The interplay of demography and selection during maize domestication and expansion

Li Wang[1,2], Timothy M. Beissinger[3,5,6], Anne Lorant[3], Claudia Ross-Ibarra[3], Jeffrey Ross-Ibarra[3,4] and Matthew B. Hufford[1]

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA
[2]Genome Informatics Facility, Iowa State University, Ames, IA, USA
[3]Department of Plant Sciences, University of California Davis, Davis, CA, USA
[4]Center for Population Biology and Genome Center, University of California Davis, Davis, CA, USA
[5]USDA-ARS Plant Genetics Research Unit, Columbia, MO, USA
[6]Divisions of Plant and Biological Sciences, University of Missouri, Columbia, MO, USA

Abstract

The history of maize has been characterized by major demographic events including changes in population size associated with domestication and range expansion as well as gene flow with wild relatives. The interplay between demographic history and selection has shaped diversity across maize populations and genomes. Here, we investigate these processes based on high-depth resequencing data from 31 maize landraces spanning the pre-Columbian distribution of maize as well as four wild progenitor individuals (*Zea mays* ssp. *parviglumis*) from the Balsas River Valley in Mexico. Genome-wide demographic analyses reveal that maize domestication and spread resulted in pronounced declines in effective population size due to both a protracted bottleneck and serial founder effects, while, concurrently, *parviglumis* experienced population growth. The cost of maize domestication and spread was an increase in deleterious alleles in the domesticate relative to its wild progenitor. This cost is particularly pronounced in Andean maize, which appears to have experienced a more dramatic founder event when compared to other maize populations. Introgression from the wild teosinte *Zea mays* ssp. *mexicana* into maize in the highlands of Mexico and Guatemala is found found to decrease the prevalence of deleterious alleles, likely due to the higher long-term effective population size of wild maize. These findings underscore the strong interaction between historical demography and the efficiency of selection species- and genome-wide and suggest domesticated species with well-characterized histories may be particularly useful for understanding this interplay.

1 Introduction

Genomes are shaped over the course of their evolutionary history through a complex interaction of demography and selection. Neutral processes that comprise a species' demographic history, such as stochastic changes in population size and migration events, influence both the pool of diversity upon which selection can act and its efficiency. Selection and genetic drift then jointly determine the fate of this diversity.

After the development of agriculture, both crops and humans have experienced profound demographic shifts that left clear signatures in genome-wide patterns of diversity [1, 2]. Early agriculturalists sampled a subset of the diversity present in crop wild relatives, resulting in an initial demographic bottleneck for many domesticates [3]. Subsequent to domestication, humans and their crops experienced a process of global expansion facilitated by the rise of agriculture [4]. In many cases expansion was accompanied by gene flow with close relatives, a demographic process that further altered patterns of diversity [5, 6].

Recent interest in the effects of demography on functional variation has led to a growing body of theory that is increasingly supported by empirical examples. To date, the relationship between demography and selection has been most thoroughly explored in the context of deleterious alleles. While theory suggests mutation load may be insensitive to demography over long periods [7, 8], empirical results are consistent with load being shaped by demography over shorter timescales [9, 10, 11, 12, 13]. For example, evidence in both plant and animal species has revealed increased mutation load in populations that have undergone recent, sudden declines in effective population size (N_e) [10, 11, 12, 14] Similarly, in geographically expanding populations, repeated sub-sampling of diversity (*i.e.*, serial founder effects) can occur during migration away from a center of origin [15, 16], a phenomenon shown to have decreased genetic diversity and increased counts of deleterious alleles in human populations more distant from Africa [17, 18]. Finally, gene flow may also affect genome-wide patterns of deleterious variants, particularly when occurring between populations with starkly contrasting N_e. For instance, during the Out-of-Africa migration, modern humans inter-mated with Neanderthals, a close relative with substantially lower N_e and higher mutation load [9]. The higher mutation load in Neanderthals presented a cost of gene flow, and subsequent purifying selection appears to have limited the amount of Neanderthal introgression near genes in the modern human genome [9, 19].

The domesticated plant maize (*Zea mays* ssp. *mays*) has a history of profound demographic shifts accompanied by selection for agronomic performance and adaptation to novel environments, making it an ideal system in which to study the interaction between demography and selection. Maize was domesticated in a narrow region of southwest Mexico from the wild plant teosinte (*Zea mays* ssp. *parviglumis*; [20, 21, 22]) and experienced an associated genetic bottleneck that removed a substantial proportion of the diversity found in its progenitor [23, 24]. Archaeological evidence suggests that after initial domestication, maize spread across the Americas, reaching the southwestern US by approximately 4,500 BP [25] and coastal South America as early as 6,700 BP [26]. Gene flow into maize from multiple teosinte species has been documented in geographical regions outside of its center of origin [5, 27]. To date, genetic studies of demography and selection in maize have primarily focused on initial domestication [28], only broadly considering the effects of subsequent population size change on diversity [2] and largely disregarding the spatial effects of geographic expansion and gene flow (but see [29]). Fur-

thermore, the effect of maize demography on the prevalence of deleterious alleles has yet to receive in-depth attention.

Here, we investigate the genome-wide effects of demographic change in maize during domestication and subsequent expansion using high-depth resequencing data from a panel of maize landraces. We present evidence for a protracted domestication bottleneck, further loss of diversity during crop expansion, and gene flow between maize and its wild relatives outside of its center of origin. We then explore how this demographic history has shaped genome-wide patterns of deleterious alleles.

2 Results

Maize population size change during domestication and expansion

We resequenced 31 open-pollinated maize landraces representing six geographical regions that span the pre-Colombian range of maize cultivation (Figure 1) as well as four wild *parviglumis* individuals from a single population located in the Balsas River Valley in Mexico. Median sequencing depth was 29X, with a range of 24-53X. Landrace accessions were selected to broadly reflect the diversity of maize in the Americas and to be representative of defined ecogeographic regions based on consultation with experts on landrace germplasm (Major Goodman, personal communication) and on descriptions in the "Races of Maize" handbooks [30].

We first estimated historical changes in effective population size (N_e) of maize and *parviglumis* using the multiple sequentially Markovian coalescent (MSMC) [31]. Consistent with archaeological evidence [21], we find that the demographic histories of the various maize populations begin to diverge from one another approximately 10,000 years before present (BP; Figure 1B). Surprisingly, our single population of *parviglumis* diverges from maize much earlier, around $75,000$ generations BP. All maize populations show a gradual decline in diversity concomitant with divergence from *parviglumis*, but the slope becomes more pronounced around the time of domestication. This period of declining N_e continues until the recent past ($\approx 1,100 - 2,400$ generations BP) and is followed by extremely rapid population growth, suggesting recovery from domestication post-dated expansion of maize across the Americas. In contrast to our results in maize, *parviglumis* shows an increase in N_e which also lasts until the recent past ($\approx 1,200 - 1,800$ generations BP). To determine if linked selection associated with domestication could bias estimates of N_e in maize (see [32]) we masked previously identified domestication candidates [24] and observed nearly identical results (Figure S1A).

One explanation for the prolonged population size reduction in maize following the onset of domestication would be repeated colonization bottlenecks during spread across the Americas. Genome-wide levels of heterozygosity across our maize samples are consistent with this idea, showing a strong negative correlation ($R^2 = 0.3636, p = 0.0004$; Figure 1C) with distance from the center of maize domestication in the Balsas River Basin. To confirm this trend, we performed a similar analysis with a much larger sample of published genotyping data ($n = 3520$; Figure S1) [33] and observed similar results.

While the gradual decrease in genetic diversity seen with distance from the Balsas indicates serial founder effects, our analyses also point to a more extreme founder event in the Andean highlands of South America. Andean landraces show a deeper bottle-

neck in our MSMC analysis (Figure 1), have the lowest overall diversity (Figure S2), and show both a distinct reduction of low frequency alleles and a greater proportion of derived homozygous alleles compared to other populations (Figure S2). To shed light on the timing of this extreme founder event, we assessed evidence for recent inbreeding. Inbreeding coefficients in Andean samples were quite low and not statistically different from other populations (all $F < 0.002$ and $p > 0.05$ based on a Wilcoxon test). Likewise, no significant differences could be found across populations in the number of runs of homozygosity (ROH) longer than $1cM$ ($p > 0.05$ in all cases, Wilcoxon test), further suggesting a lack of recent inbreeding. However, when ROH were limited to those shorter than $0.5cM$, a length of homozygosity that would be associated with a more ancient founder event, Andean samples demonstrated greater cumulative ROH genome-wide compared to all ($p < 0.05$, Wilcoxon test) but the South American lowland population ($p = 0.214$, Wilcoxon test). Together, these lines of evidence are consistent with an unusually strong founder event during colonization of the Andes.

Introgression from wild maize in highland populations

Adaptive introgression from the wild teosinte taxon *Zea mays* ssp. *mexicana* (hereafter, *mexicana*) has previously been observed in maize in the highlands of Mexico [5]. Our broad sampling allowed us to investigate whether introgressed *mexicana* haplotypes have spread to highland maize populations outside of Mexico, potentially playing a role in adaptation in other regions. In order to test this hypothesis, we calculated Patterson's D statistic [34] across all maize populations. All individuals from both the Mexican and Guatemalan highlands exhibited highly significant evidence for shared ancestry with *mexicana* (Figure S5A). Maize from the southwestern US also showed more limited evidence of introgression, consistent with findings from ancient DNA suggesting this region was originally colonized by admixed maize from the highlands of Mexico [35]. In contrast, the distribution of z-scores for South American populations overlapped zero, providing no evidence for spread of *mexicana* haplotypes to this region.

We localized introgression to chromosomal regions through genome-wide calculation of the $\hat{f}_d$ statistic [36]. Megabase-scale regions of introgression were identified in both Mexican and Guatemalan highland populations that correspond to those reported by [5] on chromosomes 4 and 6 (Figure 2; Figure S5). On chromosome 3, a large, previously unidentified region of introgression can be found in the Mexican and southwestern US highlands (Figure 2; Figure S5). This region overlaps a putative chromosomal inversion associated with flowering time in maize landraces [37] and in the maize nested association mapping population [38] and may be an example of *mexicana* contribution to modern maize lines.

The influence of demography on accumulation of deleterious alleles

Population-specific changes in historical N_e should influence the efficiency of purifying selection and alter genome-wide patterns of deleterious variants [10]. Introgression from a species with substantially different N_e may also influence the abundance and distribution of deleterious alleles in the genome [9]. Below we evaluate the effects of major demographic events during the pre-Colombian history of maize on patterns of deleterious alleles.

Domestication and deleterious alleles

We first compared counts of deleterious alleles in Mexican lowland maize individuals to four *parviglumis* individuals from a single population in the Balsas River Valley. Maize from the Mexican lowlands has not experienced substantial introgression from wild relatives and is near the center of maize origin [22], and thus best reflects the effects of domestication alone. After identifying putatively deleterious mutations using Genomic Evolutionary Rate Profiling (GERP) [39], we calculated the number of derived deleterious alleles per genome under both an additive and a recessive model across four levels of mutation severity (see Methods for details). Maize showed significantly more deleterious alleles than teosinte under both additive ($< 10\%$ more; $p = 0.0079$, Wilcoxon test; Figure S6) and recessive ($< 20 - 30\%$ more; $p = 0.0079$; Figure 3) models across all categories (Figure S7). Additionally, maize contained 81% more fixed deleterious alleles than teosinte ($48,890$ vs. $26,947$) and 3% fewer segregating deleterious alleles ($464,653$ vs. $478,594$), effects expected under a domestication bottleneck (Figure 3; [7]). GERP load (GERP score $\times$ frequency of deleterious alleles), a more direct proxy of the individual-level burden of deleterious mutations, revealed a similar trend (additive model: maize median = 23.635, teosinte median = 22.791, $p = 0.008$, Wilcoxon test; recessive model: maize median = 14.922, teosinte median = 12.231, $p = 0.008$). Maize, like other domesticates [12, 14, 40, 41], thus appears to have a higher burden of deleterious alleles compared to its wild progenitor *parviglumis*.

While the elevated burden we observe in maize relative to *parviglumis* may be driven primarily by the domestication bottleneck, positive selection on causal variants underlying domestication phenotypes may also fix nearby deleterious variants through genetic hitchhiking, which would result in a higher number of deleterious alleles in regions linked to domestication loci [40, 42]. To test this hypothesis, we first confirmed that 420 previously identified domestication candidates [24] showed evidence of selection in our data (Figure S8), and then assessed the distribution of deleterious alleles in and near ($5kb$ upstream and downstream) these genes by calculating the number of deleterious alleles per base pair under both recessive and additive models. No significant difference was found in the prevalence of deleterious alleles near domestication and random sets of genes (Figure S9), suggesting the increased mutation burden of deleterious alleles in maize has been driven primarily by the genome-wide effects of the domestication bottleneck rather than linkage associated with selection on specific genes.

The effect of the Andean founder event on deleterious alleles.

