 2019).

We evaluated the relative effects of geographical distance and environment on genetic differentiation using R package Sunder (Botta *et al.*, 2014). This method uses the decay in covariance in allele frequencies between populations as a function of the geographical and ecological distances (here, the environmental distance matrix). We implemented three models, *G*, where geographic distance alone explains the decline in covariance in allele frequencies; *E*, where environmental distance explains the decline; and *G+E*, where both distances explain equally the decline in the allele frequencies. We ran the analysis using 10 million iterations, sampling every 1,000 iterations. The initial state and upper bounds of the Dirichlet prior parameters distribution were implemented as suggested by the software authors. The convergence was assessed by plots of parameters' traces, and we used 10 % of the data as a validation set for the cross-validation step. The environmental distance matrix was generated using euclidean distances estimated from the bioclim dataset obtained for the SDM analyses (see below).

Demographic analyses

We approximated changes in the effective population size (N_e) of teosinte populations using R package VarEff v1.2 (Nikolic & Chevalet, 2014). This package estimates the effective population sizes from present to ancestral time by the simulation of demographic steps of constant size using a coalescent approach. The model calculates posterior probabilities for each step through likelihood approximation and plots the results in terms of N_e through time. We used previously reported microsatellite mutation rate range for teosinte (average=7.7×10^{-4} per generation, 95% HPD $1.1 \times 10^{-3} - 5.2 \times 10^{-4}$) (Vigouroux *et al.*, 2002) to approximate the time in generations for the fluctuation on N_e. For demographic analyses, we used the genetic clusters obtained with Structure. We ran the program for 100,000 steps with a batch length of 10 steps and sampling every 10 steps with an acceptance ratio of 0.25. We used the Two-Phase model (Di Rienzo *et al.*, 1994) with the proportion of mutational events (*c*) fixed at 0.15. As both subspecies are recognized as annual plants (Aguirre-Liguori, Aguirre-Planter & Eguiarte, 2016) we considered one year as generation time.

Spatial Distribution Modeling

We gathered a total of 375 georeferenced points including this study and previous published records of *Z. m. parviglumis* and *Z. m. mexicana* (www.conabio.org). In order to reduce bias due to autocorrelation caused by local overrepresentation, we generated a 1/10 degree grid and then selected one random point using the R package “raster”. The final data set included 62 points for *Z. m. mexicana* and 69 points for *Z. m. parviglumis*. We obtained 19 current environmental layers from the Worldclim database (http://worldclim.org/bioclim, Hijmans *et al.*, 2005). To evaluate the influence of past climate conditions in teosinte distribution, we included the past

Bioclimatic variables for the Mid-Holocene (MID), Last Glacial Maximum (LGM), and Last Interglacial (LIG) (Otto-Bliesner *et al.*, 2006). To reduce variable redundancy, we conducted a paired Spearman's rank correlation test and excluded variables with values >|0.75|. The final set included the following ten variables: BIO2 (Mean Diurnal Range), BIO3 (Isothermality), BIO4 (Temperature Seasonality), BIO8 (Mean Temperature of Wettest Quarter), BIO12 (Annual Precipitation), BIO14 (Precipitation of Driest Month), BIO15 (Precipitation Seasonality), BIO17 (Precipitation of Driest Quarter), BIO18 (Precipitation of Warmest Quarter), BIO19 (Precipitation of Coldest Quarter).

To generate the Species Distribution Models (SDMs) we used the platform biomod2 v 3.3 (Thuiller *et al.*, 2019) in R. The SDMs were constructed by means of ensemble forecasting models (Araújo & New, 2007), using the committee-averaging criteria. We implemented four modelling algorithms in our analysis: 1) Generalised Linear Model (GLM; McCullagh & Nelder, 1989); 2) Generalized Boosted Model (GBM; Friedman 1991); 3) MaxEnt (Phillips, Anderson & Schapire, 2006); and 5) Random Forest (RF; Breiman, 2001). Occurrence data was split into 70% for model calibration and 30% for model evaluation. Using this 70%-30% criteria, five random-sampled data sets were run for each algorithm. Then, we generated five pseudo-absence datasets of 5,000 points selected at random. Model performance was assessed using the true skill statistic (TSS) and the area under the receiver operating characteristic curve (AUC), and models that did not meet thresholds of TSS>0.8 and 0.95>AUC were discarded. Final ensembles were transformed to binary points selecting the threshold that maximized the TSS score.

Results

Overall, we detected a total of 355 alleles for the 22 microsatellite loci analyzed with an allelic

range from six to 30 alleles per locus (allelic frequencies for all individuals and markers are reported in Table S3). The overall genetic diversity was H_E= 0.818 (s.d.= 0.077) and H_O= 0.565 (s.d.= 0.104) (values reported in Table S4). At subspecies level, *Z. m. parviglumis* showed an average H_E= 0.666 (s.d. = 0.112), H_O= 0.538 (s.d. = 0.103), and allelic richness, *Ar*= 3.948 (s.d.= 0.683), while *Z. m. mexicana* had H_E= 0.714 (s.d. = 0.047), H_O= 0.585 (s.d=0.044), and *Ar*= 4.198 (s.d.=0.760). Furthermore, we found a wide range of genetic diversity among localities, ranging from H_E= 0.364 (s.d. = 0.172) to H_E= 0.769 (s.d. = 0.115), and an allelic richness from *Ar*= 2.335 (s.d. = 1.172) to *Ar*= 4.749 (s.d.= 1.218) for pVP and pCH respectively (Fig. 2a, Table S4). We did not find a statistical correlation between genetic diversity and sample size per locality (r^2<0.01, p>0.2).

We found an average F_{IS} value of 0.1796, while *Z. m. parviglumis* displayed F_{IS} = 0.3374 and *Z. m. mexicana* F_{IS}= 0.2621 (Fig. 2b, Table S4). In general, we found differences in F_{IS} values among localities (Fig. 2b), ranging from F_{IS}= 0.0863 (pTC) to F_{IS}= 0.3025 (pHU), with a higher fluctuation for *Z. m. parviglumis* than for *Z. m. mexicana*. Only the *Z. m. parviglumis* localities pGU and pTC were in H-W proportions. According to the F_{IS} values, Micro-Checker identified the presence of at least one null allele in each locality. However, we did not find any locus with a consistent null-allele signal across populations. Moreover, the global F_{ST} obtained with the ENA correction was not significantly different from the standard (not corrected) F_{ST} value, so we kept all loci for all posterior analyses.

Selfing estimated with RMES software showed fluctuating sg^2 values for both teosinte subespecies (Fig. 3, Table S4); *Z. m. parviglumis* displayed higher sg^2 values with an average sg^2 = 0.078, and values ranging from zero (no selfing) to 0.309 per locality, while *Z. m. parviglumis* displayed an average sg^2 = 0.052 with a range from zero to 0.133. In both subespecies, only one

third of selfing estimates were significant. Also, we did not observe a spatial pattern of localities having significant selfing values.

