



Published in final edited form as:

*New Phytol.* 2013 July ; 199(1): 264–276. doi:10.1111/nph.12247.

## Genome size variation in wild and cultivated maize along altitudinal gradients

Concepción M. Díez<sup>1</sup>, Brandon S. Gaut<sup>1</sup>, Esteban Meca<sup>2</sup>, Enrique Scheinvar<sup>3</sup>, Salvador Montes-Hernandez<sup>4</sup>, Luis E. Eguiarte<sup>3</sup>, and Maud I. Tenailon<sup>5</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, UC Irvine, Irvine, CA 92697, USA

<sup>2</sup>Department of Mathematics, UC Irvine, Irvine, CA 92607, USA

<sup>3</sup>Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, CP 04510, Mexico City, Mexico

<sup>4</sup>Campo Experimental Bajío, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, CP 38110, Celaya, Guanajuato, Mexico

<sup>5</sup>CNRS, UMR de Génétique Végétale, INRA/CNRS/Univ Paris-Sud/AgroParisTech, Ferme du Moulon, Gif-sur-Yvette, France

### Summary

- It is still an open question as to whether genome size (GS) variation is shaped by natural selection. One approach to address this question is a population-level survey that assesses both the variation in GS and the relationship of GS to ecological variants.
- We assessed GS in *Zea mays*, a species that includes the cultivated crop, maize, and its closest wild relatives, the teosintes. We measured GS in five plants of each of 22 maize landraces and 21 teosinte populations from Mexico sampled from parallel altitudinal gradients.
- GS was significantly smaller in landraces than in teosintes, but the largest component of GS variation was among landraces and among populations. In maize, GS correlated negatively with altitude; more generally, the best GS predictors were linked to geography. By contrast, GS variation in teosintes was best explained by temperature and precipitation.
- Overall, our results further document the size flexibility of the *Zea* genome, but also point to a drastic shift in patterns of GS variation since domestication. We argue that such patterns may reflect the indirect action of selection on GS, through a multiplicity of phenotypes and life-history traits.

---

© 2013 The Authors © 2013 New Phytologist Trust

Author for correspondence: Brandon Gaut Tel: +1 949 824 2564 bgaut@uci.edu.

Supporting Information

Additional supporting information may be found in the online version of this article.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

## Keywords

altitudinal gradient; climatic variables; DNA content; genome size; natural selection; *Zea mays*

---

## Introduction

Nuclear DNA content, or genome size (GS), varies > 1000-fold among angiosperms (Muñoz Díez *et al.*, 2012). Beyond its structural impacts on the genome, DNA content variation may have ecological and evolutionary significance (Biemont, 2008), because GS correlates with a range of phenotypes, including flowering time, flower size, seed mass, leaf size and photosynthetic rate (e.g. Meagher & Vassiliadis, 2005; Beaulieu *et al.*, 2007a,b, 2008). Moreover, several studies have reported within-species correlations between GS and ecological variables, such as altitude, latitude and temperature (Knight *et al.*, 2005). These phenotypic and ecological correlates suggest that GS affects the properties of species, such as regional abundances (Herben *et al.*, 2012), colonization rates and invasiveness (Lavergne *et al.*, 2010).

It thus seems likely that GS is shaped by natural selection, but nonselective alternatives have also been proposed to explain GS variation. For example, the skewed distributions of eukaryotic GSs can be explained by a purely mechanistic model in which GS evolves stochastically at a rate proportional to size (Oliver *et al.*, 2007). In vertebrates, recombination rates, rather than natural selection, seem to drive changes in GS (Nam & Ellegren, 2012). It has also been argued that correlation between GS and population or biological parameters may be blurred by the phylogenetic signal. After correcting for this signal, Whitney *et al.* (2010) reported that GS in angiosperms does not correlate with effective population size ( $N_e$ ). Because the efficacy of selection is expected to scale with  $N_e$ , the lack of relationship between GS and  $N_e$  may indicate that selection has had little impact on the broad-scale evolution of GS in higher plants.

Population-level analyses offer the best opportunity to infer selection on GS (Petrov, 2001), but most analyses of the evolutionary processes acting on GS have taken place on an interspecies scale. Within the plant kingdom, GS has perhaps been best studied within the genus *Zea*. The genus includes the species *Zea mays*, which is typified by the domesticated subspecies maize (*Zea mays ssp. mays*) and two prominent wild subspecies – *Zea mays ssp. mexicana* and *Zea mays ssp. parviglumis*. These two wild taxa, both of which are native to Mexico and collectively referred to as ‘teosinte’, are geographically and ecologically distinct. Subspecies *mexicana* is restricted primarily to the highlands in the states of Chihuahua, Durango and Puebla, as well as the Central Plateau that surrounds Mexico City (Iltis & Doebley, 1980; Fukunaga *et al.*, 2005). At an average elevation of 2135 m (Hufford *et al.*, 2012), *mexicana* exhibits adaptation to variable temperature conditions and high altitudes. By contrast, *ssp. parviglumis* is found at lower elevations in the Balsas River Basin and the state of Jalisco, tropical regions with relatively stable temperature regimes and lower average altitudes (1095 m; Hufford *et al.*, 2012). The one location in which the two teosinte taxa overlap is the Balsas River Basin (Fukunaga *et al.*, 2005), which is the presumed

location of the domestication of maize from *ssp. parviglumis* c. 9000 yr ago (Matsuoka *et al.*, 2002; van Heerwaarden *et al.*, 2011).

Although there is still debate as to whether GS commonly varies within species (Knight *et al.*, 2005; Biemont, 2008), there is little doubt that GS varies within *Zea mays sensu lato*. Among cultivated accessions, which include both open-pollinated land-races and inbred lines, GS varies by at least 30% (Muñoz Díez *et al.*, 2012). This variation exhibits inconsistent patterns with geography. In maize, for example, GS and indirect measures of GS (i.e. the number of chromosomal knobs and C-bands) typically decrease with increasing latitude (Rayburn *et al.*, 1985; but see Rayburn & Auger, 1990), but may be less consistent as a function of altitude (Bennett & Smith, 1976; Rayburn & Auger, 1990). Fewer studies have examined correlates between GS and ecological variables in *ssp. mexicana* and *ssp. parviglumis*, but at least one study has found a positive correlation between GS and altitude (Laurie & Bennett, 1985). To the extent that the pattern holds, a negative correlation between GS and altitude/latitude may reflect the need for rapid growth and early flowering in the shorter growing seasons typical of cooler regions (Rayburn *et al.*, 1994; Poggio *et al.*, 1998).

Although it is clear that GS varies within and among the sub-species of *Zea mays*, there has not yet been a systematic survey of GS variation among taxa and populations. Here, we investigate GS and ecological correlates in large samples of both cultivated landraces of maize and in annual Mexican teosintes. In doing so, we have overcome some of the common limitations of previous studies. The first of these is the sampling; most previous surveys of *Zea* have assessed GS from c. 20 total isolates. Here, we survey 22 maize landraces and 21 populations of annual teosinte, assessing GS variation from five individuals within each landrace and population. A second limitation is that the geographic scale of sampling has often been so broad that the effects of environmental factors have been blurred. Our sampling along parallel altitudinal gradients is designed specifically to reduce this limitation. Finally, environmental information has been limited mostly to altitude or latitude, without consideration of additional ecological variables that may impact more directly on GS, such as temperature and rainfall.

With GS and environmental data for multiple taxa and specific locations, we address three sets of questions. First, what is the distribution of GS among taxa, populations and plants? Is there convincing evidence that GS varies within taxa in a consistent fashion? Second, does GS correlate with bioclimatic variables? If so, are the correlates consistent across parallel gradients and also among taxa, suggesting a clear relationship between GS and environment? Or might relationships vary between the domesticate and its wild ancestors, as suggested by previous studies? Finally, what does this information reveal about the potential evolutionary and ecological forces that act on GS? Ultimately, we seek to determine whether variation in GS is shaped by, and limited by, ecological and evolutionary factors.

## Materials and Methods

### Plant materials and GS measurements

We sampled teosintes from two altitudinal gradients, hereafter called the ‘Balsas’ and ‘Jalisco’ gradients. ‘Balsas’ was located in the states of Guerrero, Morelos and Estado de México (Fig. 1). The Guerrero and Morelos states include part of the Balsas River Basin, whereas Estado de México represents part of the highlands of Central Mexico. The ‘Jalisco’ gradient was located in the Central Western region of Mexico, including portions of the states of Jalisco, Michoacán and Guanajuato that represent portions of the Bajío and Río Lerma Santiago Basins (Fig. 1). The two gradients ranged between 343 and 2581 m (Fig. 1, Table 1). We sampled 21 total teosinte populations from these gradients, with 10 and 11 populations from the Balsas and Jalisco gradients, respectively. At least 30 adult plants and their seeds (20–50 per adult) were collected from each population. Each gradient encompassed geographic regions typical of both *parviglumis* and *mexicana*; we made a provisional designation of subspecies status based on distribution maps (Hufford *et al.*, 2012).