The extreme founder event observed in the Andes could potentially alter genome-wide patterns of deleterious variants beyond the effects of domestication alone. Under a recessive model, maize from the Andes contains significantly more deleterious alleles than any other population (Figure 3; Figure S7; all p values < 0.02, Wilcoxon test), and this difference becomes more extreme when considering more severe (*i.e.*, higher GERP scores) mutations (Figure S7). In contrast, we observe no significant differences under an additive model (Figure S6; Figure S7). The Andean founder event therefore appears to have resulted in a higher mutation burden of deleterious alleles than seen in other populations of maize. This result is further supported by a higher proportion of fixed deleterious alleles within the Andes and fewer segregating deleterious alleles (Figure S10; Figure 3), a result comparable to the differences observed in load between maize and *parviglumis*.

Introgression decreases the prevalence of deleterious alleles.

Highly variable rates of *mexicana* introgression were detected across our landrace populations (Figure 2; Figure S4; Figure S5). To explore the potential effects of introgression on the genomic distribution of deleterious alleles, we fit a linear model in which the number of deleterious sites is predicted by introgression (represented by $\hat{f}_d$) and gene density (exonic base pairs per cM) in 10kb non-overlapping windows in the Mexican highland population where we found the strongest evidence for *mexicana* introgression. Gene density was included as a predictor in the regression to control for the positive correlation observed between gene density and both introgression ($p = 3.48e-08$) and deleterious alleles ($p \approx 0$). When identifying deleterious alleles under both additive and recessive models, we found a strong negative correlation with introgression (*i.e.,* fewer deleterious alleles in introgressed regions; $p \approx 0$ under both models). These findings likely reflect the larger ancestral N_e and more efficient purifying selection in *mexicana*.

3 Discussion

Demographic studies in domesticated species have focused largely on identifying progenitor population(s) and quantifying the effect of the domestication bottleneck on genetic diversity [24, 43, 44]. It is likely, however, that the demographic history of domesticates is generally more complex than a simple bottleneck followed by recovery [45, 46]. Many crops and domesticated animals have expanded from defined centers of origin to global distributions, experiencing population size changes and gene flow from closely related taxa throughout their histories [47]. With this in mind, we have characterized maize demography from domestication through initial expansion in order to provide a more complete assessment of the influence of demography on deleterious variants.

Historical changes in maize population size

Early models of maize demography suggested the ratio of the domestication bottleneck size and duration was between $\approx 2.5:1$ and $\approx 5:1$, but little statistical support was found for specific estimates of these individual parameters [23, 28]. Most recently, Beissinger *et al.* [2] fit a model assuming a bottleneck followed by instantaneous exponential recovery. While our results concur with the most recent model in finding a similar bottleneck size ($\approx 10\%$ compared to $\approx 5\%$ in Beissinger *et al.*) and that the modern N_e of maize is quite large, the flexibility of MSMC also allowed us to estimate the duration of the bottleneck. We find that the domestication bottleneck may have lasted much longer than previously believed, spanning $\approx 9,000$ generations and only beginning to recover in the recent past (Figure 1B). Analysis of bottlenecks in domesticated African rice and grapes have also suggested a duration of several thousand generations [45, 46], indicating that demographic bottlenecks during crop evolution may have generally occurred over substantial periods of time.

In addition to a strong bottleneck during domestication, our finding that levels of diversity decline in populations increasingly distant from the center of maize domestication are suggestive of serial founder effects during the spread of maize across the Americas (Figure 1C; Figure S1). Serial founder effects are the result of multiple sampling events during which small founder populations are repeatedly drawn from ancestral pools, leading to a stepwise increase in genetic drift and a concomitant decrease in ge-

netic diversity. During maize range expansion, serial founder effects would have occurred if seed carried to each successive colonized region was limited to a sample of whole ears that contained a fraction of the diversity found in the source population [29]. Consistent with serial founder effects, other researchers have found a correlation between geographic and genetic distance in maize landraces [22, 48], though this was previously attributed to limited seed dispersal across the species range leading to isolation by distance (IBD). Neutral expectations of allele frequencies across populations under serial founder effects differ substantially from those predicted under equilibrium conditions [15]. Theory has been developed that predicts the probability that an allele from an origin survives a series of founder effects and reaches high frequency once an expansion is complete [15]. Many of the world's crops have experienced such histories of expansion and may be most correctly modeled within this theoretical framework. Studies attempting to identify loci underlying crop adaptation during post-domestication expansion may therefore more accurately compare empirical data to neutral expectations under a serial founder effects demography [15] in order to identify extreme changes in allele frequency driven by selection.

While a history of serial founder effects partially explains the variation in diversity across maize landraces, there are deviations from this model. For example, our combined results showing increased runs of homozygosity (Figure S3), lower nucleotide diversity (Figure S2), and smaller effective population size (Figure 1) in the Andes all suggest a pronounced, ancient founder event and are in agreement with previous work modeling demography in this region [49]. The founder event in the Andes may reflect initially limited cultivation due to the poor performance of maize in this region relative to established root and tuber staples [50]; maize cultivation may have only become widespread after an initial lag period necessary for adaptation. Additionally, we observe somewhat higher than expected nucleotide diversity in maize landraces from the highlands of Mexico and Guatemala (Figure 1C), which may be linked to the introgression we have detected from *mexicana*.

In striking contrast to the bottleneck we observe in maize, the effective population size in *parviglumis* increases steadily from the time of initial maize domestication until the recent past. Multiple population genetic studies have reported negative genome-wide values of Tajima's D in *parviglumis* from the Balsas River Valley [2, 23, 51], findings characteristic of an expanding population. Likewise, analyses of pollen content in sediment cores from Mexico suggest herbaceous vegetation and grasslands have expanded over the last 10,000 years due to changing environmental conditions during the Holocene and human management of vegetation with fire [52, 53]. While our *parviglumis* samples are drawn from a single population in the Balsas, these data collectively suggest *parviglumis* from this region has experienced expansion over the last several millennia.

Consistent with archaeological evidence of the timing of initial maize domestication [21], we find that maize demographies begin to diverge $\approx 10,000$ generations before present, a time that appears to coincide with a steeper decline in maize N_e as well. In contrast, we estimate the timing of the split between maize and our single population of *parviglumis* to be $\approx 75,000$ generations before present, likely reflecting strong population structure in *parviglumis* and divergence of our sampled *parviglumis* from the maize progenitor population. Beissinger *et al.* [2], using samples from additional populations, also find an estimate of maize-*parviglumis* divergence older than the probable onset of domestication, suggesting that currently available sequences of *parviglumis* may not sample well from the populations directly ancestral to domesticated maize.

The prevalence of gene flow during maize diffusion

Increasingly, range-wide analyses of crops and their wild relatives have identified evidence of gene flow during post-domestication expansion from newly sympatric populations of their progenitor taxa and closely related species [54, 55, 56]. Consistent with previous results from genotyping data [5, 22, 57], we find strong support for introgression from *mexicana* to maize in the highlands of Mexico. While *mexicana* is not currently found in the highlands of Guatemala, we also find strong evidence for *mexicana* introgression in maize from this region, suggesting either *mexicana* was at one time more broadly distributed, or, perhaps more likely, that highland maize from Mexico was introduced to the Guatemalan highlands. Support is also found for *mexicana* introgression in the southwestern US at specific chromosomal regions such as a putative inversion polymorphism on chromosome 3 (Figure 2). These results confirm previous findings suggesting maize from the highlands of Mexico originally colonized the southwestern US [35]. The more limited signal of *mexicana* introgression here may be due to subsequent gene flow from lowland maize as suggested by [35]. Very little evidence is found for *mexicana* haplotypes extending into South America, as highland-adapted haplotypes would likely have been maladaptive and removed by selection as maize traversed the lowland regions of Central America ([49]).

Impacts of demography on accumulation of deleterious variants

The availability of high-density SNP data from range-wide samples of a species allows for an in-depth assessment of the influence of demography on the prevalence of deleterious alleles. For example, recent studies in both humans and dogs have revealed that historical changes in population size [10, 12, 18] and introgression [9] may contribute substantially to variation in patterns of deleterious variants among populations and across the genome. Previous work in maize has characterized genome-wide trends in deleterious alleles across modern inbred maize lines, revealing that inbreeding during the formation of modern lines has likely purged many recessive deleterious variants [58] and that complementation of deleterious alleles likely underlies the heterosis observed in hybrid breeding programs [58, 59]. Additionally, [2] revealed that purifying selection has removed a greater extent of pairwise diversity (θ_π) near genes in *parviglumis* than in maize due to the higher historical N_e in *parviglumis*, but that this trend is reversed when considering younger alleles due to the recent dramatic expansion in maize population size. To date, however, few links have been made between the historical demography of maize domestication and expansion and the prevalence of deleterious alleles (but see [60] for a comparison of the frequencies of some coding changes). Our analysis reveals that demography has played a pivotal role in determining both the geographic and genomic landscapes of deleterious alleles in maize.

Population size and deleterious variants.

Previous studies have suggested a "cost of domestication" in which a higher burden of deleterious alleles is found in domesticates compared to their wild progenitors [12, 40, 42, 61, 62]. Consistent with these previous studies, we detect an excess of deleterious alleles in maize relative to *parviglumis* (Figure 3; Figure S6; Figure S7), which could be caused by two potential factors. First, reduced population size during the domestication bottleneck could result in deleterious alleles drifting to higher allele frequency. Second, hitchhiking caused by strong positive selection on domestication genes could cause linked

deleterious alleles to rise in frequency [12, 61]. While we find support for the former in maize, we see little evidence of the latter. In addition to the cost of domestication, we find a cost of geographic expansion that is likely driven by serial founder effects. The increase in deleterious alleles during expansion is most pronounced in the Andes and may be symptomatic of the extreme founder event we propose above.

Differences in the number of deleterious alleles between maize and *parviglumis* and non-Andean and Andean maize are more dramatic under a recessive model than an additive model. This trend may indicate that the bulk of deleterious alleles in maize are at least partially recessive, such that heterozygous sites contribute less to a reduction in individual fitness. Previous work in human populations has shown that the majority of deleterious mutations are recessive or partially recessive [63], and data from knock-out mutations in yeast have revealed that large-effect mutations tend to be more recessive [64]. Likewise, both theory and empirical evaluation across a number of organisms suggest that mildly deleterious mutations are likely to be partially recessive [65]. In maize, Yang *et al.* [58] have found that most deleterious alleles are at least partially recessive and note a negative correlation between the severity of a deleterious variant and its dominance. Our results thus match nicely both with previous empirical data and theoretical expectations showing that recent population bottlenecks should only show strong differences in load under a recessive model [7].

Introgression and deleterious variants.

Very few studies have investigated the effects of introgression from a taxon with substantially different N_e on the genomic landscape of deleterious variants. The best example is found in the human literature, where confirmation has been found that introgression from Neanderthals with low ancestral N_e increased the overall mutation load in modern humans [9, 19]. We report here the opposite pattern in maize, as introgression appears to have reduced the proportion of deleterious variants. Nonetheless, the underlying interpretation is similar: the introgressing taxon *mexicana* has had a larger ancestral N_e than maize [27], and introgressed haplotypes have thus experienced more efficient long-term purging of deleterious alleles.

4 Conclusions

We have demonstrated that demography during the domestication and expansion of maize across the Americas has profoundly influenced putative functional variation across populations and within individual genomes. More generally, we have learned that population size changes and gene flow from close relatives with contrasting effective population size will influence the distribution of deleterious alleles in species undergoing rapid shifts in demography. The significance of our results extends far beyond maize. For example, invasive species that have recently experienced founder events followed by expansion, endangered species subject to precipitous declines in N_e, species with a history of post-glacial expansion, and new species expanding their range will all likely show clear genetic signals of the interplay between demography and selection. This interaction bears importantly on the adaptive potential of both individual populations and species. By fully characterizing this relationship we can better understand how the current evolutionary trajectory of a species has been influenced by its history.

5 Materials and Methods

Samples, whole genome resequencing, and read mapping

A total of 31 maize landrace accessions were obtained from the US Department of Agriculture (USDA)'s National Plant Germplasm System and through collaborators (Figure S12). Samples were chosen from four highland populations (Andes, Mexican Highlands, Guatemalan Highlands and southwestern US Highlands) and two lowland populations (Mexican and South American lowlands) (Figure 1A). In addition, four open-pollinated *parviglumis* samples were selected from a single population in the Balsas River Valley in Mexico. DNA was extracted from leaves using a standard cetyltrimethyl ammonium bromide (CTAB) protocol [66]. Library preparation and Illumina HiSeq 2000 sequencing (100-bp paired-end) were conducted by BGI following their established protocols. BWA v. 0.7.5.a [67] was used to map reads to the maize B73 reference genome v3 [68] with default settings. The duplicate molecules in the realigned bam files were removed with MarkDuplicates in Picardtools v. 1.106 (http://broadinstitute.github.io/picard) and indels were realigned with GATK v. 3.3-0 [69]. Sites with mapping quality less than 30 and base quality less than 20 were removed and only uniquely mapped reads were included in downstream analyses.