We obtained an overall D_{EST} value of D_{EST}= 0.147 between subspecies, however we found contrasting values of genetic structure among teosinte localities (Fig. 4). We observed the highest values between the *Z. m. parviglumis* pHU and pVP (D_{EST}= 0.796), and the lowest values among the Michoacán localities (mAC1, mAC2, mAC3, mSAM and mHU30; mean D_{EST}=0.136) of *Z. m. mexicana*. Notably, the Western-Jalisco localities (pVP and pTP) along with Oaxaca (pOX) were the most differentiated from the rest of the localities. Moreover, mJM was placed among the *Z. m. parviglumis* localities.

The Structure analysis indicated the presence of four genetic groups (Fig. 5a). The analysis divided the localities in subspecies and each subspecies in two groups. *Z. m. parviglumis* displayed a subdivision between the Western-Jalisco localities (pVP and pTP) and the rest of the localities, while in *Z. m. mexicana*, the subdivision was between the localities of Michoacán and the rest of *Z. m. mexicana* localities. Also, the analysis showed admixed individuals between subspecies (Fig. 5a). It is worth noting that we obtained similar results when analyzing each subspecies separately (data not shown). The recent-migration analysis conducted with BayesASS between the Structure clusters yielded low migration rates (Fig. S3), and the highest rates (0.009) were obtained between Western-Jalisco and *Z. m. parviglumis* clusters.

The PCA displayed the Western-Jalisco localities (pVP and pTP) as the most genetically divergent individuals (Fig. 5b), in accordance to D_{EST} and Structure results. In addition, the PCA showed genetic closeness among individuals of both subspecies, in particular from the eastern localities. On the other hand, Monmonier analysis revealed significant boundaries between the Western-Jalisco and the rest of the localities (data not shown). However, due to the possible bias

generated by these genetic divergent localities (as displayed by D_{EST} and PCA), we performed the Monmonier analysis again excluding pVP and pTP, and we found significant boundaries among the Michoacán and the neighbor *Z. m. parviglumis* localities (Fig. 5c); moreover, the analysis did not detect significant boundaries within subspecies, nor between subspecies among eastern localities.

The Mantel tests between genetic distance and geographic and ecological distances displayed significant and positive correlations for both distance matrices in both subspecies (r>0.34, p<0.05, Fig. S1). Nevertheless, the Sunder analysis showed a higher probability for the environmental (E) model than for the geographic (G) and the geographic+environment (G+E) models in both subespecies, indicating that environmental conditions alone have a stronger influence on the change of allelic frequencies.

The VarEff tests displayed demographic changes in the four Structure clusters (Fig. 6 and Table 1). However, these analyses showed some differences in the trends and values of effective population size (N_e) among clusters. For *Z. m. parviglumis*, the Western-Jalisco cluster showed a constant past population size, followed by steady growth, and according to its low genetic diversity, this cluster had the lowest N_e values (Table 1). The rest of the localities of *Z. m. parviglumis* displayed a past steady N_e followed by ancient demographic contraction and a posterior demographic growth (Fig. 6). In the case of *Z. m. mexicana*, the Michoacán cluster displayed a past constant N_e, followed by the recent contraction observed for all clusters; moreover, the rest of the localities of *Z. m. mexicana* had a low constant N_e followed by demographic growth.

Despite the differences in the demographic trends among clusters, we observed a recent and drastic N_e contraction 130-1,500 generations ago, resulting in a very reduced current

effective size in all clusters (Table 1). Moreover, except for the Michoacán cluster, teosinte populations exhibited demographic growth after the Mid-Holocene (~6,000 years ago) and, as displayed in Table 1, the estimated N_e suggested large population sizes in the past (as large as $<11x10^6$ for the Michoacán cluster).

SDM displayed allopatric current potential distribution areas among subspecies (Fig. 7), showing only ~3.5% overlap in the boundaries of each subspecies distribution. Furthermore, both subspecies exhibited changes in their distribution areas across four different time periods. In the LIG (~120kys) period we estimated a small distribution area, while for the LGM (~22kys) we observed an increase in the potential distribution area in both subspecies. Moreover, for the MID (~6kys) we observed an acute reduction in the potential distribution, followed by an increment to the present (Fig. 7). The reduction in the predicted distribution area for the MID may be explained by the contrasting environmental conditions at this period (Fig. S2).

Discussion

Our analyses in current populations of *Z. m. parviglumis* and *Z. m. mexicana* revealed that local factors like selfing rates, isolation and historical processes as climate change, have shaped the levels and distribution of their genetic diversity. At the local scale, the number of alleles and heterozygosity were consistent with previous reports on teosinte (Matsuoka *et al*., 2002; Fukunaga *et al*., 2005). Nevertheless, the low diversity detected in the Western-Jalisco cluster could be related to genetic drift due to genetic isolation, as this was confirmed by the low N_e estimated for these localities (Table 1). Besides, F_{IS} values showed that most teosinte localities had from moderate to high levels of heterozygote deficiency and only two *Z. m. parviglumis* localities were in H-W proportions, suggesting ongoing evolutionary process like genetic drift and inbreeding affecting allele frequencies. Even though positive F_{IS} values may

indicate within-localities inbreeding (Wright, 1965), the generalized heterozygote deficiency signal that we recovered, could also be caused by Wahlund effect (excess of observed homozygotes resulting from sampling a structured population) produced by local genetic structure. Moreover, the RMES analysis showed no correlation between selfing and F_{IS} values and we found that neighbouring localities did not have similar selfing values. For instance, localities from the Michoacán cluster had the lowest D_{EST} values (suggesting some degree of genetic connectivity) while their selfing values sg^2 ranged from zero to 0.233 (Fig. 3). A previous study reported high levels of outcrossing at local scales (Van Heerwaarden *et al*., 2010); however, the authors also suggested that the fine-scale environmental features, like topography and wind, may influence pollen and seed dispersal generating local genetic structure. The lack of spatial arrangement of F_{IS} and sg^2, could be associated with limited seed dispersion that promote local selfing and genetic structure. In this sense, the teosinte is considered as a Pleistocene species adapted to seed dispersal by large herbivores, and although livestock grazes on their plants in the present, their range of dispersion is limited by the human agricultural niche construction (mainly land crops) (Webster, 2011), restricting dispersion to a short distances. This, in turn, leads to the foundation of small and fragmented populations genetically isolated.