We also analyzed 22 landraces of maize (*Zea mays* ssp. *mays* L.). These were sampled along three altitudinal gradients. The first was the ‘Balsas’ gradient, consisting of nine landraces; the second was a gradient in the state of Oaxaca (with eight landraces); and the third, for which there was sampling of four landraces, was the ‘Jalisco’ gradient (Fig. 1, Table 1). Our sampling included two landraces (MEXI05 and YUCA117) that have been reported previously as having a small (Vielle-Calzada *et al.*, 2009) and a large (J. P. Vielle-Calzada, pers. comm.) genome, respectively. In total, we assessed 22 primitive landraces with the altitudinal range of collection sites varying from 10 to 3073 m (Table 1). All landrace seeds were provided by Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT).

Five plants per population or landrace were germinated and grown under controlled glasshouse conditions between February and August of 2011 at UC Irvine, CA, USA (33°38′49″N, 117°50′43″W; altitude, 25 m). Leaf samples were sent to Plant Cytometry Services (Schijndel, the Netherlands) where relative DNA measurements were performed by flow cytometry using 4′,6-diamidino-2-phenylindole (DAPI) staining and *Ilex crenata* ‘Fastigiata’ (2C = 2.2 pg) as an internal standard. This technique is particularly well suited for the detection of small DNA variation among samples. In order to limit technical error, we included the reference maize inbred line B73 in every batch as a second internal control. In addition, the GS measurements were made on three technical replicates per plant on a single leaf harvested at the same level of maturity for each plant. The GS measurement for each technical replicate was calculated as the ratio of the measure to that of maize reference B73; the estimate for each plant was the average among three technical replicates. Because we are interested in ranges of GS variation among *Zea* individuals, for analyses we expressed GS values as a ratio relative to B73, where B73 has a GS value of 1.0. However, we also provided estimates of GS, in picograms (pg) per 2C nucleus, for comparison with other species (Fig. 2, Supporting Information Table S1); these picogram estimates were calculated assuming that the B73 reference has a GS of 5.64 pg/2C, as based on propidium iodide (PI) staining against the *I. crenata* standard.

### Statistical analysis of GS within and among populations

The GS data satisfied the homogeneity of variances (O'Brien test,  $P > 0.05$ ), but did not fulfill the normality of the variance (Shapiro–Wilk test,  $P < 0.0001$ ) required to apply an ANOVA. For this reason, the distribution of GS variance as a result of differences among populations (Pop) and among plants within populations (Pop : Plant) was determined by a nonparametric multivariate analysis of variance (PERMANOVA; Anderson, 2001), using the function 'adonis' of the package 'vegan 1.17-6' (Oksanen, 2007) in R statistical software (Team, 2008). The 'adonis' function performs an ANOVA with matrices of Euclidean distances, testing significance via the analysis of 999 permuted matrices. We also compared variation in GS between gradients (Gradient) and among populations within gradients (Gradient : Pop) using PERMANOVA, based on landrace samples from the Balsas and Oaxaca gradients and teosinte samples from the Balsas and Jalisco gradients.

To assess the extent of spatial autocorrelation among the GS of populations and landraces, we estimated Moran's  $I$  statistic for all possible 'neighborhoods' defined by distances between locations that varied from 15 km up to 1000 km, with 5-km increases between the two extremes. The significance of the  $I$  statistic was determined by 10 000 random permutations using `spdep v. 0.5-56` (Bivand *et al.*, 2012) in R statistical software.

### Comparison of GS with bioclimatic variables

For the collection site of each natural population and landrace, we recorded passport data, including geographical coordinates and altitude. These data were used to obtain climatic information for each natural population and landrace from the current bioclimatic grid layers from WorldClim V.1.4 (30 arc-seconds c. 1 km resolution; <http://worldclim.org/bioclim>; Hijmans *et al.*, 2005). The 19 bioclimatic variables represent summaries of means and variations in temperature and precipitation, and characterize the dimensions of climate considered to be particularly relevant in the determination of species distribution at a regional scale (Hijmans *et al.*, 2005).

Given the history in the literature of comparing GS with elevation, we first focused on this relationship, employing a linear quantile regression (QR) model to explore the effect of elevation on percentiles of GS values (Knight & Ackerly, 2002; Knight *et al.*, 2005). QR estimates the relationship between two variables for all portions (quantiles) of a probability distribution, rather than just the mean, as in standard regression analyses (Cade & Noon, 2003). As a result, QR models can provide insights into predictive relationships at the extremes of the distribution of a response variable. We estimated the QR functions from the 15th to the 95th quantile using the package 'quantreg' (Koenker, 2008) in R statistical software. Confidence intervals were calculated by 1000 bootstrap resamplings.

To assess the relationships among GS and bioclimatic variables, we applied Spearman Rank correlation tests and principal component analysis (PCA) using R. We also integrated the bio-climatic variables into a linear model based on partial least-squares regression (PLSR). The goal of PLSR was to assess the predictive value of the models relative to observed GS values. The PLSR method combines aspects of PCA and multiple regression, and is particularly useful when there are many predictor variables that may be nonindependent

(Maestre, 2004). To determine the number of components to be included in the PLSR model, we performed a standard cross-validation procedure that also estimates the root-mean-square error of prediction (RMSEP; Maestre, 2004; Mevik & Wehrens, 2007). We retained only those components which contribute substantially to decrease RMSEP, which also happened to explain > 5% of the original variance ( $R^2$ ) in the response variable (Carrascal *et al.*, 2009). The sum of the  $R^2$  of the retained components provided the total explanatory capacity of the PLSR model (Mevik & Wehrens, 2007). PLSR analysis was performed using the package ‘pls’ (Mevik & Wehrens, 2007) in the R statistical package.

## Results

### GS variation among and within populations

We measured the GS of five plants per population belonging to 21 natural populations of teosinte and 22 maize landraces, most of which were collected along altitudinal gradients in Mexico. We sampled five plants per population or landrace, each with three technical replicates, resulting in a dataset of 645 GS measurements. For each plant, the GS value is reported as the average of three technical replicates based on comparisons with the maize inbred line B73, which has a normalized value of 1.0 in this study.

GS values varied widely among individuals, from 0.948 for an individual from the landrace ‘Palomero Legítimo’ (MEXI211) to 1.299 for an individual from the landrace ‘Olot colorado’ (OAXA522; Fig. 2). The average GS value per plant was 1.111, but these averages varied between wild and cultivated samples; the average GS of teosintes (1.129) was significantly larger ( $P < 0.001$ , Kruskal–Wallis test) than the average GS of cultivated maize (1.095). The overall coefficient of variation was nearly identical between the wild and cultivated samples, however, at 6.07% and 6.06%, respectively.

The distribution of GS among and within populations was evaluated independently in the wild and cultivated groups by a PERMANOVA test. The major proportion of the variance was significantly ( $P < 0.001$ ) distributed among populations in both groups (Table 2). This proportion was similar, but higher, in the cultivated ( $R^2 = 84.4\%$ ) than in the wild ( $R^2 = 73.2\%$ ) group. Conversely, the proportion of the variance within populations in the wild group ( $R^2 = 23.8\%$ ;  $P < 0.001$ ) was higher than that in the cultivated group ( $R^2 = 13\%$ ;  $P < 0.001$ ). The residual accounted for the technical error among the three biological replicates for each plant, and was *c.* 3% ( $R^2$ ) in both groups.

One source of differences among wild populations could be a difference between ssp. *mexicana* and *parviglumis*. We therefore estimated the variance attributable to subspecies (Table S2). It was significant but quite low ( $R^2 = 2.2\%$ ;  $P < 0.001$ ), with a much larger proportion of the variance attributable to populations ( $R^2 = 70.9\%$ ;  $P < 0.001$ ). Given that *mexicana* and *parviglumis* do not differ substantially by this analysis, many of our analyses combine samples from the two teosinte taxa into a single dataset. However, we do accentuate taxonomic differences when appropriate.

We sampled separate elevation gradients for both the landrace and teosinte samples with the thought that these gradients provide a measure of pseudo-replication. We therefore



contrasted the Balsas and Jalisco gradients for the teosinte samples and the two best sampled gradients (Balsas and Oaxaca) for the landraces. For both maize and teosinte, the variance caused by differences between the two gradients was significant ( $P < 0.001$ ), but far less for the wild samples ( $R^2 = 4.6\%$ ) than for the cultivated samples ( $R^2 = 18.5\%$ ). The highest component of variance again reflected differences among populations ( $R^2 = 68.6\%$ ) or among landraces ( $R^2 = 66.7\%$ ).

Finally, GS variation showed some spatial structure, because population GS values were spatially autocorrelated for both cultivated and wild samples (Fig. S1). Moran's  $I$  was significant for neighborhoods of 40 km for all GS data, 65 km for cultivated data alone and 75 km for teosintes data, with the  $I$  statistic remaining positive up to neighborhood sizes of 120 km (Fig. S1). To sum, our GS estimates represent significant differences among populations and among landraces, with lower GS variances attributable to either gradients or individuals within populations.