Demography of maize domestication and diffusion

The recently developed method MSMC [31] was utilized to infer effective population size changes in both *parviglumis* and maize. SNPs were called via HaplotypeCaller and filtered via VariantFiltration in GATK [69] across all samples. SNPs with the following metrics were excluded from the analysis: QD $<$ 2.0; FS $>$ 60.0; MQ $<$ 40.0; MQRankSum $<$ -12.5; ReadPosRankSum $<$ -8.0. Vcftools v. 0.1.12 [70] was used to further filter SNPs to include only bi-allelic sites. SNPs were phased using BEAGLE v. 4.0 [71] with SNP data from the maize HapMap2 panel [60] used as a reference. Only sites with depth between half and twice of the mean depth were included in analyses. In addition, the software SNPable (http://lh3lh3.users.sourceforge.net/snpable.shtml) was used to mask genomic regions in which reads were not uniquely mapped. The mappability mask file for MSMC was generated by stepping in 1 *bp* increments across the maize genome to generate 100 *bp* single-end reads, which were then mapped back to the maize B73 reference genome [68]. Sites with the majority of overlapping 100-mers mapped uniquely without mismatch were determined to be "SNPable" sites and used for the MSMC analyses. For effective population size inference in MSMC, we used $5 \times 4 + 25 \times 2 + 5 \times 4$ as the pattern parameter and the value m was set as half of the heterozygosity in *parviglumis* and maize populations, respectively.

In order to explore the trend of genetic diversity away from the domestication center, the correlation between the percentage of polymorphic sites and the geographic distance to the Balsas River Valley (latitude: 18.099138; longitude: -100.243303) was examined via linear regression. Geographical distance in kilometers was calculated based on great circle distance using the haversine transformation [17].

Population structure, genetic diversity and inbreeding coefficients

We first evaluated population structure by principal components analysis (PCA) with ngsCovar [72] in ngsTools [73] based on the matrix of posterior probabilities of SNP

genotypes produced in ANGSD v. 0.614 [74], and then utilized NGSadmix v. 32 [75] to investigate the admixture proportion of each accession. The NGSadmix analysis was conducted based on genotype likelihoods for all individuals, which were generated with ANGSD (options -GL 2 -doGlf 2 -SNP_pval $1e-6$), and K from 2 to 10 was set to run the analysis for sites present in a minimum of 77 % of all individuals (24 in 31). A clear outlier in the Mexican Highland population was detected, RIMMA0677, a sample from relatively low altitude, which was suspected to contain a divergent haplotype. A neighbor joining tree of SNPs within an inversion polymorphism on chromosome 4 that includes a diagnostic highland haplotype [5] was constructed with the R package phangorn [76]. The sample RIMMA0677 was not clustered with other highland samples, but embedded within lowland haplotypes (Figure S11), so it was removed from further analyses.

The genetic diversity measures Watterson's θ and θ_π were calculated in ANGSD [74] for each population. The neutrality test statistic Tajima's D was calculated with an empirical Bayes approach [77] implemented in ANGSD by first estimating a global site frequency spectrum (SFS) then calculating posterior sample allele frequencies using the global SFS as a prior. The three statistics were summarized across the genome using 10-kb non-overlapping sliding windows.

Inbreeding coefficients for each individual were estimated with ngsF [78] with initial values of F_{IS} set to be uniform at 0.01 with an epsilon value of $1e-5$.

In addition, SNPs were polarized using the *T. dactyloides* genome to assess the frequency of derived homozygous sites in each maize landrace population. *T. dactyloides* short reads were downloaded from the NCBI SRA database (SRR447804 - SRR447807), mapped to the B73 reference genome v3 ([68]) with BWA [67] and incorporated into SNP calling as described above.

Runs of Homozygosity

SNPs were down-sampled to contain one SNP in a 2-*kb* window to identify segments representing homozygosity by descent (*i.e.*, autozygosity) rather than by chance. PLINK v. 1.07 (http://pngu.mgh.harvard.edu/purcell/plink/) [79] was applied to identify segments of ROH in a window containing 20 SNPs, among which the number of the maximum missing SNPs was set to 2 and the number of the maximum heterozygous sites was set to 1. The shortest length of final ROHs was set to be 300 *kb*. Physical distances were converted into genetic distances based on a recent genetic map [80].

Detection of Introgression

To assess per-genome evidence of population admixture between maize landraces and teosinte, we calculated the D statistic using ANGSD [74]. The statistic was calculated using trees of the form (((X, low),*mexicana*),*T. dactyloides*). One accession from the Mexican Lowland population was randomly sampled as the "low" taxon, and each sample from all other populations except the Mexican Lowland was set as "X". The *mexicana* accession TIL25 from the maize HapMap2 project [60] was treated as the third column species. The D statistic was calculated in a 1*kb*-block and then jackknife bootstrapping was conducted to estimate significance.

In addition, the $\hat{f}_d$ statistic [36] was calculated based on a similar tree form $(((P_1, P_2), P_3), O)$, but using allele frequencies across multiple individuals for each position on the tree. We fixed P_1 as the Mexican Lowland population, P_3 as two lines of *mexicana* (TIL08 and

TIL25) and *T. dactyloides* as the outgroup. P_2 was set to each of the four highland populations and the South American Lowland population.

The $\hat{f}_d$ statistic was calculated in 10-kb non-overlapping windows across the genome with the python script egglib_sliding_windows.py (https://github.com/johnomics/Martin_Davey_Jiggins_evaluating_introgression_statistics), which makes use of the EggLib library [81]. The input file was generated by first identifying genotypes using ANGSD (-doMajorMinor 1 -doMaf 1 -GL 2 -doGeno 4 -doPost 1 -postCutoff 0.95 -SNP_pval $1e-6$) followed by format adjustments with a custom script (Please see "Data and analysis pipeline accessibility").

Estimating burden of deleterious mutations

We estimated the individual burden of deleterious alleles based on GERP scores [82] for each site in the maize genome, which reflects the strength of purifying selection based on constraint in a whole genome alignment of 13 plant species [83]. The alignment and estimated GERP scores are available at iplant (/iplant/home/lilepisorus/GERPv3). Scores above 0 may be interpreted as historically subject to purifying selection, and mutations at such sites are likely deleterious. We identified *Sorghum bicolor* alleles in the multiple species alignment as ancestral and defined the non-*Sorghum* allele as the deleterious allele. Only biallelic sites were included for our evaluation. Inclusion of the maize B73 reference genome when calculating GERP scores [83] introduces a bias toward underestimation of the burden of deleterious alleles in maize versus teosinte populations. Therefore, we corrected the GERP scores of sites where the B73 allele is derived following [7]. Briefly, we divided SNPs where the B73 allele is ancestral into bins of 1% derived allele frequency based on maize HapMap3 [84] and used this frequency distribution to estimate the posterior probability of GERP scores for SNPs where the B73 allele is derived.

The sum of GERP scores multiplied by deleterious allele frequency for each SNP site was used as a proxy of individual burden of deleterious alleles under an additive model ($HET * 0.5 + HOM * 1$). This burden was calculated under a recessive model as the sum of GERP scores multiplied by one for each deleterious homozygous site ($HOM * 1$). For a better understanding of the variation of individual burden among sites under varied selection strength, we partitioned the deleterious SNPs into four categories (-2 <GERP $\leq$ 0, nearly neutral; 0 < GERP $\leq$ 2, slightly deleterious; 2 < GERP $\leq$ 4, moderately deleterious; GERP > 4, strongly deleterious) and recapitulated the above statistics.

Data and analysis pipeline accessibility

The pipeline and custom scripts utilized in this paper are documented in the following GitHub repository: https://github.com/HuffordLab/Wang_Private/tree/master/demography/analyses The WGS raw reads have been deposited in NCBI SRA (SRP065483).

6 Acknowledgments

This study was supported by the US Department of Agriculture (USDA #2009-65300-05668), the USDA Agricultural Research Service, the National Science Foundation (NSF IOS #1546719), USDA Hatch project (CA-D-PLS-2066-H), and startup funds from Iowa State University. Additionally, we thank Dr. Andrew Severin and Dr. Arun Seetharam for bioinformatic support.

References

[1] Jun Z. Li, Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, Gregory S. Barsh, Marcus Feldman, Luigi L. Cavalli-Sforza, and Richard M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.

[2] Timothy M Beissinger, Li Wang, Kate Crosby, Arun Durvasula, Matthew B Hufford, and Jeffrey Ross-Ibarra. Recent demography drives changes in linked selection across the maize genome. *Nature plants*, 2:16084, 2016.

[3] John F. Doebley, Brandon S. Gaut, and Bruce D. Smith. The molecular genetics of crop domestication. *Cell*, 127:1309–1321, 2006.

[4] Christopher R. Gignoux, Brenna M. Henn, and Joanna L. Mountain. Rapid, global demographic expansions after the origins of agriculture. *Proceedings of the National Academy of Sciences*, 108:6044–6049, 2011.

[5] Matthew B Hufford, Pesach Lubinksy, Tanja Pyhäjärvi, Michael T Devengenzo, Norman C Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild introgression in maize. *PLoS Genet*, 9:e1003477, 2013.

[6] Kay Prufer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Helene Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Paabo. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505:43–49, 2014.

[7] Yuval B Simons, Michael C Turchin, Jonathan K Pritchard, and Guy Sella. The deleterious mutation load is insensitive to recent population history. *Nature genetics*, 46:220–224, 2014.

[8] Ron Do, Daniel Balick, Heng Li, Ivan Adzhubei, Shamil Sunyaev, and David Reich. No evidence that selection has been less effective at removing deleterious mutations in europeans than in africans. *Nature genetics*, 47:126–131, 2015.

[9] Kelley Harris and Rasmus Nielsen. The genetic cost of neanderthal introgression. *Genetics*, 203:881–891, 2016.

[10] Wenqing Fu, Rachel M Gittelman, Michael J Bamshad, and Joshua M Akey. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *The American Journal of Human Genetics*, 95:421–436, 2014.

[11] M Zhang, L Zhou, R Bawa, H Suren, and JA Holliday. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Molecular Biology and Evolution*, 33:2899–2910, 2016.

[12] Clare D Marsden, Diego Ortega-Del Vecchyo, Dennis P OBrien, Jeremy F Taylor, Oscar Ramirez, Carles Vilà, Tomas Marques-Bonet, Robert D Schnabel, Robert K Wayne, and Kirk E Lohmueller. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences*, 113:152–157, 2016.

[13] Yuval B Simons and Guy Sella. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current Opinion in Genetics & Development*, 41:150–158, 2016.

[14] Qingpo Liu, Yongfeng Zhou, Peter L Morrell, and Brandon S Gaut. Deleterious variants in asian rice and the potential cost of domestication. *bioRxiv*, page 057224, 2016.

[15] Montgomery Slatkin and Laurent Excoffier. Serial founder effects during range expansion: a spatial analog of genetic drift. *Genetics*, 191:171–181, 2012.

[16] Frdric Austerlitz, Bernard Jung-Muller, Bernard Godelle, and Pierre-Henri Gouyon. Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, 51:148–164, 1997.

[17] Sohini Ramachandran, Omkar Deshpande, Charles C Roseman, Noah A Rosenberg, Marcus W Feldman, and L Luca Cavalli-Sforza. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in africa. *Proceedings of the National Academy of Sciences of the United States of America*, 102:15942–15947, 2005.

[18] Brenna M Henn, Laura R Botigué, Carlos D Bustamante, Andrew G Clark, and Simon Gravel. Estimating the mutation load in human genomes. *Nature Reviews Genetics*, 16:333–343, 2015.

[19] Ivan Juric, Simon Aeschbacher, and Graham Coop. The strength of selection against neanderthal introgression. *PLoS Genetics*, 12:1–25, 2016.

[20] Yoshihiro Matsuoka, Yves Vigouroux, Major M. Goodman, Jesus Sanchez G., Edward Buckler, and John Doebley. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences*, 99:6080–6084, 2002.

[21] Dolores R. Piperno, Anthony J. Ranere, Irene Holst, Jose Iriarte, and Ruth Dickau. Starch grain and phytolith evidence for early ninth millennium b.p. maize from the central balsas river valley, mexico. *Proceedings of the National Academy of Sciences*, 106:5019–5024, 2009.

[22] Joost van Heerwaarden, John Doebley, William H Briggs, Jeffrey C Glaubitz, Major M Goodman, Jose de Jesus Sanchez Gonzalez, and Jeffrey Ross-Ibarra. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences*, 108:1088–1092, 2011.

[23] Stephen I Wright, Irie Vroh Bi, Steve G Schroeder, Masanori Yamasaki, John F Doebley, Michael D McMullen, and Brandon S Gaut. The effects of artificial selection on the maize genome. *Science*, 308:1310–1314, 2005.

[24] Matthew B Hufford, Xun Xu, Joost Van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A Cartwright, Robert J Elshire, Jeffrey C Glaubitz, Kate E Guill, Shawn M Kaeppler, et al. Comparative population genomics of maize domestication and improvement. *Nature genetics*, 44:808–811, 2012.

[25] William L Merrill, Robert J Hard, Jonathan B Mabry, Gayle J Fritz, Karen R Adams, John R Roney, and Art C MacWilliams. The diffusion of maize to the southwestern united states and its impact. *Proceedings of the National Academy of Sciences*, 106:21019–21026, 2009.

[26] Alexander Grobman, Duccio Bonavia, Tom D. Dillehay, Dolores R. Piperno, Jos Iriarte, and Irene Holst. Preceramic maize from paredones and huaca prieta, peru. *Proceedings of the National Academy of Sciences*, 109:1755–1759, 2012.

