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 interplay of demography and selection. Neutral processes that comprise a species' demography (*e.g.,* stochastic changes in population size and migration events) influence both the pool of diversity upon which selection can act and its efficiency. Selection, in combination with genetic drift, then plays a role in the ultimate fate of this diversity.

Both humans and their associated agricultural species have experienced recent demographic shifts that have 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 in many domesticates [3]. Subsequent to domestication, both humans and many of their domesticated crops have experienced a process of global expansion facilitated by the invention of agriculture itself [4] and the warming and stabilization of global climate in the early Holocene [5]. Expansion has often been accompanied by gene flow with close relatives, a demographic process that has further altered patterns of diversity [6, 7].

Recent interest in the effects of demography on functional variation has led to a growing body of theory on the interplay of selection and demography. To date, this relationship has been most thoroughly investigated in the context of deleterious alleles. While theory suggests that over longer periods mutation load may be insensitive to demography [8, 9], empirical data are consistent with changes in population size and gene flow affecting the distribution of deleterious variation over shorter timescales [10, 11, 12, 13, 14]. For example, though typically removed by purifying selection, deleterious alleles can rise to appreciable frequency or fix when the product of the effective population size N_e and the selection coefficient s is such that $N_e s < 1$. In this scenario, the fate of deleterious alleles is primarily determined by genetic drift. Examples from both plant and animal species have confirmed that populations that have undergone recent bottlenecks and exhibit low N_e demonstrate higher mutation load relative to populations with higher N_e [11, 12, 13, 15]. Similarly, in geographically expanding populations repeated sub-sampling of diversity can occur during migration away from the center of origin, a phenomenon known as serial founder effects [16, 17]. Such effects have been proposed as an explanation for the observed decline in genetic diversity in human populations with increasing geographic distance from Africa [18], and reductions in N_e in populations distant from Africa have been linked to increased counts of deleterious alleles [19]. Finally, gene flow may affect genome-wide patterns of deleterious variants, particularly when occurring between populations with starkly contrasting N_e. During the Out-of-Africa migration, modern humans inter-mated with Neanderthals, a close relative with substantially lower N_e and higher mutation load [10]. 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 [10, 20].

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 interplay between demography and selection. Maize was domesticated in a narrow region of southwest Mexico from the wild plant teosinte (*Zea mays* ssp. *parviglumis*; [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 from multiple teosinte species has also been documented in maize in regions outside of its center of origin [6, 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 [2] and largely disregarding the spatial effects of geographic expansion and gene flow (but see [29]).

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 representing the pre-Columbian distribution of maize. We present novel findings regarding the extent and timing of maize population size change during this period and document patterns of gene flow between maize and its wild relatives during its expansion across the Americas. We then characterize the interplay between demography and selection in maize by linking population size change and gene flow to patterns of deleterious alleles.

2 Results

2.1 Maize population size change during domestication and diffusion

We resequenced to high depth (24-53X, with median depth of 29X) 31 out-crossed maize landraces from six geographical regions spanning the pre-Columbian range of maize cultivation (Figure 1) and four *parviglumis* individuals from a single population located in the Balsas River Valley in Mexico.

We first estimated historical changes in effective population size (N_e) in maize and *parviglumis* using the multiple sequentially Markovian coalescent (MSMC) [30]. Our analysis revealed a sustained reduction in N_e in all maize populations that commenced between $\sim 20,000 - 10,000 BP$ and endured until the recent past ($\sim 1,100 - 2,400 BP$; Figure 1), suggesting that recovery from domestication post-dated expansion of maize across the Americas. While the timing of the onset of population size reduction varied among runs, timing of the recovery was stable (Figure S1A-B). In contrast to our results in maize, *parviglumis* showed an increase in N_e concurrent with the onset of maize domestication, also lasting until the recent past ($\sim 1,200 - 1,800 BP$). To determine if linked selection associated with domestication could bias estimates of N_e in our MSMC analysis (see [31]) we masked previously identified domestication candidates [24] and observed nearly identical results (Figure S1C).

After domestication, maize spread quickly across the Americas, reaching the full extent of its pre-Columbian distribution by $5,300 - 4,500$ BP [25, 32]. Genome-wide levels of heterozygosity across our maize samples are consistent with a history of serial founder effects, showing a strong negative correlation ($R^2 = 0.3636, p = 0.0004$; Figure 1) with distance from the center of maize domestication in the Balsas River Basin; analysis of genotyping data from more than 3,500 maize landraces shows a similar result (Figure S1).