### The relationship between GS and elevation

We have established that populations and landraces vary significantly in GS. But does this variation follow a discernible pattern with regard to environment? We first approached this question with an analysis of elevation, both because elevation has been used as a proxy for a suite of environmental variables and because its relationship with GS has long been investigated (Rayburn & Auger, 1990; Poggio *et al.*, 1998; Knight & Ackerly, 2002; Knight *et al.*, 2005).

We examined elevation by applying linear QR to both cultivated and wild samples (Fig. 3). In the cultivated group, GS and elevation showed a significant negative relationship ( $P < 0.001$ ) for all six quantiles considered (0.15, 0.30, 0.45, 0.60, 0.75, 0.90). This negative relationship was consistent across the Oaxaca and Balsas gradients and thus appears to be a general feature of our sample of maize landraces. It should also be noted that the slope of the 95% quantile was noticeably more negative than that of other quantiles (Fig. 3).

By contrast, the pattern for the wild samples, taken as a whole, was not clear. The lower quantiles (0.15 and 0.30) exhibited a significant positive relationship ( $P < 0.05$ ) between elevation and GS, but the upper quantiles (0.45, 0.60, 0.75, 0.90) showed nonsignificant ( $P > 0.05$ ) correlations. When we partitioned the samples by gradient, the picture became even murkier (Fig. 3). There was a consistent, significant and positive relationship at all quantiles for the Balsas gradient, but the exact opposite held for the Jalisco gradient. When individual subspecies were considered, the positive correlation held in both subspecies for the 50% quantile, but the slopes for all other quantiles differed in sign between sub-species (Fig. S2). Thus, the correlation with GS and elevation varied as a function of the teosinte sample.

### Correlations between GS and bioclimatic variables

In theory, elevation is a proxy for several correlated variables, and hence we investigated correlations between GS and bioclimatic variables directly. Several geographic and climatic variables correlated significantly with GS based on Spearman Rank correlations (Table 3). In maize, for example, the approach recapitulated the negative correlation with elevation.

Several additional correlations were significant for the entire landrace sample, including variables representing longitude, temperature (BIO5 and BIO6) and precipitation (e.g. BIO13). Significance based on the total maize sample seems to be primarily a function of the Balsas gradient, because correlations were generally not significant for the Oaxaca gradient, even though the two gradients had a similar number of landraces sampled ( $n = 8$  vs 9 populations for Balsas vs Oaxaca). The only correlations that were both shared and significant between gradients were those with elevation (Table 3; Fig. 2) and seasonal precipitation (BIO15).

The teosintes exhibited different trends from maize (Table 3). For the complete sample, only a few variables were strongly correlated with GS – i.e. longitude and precipitation in the warmest and coldest quarters (BIO17 and BIO19). Otherwise, the results depended heavily on sample partitions. For example, for the two gradients, the patterns of correlations with GS were significant, but in opposite directions, for several variables (e.g. BIO3, BIO6, BIO11, BIO14, BIO16 and BIO18; Table 2). When the samples were partitioned by subspecies (Table S3), the correlations become slightly more regular, with the most strong and consistent correlations between GS and longitude, temperature seasonality (BIO4) and levels of precipitation in the coldest and driest quarters (BIO17 and BIO19).

Pairwise correlations suffer from several weaknesses (see Discussion), and so we turned to PLSR, which is a technique to build a linear model that decomposes the data into (in this case) orthogonal components. Once the model is estimated, it can then be used to predict the dependent variable, GS, from potentially nonindependent predictor variables. We applied PLSR separately to cultivated and wild GS measurements with the 22 predictor variables (Table 3). For cultivated data, we included GS data from the two main gradients sampled (Balsas and Oaxaca); for the teosinte data, both gradients were included. PLSR identified four and five components for cultivated and wild samples, respectively, which explained > 5% of the variance in GS, which we retained in the PLSR model. Once the PLSR model was determined, it was used to demonstrate the fit of the predicted values to observed values (Fig. 4). Together, the significant components explained 78.6% and 62.7% of the observed variance in GS for cultivated and wild samples, respectively. In other words, GS can be predicted partially on the basis of climatic and geographic variables.

PLSR is usually implemented to build a regression model, but it can also provide insights into the weight and significance of predictor variables to the model. Table 3 presents the coefficients (or loadings) of each predictor variable, as summed across orthogonal components. The magnitude of each coefficient provides insight into its importance in the model; the significance of coefficients was calculated by a jackknife procedure. The weights of coefficients yielded three unique insights. First, in maize, the suite of most important variables is related to geography (elevation, longitude and, to a lesser extent, latitude). Because we were puzzled by the relationship between geographic measures and GS, we reapplied the PLSR model without latitude and longitude (Table S4), with a commensurate reduction in predictive value (Fig. S3). Second, the most highly weighted predictors in teosintes are related to temperature and precipitation. Finally, and perhaps most surprisingly, the variables that predict GS have shifted dramatically between cultivated landraces in



Mexico and wild teosinte populations, even though the landrace and wild samples originate from similar regions and sampling strategies.

### PCA reveals distinct groups of teosintes

We conducted a PCA to group populations on the basis of the complete dataset – that is, GS data plus the 22 predictor variables for both maize and teosintes (Figs 5, S4). The PCA approach is helpful because variables may be represented diagrammatically as vectors of particular magnitude and direction in the space defined by principal components. For example, it is possible to infer that variables related to temperature seasonality (BIO1, BIO5, BIO8, BIO9, BIO10, BIO11 and BIO15) have similar (correlated) effects on the grouping of populations because their associated vectors have similar direction (Fig. 5).

In our analysis, the first principal component (PC1) explained 12.7% of the total variance, and the second component (PC2) explained 3.4% of the variance. Together, they grouped particular subsets of populations. For example, the Balsas teosinte populations formed a distinct group relative to the remaining teosintes. By contrast, the PCA produced no obvious grouping of landraces, but instead illustrated the remarkable ecological flexibility of cultivated maize, because landraces were distributed across the space defined by the first two principal components. We note, however, that most of the landraces from the Jalisco gradient grouped near the wild populations collected in the same area, as is expected if bioclimatic variables dictate landrace usage.

## Discussion

### GS variation within species

Our study was designed to investigate variation in GS, the relationship of GS with climatic variables and, ultimately, the evolutionary processes that shape GS variation. With regard to the first, there is still debate as to whether GS varies within species and, if so, the extent of that variation (Knight *et al.*, 2005). In order to overcome technical difficulties that may create false variation among individuals (Nardon *et al.*, 2003; Biemont, 2008), we used plants grown and harvested under similar conditions, estimated GS by the same technique on all samples and employed extensive technical replication.

Our analyses clearly indicate that GS varies both among and within subspecies of *Z. mays*. Among subspecies, our sample of maize landraces has a moderately (*c.* 3%), but significantly, lower average GS than our sample of teosintes. The GS of elite inbred lines of maize is smaller still (Laurie & Bennett, 1985), suggesting that average GSs have decreased through the processes of domestication and subsequent crop improvement, perhaps because of strong selection (Rayburn *et al.*, 1994). Average GS also varies significantly between the wild *parviglumis* and *mexicana* subspecies, but this variation pales in comparison with the amount of variation among populations. Indeed, among-population variation is the largest component of GS size variation, larger even than differences between subspecies or altitudinal gradients (Table 2).

Given GS variation, it is interesting to speculate on the genomic components that contribute to size variation – that is, features such as genes, transposable elements (TEs), centromeric

repeats, telomeres, rDNAs and heterochromatic knobs. There is already substantial evidence that maize and teosinte genomes vary both in gene content (Swanson-Wagner *et al.*, 2010) and TE complement (Wang & Dooner, 2006). Maize accessions are also known for varying in their number of heterochromatic knobs (McClintock, 1978). The measurement of knobs and TE repeats has been enhanced by the use of high-throughput sequencing (HTS; Tenaillon *et al.*, 2011). For example, a study of 38 teosinte and maize inbred lines used HTS data to show that intraspecies differences in GSs are driven primarily by the amount of knob repeats (Chia *et al.*, 2012). Given that knobs constitute only 8% of the maize genome (Ananiev *et al.*, 1998), whereas TEs represent c. 85% (Schnable *et al.*, 2009), this observation is somewhat surprising. Even more surprising is the observation that the overall TE content of these lines actually correlates *negatively* with GS (Chia *et al.*, 2012).

These observations seem to imply some sort of genomic ‘trade-off’ between heterochromatic knobs and TEs, as if they balance actively to maintain limits on GS. Although it is difficult to imagine a precise molecular mechanism for this tradeoff, there are some hints that optimal GSs may indeed exist. For example, it has long been suggested that large genomes are disadvantageous (Knight *et al.*, 2005) and stabilizing selection has been shown to act on GS (Smarda *et al.*, 2010) over short time scales (Smarda *et al.*, 2007). There are hints of similar GS limitations in maize data (Poggio *et al.*, 1998). Take, for example, our QR analysis between GS and elevation (Fig. 3). Although there is a clear and consistent positive association for both maize gradients, the higher quantiles have lower slopes, as might be expected if there are limits to GS. Moreover, the visual distribution of points suggests a quadratic relationship between GS and elevation; that is, the curvature of points suggests that GS reaches a maximum size at moderate elevations, with upper and lower limits (e.g. Knight & Ackerly, 2002). We tested a quadratic QR on these data, but the quadratic fits no better than the linear relationship (data not shown). Nonetheless, the pattern of observations is superficially consistent with a limit on maize GS as a function of elevation or variables correlated with elevation. We note that such limits on GS in maize have been argued previously based on the abundance of B chromosomes at different altitudes (Lia *et al.*, 2007).