[27] Jeffrey Ross-Ibarra, Maud Tenaillon, and Brandon S Gaut. Historical divergence and gene flow in the genus zea. *Genetics*, 181:1399–1413, 2009.

[28] Maud I. Tenaillon, Jana U'Ren, Olivier Tenaillon, and Brandon S. Gaut. Selection versus demography: A multilocus investigation of the domestication process in maize. *Molecular Biology and Evolution*, 21:1214–1225, 2004.

[29] Jacob Van Etten and Robert J Hijmans. A geospatial modelling approach integrating archaeobotany and genetics to trace the origin and dispersal of domesticated plants. *PLoS One*, 5:e12060, 2010.

[30] Races of maize.

[31] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46:919–925, 2014.

[32] Daniel R Schrider, Alexander G Shanku, and Andrew D Kern. Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 204:1207–1223, 2016.

[33] S Hearne, C Chen, E Buckler, and S Mitchell. Unimputed gbs derived snps for maize landrace accessions represented in the seed-maize gwas panel, 2014.

[34] Eric Y Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, 28:2239–2252, 2011.

[35] Rute R da Fonseca, Bruce D Smith, Nathan Wales, Enrico Cappellini, Pontus Skoglund, Matteo Fumagalli, José Alfredo Samaniego, Christian Carøe, María C Ávila-Arcos, David E Hufnagel, et al. The origin and evolution of maize in the american southwest. *Nature plants*, 1:14003, 2015.

[36] Simon H Martin, John W Davey, and Chris D Jiggins. Evaluating the use of abba–baba statistics to locate introgressed loci. *Molecular biology and evolution*, 32:244–257, 2015.

[37] J Alberto Romero Navarro, Martha Willcox, Juan Burgueño, Cinta Romay, Kelly Swarts, Samuel Trachsel, Ernesto Preciado, Arturo Terron, Humberto Vallejo Delgado, Victor Vidal, et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nature genetics*, 49:476–480, 2017.

[38] Edward S. Buckler, James B. Holland, Peter J. Bradbury, Charlotte B. Acharya, Patrick J. Brown, Chris Browne, Elhan Ersoz, Sherry Flint-Garcia, Arturo Garcia, Jeffrey C. Glaubitz, Major M. Goodman, Carlos Harjes, Kate Guill, Dallas E. Kroon, Sara Larsson, Nicholas K. Lepak, Huihui Li, Sharon E. Mitchell, Gael Pressoir, Jason A. Peiffer, Marco Oropeza Rosas, Torbert R. Rocheford, M. Cinta Romay, Susan Romero, Stella Salvo, Hector Sanchez Villeda, H. Sofia da Silva, Qi Sun, Feng Tian, Narasimham Upadyayula, Doreen Ware, Heather Yates, Jianming Yu, Zhiwu Zhang, Stephen Kresovich, and Michael D. McMullen. The genetic architecture of maize flowering time. *Science*, 325(5941):714–718, 2009.

[39] Gregory M Cooper, Eric A Stone, George Asimenos, Eric D Green, Serafim Batzoglou, and Arend Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15:901–913, 2005.

[40] Sebastien Renaut and Loren H Rieseberg. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Molecular biology and evolution*, 32:2273–2283, 2015.

[41] Torsten Günther and Karl J Schmid. Deleterious amino acid polymorphisms in arabidopsis thaliana and rice. *Theoretical and Applied Genetics*, 121:157–168, 2010.

[42] Thomas JY Kono, Fengli Fu, Mohsen Mohammadi, Paul J Hoffman, Chaochih Liu, Robert M Stupar, Kevin P Smith, Peter Tiffin, Justin C Fay, and Peter L Morrell. The role of deleterious substitutions in crop genomes. *Molecular Biology and Evolution*, 33:2307–2317, 2016.

[43] Qihui Zhu, Xiaoming Zheng, Jingchu Luo, Brandon S Gaut, and Song Ge. Multilocus analysis of nucleotide variation of oryza sativa and its wild relatives: severe bottleneck during domestication of rice. *Molecular Biology and Evolution*, 24:875–888, 2007.

[44] Hon-Ming Lam, Xun Xu, Xin Liu, Wenbin Chen, Guohua Yang, Fuk-Ling Wong, Man-Wah Li, Weiming He, Nan Qin, Bo Wang, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature genetics*, 42:1053–1059, 2010.

[45] Yongfeng Zhou, Melanie Massonnet, Jaleal Sanjak, Dario Cantu, and Brandon S Gaut. The evolutionary genomics of grape (vitis vinifera ssp. vinifera) domestication. *bioRxiv*, page 146373, 2017.

[46] Rachel S Meyer, Jae Young Choi, Michelle Sanches, Anne Plessis, Jonathan M Flowers, Junrey Amas, Katherine Dorph, Annie Barretto, Briana Gross, Dorian Q Fuller, Isaac Kofi Bimpong, Marie-Noelle Ndjiondjop, Khaled M Hazzouri, Glenn B Gregorio, and Michael D Purugganan. Domestication history and geographical adaptation inferred from a snp map of african rice. *Nat Genet*, 48:1083–1088, 2016.

[47] Brandon S Gaut, Concepción M Díez, and Peter L Morrell. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends in Genetics*, 31:709–719, 2015.

[48] Yves Vigouroux, Jeffrey C Glaubitz, Yoshihiro Matsuoka, Major M Goodman, Jesús Sánchez, and John Doebley. Population structure and genetic diversity of new world maize races assessed by dna microsatellites. *American Journal of Botany*, 95:1240–1253, 2008.

[49] Shohei Takuno, Peter Ralph, Kelly Swarts, Rob J Elshire, Jeffrey C Glaubitz, Edward S Buckler, Matthew B Hufford, and Jeffrey Ross-Ibarra. Independent molecular basis of convergent highland adaptation in maize. *Genetics*, 200:1297–1312, 2015.

[50] Deborah M. Pearsall. *Plant domestication and the shift to agriculture in the Andes*, pages 105–120. Springer New York, New York, NY, 2008.

[51] Jeffrey Ross-Ibarra, Maud Tenaillon, and Brandon S. Gaut. Historical divergence and gene flow in the genus zea. *Genetics*, 181:1399–1413, 2009.

[52] Dolores R Piperno, J Enrique Moreno, José Iriarte, Irene Holst, Matthew Lachniet, John G Jones, Anthony J Ranere, and Ronald Castanzo. Late pleistocene and holocene environmental history of the iguala valley, central balsas watershed of mexico. *Proceedings of the National Academy of Sciences*, 104:11874–11881, 2007.

[53] A. Correa-Metrio, S. Lozano-Garca, S. Xelhuantzi-Lpez, S. Sosa-Njera, and S. E. Metcalfe. Vegetation in western central mexico during the last 50000 years: modern analogs and climate in the zacapu basin. *Journal of Quaternary Science*, 27:509–518, 2012.

[54] Ana M Poets, Zhou Fang, Michael T Clegg, and Peter L Morrell. Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome biology*, 16:1, 2015.

[55] Jessen V Bredeson, Jessica B Lyons, Simon E Prochnik, G Albert Wu, Cindy M Ha, Eric Edsinger-Gonzales, Jane Grimwood, Jeremy Schmutz, Ismail Y Rabbi, Chiedozie Egesi, et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature biotechnology*, 34:562–570, 2016.

[56] Benpeng Miao, Zhen Wang, and Yixue Li. Genomic analysis reveals hypoxia adaptation in the tibetan mastiff by introgression of the grey wolf from the tibetan plateau. *Molecular Biology and Evolution*, 34:734–743, 2016.

[57] John Doebley, Major M. Goodman, and Charles W. Stuber. Patterns of isozyme variation between maize and mexican annual teosinte. *Economic Botany*, 41(2):234–246, 1987.

[58] Jinliang Yang, Sofiane Mezmouk, Andy Baumgarten, Edward S Buckler, Katherine E Guill, Michael D McMullen, Rita H Mumm, and Jeffrey Ross-Ibarra. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *bioRxiv*, page 086132, 2016.

[59] Justin P. Gerke, Jode W. Edwards, Katherine E. Guill, Jeffrey Ross-Ibarra, and Michael D. McMullen. The genomic impacts of drift and selection for hybrid performance in maize. *Genetics*, 201(3):1201–1211, 2015.

[60] Jer-Ming Chia, Chi Song, Peter J Bradbury, Denise Costich, Natalia de Leon, John Doebley, Robert J Elshire, Brandon Gaut, Laura Geller, Jeffrey C Glaubitz, et al. Maize hapmap2 identifies extant variation from a genome in flux. *Nature genetics*, 44:803–807, 2012.

[61] Jian Lu, Tian Tang, Hua Tang, Jianzi Huang, Suhua Shi, and Chung-I Wu. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics*, 22:126–131, 2006.

[62] Mikkel Schubert, Hákon Jónsson, Dan Chang, Clio Der Sarkissian, Luca Ermini, Aurélien Ginolhac, Anders Albrechtsen, Isabelle Dupanloup, Adrien Foucal, Bent Petersen, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences*, 111:E5661–E5669, 2014.

[63] Ruth McQuillan, Niina Eklund, Nicola Pirastu, Maris Kuningas, Brian P McEvoy, Tõnu Esko, Tanguy Corre, Gail Davies, Marika Kaakinen, Leo-Pekka Lyytikäinen, et al. Evidence of inbreeding depression on human height. *PLoS Genet*, 8:e1002655, 2012.

[64] Aneil F Agrawal and Michael C Whitlock. Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics*, 187:553–566, 2011.

[65] Federico Manna, Guillaume Martin, and Thomas Lenormand. Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics*, 189:923–937, 2011.

[66] Jeff J Doyle. A rapid dna isolation procedure for small quantities of fresh leaf tissue. *Phytochem bull*, 19:11–15, 1987.

[67] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrowswheeler transform. *Bioinformatics*, 26:589–595, 2010.

[68] Patrick S Schnable, Doreen Ware, Robert S Fulton, Joshua C Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A Graves, et al. The b73 maize genome: complexity, diversity, and dynamics. *Science*, 326:1112–1115, 2009.

[69] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43:491–498, 2011.

[70] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27:2156–2158, 2011.

[71] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81:1084–1097, 2007.

[72] Matteo Fumagalli, Filipe G Vieira, Thorfinn Sand Korneliussen, Tyler Linderoth, Emilia Huerta-Sánchez, Anders Albrechtsen, and Rasmus Nielsen. Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, 195:979–992, 2013.

[73] Matteo Fumagalli, Filipe G Vieira, Tyler Linderoth, and Rasmus Nielsen. ngstools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, 30:1486–1487, 2014.

[74] Thorfinn S Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, 15:356, 2014.

[75] Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195:693–702, 2013.

[76] Klaus Peter Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27:592–593, 2011.

[77] Thorfinn Sand Korneliussen, Ida Moltke, Anders Albrechtsen, and Rasmus Nielsen. Calculation of tajimas d and other neutrality test statistics from low depth next-generation sequencing data. *BMC bioinformatics*, 14:289, 2013.

[78] Filipe G Vieira, Matteo Fumagalli, Anders Albrechtsen, and Rasmus Nielsen. Estimating inbreeding coefficients from ngs data: Impact on genotype calling and allele frequency estimation. *Genome research*, 23:1852–1861, 2013.

[79] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81:559–575, 2007.

[80] Funda Ogut, Yang Bian, Peter J Bradbury, and James B Holland. Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity*, 114:552–563, 2015.

[81] Stéphane De Mita and Mathieu Siol. Egglib: processing, analysis and simulation tools for population genetics and genomics. *BMC genetics*, 13:27, 2012.

[82] Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Comput Biol*, 6:e1001025, 2010.

[83] Eli Rodgers-Melnick, Peter J Bradbury, Robert J Elshire, Jeffrey C Glaubitz, Charlotte B Acharya, Sharon E Mitchell, Chunhui Li, Yongxiang Li, and Edward S Buckler. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences*, 112:3823–3828, 2015.

[84] Robert Bukowski, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, Dawen Xu, Bicheng Yang, Chuanxiao Xie, et al. Construction of the third generation zea mays haplotype map. *bioRxiv*, page 026963, 2015.