On a global scale, we found relatively high levels of genetic structure, and accordingly, low estimates of gene flow. Furthermore, we observed that the Oaxaca locality (pOX) and Western-Jalisco cluster (pVP and pTP) were the most differentiated localities. While this was expected for pOX (as this locality was the most geographically distant), the Western-Jalisco cluster was highly differentiated even in the presence of relatively close localities (for example, pGU, pEJ, and pTO <80 km; Fig. 1 and 4). Moreover, the mJM locality is situated in a neighboring area near the Michoacán cluster localities (Fig. 5); yet, it was highly structured from

the other *Z. m mexicana* localities. Our results suggest that spatial distribution alone is not the determinant factor for the geographical distribution of genetic diversity on teosinte populations. In particular, the Sunder analysis showed that the environment provides a better explanation for values of genetic differentiation in teosinte than geographic distance. This result supports the signatures of local adaptation detected by Aguirre-Liguori *et al*., (2019) using a large SNPs data set. Hence, our results indicate that adaptation to local conditions is promoting the genetic divergence among teosinte populations and subspecies, despite the presence of dispersal and gene flow. This hypothesis is supported by the differences in the environmental conditions encountered in the distribution of the genetic clusters (Fig. S4), and the reduced proportion of suitable areas shared by subspecies, as shown by current SDM projections (Fig. 7). Another factor that may reduce the homogenization of genetic diversity via gene flow is the variable outcrossing rates observed that preclude admixture with external populations. Moreover, we observed that our genetic clusters of *Z. m. mexicana* correspond to the "geographic races" Chalco (main *Z. m. mexicana* cluster) and Central Plateau (Michoacán cluster) proposed by Gonzalez et al. (2018) based exclusively on environmental features.

On the other hand, the drastic reduction in the predicted area during MID may be produced by the marked differences on seven of the ten bioclimatic variables in that period, (mainly rain regimes, Fig. S2). Furthermore, the observed changes in the historical effective population sizes (N_e) concurred with SDM analyses. For three of the four genetic clusters, we found a demographic increase with a peak of the N_e posterior to MID (Fig. 6), indicating that demographic growth may have been promoted by a period of suitable conditions after a period of non-favorable conditions. Although we observed a fluctuation of suitable distribution areas across the four analyzed periods (a very reduced area in LIG, an increase in the suitable area in

the LGM, a drastic reduction in MID, and a new increase in the current-time), microsatellite markers showed resolution for the last ~20,000 generations, thus limiting the analyses to more recent events than the LGM. Interestingly, the Michoacán cluster displayed an ancient constant N_e, suggesting a negligible impact from the MID environmental shift. This trend could be related to the possible existence of a post-glacial refugium (not detected by our projections) that, coupled with the genetic connectivity observed in this area, allowed large N_e across time. Another possible explanation is related to the particular environmental requirements of this cluster (Fig. S4). If that is the case, projections performed in the complete database of *Z. m. mexicana* may result in a biased distribution modeling of the Michoacán cluster. To evaluate this hypothesis, however, it will be necessary to perform extensive sampling and posterior assignment analysis to collect enough georeferenced data and to obtain reliable projections.

Finally, the low N_e values observed as a consequence of a recent demographic contraction should be interpreted with caution. The severe bottleneck observed may be the outcome of the exploitation of teosinte as a food source in the pre-Columbian Mesoamerican populations (Webster, 2011), and more recently, the result of displacement of teosinte due to maize production. However, in some of our estimations, the effective size contraction continued as recently as ten generations ago, resulting in an N_e of almost zero. This is likely a result of an inaccurate adjustment of N_e caused by a complex demographic history and a very recent and severe demographic contraction, causing the algorithm not to achieve an accurate estimation in very recent time scale.

Conclusions

Our results showed that historical and contemporary factors at local and global scale shaped the genetic diversity of teosinte. While local and contemporary factors cause differences in mating patterns and genetic structure, the historical factors shaped the amount and distribution of genetic diversity of teosinte populations. In this sense we remark the importance of including data from different temporal and spatial scales in order to obtain a better picture of the mechanisms that affect genetic diversity. Also, we found evidence that suggests that the genetic lineages found could be the result of local adaptation as previously suggested by other authors (Aguirre-Liguori *et al.*, 2017; Fustier *et al.*, 2017; Aguirre-Liguori *et al.*, 2019). This adaptability to specific environmental and agroecological conditions permits the wide ecological conditions inhabited by teosinte.

Even though microsatellites are considered neutral markers, we recovered strong signals of genetic structure that suggest both an ongoing process of genetic drift, and local adaptation to environmental conditions that maintain different genetic lineages despite possible introgression. We also found that teosinte populations have undergone multiple genetic bottlenecks, and that, along with the detected genetic structure, make teosinte populations prone to local extinction. As a consequence, we may be losing an important genetic reservoir for the permanence of teosinte, but also genetic variants that could be relevant to maize improvement and adaptation to global environmental change.

Acknowledgments

The original sampling of the populations and of the overall study was designed with the help and enthusiasm of Drs. Brandon S. Gaut, Maud I. Tenaillon, Valeria Souza and Salvador

Montes-Hernández. We thank L. Espinosa-Asuar and M. Rosas-Barrera for their technical assistance, Alberto Villasante Barahona for his help in generating the microsatellite datasets and Silvia Barrientos for her support during this investigation. We also thank Dr. Gabriela Castellanos-Morales for her valuable comments for the improvement of the manuscript. This work was funded by grants CB2011/167826 (CONACYT Investigación Científica Básica), CN 10 393 (UC MEXUS CONACYT), and M12 A03 ECOS Nord France CONACYT-ANUIES.

References

Aguirre-Liguori JA, Aguirre-Planter E, Eguiarte LE. 2016. Genetics and ecology of wild and cultivated maize: domestication and introgression. In Lira R, Casas A, Blancas J, eds. *Ethnobotany of Mexico*. New York: Springer, 403-416.

Aguirre-Liguori JA, Tenaillon MI, Vázquez Lobo A, Gaut BS, Jaramillo Correa JP, Montes Hernandez S, Souza V, Eguiarte LE. 2017. Connecting genomic patterns of local adaptation and niche suitability in teosintes. *Molecular Ecology* 26: 4226–4240.

Aguirre-Liguori JA, Gaut BS, Jaramillo-Correa JP, Tenaillon MI, Montes-Hernández S, García-Oliva F, Hearne SJ, Eguiarte LE. 2019. Divergence with gene flow is driven by local adaptation to temperature and soil phosphorus concentration in teosinte subspecies (*Zea mays parviglumis* and *Zea mays mexicana*). *Molecular Ecology* 28.

Araújo MB, New M. 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* 22: 42–47.

Balloux F, Lugon Moulin N. 2002. The estimation of population differentiation with microsatellite markers. *Molecular Ecology* 11: 155–165.

Berger WH. 1990. The Younger Dryas cold spell: a quest for causes. *Palaeogeography, Palaeoclimatology, Palaeoecology* 89: 219–237.

Botta F, Eriksen C, Fontaine MC, Guillot G. 2015. Enhanced computational methods for quantifying the effect of geographic and environmental isolation on genetic differentiation. *Methods in Ecology and Evolution* 6: 1270–1277.

Breiman L. 2001. Random forests. *Machine Learning* 45: 5–32.

Buckler ES, Gaut BS, McMullen MD. 2006. Molecular and functional diversity of maize. *Current Opinion in Plant Biology* 9: 172–176.