By contrast, there is no consistent evidence for constraints on GS in teosintes on the basis of QRs with elevation, whether the wild samples are treated as a whole (Fig. 3), in different gradients (Fig. 3) or in different subspecies (Fig. S2).

### Climatic variables and GS

A related question is whether GS varies predictably as a function of climatic factors, suggesting ecological constraints on GS. Our study has uncovered a wealth of pairwise correlations between GS and bioclimatic variables (Table 3), but the pairwise correlations themselves are inconsistent among subspecies and gradients (Tables 3, S3), making it difficult to interpret their meaning and importance. Pairwise correlations may be inconsistent for a variety of reasons. One reason is sampling, both in terms of the depth of sampling and the climatic range over sampled locations. Although we have sampled extensively, our sampling could very well be too sparse to be able to identify consistent patterns between gradients or subspecies. Another sampling issue is climatic range; without

sufficient variation in climate, it may be difficult to identify correlations between individual variables and GS. Our sampling of altitudinal gradients was nonetheless intended to cover climatic ranges within the limited distribution of the annual teosintes.

Another potential cause for inconsistencies in pairwise correlations may be that there are real but multifaceted relationships between environmental factors and GSs. Accordingly, we applied PLSR to our data. PLSR is designed to model relationships when there are many potentially nonindependent predictor variables and a limited number of observations (Mevik & Wehrens, 2007). The resulting models are predictive, explaining 78.6% and 62.7% of GS variation for landraces and teosintes, respectively (Fig. 5). Although PLSR is not specifically designed to identify the contribution of particular variables to the model, it does provide insight into the suites of variables that contribute to GS (Table 3). For example, the PLSR indicates that several variables are predictors of GS in the teosintes, including temperature-related variables (BIO2, BIO3 and BIO4) and precipitation in the wettest quarter (BIO16). Both BIO4 and BIO16 are also determinants of the climatic distribution of *parviglumis* and *mexicana* (Hufford *et al.*, 2012).

In addition to elevation, it is notable that PLSR models also detect latitude and longitude as strong predictors of GS in maize (Table 3). We assume that these geographic coordinates somehow encapsulate a complex combination of climatic (and perhaps even sociological) factors that are otherwise unrepresented by the bioclimatic data. For example, higher latitudes may correlate with cooler growing seasons, as is often assumed for elevation. In this particular study, longitude is related to distance from the coast, and thus may capture aspects of climate (such as humidity) or even historical factors that vary from coastal areas to the highlands.

### Does selection drive variation in GS?

We ultimately seek to determine whether GS variation is influenced by natural selection. To that end, as already noted, we find that within-population variation in GS is small relative to variation among populations, and that such variation is spatially autocorrelated. Although this pattern does not itself implicate the action of natural selection (population divergence can also be a function of genetic drift and isolated populations), it is a necessary precursor to infer that natural selection affects GS.

Although stabilizing and directional selection have been shown to affect GS in experimental settings (Rayburn *et al.*, 1994; Cullis, 2005; Smarda *et al.*, 2010), the appropriate null hypothesis for a survey like ours is that natural selection does not affect GS. This null hypothesis is difficult to reject, but we nonetheless suggest that the predictive performance of the PLSR models, which demonstrate dependences between GS and both geographical (elevation, longitude) and environmental (temperature and precipitation) factors, tends to discount it (Fig. 4, Table 3).

We also designed our sampling under a secondary hypothesis: if natural selection has shaped GS, we assume a priori that there should be consistent relationships between GS and environmental factors across taxa and gradients. This secondary hypothesis is not supported. For example, the environmental factors that correlate most noticeably with GS in teosintes

(e.g. BIO2, BIO3, BIO4, BIO16; Table 3) do not correlate consistently with GS in maize, and vice versa (Table 3, Fig. 3).

We are left to conclude that, if GS is affected by selection, the relationship is defined by at least two salient features. The first is complexity, born of a shifting mix of environmental factors and phenotypes. The phenotypes that may affect GS include growth rate, flowering time, flower size, seed mass, leaf size and photo-synthetic rate (e.g. Rayburn *et al.*, 1994; Meagher & Vassiliadis, 2005; Beaulieu *et al.*, 2007a,b, 2008). Each phenotype may be under different selection pressures in different biotic and abiotic environments; in turn, selection on multiple phenotypes may affect GS in synergistic or opposing fashions. It should also be noted that chromosomal knobs may be under selection through the effects of genetic factors, such as Abnormal chromosome 10 (Ab10), which may be less common in regions with short growing seasons. Ab10 is significant because it is associated with meiotic drive for increased knob heterochromatin (Buckler *et al.*, 1999). The Ab10-carrying genotype may decrease with altitude because of fitness costs associated with genes in tight linkage (Buckler *et al.*, 1999), generating an indirect effect on both knob frequency and GS.

The second is that shifts in the relationship can be very rapid. We have shown, for example, that GS in our sample of maize, which was domesticated *c.* 9000 yr ago, exhibits different relationships to both climate and geographic factors than does GS in teosintes (Fig. 3). It is possible that strong artificial selection on maize has fundamentally exaggerated, altered or skewed relationships between GS and phenotypes. This point may be best illustrated by our PCA (Fig. 4): given that maize was domesticated from *parviglumis* in the lowlands of the Balsas regions (Matsuoka *et al.*, 2002; van Heerwaarden *et al.*, 2011), we expect that the first domesticates of maize would have clustered with the Balsas teosinte group. Thereafter, maize was quickly distributed to the Mexican highlands and beyond (Fukunaga *et al.*, 2005), across a range of climates (Hufford *et al.*, 2012), so that maize samples of different GSs are now distributed throughout the climactic space defined by the PCA (Figs 4, S4). A final note is that, if our inferences are correct – that is, that the relationship between GS and selection is defined by complexity and rapid shifts – it may not be surprising that global analyses, like those of Whitney *et al.* (2010), fail to find convincing evidence of a strong effect of natural selection despite the fact that experimental analyses clearly do (Rayburn *et al.*, 1994; Cullis, 2005; Smarda *et al.*, 2010).

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We would like to thank S. Taba at CIMMYT and J.-P. Vielle-Calzada for providing information and seeds on landraces. We would like to acknowledge the support of a UC-MEXUS grant #49298 to B.S.G. and L.E.E., as well as the Agence Nationale de la Recherche, ANR-12-ADAP-002, to M.I.T. C.M.D. was supported by a fellowship funded by the project P09-AGR-5010 of the Consejería Economía, Innovación y Empleo de la Junta de Andalucía, Spain. E. M. was supported by the Balsells Fellowship at UC Irvine and from NIH Training Grant P50GM76516.