Figures

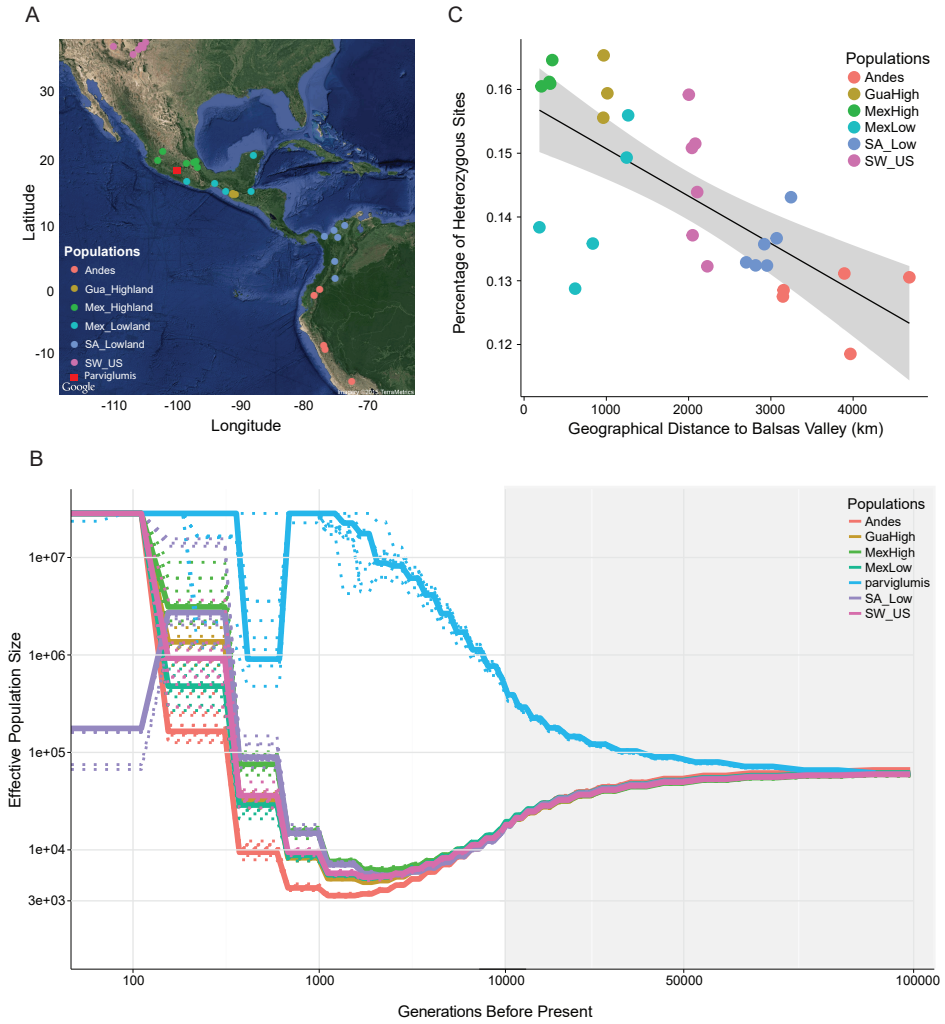

Figure 1: Maize domestication and expansion. A.Sampling locations. B. Estimates of effective population size over time. Dashed lines represent bootstrapping results. The x axis is $log10$ scaled when time is less than 10,000 generations BP and linear when greater than 10,000 generations BP as indicated by the grey background. C. The percentage of polymorphic sites versus distance from the maize domestication center. Abbreviations for populations: GuaHigh, Guatemalan Highlands; MexHigh, Mexican Highlands; MexLow, Mexican Lowlands; SA_Low, South American Lowlands; SW_US, Southwestern US Highlands.

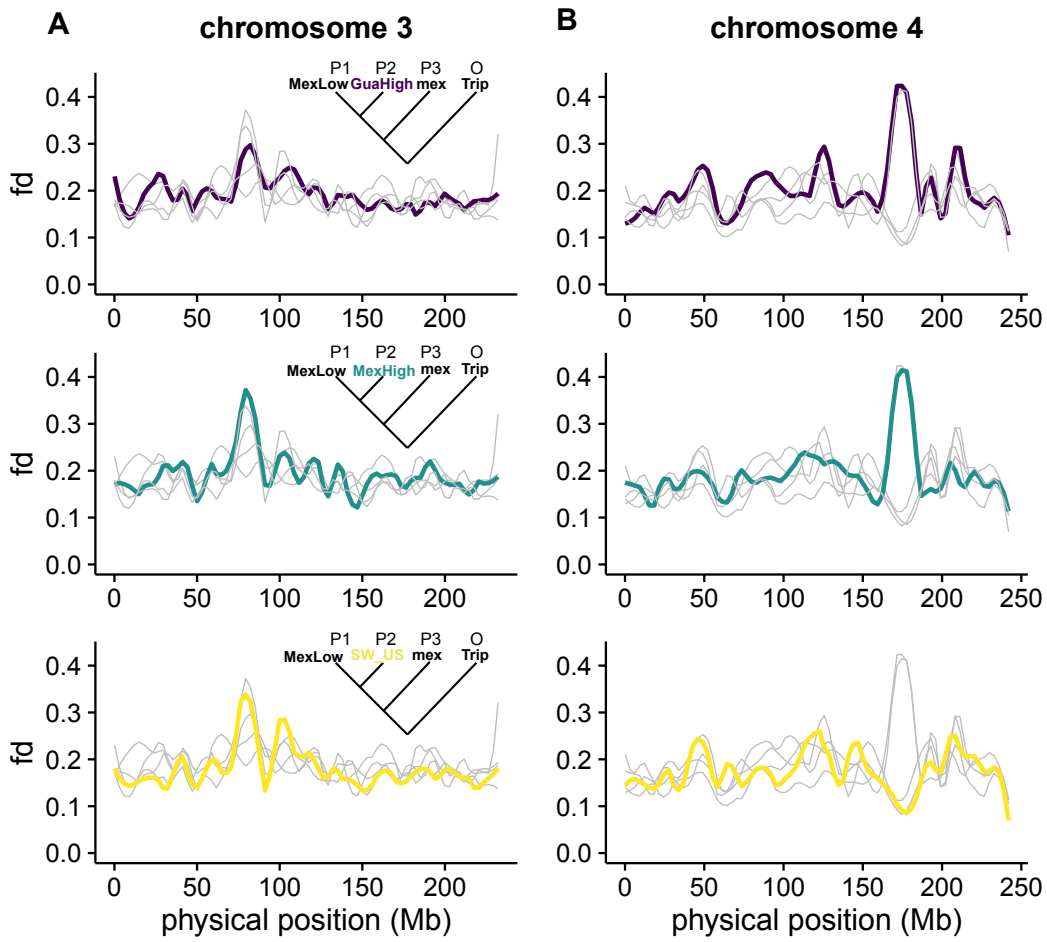

Figure 2: Introgression from *mexicana* into maize landraces. (A) chromosome 3 and (B) chromosome 4. The statistic $\hat{f}_d$ was calculated based on the tree, in which P2 is varied across highland and the South American lowland populations. Shown are results for the indicated three highland populations; other populations are drawn in grey. mex: *mexicana*; Trip: *Tripsacum*.

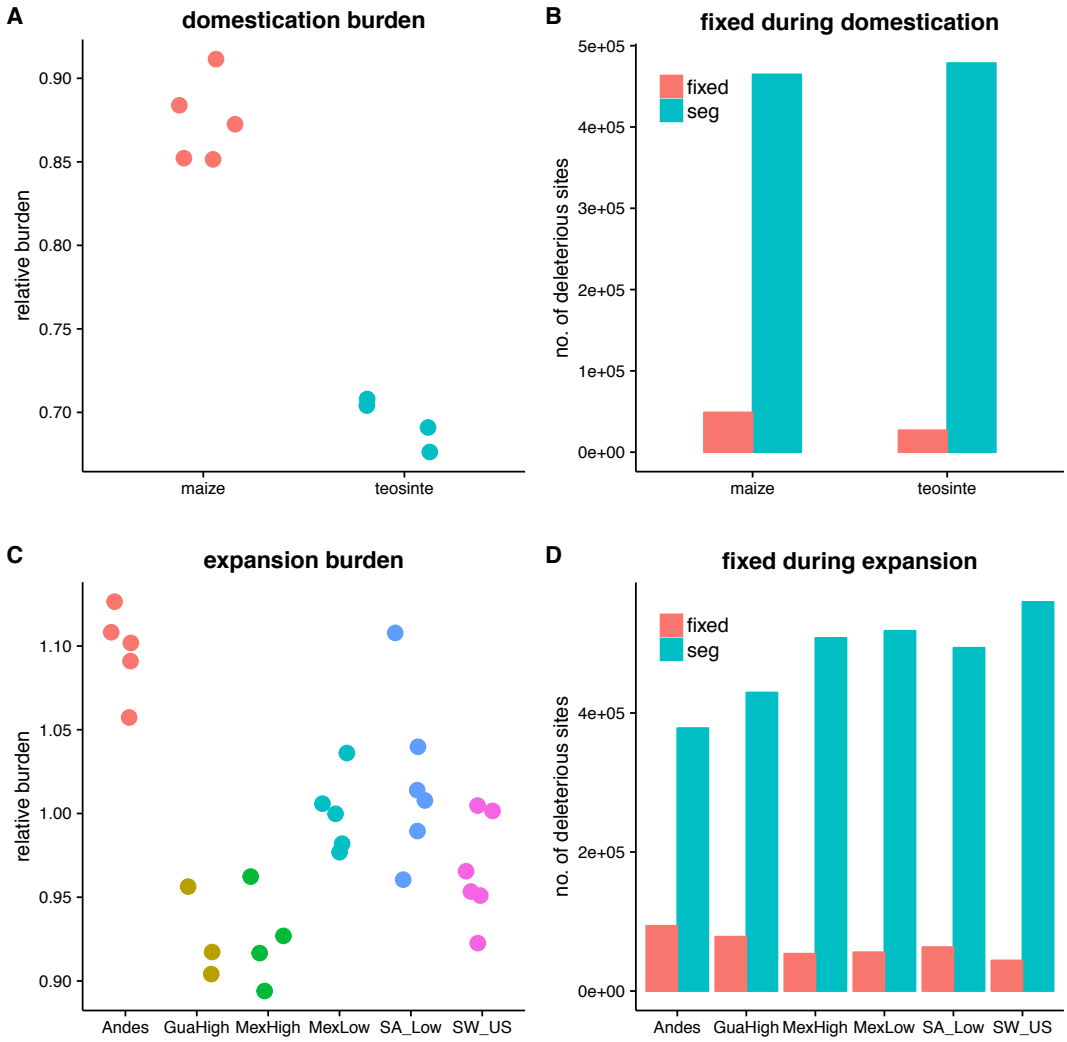

Figure 3: Burden of deleterious mutations during maize domestication and expansion. Comparison of counts of deleterious sites at the individual level (A) between *parviglumis* and maize and (C) among maize populations under a recessive model; comparison of fixed vs segregating (seg) deleterious sites at the population level (B) between *parviglumis* and maize and (D) among maize populations.

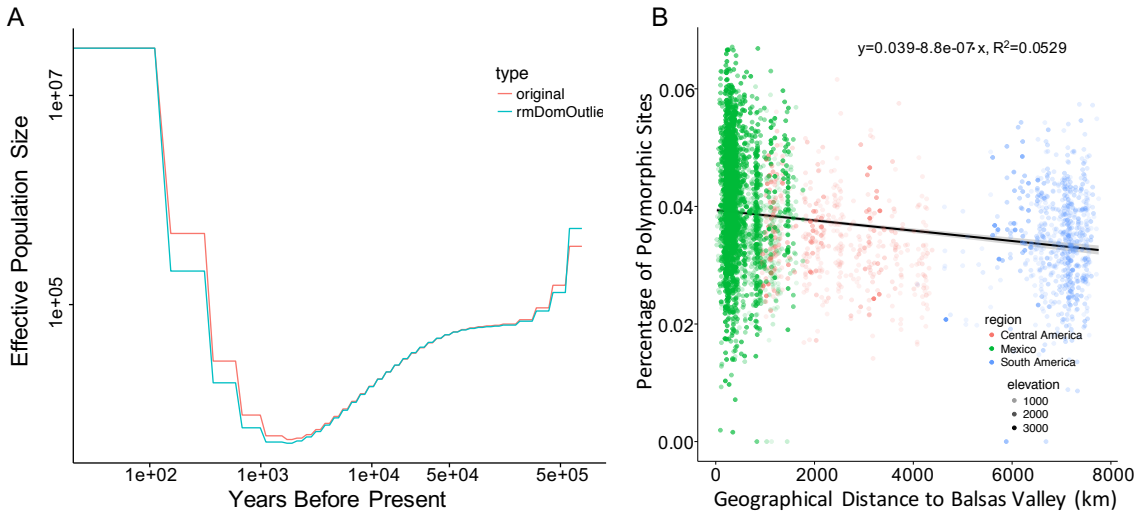

Figure S1: Demography of maize populations. A. MSMC results before and after masking candidate regions under selection during domestication. B. Percentage of heterozygous sites versus distance from the Balsas Valley in 3520 samples from the SeeDs data set.