Castellanos-Morales G, Ruiz-Mondragón KY, Hernández-Rosales HS, Sánchez-de la Vega S, Gámez N, Aguirre-Planter E, Montes-Hernández S, Lira-Saade R, Eguiarte LE. 2019. Tracing back the origin of pumpkins (*Cucurbita pepo* ssp. pepo L.) in Mexico. *Proceedings of the Royal Society B* 286.1908: 20191440.

Chapuis MP, Estoup A. 2007. Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* 24: 621–631.

David P, Pujol B, Viard F, Castella V, Goudet J. 2007. Reliable selfing rate estimates from imperfect population genetic data. *Molecular Ecology* 16: 2474–2487.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. 1994. Mutational processes of simple-sequence repeat loci in human populations. *Proceedings of the National Academy of Sciences* 91: 3166–3170.

Doebley JF, Iltis HH. 1980. Taxonomy of *Zea* (Gramineae) I. A subgeneric classification with key to taxa. *American Journal of Botany*. 67:982-993.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–1.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4: 359–361.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14: 2611–2620.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164: 1567–1587.

Friedman JH. 1991. Multivariate adaptive regression splines. *The Annals of Statistics* 19: 1–67.

Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, Sanchez J, Liu K, Buckler ES, Doebley J. 2005. Genetic diversity and population structure of teosinte. *Genetics* 169: 2241–2254.

Fustier MA, Brandenburg JT. Boitard S, Lapeyronnie J, Eguiarte LE, Vigouroux Y, Manicacci D, Tenaillon MI. 2017. Signatures of local adaptation in lowland and highland teosintes from whole genome sequencing of pooled samples. *Molecular Ecology* 26: 2738–2756.

Gerlach G, Jueterbock A, Kraemer P, Deppermann J, Harmand P. 2010. Calculations of population differentiation based on GST and D: forget GST but not all of statistics!. *Molecular Ecology* 19: 3845–3852.

Gonzalez JD, Corral JA, García GM, Ojeda GR, De la Cruz Larios L, Holland JB, Medrano RM, Romero GE. 2018. Ecogeography of teosinte. *PloS one* 13:2.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ. 2011. Current trends in microsatellite genotyping. *Molecular Ecology Resources* 11: 591–611.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24: 1403–1405.

Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. 2011. Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177: 309–334.

Hijmans RJ. 2019. raster: Geographic Data Analysis and Modeling. R package version 2.9-23. https://CRAN.R-project.org/package=raster.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 25: 1965–1978.

Hufford MB, Bilinski P, Pyhäjärvi T, Ross-Ibarra J. 2012. Teosinte as model system for population and ecological genomics. *Trends in Genetics* 28:606-15.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27: 209–220.

Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J. 2002. Microsatellites in Zea–variability, patterns of mutations, and use for evolutionary studies. *Theoretical and Applied Genetics* 104: 436–450.

McCullagh P, Nelder JA. 1989. Generalized Linear Models. London: Chapman and Hall, 551pp.

Moeller DA, Tenaillon MI, Tiffin P. 2007. Population structure and its effects on patterns of nucleotide polymorphism in teosinte (*Zea mays ssp. parviglumis*). *Genetics* 176: 1799–1809.

Monmonier MS. 1973. Maximum difference barriers: an alternative numerical regionalization method. *Geographical Analysis* 5: 245–261.

Natacha N, Chevalet C. 2014. VarEff: Estimation of effective population size and variation in the time scale. R package version 1.2. https://qgp.jouy.inra.fr.

Nikolic N, Chevalet C. 2014. Detecting past changes of effective population sizes. *Evolutionary Applications* doi: 10.111/eva.12170.

Oksanen J,Guillaume B, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2019. vegan: Community Ecology Package. R package version 2.5-5. https://CRAN.R-project.org/package=vegan

Orozco-Ramírez Q, Ross-Ibarra J, Santacruz-Varela A, Brush S. 2016. Maize diversity associated with social origin and environmental variation in Southern Mexico. *Heredity* 116: 477.

Otto-Bliesner BL, Marshall SJ, Overpeck JT, Miller GH, Hu A. 2006. Simulating Arctic climate warmth and icefield retreat in the last interglaciation. *Science* 311: 1751–1753.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J. 2013. Complex patterns of local adaptation in teosinte. *Genome Biology and Evolution* 5: 1594–1609.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Ross-Ibarra J, Tenaillon M, Gaut BS. 2009. Historical divergence and gene flow in the genus *Zea*. *Genetics* 181: 1399–1413.

Rousset F. 2008. genepop'007: a complete re implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* 8: 103–106.

Sokal RR, Jacquez GM, Wooten MC. 1989. Spatial autocorrelation analysis of migration and selection. *Genetics* 121: 845–855.

Selkoe KA, Toonen RJ. 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* 9: 615–629.

Thuiller W, Georges D, Engler R, Breiner F. 2019. biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.3-7.1. https://CRAN.R-project.org/package =biomod2.

Van Heerwaarden J, Ross Ibarra J, Doebley J, Glaubitz JC, Sánchez-Gónzalez, Gaut BS. Eguiarte LE. 2010. Fine scale genetic structure in the wild ancestor of maize (*Zea mays ssp. parviglumis*). *Molecular Ecology* 19: 1162–1173.

Van Oosterhout C, Hutchinson WF, Wills DP, Shipley P. 2004. MICRO CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* 4: 535–538.

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J. 2002. Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution*, 19: 1251–1260.

Waples, RS. 2014. Testing for Hardy–Weinberg proportions: have we lost the plot?. *Journal of Heredity* 106: 1–19.

Webster DL. 2011. Backward Bottlenecks: Ancient Teosinte/Maize Selection. *Current Anthropology* 52: 77–104.

Weir BS, Cockerham CC. 1984. Estimating F statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Wilson GA, Rannala B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* 163: 1177–1191.

Wright S. 1965. The interpretation of population structure by F statistics with special regard to systems of mating. *Evolution* 19: 395–420.

Wu CH, Drummond AJ. 2011. Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo. *Genetics* 188: 151–164.