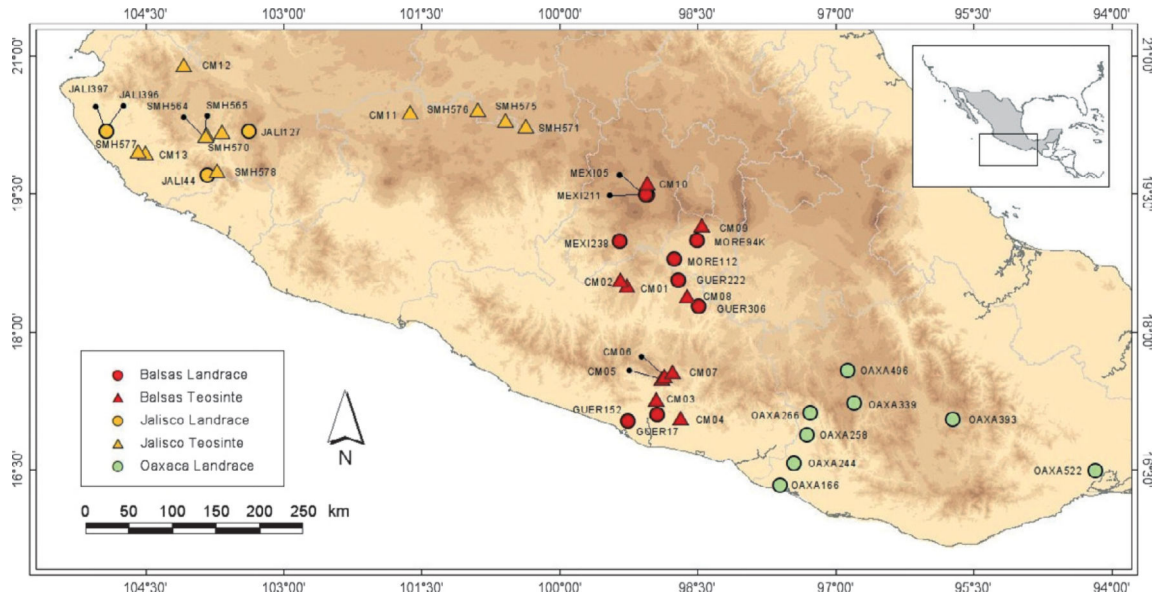
## References

- Ananiev EV, Phillips RL, Rines HW. A knob-associated repeat in maize capable of forming fold-back DNA segments: are chromosome knobs megatransposons? *Proceedings of the National Academy of Sciences, USA*. 1998; 95:10785–10790.
- Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*. 2001; 26:32–46.
- Beaulieu JM, Leitch IJ, Knight CA. Genome size evolution in relation to leaf strategy and metabolic rates revisited. *Annals of Botany*. 2007a; 99:495–505. [PubMed: 17204534]
- Beaulieu JM, Leitch IJ, Patel S, Pendharkar A, Knight CA. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist*. 2008; 179:975–986. [PubMed: 18564303]
- Beaulieu JM, Moles AT, Leitch IJ, Bennett MD, Dickie JB, Knight CA. Correlated evolution of genome size and seed mass. *New Phytologist*. 2007b; 173:422–437. [PubMed: 17204088]
- Bennett MD, Smith JB. Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London*. 1976; 274:227–274. [PubMed: 6977]
- Biemont C. Genome size evolution: within-species variation in genome size. *Heredity*. 2008; 101:297–298. [PubMed: 18665185]
- Bivand, R.; Altman, M.; Anselin, ML.; Assunção, R.; Olaf, B. [1 May 2012] spdep: Spatial dependence: weighting schemes, statistics and models, R package version 0.5-46. 2012. URL <http://cran.r-project.org/web/packages/spdep/index.html>
- Buckler EST, Phelps-Durr TL, Buckler CS, Dawe RK, Doebley JF, Holtsford TP. Meiotic drive of chromosomal knobs reshaped the maize genome. *Genetics*. 1999; 153:415–426. [PubMed: 10471723]
- Cade BS, Noon BR. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*. 2003; 1:412–420.
- Carrascal LM, Galvan I, Gordo O. Partial least squares regression as an alternative to current regression methods used in ecology. *Oikos*. 2009; 118:681–690.
- Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*. 2012; 44:803–807. [PubMed: 22660545]
- Cullis CA. Mechanisms and control of rapid genomic changes in flax. *Annals of Botany*. 2005; 95:201–206. [PubMed: 15596467]
- Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, Sanchez GJ, Liu K, Buckler ES, Doebley J. Genetic diversity and population structure of teosinte. *Genetics*. 2005; 169:2241–2254. [PubMed: 15687282]
- van Heerwaarden J, Doebley J, Briggs WH, Glaubitz JC, Goodman MM, de Jesus Sanchez Gonzalez J, Ross-Ibarra J. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences, USA*. 2011; 108:1088–1092.
- Herben T, Suda J, Klimesova J, Mihulka S, Riha P, Simova I. Ecological effects of cell-level processes: genome size, functional traits and regional abundance of herbaceous plant species. *Annals of Botany*. 2012; 110:1357–1367. [PubMed: 22628380]
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 2005; 25:1965–1978.
- Hufford MB, Martinez-Meyer E, Gaut BS, Eguiarte LE, Tenaillon M. Past and present distributions in *Zea mays*: a chance to revisit maize history. *PLoS ONE*. 2012; 7:e47659. [PubMed: 23155371]
- Iltis HH, Doebley JF. Taxonomy of *Zea* (Gramineae). 2. Subspecific categories in the *Zea mays* complex and a generic synopsis. *American Journal of Botany*. 1980; 67:994–1004.
- Knight CA, Ackerly DD. Variation in nuclear DNA content across environmental gradients: a quantile regression analysis. *Ecology Letters*. 2002; 5:66–76.
- Knight CA, Molinari NA, Petrov DA. The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*. 2005; 95:177–190. [PubMed: 15596465]

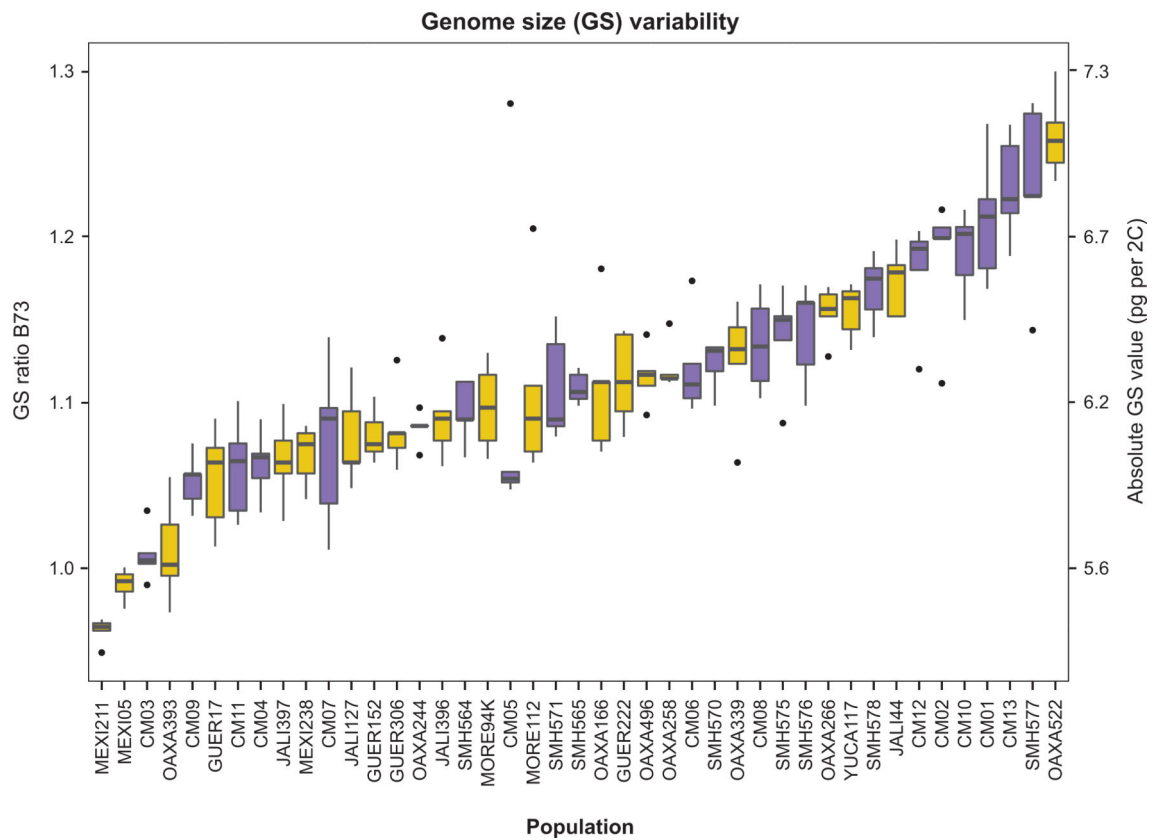
- Koenker, R. [18 May 2012] quantreg version 4.76. 2008. URL <http://cran.r-project.org/web/packages/quantreg/index.html>
- Laurie DA, Bennett MD. Nuclear DNA content in the genera *Zea* and *Sorghum* – intergeneric, interspecific and intraspecific variation. *Heredity*. 1985; 55:307–313.
- Lavergne S, Muenke NJ, Molofsky J. Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Annals of Botany*. 2010; 105:109–116. [PubMed: 19887472]
- Lia VV, Confalonieri VA, Poggio L. B chromosome polymorphism in maize landraces: adaptive vs. demographic hypothesis of clinal variation. *Genetics*. 2007; 177:895–904. [PubMed: 17954923]
- Maestre FT. On the importance of patch attributes, environmental factors and past human impacts as determinants of perennial plant species richness and diversity in Mediterranean semiarid steppes. *Diversity and Distributions*. 2004; 10:21–29.
- Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez J, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences, USA*. 2002; 99:6080–6084.
- McClintock, B. Significance of chromosome constitutions in tracing the origin and migration of races of maize in the Americas. In: Walden, DB., editor. *Maize breeding and genetics*. John Wiley and Sons; New York, NY, USA: 1978. p. 159-184.
- Meagher TR, Vassiliadis C. Phenotypic impacts of repetitive DNA in flowering plants. *New Phytologist*. 2005; 168:71–80. [PubMed: 16159322]
- Mevik B-H, Wehrens R. The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software*. 2007; 18:1–24.
- Muñoz Díez, C.; Vitte, C.; Ross-Ibarra, J.; Gaut, BS.; Tenaillon, M. Using nextgen sequencing to investigate genome size variation and transposable element content. In: Grandbastien, M-A.; Casacuberta, J., editors. *Plant transposons (Topics in current genetics. Vol. 24. Springer-Verlag; Germany: 2012. p. 41-58.*
- Nam K, Ellegren H. Recombination drives vertebrate genome contraction. *PLoS Genetics*. 2012; 8:e1002680. [PubMed: 22570634]
- Nardon C, Weiss M, Vieira C, Biémont C. Variation of the genome size estimate with environmental conditions in *Drosophila melanogaster*. *Cytometry Part A*. 2003; 55:43–49.
- Oksanen, J. [29 May 2012] Multivariate analysis of ecological communities in R: vegan tutorial. 2007. p. 39 URL <http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf>
- Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM. The mode and tempo of genome size evolution in eukaryotes. *Genome Research*. 2007; 17:594–601. [PubMed: 17420184]
- Petrov DA. Evolution of genome size: new approaches to an old problem. *Trends in Genetics*. 2001; 17:23–28. [PubMed: 11163918]
- Poggio L, Rosato M, Chiavarino AM, Naranjo CA. Genome size and environmental correlations in maize (*Zea mays* ssp. *mays*, Poaceae). *Annals of Botany*. 1998; 82:107–115.
- Rayburn AL, Auger JA. Genome size variation in *Zea mays* ssp. *mays* adapted to different altitudes. *Theoretical and Applied Genetics*. 1990; 79:470–474. [PubMed: 24226450]
- Rayburn AL, Dudley JW, Biradar DP. Selection for early flowering results in simultaneous selection for reduced nuclear-DNA content in maize. *Plant Breeding*. 1994; 112:318–322.
- Rayburn AL, Price HJ, Smith JD, Gold JR. C-Band heterochromatin and DNA content in *Zea mays*. *American Journal of Botany*. 1985; 72:1610–1617.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009; 326:1112–1115. [PubMed: 19965430]
- Smarda P, Bures P, Horova L. Random distribution pattern and nonadaptivity of genome size in a highly variable population of *Festuca pallens*. *Annals of Botany*. 2007; 100:141–150. [PubMed: 17565968]
- Smarda P, Horova L, Bures P, Hralova I, Markova M. Stabilizing selection on genome size in a population of *Festuca pallens* under conditions of intensive intraspecific competition. *New Phytologist*. 2010; 187:1195–1204. [PubMed: 20561203]



- Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*. 2010; 20:1689–1699. [PubMed: 21036921]
- Team, RDC. [1 March 2012] R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2008. URL <http://www.r-project.org/foundation/>
- Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. Genome size and transposable element content as determined by high-throughput sequencing in Maize and *Zea luxurians*. *Genome Biology and Evolution*. 2011; 3:219–229. [PubMed: 21296765]
- Vielle-Calzada JP, Martinez de la Vega O, Hernandez-Guzman G, Ibarra-Laclette E, Alvarez-Mejia C, Vega-Arreguin JC, Jimenez-Moraila B, Fernandez-Cortes A, Corona-Armenta G, Herrera-Estrella L, et al. The Palomero genome suggests metal effects on domestication. *Science*. 2009; 326:1078. [PubMed: 19965420]
- Wang Q, Dooner HK. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proceedings of the National Academy of Sciences, USA*. 2006; 103:17644–17649.
- Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, et al. A role for nonadaptive processes in plant genome size evolution? *Evolution*. 2010; 64:2097–2109. [PubMed: 20148953]

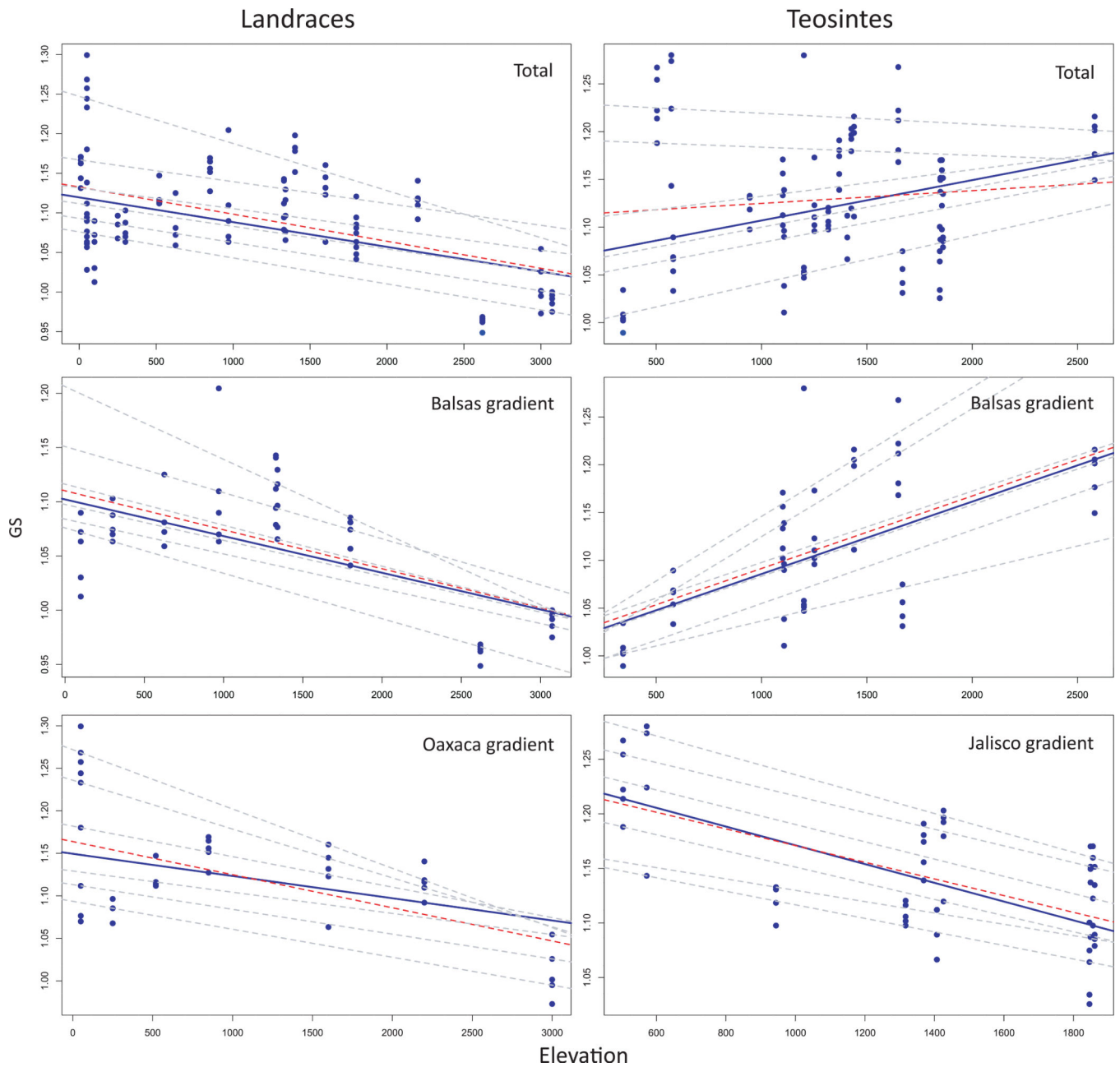


**Fig. 1.** Map of the collection locations of cultivated maize and wild teosinte samples in Mexico. Elevation gradients are denoted by separate colors. Wild and landrace samples are denoted by triangles and circles, respectively. The YUCA117 sample is not represented on the map and is not included in any analyses that compare gradients.