Supporting Information

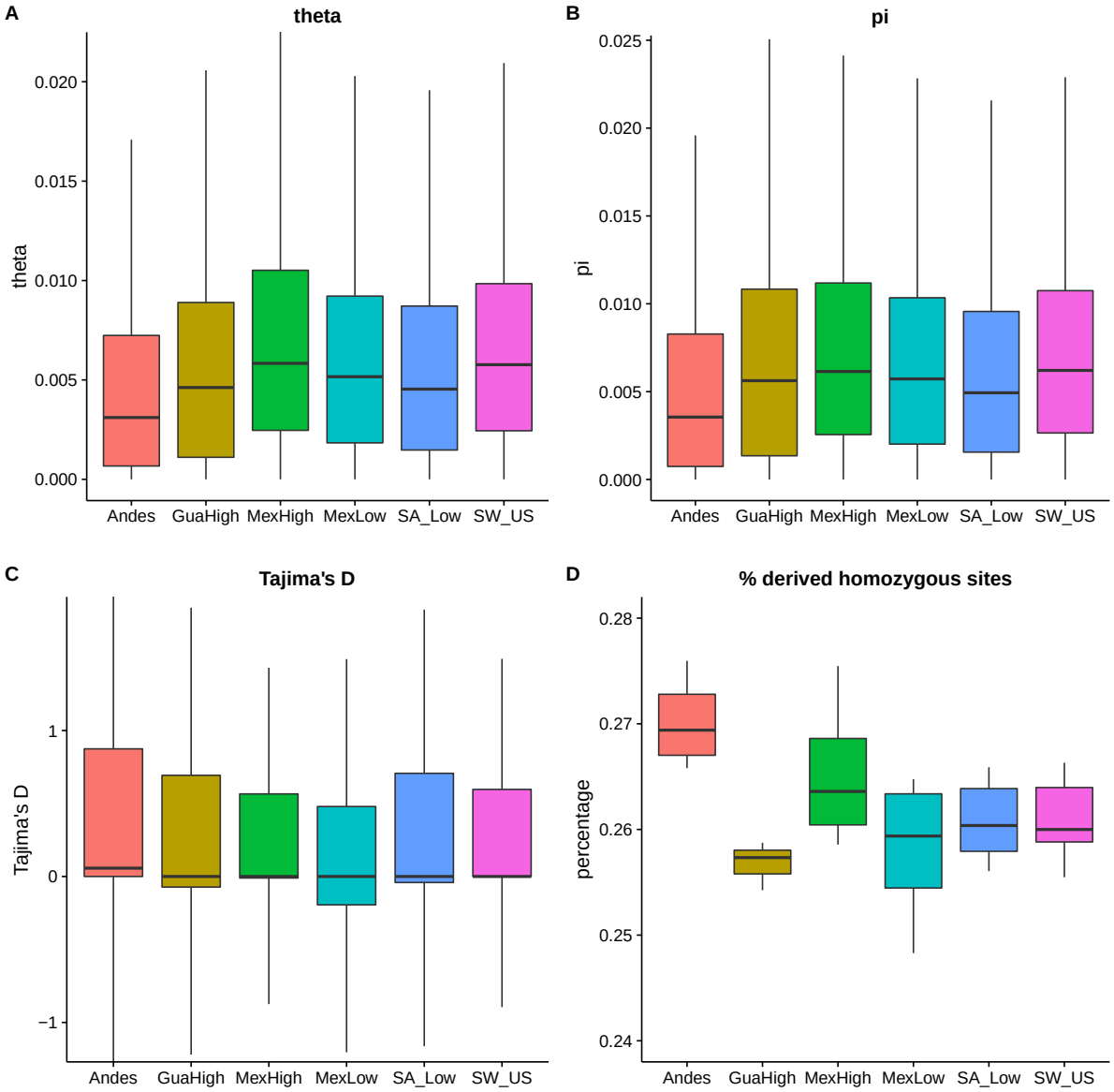

Figure S2: Boxplot of multiple population genetic statistics. Watterson's *theta* (A), θ_π)(B) and Tajima's D (C) are based on values in 10-kb non-overlapping windows across the genome. Percentage of derived homozygous sites was calculated for each individual and reported per population

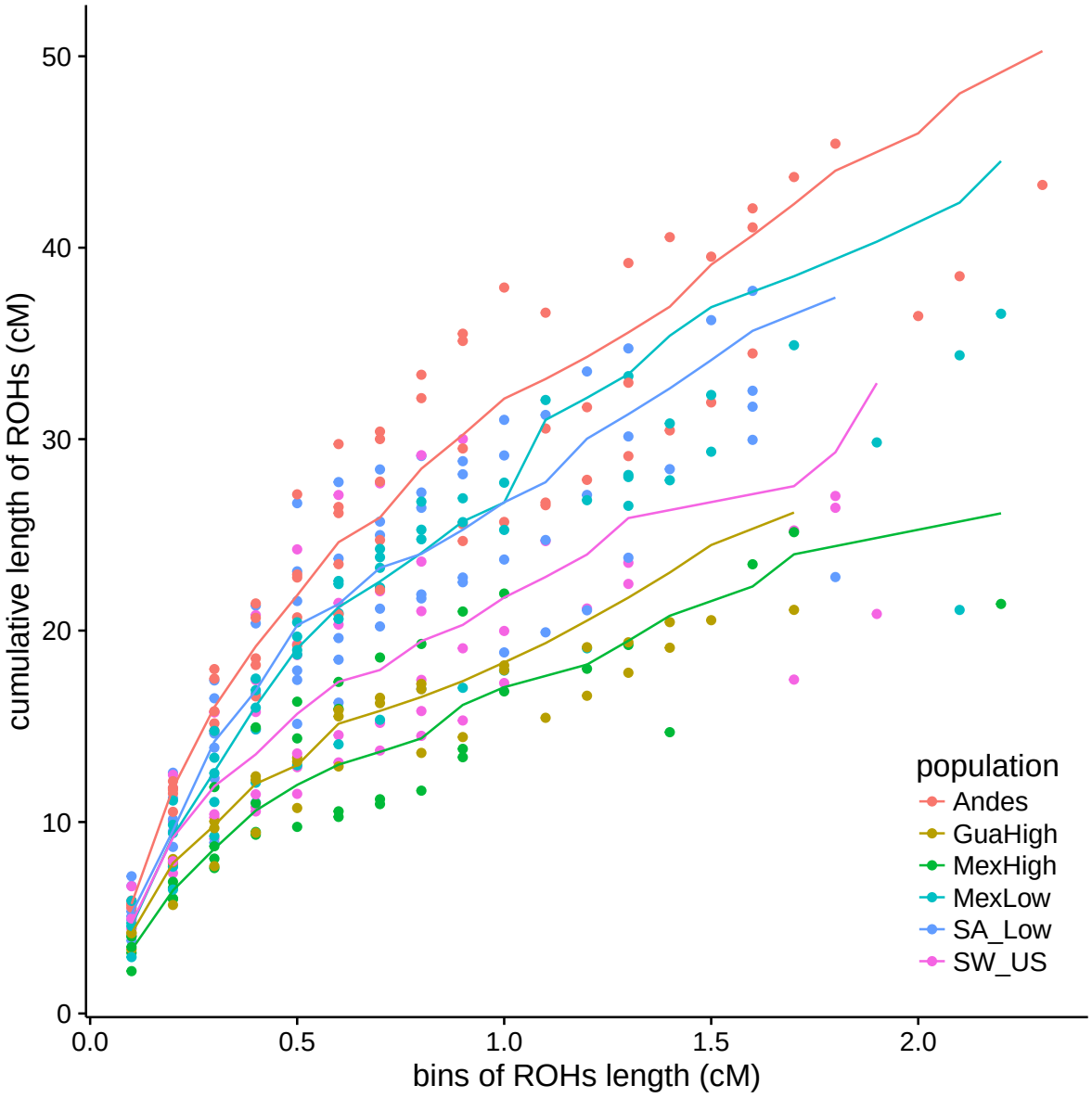

Figure S3: Cumulative length of ROHs in cM among populations. The lines indicate median level in each population. ROH: runs of homozygosity.

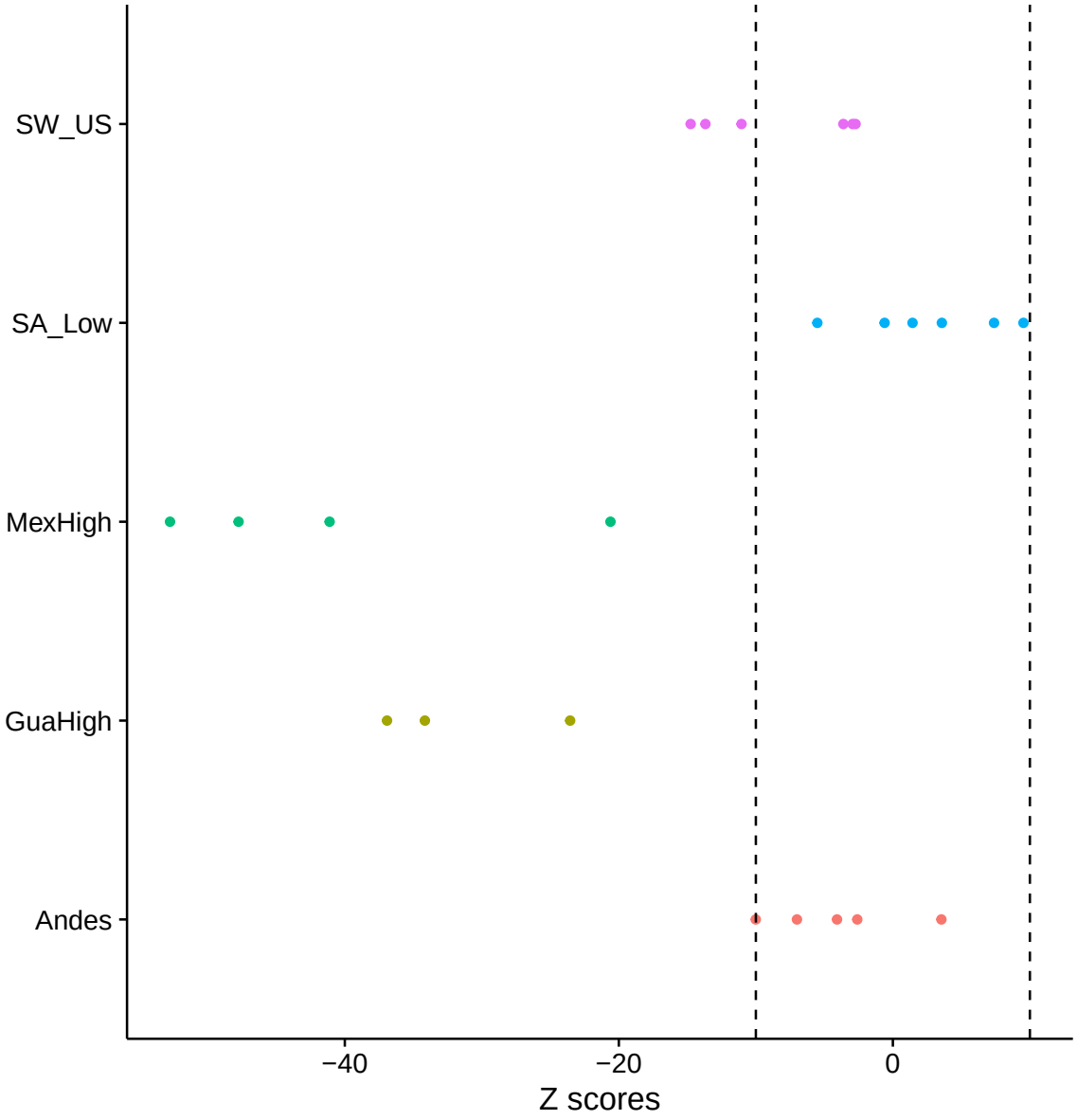

Figure S4: Introgression from *mexicana* into maize landraces. Evidence of introgression from *mexicana* into Mexican highland, Guatemalan highland and Southwestern US highland maize populations. The dashed lines correspond to Z scores equal to -10 and 10.