Figures

Figure 1

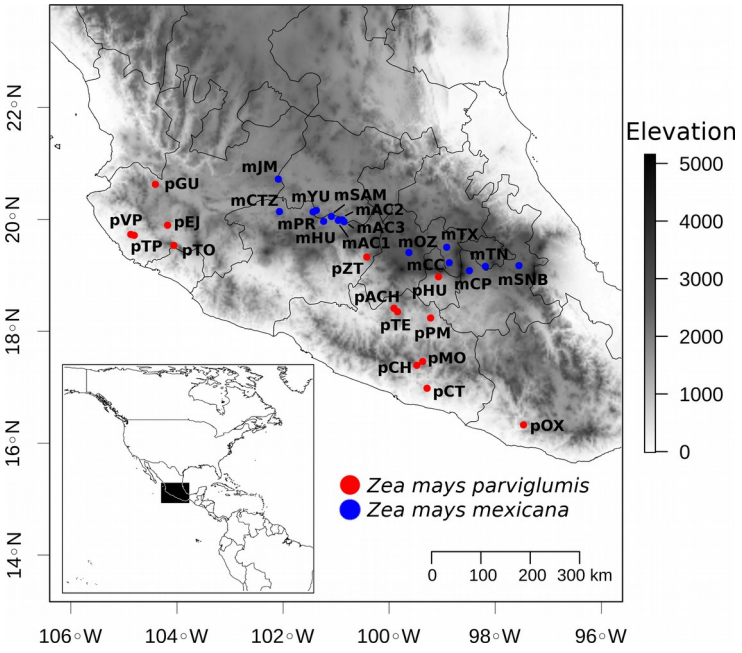


Figure 2

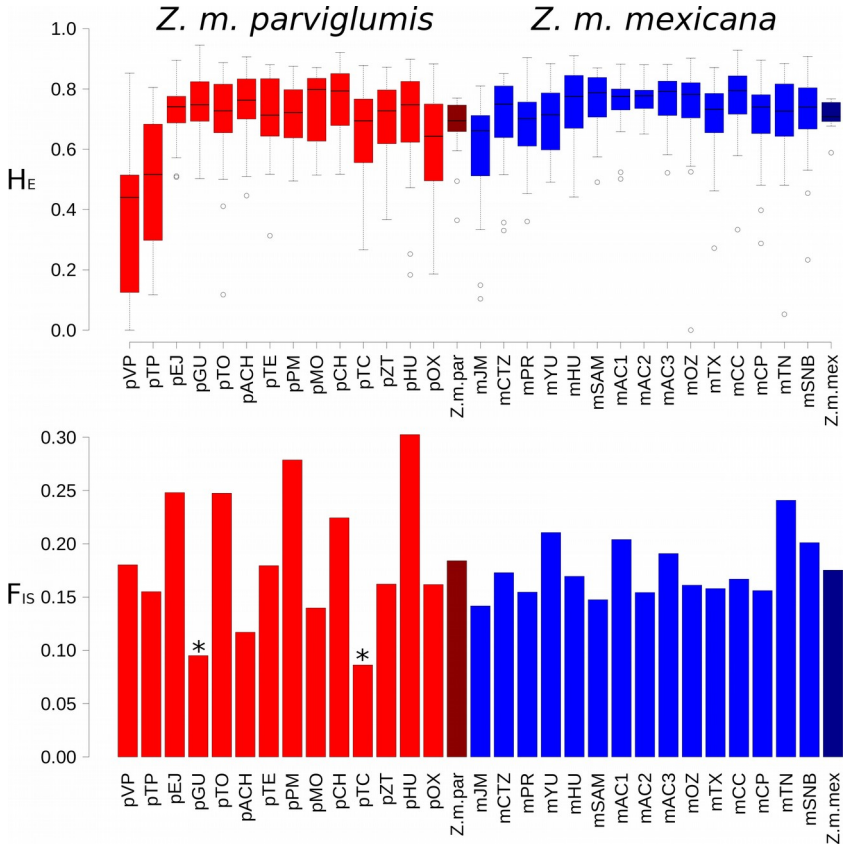


Figure 3

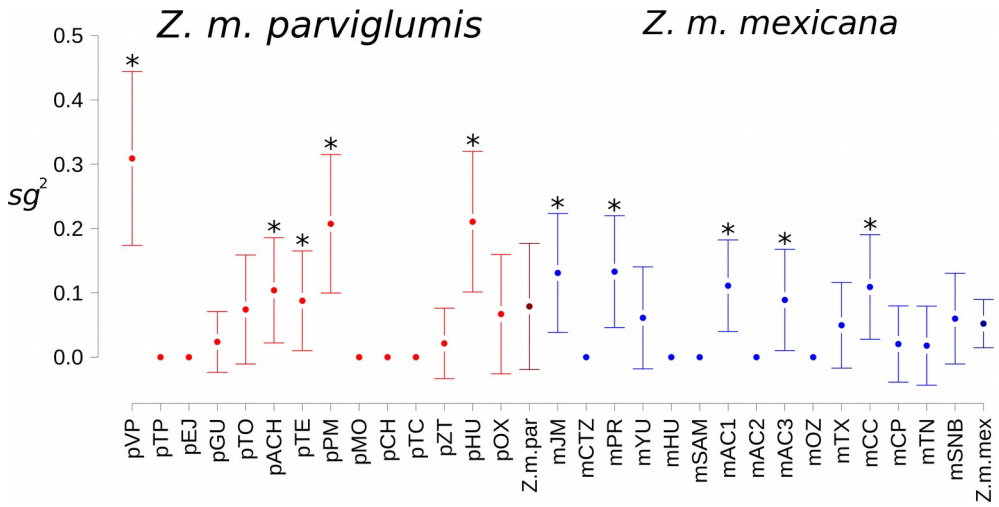


Figure 4

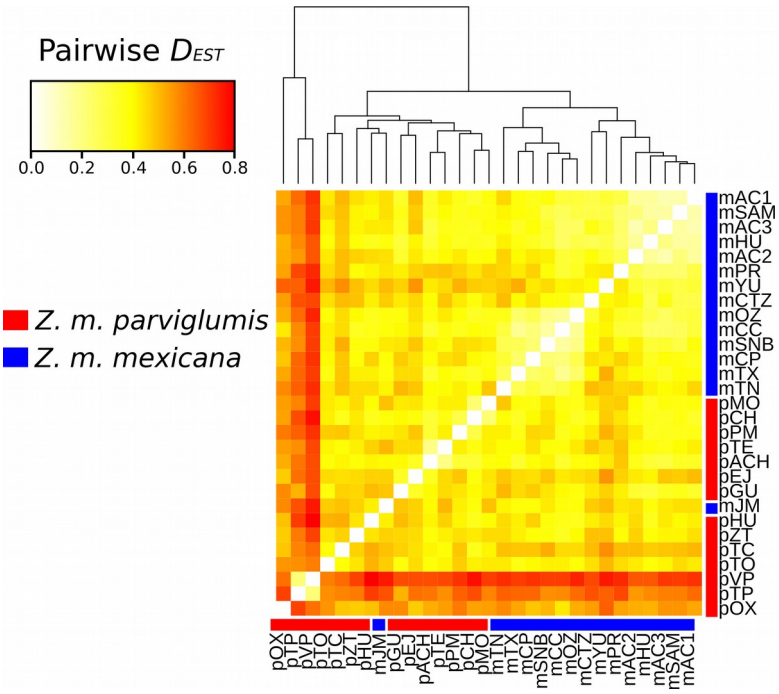


Figure 5

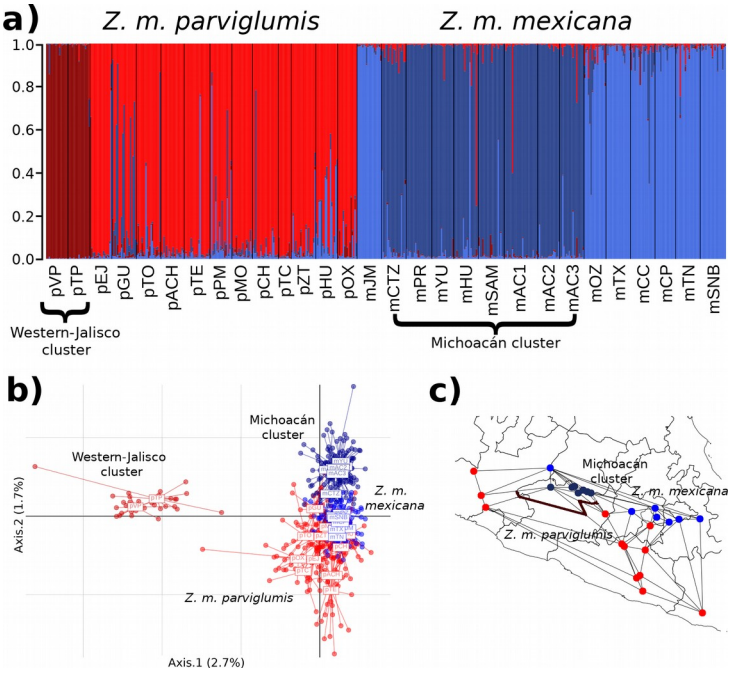


Figure 6

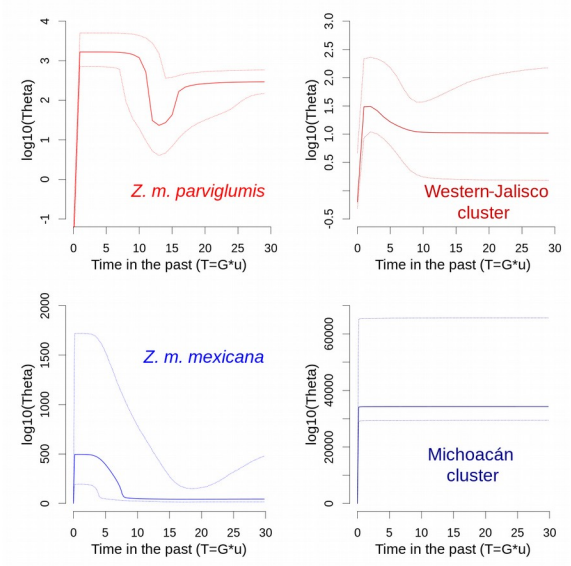
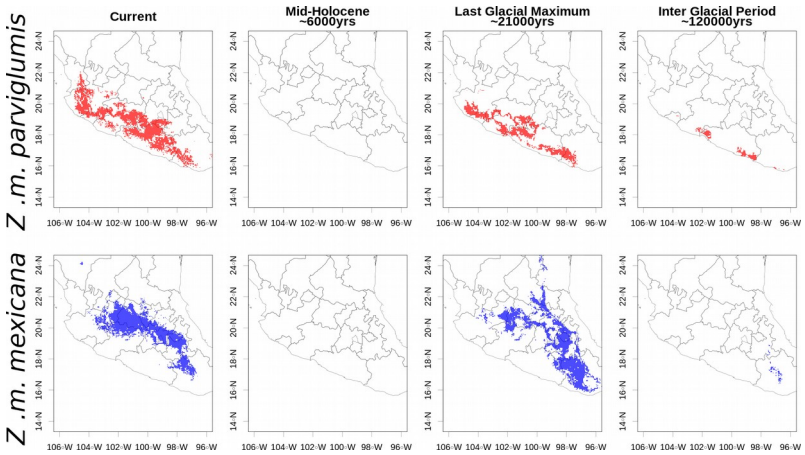


Figure 7



Figures Captions

Figure 1. Map depicting teosinte sampled localities. In red *Zea mays parviglumis*, in blue *Zea mays mexicana*.

Figure 2. a) Expected heterozygosity per locality for *Zea mays parviglumis* (red), and *Zea mays mexicana* (blue). Z.m.par and Z.m.mex corresponds to the average value for each subespecies. b) F_{IS} values per locality, asterisks denote not-significant F_{IS} values at p>0.05. Z.m.par and Z.m.mex corresponds to the average value for each subespecies.

Figure 3. Selfing rates (sg^2) estimated by RMES for each teosinte locality. Z.m.par and Z.m.mex corresponds to the average value for each subespecies. Bars represent standard deviation, asterisks denote values significantly different from zero.

Figure 4. Pairwise D_{EST} values between localities. Localities were clustered using Euclidean distance with complete linkage method. The temperature color represent values from zero (white) to 0.8 (red).

Figure 5. a) Bayesian assignment plot obtained with Structure for K=4. Localities were ordered from left to right according to their latitude. b) Principal Component Analysis of individuals using localities as grouping factor, dots represent individuals. c) Monmonier analysis among localities, excluding the Western-Jalisco localities (pTP and pVp, see text).

Figure 6. Changes in effective population sizes (N_e) in the past estimated with VarEff for the clusters obtained with Structure. The x-axis represents time as the product of generation time per mutation rate. The y-axis represents the effective size scaled as the logarithm base 10 (log10) of Theta (4Ne*u).

Figure 7. Species Distribution Ensemble Models for *Zea mays parviglumis* (red) and *Zea mays mexicana* (blue) considering current conditions, mid-holocene (MID 6,000 years), the last glacial maximum (LGM 21,000 years) and last interglacial period (LIG 120,000 years).

Tables

Table 1. Effective population sizes (N_e) and time of change estimated using VarEff for the clusters obtained with Structure and the range of mutation rates reported by Vigouroux et al. (2012). Time is expressed in generations. Values in parenthesis represent the 95% HPD intervals.