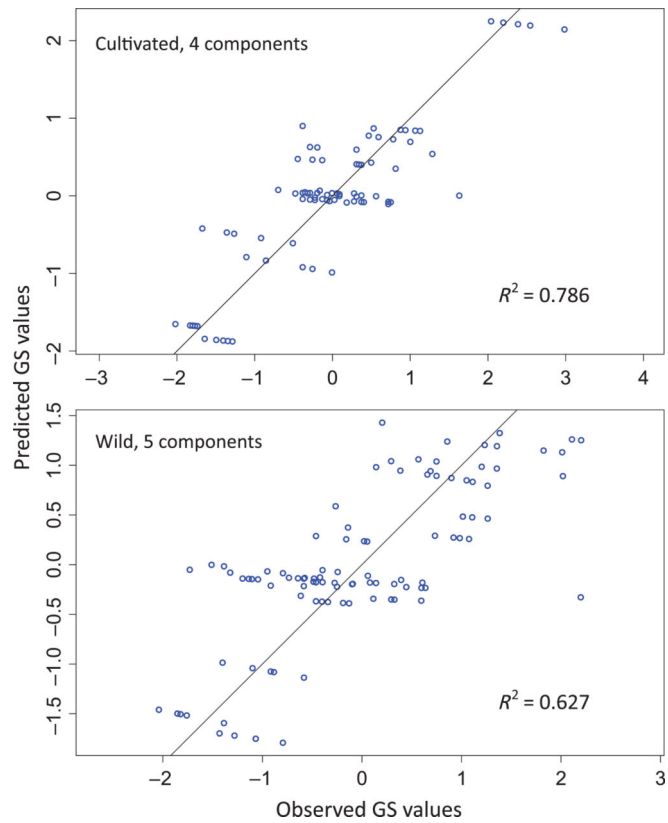


**Fig. 2.**

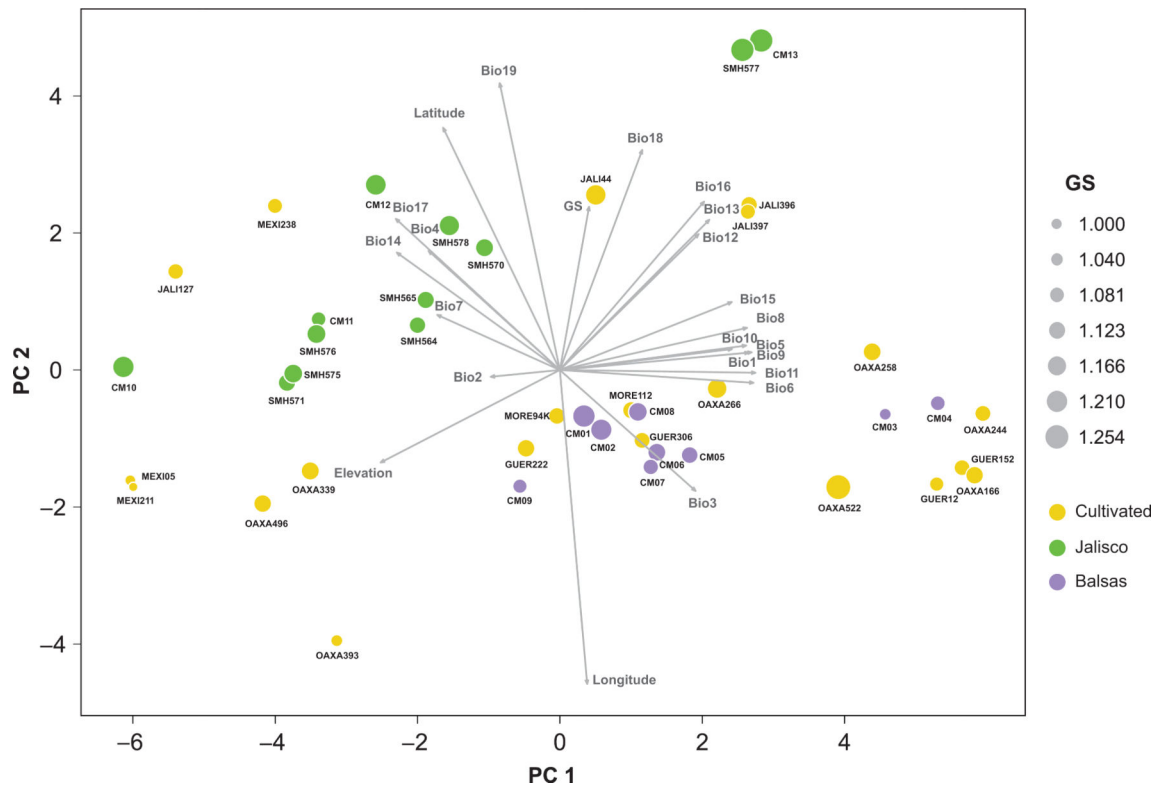
Genome size (GS) estimates for each maize landrace and each teosinte population sample measured either as a ratio over the B73 maize reference genome or in picograms per 2C nucleus (pg/2C). Populations are arranged by mean GS values. Cultivated, yellow boxes; wild, blue boxes. The boxes indicate the first quartile (leftmost line), the median (central line) and the third quartile (rightmost line). The whiskers represent the standard deviation, with the outliers represented by dots.



**Fig. 3.** Scatter plot and quantile regression showing the relationship between genome size (GS) and elevation for the cultivated maize landraces and wild teosinte groups of samples. The dashed gray lines correspond to the quantiles (0.15, 0.30, 0.45, 0.60, 0.75, 0.90), the median fit is in solid blue and the least-squares estimate of the conditional mean function is the dashed (red) line.



**Fig. 4.** Comparison between the observed and predicted genome size (GS) for maize (cultivated) and teosintes (wild), based on the linear model inferred using partial least-squares regression (PLSR). The  $R^2$  of the Spearman correlation for both fits is also provided.



**Fig. 5.** Principal component analysis (PCA) including the 22 cultivated maize and 21 wild populations of teosinte included in this study. The gray arrows represent the different variables considered in the model. PC1 and PC2 explain 12.7% and 3.4%, respectively, of the variance.



Table 1

Information about maize (Cultivated) and teosinte (Wild) samples included in this study, as well as the sampling locations

Name	Status	Bank accession	Common name	Latitude	Longitude	Elevation (m)	Place	State	Gradient	GS <sup>2</sup>	SD <sup>2</sup>
GUER152	Cultivated	19 406	Breve prieto	17.017	-99.533	300	Tierra Colorada	Guerrero	Balsas	1.08	0.016
GUER17	Cultivated	230	Conejo	16.95	-99.83	99	El Treinta	Guerrero	Balsas	1.054	0.032
GUER222	Cultivated	2037	Maiz ancho	18.4	-99.31	1330	Cacahuamilpa	Guerrero	Balsas	1.114	0.028
GUER306	Cultivated	22 287	Rojo itzihuini	18.133	-99.1	626	Atenango del Rio	Guerrero	Balsas	1.084	0.025
JALI127	Cultivated	14 052	Rojo sol	19.933	-103.733	1800	Tapalpa	Jalisco	Jalisco	1.078	0.029
JALI396	Cultivated	22 270	Tempranillo	19.94	-105.2	50	Tomatlan	Jalisco	Jalisco	1.092	0.029
JALI397	Cultivated	22 283	Cristalino tempranillo	19.94	-105.2	50	Tomatlan	Jalisco	Jalisco	1.065	0.026
JALI44	Cultivated	2208	Ahumado	19.48	-104.16	1402	Los Tocotes	Jalisco	Jalisco	1.172	0.02
MEX105	Cultivated	2233	Palomero toluqueño	19.283	-99.633	3073	Metepc	Mexico	Balsas	0.99	0.01
MEX1211	Cultivated	1940	Palomero legitimo	19.283	-99.65	2620	Toluca	Mexico	Balsas	0.962	0.008
MEX1238	Cultivated	2074	Maiz de Tehuacan	18.8	-99.917	1800	Llano Grande	Mexico	Balsas	1.068	0.018
MORE112	Cultivated	2050	Ancho	18.617	-99.35	970	San Gabriel de las Palmas	Morelos	Balsas	1.107	0.057
MORE94K	Cultivated	2025	Delgado	18.81	-99.12	1340	Tetlama	Morelos	Balsas	1.097	0.027
OAXA166	Cultivated	100	Costeño	16.283	-98.267	50	El Ciruelo	Oaxaca	Oaxaca	1.11	0.044
OAXA244	Cultivated	5847	Tepecintle amarillo	16.51	-98.12	250	Costatitlan (El Limón)	Oaxaca	Oaxaca	1.084	0.01
OAXA258	Cultivated	5849	Olotillo de Chaguas	16.81	-97.99	520	El Tapanco	Oaxaca	Oaxaca	1.121	0.015
OAXA266	Cultivated	5850	Tehuacanero azul	17.033	-97.95	850	Santa María Putla	Oaxaca	Oaxaca	1.154	0.016
OAXA339	Cultivated	6075	Taan gacia	17.13	-97.5	1600	San Juan Copala	Oaxaca	Oaxaca	1.125	0.037
OAXA393	Cultivated	6100	Pinto	16.967	-96.483	3000	Cuajimoloyas	Oaxaca	Oaxaca	1.01	0.031
OAXA496	Cultivated	17 443	De cajete	17.467	-97.567	2200	Yodomano	Oaxaca	Oaxaca	1.115	0.018
OAXA522	Cultivated	22 024	Orote colorado	16.433	-95.017	50	Col. Albaro Obregón	Oaxaca	Oaxaca	1.261	0.026
YUCA117	Cultivated	843	Dzit bacal	21.01	-89.54	10	Mérida	Yucatan	-	1.155	0.017
CM01	Wild	-	-	18.35	-99.841	1649	Teloloapan	Guerrero	Balsas	1.21	0.039
CM02	Wild	-	-	18.411	-99.908	1439	Aichofoa	Guerrero	Balsas	1.186	0.042
CM03	Wild	-	-	17.172	-99.541	343	Tierra Colorada	Guerrero	Balsas	1.008	0.016
CM04	Wild	-	-	16.981	-99.285	581	Teconapa	Guerrero	Balsas	1.062	0.021
CM05	Wild	-	-	17.392	-99.478	1201	Chilpancingo	Guerrero	Balsas	1.098	0.102
CM06	Wild	-	-	17.422	-99.451	1251	Mazatlán	Guerrero	Balsas	1.121	0.031