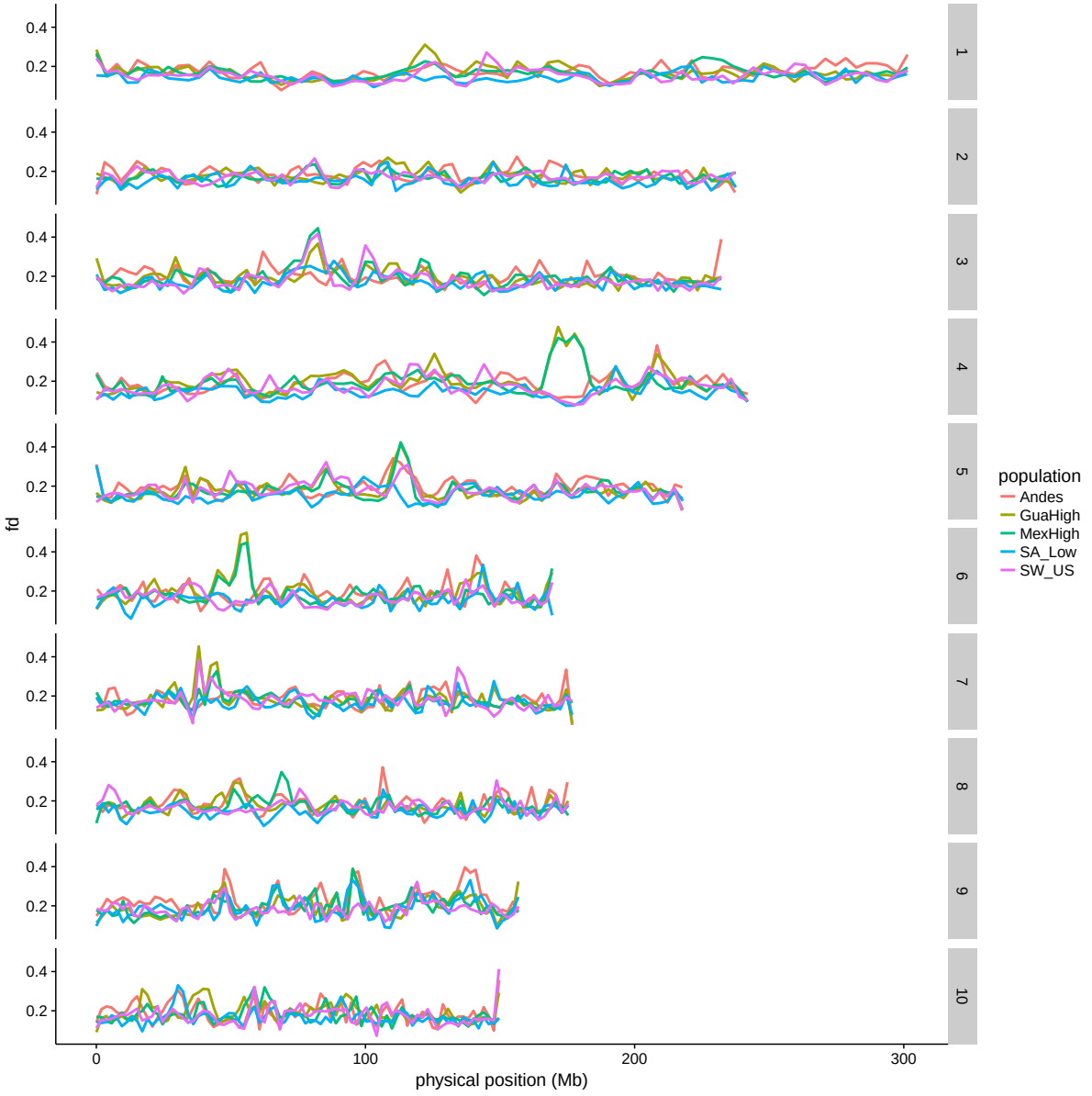

Figure S5: Loess regression of $\hat{f}_d$ in 10-kb nonoverlapping windows across all chromosomes.

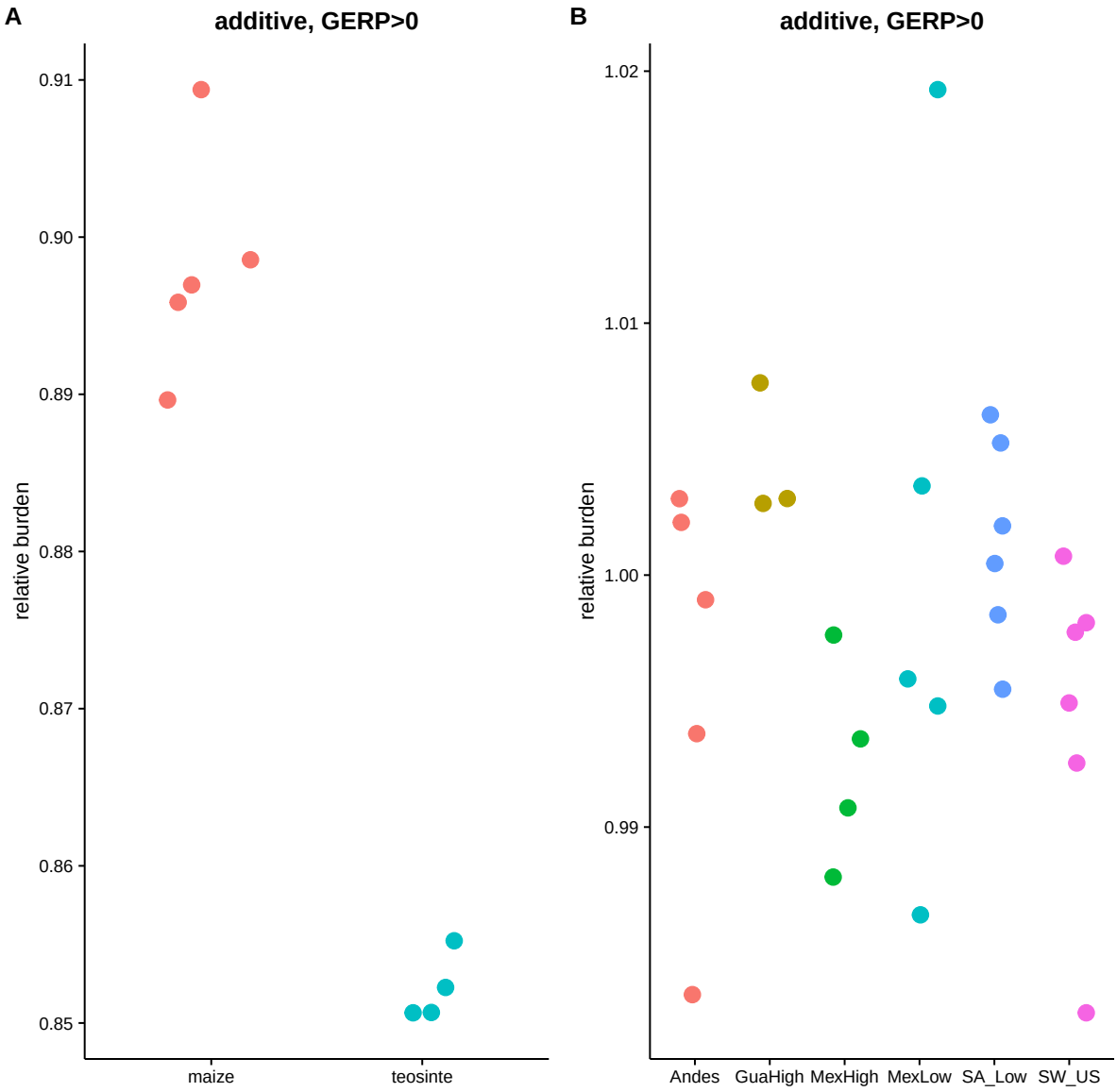

Figure S6: Relative burden of deleterious alleles under additive model between maize and teosinte (A) and among maize populations (B).

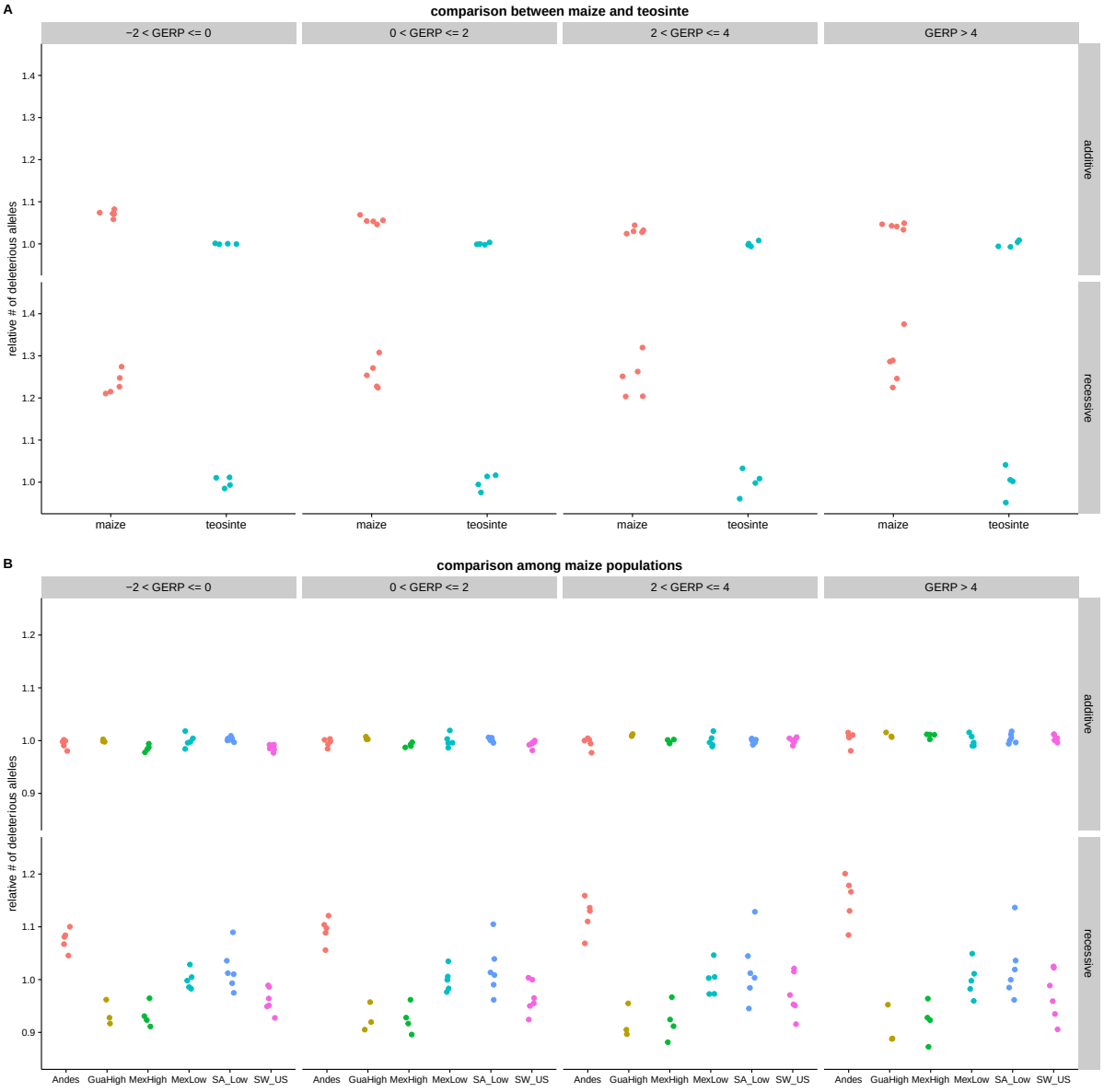

Figure S7: Relative burden of deleterious alleles under both additive and recessive models with different GERP partitions between maize and teosinte (A) and among maize populations (B).

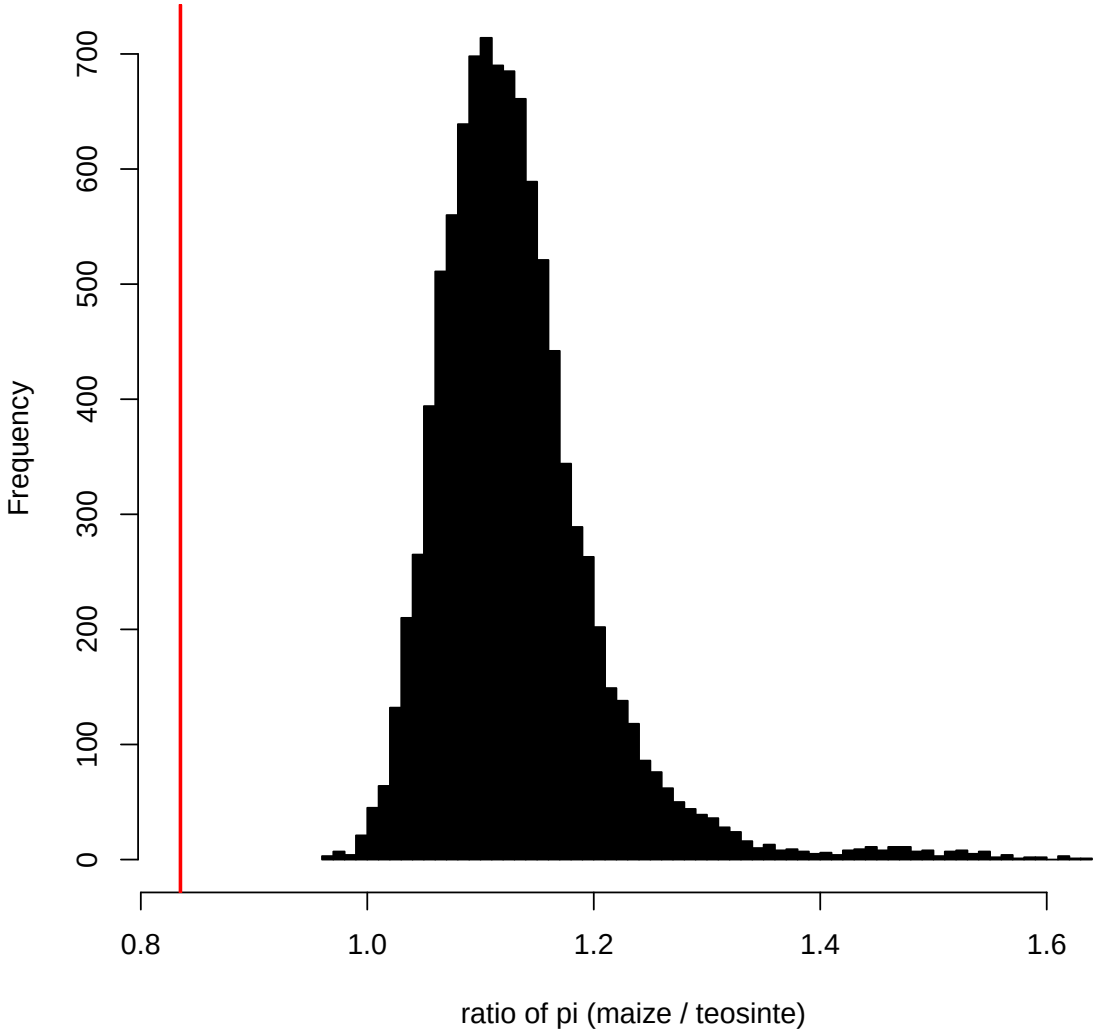

Figure S8: Distribution of ratio of θ_π between maize and teosinte in 420 domestication candidate genes (mean value was indicated with red line) compared to 10,000 replicates of genome-wide sampling of 420 random genes.