Subspecies	Cluster	Mutation Rate	Maximum N_e	Time for Maximum N_e	Current N_e	Most recent N_e reduction
Zea mays parviglumis	Western-Jalisco	$1.1x10^{-3}$	4484 (234–33345)	1760	143 (106–543)	210
		$7.7x10^{-4}$	6093 (3387–29288)	3040	193 (145–310)	250
		$5.2x10^{-4}$	11332 (5083–94130)	3680	314 (224–2356)	480
	Zea mays parviglumis	$1.1x10^{-3}$	330430 (162888–1222329)	4160	4 (2–303601)	130
		$7.7x10^{-4}$	436963 (213587–1443738)	4160	6 (2–16)	150
		$5.2x10^{-4}$	685267 (330390–2345650)	4640	6	400

					(3 –20)	
Zea mays mexicana	Michoacán	$1.1x10^{-3}$	7913662 (6910970–12600134)	<20000	57 (17–126)	1500
		$7.7x10^{-4}$	11694825 (9751364–19545146)	<20000	53 (22–139)	600
		$5.2x10^{-4}$	4093241 (1452244–22966322)	<20000	83 (21–206)	800
	Zea mays mexicana	$1.1x10^{-3}$	107089 (47895–450297)	960	2 (1–5)	200
		$7.7x10^{-4}$	142406 (64548–575959)	1280	2 (1-6)	240
		$5.2x10^{-4}$	201243 (92335–794399)	2240	3 (1–8)	450

Supplementary Materials

Supplemental Figures

Figure S1

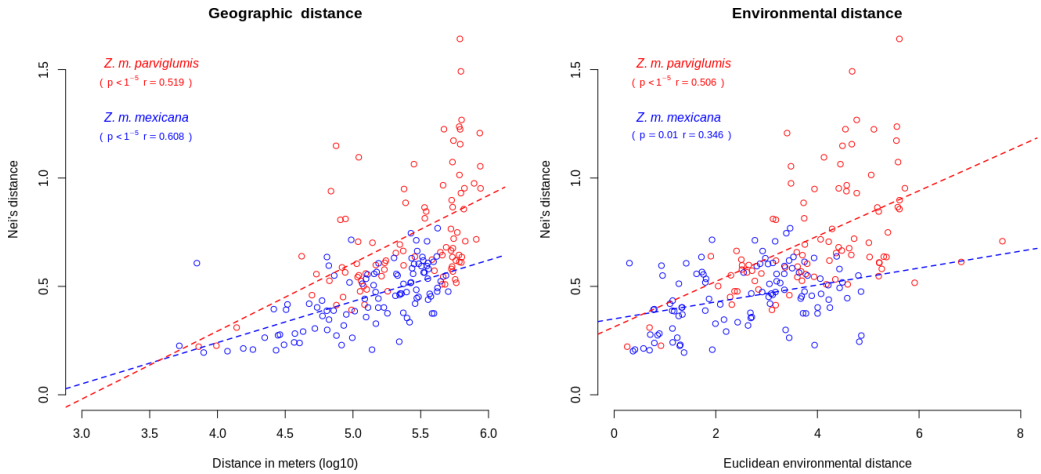
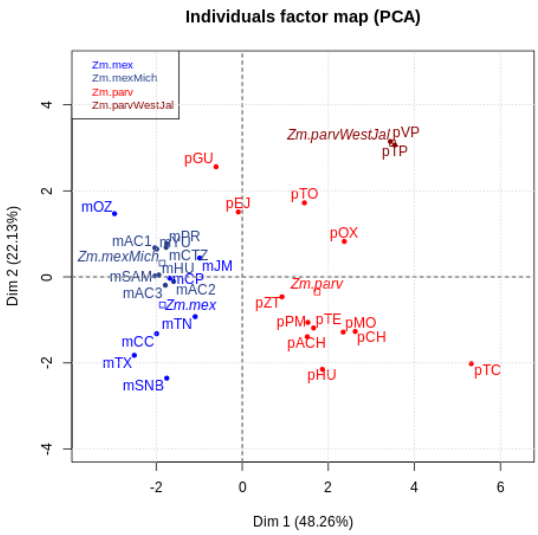


Figure S2

Figure S3

Figure S4



Supplemental figures captions

Figure S1. Mantels tests for genetic distance matrix against geographic distance, and environmental distance for *Zea mays parviglumis* (in red) and *Zea mays mexicana* (in blue).

Figure S2. Boxplots of the environmental variables included for the SDM projections for four time periods (current, MID, LGM and LIG).

Figure S3. Migration rates among Structure clusters calculated with BayesASS software.

Figure S4. Principal Components Analysis of the bioclim variables for teosinte localities analyzed in this study.

Supplemental tables

Table S1. Information of the localities analyzed in this study.

Key	Locality	Subspecies	Elevation	Latitude	Longitude
pTP	Teloloapan, Guerrero	*Z. m. parviglumis*	1649	18°20’59.4”	99°50’27.9”
pACH	Alcholoa, Guerrero	*Z. m. parviglumis*	1439	18°24’39.6”	99°54’30.3”
pTC	Teconoapan, Guerrero	*Z. m. parviglumis*	581	16°58’52.1”	99°17’7.7”
pCH	Chilpancingo, Guerrero	*Z. m. parviglumis*	1201	17°23’30.3”	99°28’39.5”

pMO	Mochitlán, Guerrero	*Z. m. parviglumis*	1107	17°27’35.7”	99°22’5.9”
pHU	Huitzuco, Guerrero	*Z. m. parviglumis*	1101	18°14’15.7”	99°13’4.8”
pTZ	Tepoztlán, Morelos	*Z. m. parviglumis*	1664	18°58’26.8”	99°4’13”
mOZ	Otzolotepec, Estado de México	*Z. m. mexicana*	2581	19°24’26.9”	99°37’37.4”
mCTZ	Churintzio, Michoacán	*Z. m. mexicana*	1846	20°8’21.8”	102°4’6.4”
pGU	Guachinango, Jalisco	*Z. m. parviglumis*	1426	20°37’38.3”	104°24’28.4”
pTP	Telpitita, Jalisco	*Z. m. parviglumis*	504	19°42’55.6”	104°48’17.6”
pEJ	Ejutla, Jalisco	*Z. m. parviglumis*	1317	19°53’46.1”	104°10’36.3”
mAC1	Acámbaro, Guanajuato	*Z. m. mexicana*	1861	19°59’30.2”	100°53’5.6”
mAC2	Acámbaro, Guanajuato	*Z. m. mexicana*	1854	19°57’29.5”	100°50’58.8”
mAC3	Acámbaro, Guanajuato	*Z. m. mexicana*	1878	19°58’58.1”	100°57’35.7”
mSAM	Santa Ana Maya, Michoacán	*Z. m. mexicana*	1849	20°3’11.5”	101°5’17.2”
mYU	Yuriria, Guanajuato	*Z. m. mexicana*	1856	20°9’38.6”	101°22’23.1”
pVP	Villa de Purificación, Jalisco	*Z. m. parviglumis*	572	19°43’56.9”	104°52’18.7”
pTO	Tolimán, Jalisco	*Z. m. parviglumis*	1369	19°32’7.3”	104°3’29.7”
mPU	Puruándiro, Michoacán	*Z. m. mexicana*	2002	20°08’01.7”	101°26’03.0”
mHU	Huandacareo, Michoacán	*Z. m. mexicana*	1844	19°57’52.6”	101°14’07.5”
pZT	Zitácuaro, Michoacán	*Z. m. parviglumis*	1383	19°19’38.6”	100°25’17.2”
mJM	Jesús María, Jalisco	*Z. m. mexicana*	2176	20°43’10.4”	102°05’23.4”
mCC	Cocotitlán, Estado de México	*Z. m. mexicana*	2252	19°13’34.5”	98°51’58.1”

mSNB	San Nicolas Buenos Aires, Puebla	*Z. m. mexicana*	2375	19°10’27.3”	97°33’10”
mTN	Tenancingo, Tlaxcala	*Z. m. mexicana*	2306	19°9’34.1”	98°11’7.5”
mCP	Calpan, Puebla	*Z. m. mexicana*	2447	19°5’2”	98°29’14.8”
mTX	Texcoco, Estado de México	*Z. m. mexicana*	2234	19°30’9.1”	98°54’52.6”
pOX	San Cristobal Honduras, Oaxaca	*Z. m. parviglumis*	1094	16°19’36.1”	97°28’0.1”

Table S2. Locus name for the 22 microsatellite loci used in this study, chromosomal position (Chr) in *Zea mays* according to reported on MaizeGDB (http://www.maizegdb.org/), primer sequences, and reference.