Name	Status <sup>1</sup>	Bank accession	Common name	Latitude	Longitude	Elevation (m)	Place	State	Gradient	G <sub>S</sub> <sup>2</sup>	SD <sup>2</sup>
CM07	Wild	–	–	17.46	–99.368	1107	Mochitlán	Guerrero	Balsas	1.075	0.051
CM08	Wild	–	–	18.238	–99.218	1101	Paso Morelos	Guerrero	Balsas	1.135	0.029
CM09	Wild	–	–	18.974	–99.07	1669	Huilotepic	Morelos	Balsas	1.052	0.017
CM10	Wild	–	–	19.407	–99.627	2581	Villa Seca	Mexico	Balsas	1.19	0.027
CM11	Wild	–	–	20.139	–102.068	1846	Churintzio	Michoacán	Jalisco	1.06	0.03
CM12	Wild	–	–	20.627	–104.408	1426	Guachimango	Jalisco	Jalisco	1.178	0.034
CM13	Wild	–	–	19.715	–104.804	504	Talpitita	Jalisco	Jalisco	1.229	0.032
SMH564	Wild	–	–	19.91	–104.173	1407	Ejutla	Jalisco	Jalisco	1.094	0.019
SMH565	Wild	–	–	19.896	–104.177	1317	Ejutla	Jalisco	Jalisco	1.108	0.01
SMH570	Wild	–	–	19.934	–104.008	944	La Labor	Jalisco	Jalisco	1.123	0.015
SMH571	Wild	–	–	19.992	–100.885	1861	San Jose de las Pilas	Guanajuato	Jalisco	1.108	0.033
SMH575	Wild	–	–	20.053	–101.088	1849	San Rafael	Michoacán	Jalisco	1.139	0.031
SMH576	Wild	–	–	20.161	–101.373	1856	Tejocote de Calera	Guanajuato	Jalisco	1.142	0.031
SMH577	Wild	–	–	19.732	–104.872	572	El Llamito	Jalisco	Jalisco	1.229	0.055
SMH578	Wild	–	–	19.535	–104.058	1369	El Rodeo	Jalisco	Jalisco	1.168	0.021

<sup>1</sup> All cultivated materials were provided by Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT); teosinte samples were sampled *in situ* by, and are available from, the authors.

<sup>2</sup> G<sub>S</sub>, average genome size among five plants and three replicates per plant, with standard deviation (SD).

**Table 2**

Nonparametric multivariate analysis of variance (PERMANOVA) for teosinte and maize landrace samples

Category	df	SS	MS	F	R <sup>2</sup>	P (> F)
Teosinte						
Populations						
Pop	20	1.1067	0.0553	252.752	0.7320	0.001
Pop: Plant	84	0.3593	0.0043	19.536	0.2376	0.001
Residuals	210	0.0460	0.0002		0.0304	
Total	314	1.5120			1	
Gradients						
Gradient	1	0.0699	0.0699	50.720	0.0462	0.001
Gradient: Pop	19	1.0368	0.0546	39.588	0.6857	0.001
Residuals	294	0.4053	0.0014		0.2680	
Total	314	1.5120			1	
Landraces						
Landraces						
Pop	21	1.2491	0.0595	335.670	0.8439	0.001
Pop : Plant	88	0.1920	0.0022	12.310	0.1297	0.001
Residuals	220	0.0390	0.0002		0.0263	
Total	329	1.4800			1	
Gradients						
Gradient	1	0.2348	0.2348	299.045	0.1852	0.001
Gradient: Pop	15	0.8461	0.0564	71.842	0.6674	0.001
Residuals	238	0.1869	0.0008		0.1474	
Total	254	1.2677			1	

**Table 3**

Environmental variables, their Spearman correlation with genome size (GS) and their coefficients in the linear model inferred by partial least-squares regression (PLSR; results are shown for gradients in maize landraces and teosintes separately)

Environmental variable <sup>1</sup>	Landraces				Teosintes			
	<i>R</i> (All) <sup>2</sup>	<i>R</i> (Balsas)	<i>R</i> (Oaxaca)	PLSR (All) <sup>3</sup>	<i>R</i> (All)	<i>R</i> (Balsas)	<i>R</i> (Jalisco)	PLSR (All)
Elevation	-0.381***	-0.406**	-0.391*	-0.462**	0.091	0.542***	-0.434***	0.250*
Latitude	-0.133	-0.405**	-0.033	-0.187**	0.241*	0.536***	-0.307*	0.160
Longitude	0.387***	0.505***	0.240	0.515***	-0.443***	-0.476***	-0.515***	0.195
BIO1 = Annual Mean Temperature	0.235*	0.430**	0.035	0.051*	-0.127	-0.541***	0.414**	0.015
BIO2 = Mean Diurnal Range (mean of monthly (max temp – min temp))	-0.012	0.103	0.059	0.182	-0.151	-0.326*	-0.466***	-0.474***
BIO3 = Isothermality (P2/P7)(× 100)	-0.028	0.104	-0.308	-0.053	-0.238*	-0.764***	0.568***	-0.593***
BIO4 = Temperature Seasonality (standard deviation × 100)	-0.147	0.073	0.024	-0.068	0.284**	0.652***	-0.355**	0.186**
BIO5 = Max Temperature of Warmest Month	0.250**	0.558***	-0.070	0.113**	-0.098	-0.345*	0.162	0.112
BIO6 = Min Temperature of Coldest Month	0.297**	0.400**	0.262	0.032	-0.100	-0.471***	0.497***	0.137
BIO7 = Temperature Annual Range	-0.087	0.084	-0.089	0.117	0.164	0.428**	-0.486***	-0.108
BIO8 = Mean Temperature of Wettest Quarter	0.228*	0.406**	0.035	0.030	-0.120	-0.542***	0.363**	-0.002
BIO9 = Mean Temperature of Driest Quarter	0.253**	0.465**	0.035	0.076*	-0.106	-0.397**	0.345**	-0.209
BIO10 = Mean Temperature of Warmest Quarter	0.228*	0.430**	0.164	0.053	-0.120	-0.493***	0.326*	0.081
BIO11 = Mean Temperature of Coldest Quarter	0.253**	0.430**	0.035	0.061**	-0.103	-0.537***	0.476***	0.043
BIO12 = Annual Precipitation	0.253**	0.400**	0.135	-0.055	0.071	-0.338*	0.476***	0.211*
BIO13 = Precipitation of Wettest Month	0.325***	0.395**	0.199	0.059	0.014	-0.406**	0.607***	-0.033
BIO14 = Precipitation of Driest Month	-0.052	-0.366*	0.262	0.328**	0.194*	0.497***	-0.738***	-0.263
BIO15 = Precipitation Seasonality	0.203*	0.446**	0.369*	0.274***	-0.034	-0.278	0.454***	-0.121
BIO16 = Precipitation of Wettest Quarter	0.248**	0.395**	0.199	-0.023	0.065	-0.425**	0.509***	0.212***
BIO17 = Precipitation of Driest Quarter	0.016	-0.311*	0.313*	0.318**	0.357***	0.498***	-0.068	0.400*
BIO18 = Precipitation of Warmest Quarter	0.247**	0.063	-0.156	-0.328*	0.105	-0.509***	0.589***	0.262

Environmental variable <sup>1</sup>	Landraces				Teosintes			
	<i>R</i> (All) <sup>2</sup>	<i>R</i> (Balsas)	<i>R</i> (Oaxaca)	PLSR (All) <sup>3</sup>	<i>R</i> (All)	<i>R</i> (Balsas)	<i>R</i> (Jalisco)	PLSR (All)
BIO19 = Precipitation of Coldest Quarter	0.014	-0.174	-0.042	0.100	0.326***	0.126	0.618***	-0.023

<sup>1</sup> Bioclimatic variables defined as: BIO1, Annual Mean Temperature; BIO2, Mean Diurnal Range (mean of monthly (max temp – min temp)); BIO3, Isothermality (BIO2/BIO7) ( $\times 100$ ); BIO4, Temperature Seasonality (standard deviation  $\times 100$ ); BIO5, Max Temperature of Warmest Month; BIO6, Min Temperature of Coldest Month; BIO7, Temperature Annual Range (BIO5–BIO6); BIO8, Mean Temperature of Wettest Quarter; BIO9, Mean Temperature of Driest Quarter; BIO10, Mean Temperature of Warmest Quarter; BIO11, Mean Temperature of Coldest Quarter; BIO12, Annual Precipitation; BIO13, Precipitation of Wettest Month; BIO14, Precipitation of Driest Month; BIO15, Precipitation Seasonality (coefficient of variation); BIO16, Precipitation of Wettest Quarter; BIO17, Precipitation of Driest Quarter; BIO18, Precipitation of Warmest Quarter; BIO19, Precipitation of Coldest Quarter.

<sup>2</sup> Spearman correlation coefficient *R* for (in parentheses) the entire sample (All) or a particular gradient (Balsas, Oaxaca, Jalisco). Significance values are denoted by asterisks: \*,  $P < 0.05$ ; \*\*,  $P < 0.01$ ; \*\*\*,  $P < 0.001$ .

<sup>3</sup> The PLSR coefficient based on the total sample. Significance levels are denoted by asterisks: \*,  $P < 0.01$ ; \*\*,  $P < 0.001$ ; \*\*\*,  $P < 0.0001$ .