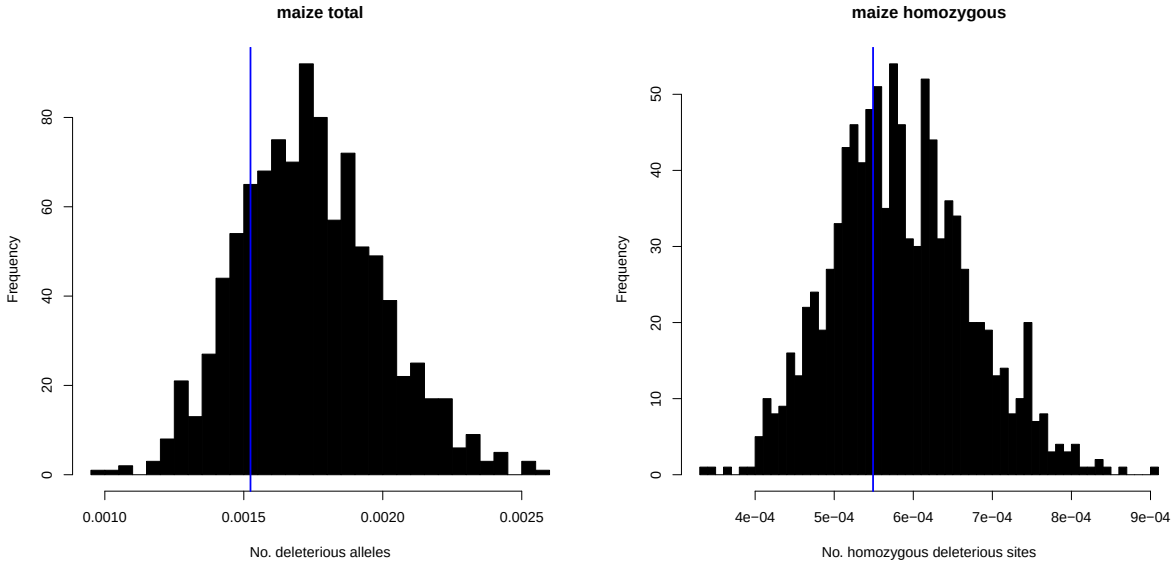

Figure S9: Distribution of number of deleterious sites per bp in 420 domestication candidate genes (indicated with blue line) compared to genome-wide random samples under an (A) additive model and (B) recessive model.

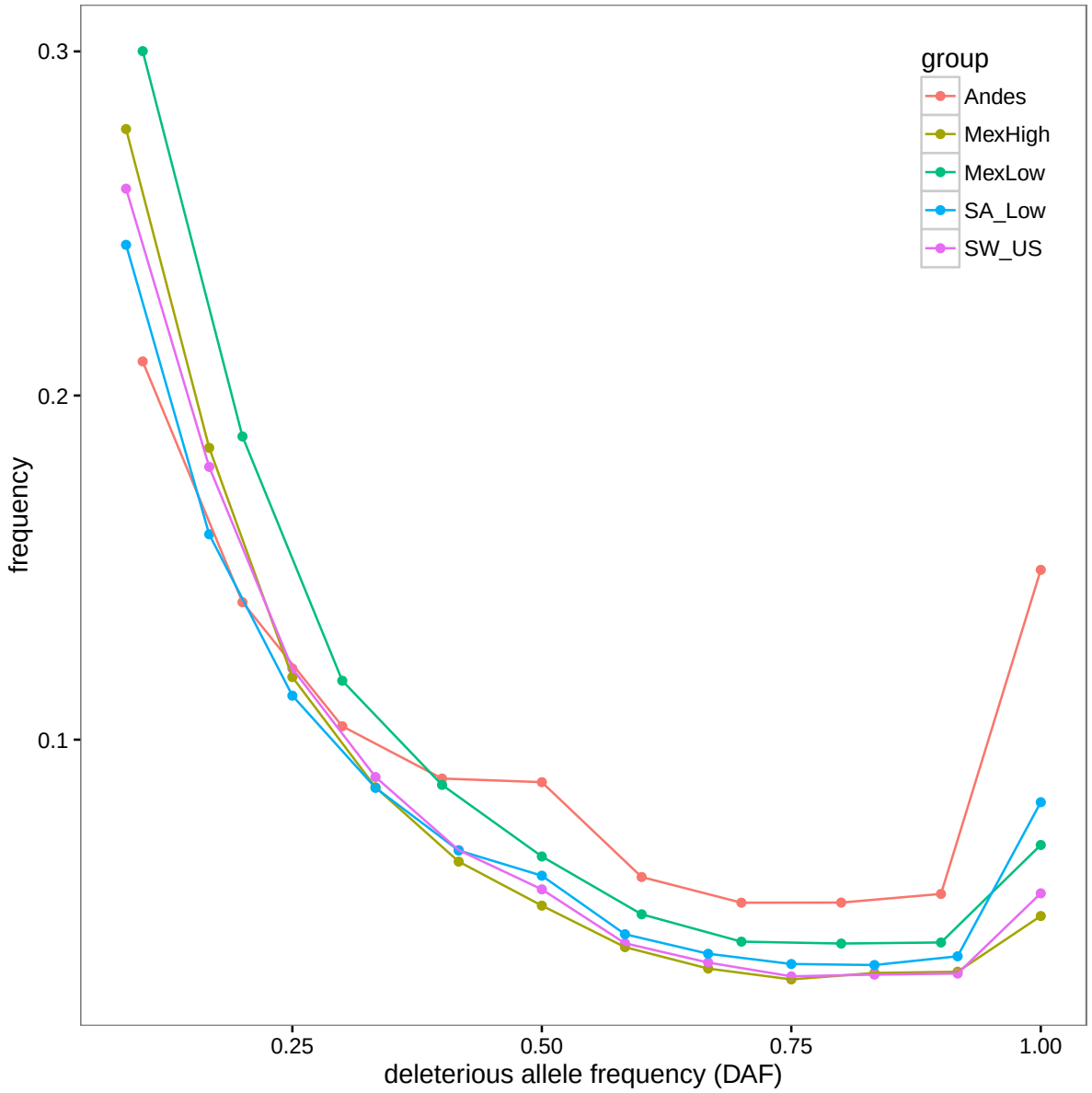

Figure S10: Site frequency spectrum of deleterious SNPs in five populations; GuaHigh is not included since the small sampling limited power for the SFS.

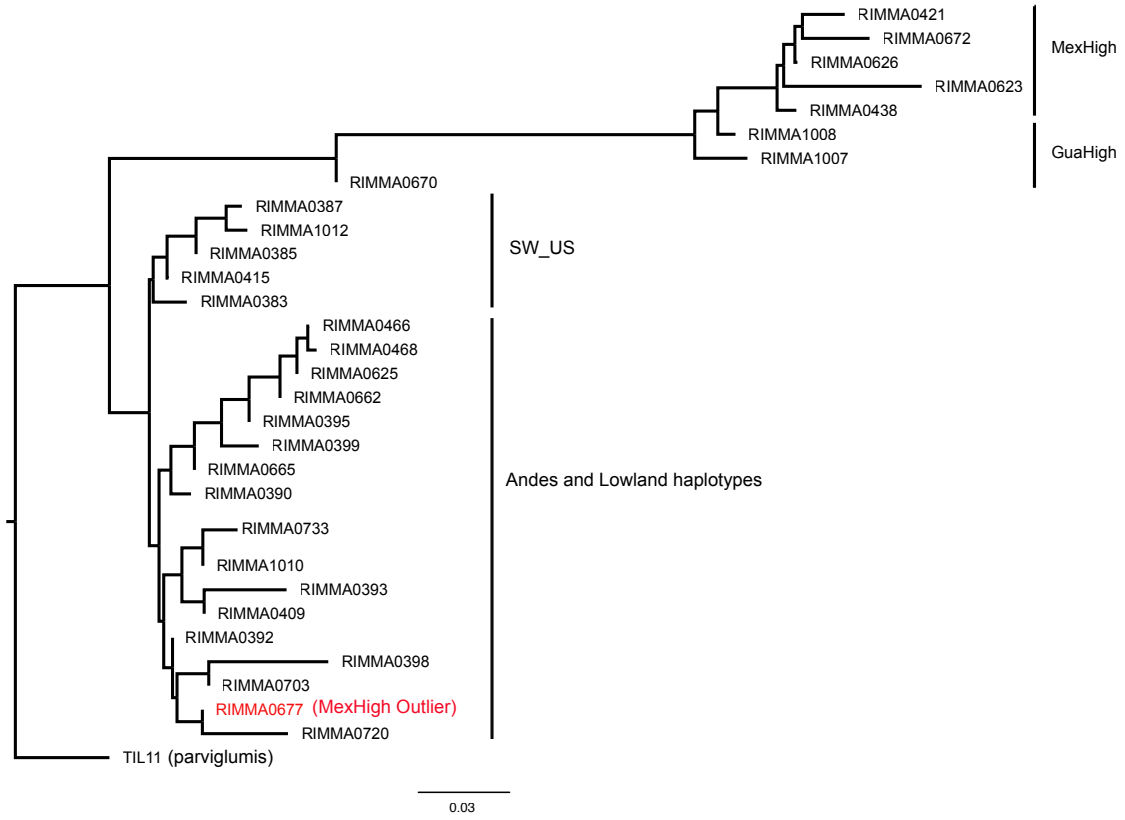

Figure S11: Neighbor Joining tree of SNPs from an inversion on chromosome 4 with a diagnostic haplotype for highland Mexican material.

RI_Group	Latitude	Longitude	RI_Accession	Elevation	Locality
Andes	-14.317	-72.917	RIMMA0466	3600	Apurimac, Peru
Andes	-9.383	-77.167	RIMMA0468	3150	Ancash, Peru
Andes	-8.700	-77.383	RIMMA0625	2820	Ancash, Peru
Andes	0.000	-78.000	RIMMA0662	2195	Ecuador
Andes	-0.917	-78.917	RIMMA0665	2931	Ecuador
GuaHigh	14.967	-91.767	RIMMA0670	2378	San Marcos, Guatemala
GuaHigh	15.033	-91.783	RIMMA1007	3049	San Marcos, Guatemala
GuaHigh	14.917	-91.333	RIMMA1008	2774	Totonicapan, Guatemala
MexHigh	19.850	-97.983	RIMMA0421	2250	Puebla, Mexico
MexHigh	19.000	-97.383	RIMMA0438	2600	Puebla, Mexico
MexHigh	20.033	-103.683	RIMMA0623	2520	Jalisco, Mexico
MexHigh	19.883	-97.583	RIMMA0626	2260	Puebla, Mexico
MexHigh	19.683	-99.133	RIMMA0672	2256	Mexico, Mexico
MexHigh	21.367	-102.850	RIMMA0677	1951	Zacatecas, Mexico
MexLow	15.433	-92.900	RIMMA0409	107	Chiapas, Mexico
MexLow	20.833	-88.517	RIMMA0703	30	Yucatan, Mexico
MexLow	15.467	-88.850	RIMMA0720	39	Guatemala
MexLow	16.567	-94.617	RIMMA0733	107	Oaxaca, Mexico
MexLow	16.850	-99.067	RIMMA1010	201	La Concordia, Guerrero
SA_Low	4.517	-75.633	RIMMA0390	353	Caldas, Colombia
SA_Low	1.750	-75.583	RIMMA0392	555	Caqueta, Colombia
SA_Low	8.317	-75.150	RIMMA0393	100	Cordoba, Colombia
SA_Low	8.500	-77.267	RIMMA0395	30	Choco, Colombia
SA_Low	9.433	-75.700	RIMMA0398	27	Magdalena, Colombia
SA_Low	10.183	-74.050	RIMMA0399	250	Magdalena, Colombia
SW_US	34.900	-107.583	RIMMA0383	2073	Acoma Pueblo, NM, USA
SW_US	36.050	-106.283	RIMMA0384	2134	San Lorenzo Pueblo, NM, USA
SW_US	36.450	-105.550	RIMMA0385	2134	Taos Pueblo, NM, USA
SW_US	35.617	-106.733	RIMMA0387	1829	Jemez Pueblo, NM, USA
SW_US	35.900	-110.667	RIMMA0415	1941	Hotevilla, Arizona, USA
SW_US	35.762	-105.933	RIMMA1012	2073	Tesuque Pueblo, NM, USA

Figure S12: Basic information regarding the sampled maize landrace accessions. NM: New Mexico.