Locus name	Chr	Primer sequence (5’-3’)	Reference
phi064	1	Fwd-CCGAATTGAAATAGCTGCGAGAACCT Rev-ACAATGAACGGTGGTTATCAACACGC	Chin et al., 1996
phi96100	2	Fwd-AGGAGGACCCCAACTCCTG Rev-TTGCACGAGCCATCGTAT	Chin et al., 1996
phi053	3	Fwd-AACCCAACGTACTCCGGCAG Rev-CTGCCTCTCAGATTCAGAGATTGAC	Chin et al., 1996
phi072	4	Fwd-ACCGTGCATGATTAATTTCTCCAGCCTT Rev-GACAGCGCGCAAATGGATTGAACT	Chin et al., 1996; Senior et al., 1996
phi109188	5	Fwd-AAGCTCAGAAGCCGGAGC Rev-GGTCATCAAGCTCTCTGATCG	Chin et al., 1996; Senior et al., 1996
phi034	7	Fwd-TAGCGACAGGATGGCCTCTTCT Rev-GGGGAGCACGCCTTCGTTCT	Chin et al., 1996; Senior et al., 1996

phi015	8	Fwd-GCAACGTACCGTACCTTTCCGA Rev-ACGCTGCATTCAATTACCGGGAAG	Chin et al., 1996; Senior et al., 1996
phi033	9	Fwd-ATCGAAATGCAGGCGATGGTTCTC Rev-ATCGAGATGTTCTACGCCCTGAAGT	Chin et al., 1996; Senior et al., 1996
phi96342	10	Fwd-GTAATCCCACGTCCTATCAGCC Rev-TCCAACTTGAACGAACTCCTC	Chin et al., 1996; Senior et al., 1996
phi427913	1	Fwd-CAAAAGCTAGTCGGGGTCA Rev-ATTGTTCGATGACACACTACGC	Senior et al., 1996
phi127	2	Fwd-ATATGCATTGCCTGGAACTGGAAGGA Rev-AATTCAAACACGCCTCCCGAGTGT	Senior et al., 1996
Phi051	7	Fwd-CGACATCGTCAGATTATATTGCAGACCA Rev-GGCGAAAGCGAACGACAACAATCTT	Senior et al., 1996
phi159819	6	Fwd-GATGGGCCCTAGACCAGCTT Rev-GCCTCTCCCATCTCT	Senior et al., 1996
phi093	4	Fwd-AGTGCGTCAGCTTCATCGCCTACAAG Rev-AGGCCATGCATGCTTGCAACAATGGATACA	Chin et al., 1996
bnlg1194	8	Fwd-GCGTTATTAAGGCAAGCTGC Rev-ACGTGAAGCAGAGGATCCAT	Chin et al., 1996
umc1141	8	Fwd-AGAGGAGAAAGAGACAGACAGGCA Rev-CAGGAACTGAATGAAAGCAACTCA	Chin et al., 1996
nc004	4	Fwd-TGCGAAGAAGCAGTAGCAAA Rev-TGGAGGTAGAAGACGCACG	Chin et al., 1996; Senior et al., 1998
bnlg1018	2	Fwd-CGAGGTTAGCACCGACAAAT Rev-CGAGTAAATGCTCTGTGCCA	Chin et al., 1996; Senior et al., 1998

umc1930	10	Fwd-TCTTCTCCAAGTGTGTTAATGCCC Rev-CACACAGTGAGTCGTTTCTTTCGT	Chin et al., 1996; Senior et al., 1998
phi037	1	Fwd-CCCAGCTCCTGTTGTCGGCTCAGAC Rev-TCCAGATCCGCCGCACCTCACGTCA	Senior et al., 1998
bnlg1523	3	Fwd-GAGCACAGCTAGGCAAAAGG Rev-CTCGCACGCTCTCTCTTCTT	Senior et al., 1998

Table S4. Values of expected heterozygosity (H_E), observed heterozygosity (H_O), fixation index (F_{IS}) and selfing (sg^2) for each teosinte locality.

Locality	H_E	H_O	F_{IS}	sg^2
pVP	0.36435	0.28632	0.1803	0.30904
pTP	0.49409	0.41591	0.1551	0
pEJ	0.72746	0.54663	0.2481	0
pGU	0.74926	0.66433	0.0951	0.02385
pTO	0.68631	0.50762	0.2475	0.07412
pACH	0.74871	0.65616	0.117	0.104
pTE	0.71214	0.58397	0.1795	0.08773
pPM	0.7025	0.50028	0.2787	0.20735
pMO	0.74663	0.64286	0.1398	0
pCH	0.76996	0.59383	0.2244	0
pTC	0.65878	0.60253	0.0863	0
pZT	0.68678	0.57298	0.1622	0.02147
pHU	0.6851	0.47605	0.3025	0.2106
pOX	0.59499	0.49343	0.1618	0.06703
mJM	0.58855	0.50467	0.1417	0.13099
mCTZ	0.69755	0.5758	0.1729	0

mPR	0.67688	0.57091	0.1547	0.13302
mYU	0.70021	0.5513	0.2106	0.06118
mHU	0.75793	0.62758	0.1695	0
mSAM	0.75529	0.64286	0.1476	0
mAC1	0.7532	0.59399	0.2041	0.1111
mAC2	0.76713	0.64706	0.1543	0
mAC3	0.75575	0.6043	0.1909	0.08903
mOZ	0.72711	0.61046	0.1612	0
mTX	0.6942	0.58386	0.158	0.04966
mCC	0.76526	0.63755	0.1669	0.10916
mCP	0.68992	0.56629	0.1561	0.02043
mTN	0.68381	0.50982	0.2409	0.01796
mSNB	0.70817	0.55983	0.2011	0.05991

References for supplemental information

Chin, E. C. L., Shu, H., Smith, J. S. C., & Senior, M. L. (1996). Maize simple repetitive DNA sequences: abundance and allele variation. Genome, 39(5), 866-873.

Senior, M. L., Murphy, J. P., Goodman, M. M., & Stuber, C. W. (1998). Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. *Crop science*, *38*(4), 1088-1098.