In addition to allowing inferences regarding species-wide demographic change, our sampling allows us to investigate the history of individual maize populations. Most notably, we observe a distinctly stronger bottleneck in maize from the Andean highlands of South America (Figure 1B-C). Andean landraces exhibit a paucity of low frequency alleles and a greater proportion of derived homozygous alleles compared to other populations (Figure S2), results consistent with an extreme founder event. To provide greater resolution regarding the timing of this founder event, we assessed population genetic evidence for signs of recent inbreeding. Inbreeding coefficients in Andean samples are 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 can 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 are limited to those shorter than $0.5cM$, a length of homozygosity that would be associated with more ancient founder events, Andean samples demonstrate significantly more cumulative ROH genome-wide when compared to all ($p < 0.05$, Wilcoxon test) but the South American lowland population ($p = 0.214$, Wilcoxon test), suggesting an extreme founder event may have occurred in the Andes during initial colonization.

2.2 Introgression from wild maize in the Guatemalan and Mexican highlands

Putatively adaptive introgression from the wild teosinte taxon *Zea mays* ssp. *mexicana* (hereafter, *mexicana*) has previously been observed in maize from the highlands of Mexico [6]. Our broad sampling of maize allowed us to investigate whether introgressed *mexicana* haplotypes have spread to highland maize populations outside of Mexico, potentially playing a role in highland adaptation in other regions. In order to test this hypothesis, we calculated Patterson's D statistic [33] across all maize populations. All individuals from both the Mexican and Guatamalan highlands exhibited highly significant evidence for shared ancestry with *mexicana* (Figure S4A). Maize from the southwestern US also showed more limited evidence of introgression, consistent with findings from ancient DNA suggesting that *mexicana* haplotypes may be found outside of Mesoamerica [34]. In contrast, the distribution of z-scores for South American populations overlapped zero, providing no evidence for *mexicana* haplotypes in this region.

We localized introgression to chromosomal regions through genome-wide calculation of the $\hat{f}_d$ statistic [35]. Megabase-scale regions of introgression were identified in both Mexican and Guatemalan highland populations that correspond to those reported by [6] on chromosomes 4 and 6 (Figure 2; Figure S4). On chromosome 3, a large, previously unidentified region of introgression can be found in the Mexican and southwestern US highlands (Figure 2; Figure S4). This region overlaps a putative chromosomal inversion associated with flowering time in the maize Nested Association Mapping (NAM) populations [36] and may be an example of *mexicana* contribution to modern temperate maize.

2.3 The influence of demography on accumulation of deleterious alleles

Population-specific changes in historical N_e should influence genome-wide patterns of deleterious variants due to the association between N_e and the efficiency of purifying selection [11]. Introgression from a species with substantially different N_e may also result in deviation at introgressed sites from the genome-wide average of deleterious variants per basepair [10]. Below we evaluate the effects of the major demographic events during the pre-Columbian history of maize on genome-wide patterns of deleterious alleles.

2.3.1 The effect of domestication on deleterious alleles

We first compared the count of deleterious sites 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, and thus best reflects the effects of domestication alone. After identifying putatively deleterious mutations using Genomic Evolutionary Rate Profiling (GERP) [37], we calculated the number of derived deleterious alleles per genome under both an additive and a recessive model across four categories of deleterious sites according to mutation severity (see Methods for details). Maize showed significantly more deleterious alleles than teosinte (Figure 3) under both additive ($< 10\%$ more; $p = 0.0079$, Wilcoxon test) and recessive ($< 20 - 30\%$ more; $p = 0.0079$) models across all deleterious GERP categories (Figure S5). 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 demography including a domestication bottleneck (Figure S6; [8]). GERP load (GERP score × 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 [13, 15, 38, 39], thus appears to have a higher burden of deleterious alleles compared to its wild progenitor *parviglumis*.

While the elevated burden of deleterious alleles we observe in maize relative to *parviglumis* may be driven primarily by the demographic effects of the domestication bottleneck, positive selection on causal variants underlying domestication phenotypes could also fix nearby deleterious variants through genetic hitchhiking, which would result in a higher number of deleterious alleles within and near domestication loci [38, 40]. To test this hypothesis, we assessed the distribution of deleterious alleles in and near ($5kb$ upstream and downstream) 420 domestication candidate genes [24], 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 S7), 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 an extreme founder event in the Andes on mutation burden of deleterious alleles. The drastic reduction in N_e in the Andes could potentially result in deleterious variants segregating in a nearly neutral fashion in this population. While we observe no significant difference between the number of deleterious alleles in individuals from the Andes versus those from other populations (Figure 3) under a completely additive model, under a recessive model maize from the Andes contains significantly more derived deleterious alleles than any other population (Figure 3; Figure S5 ; $p < 0.02$, Wilcoxon test); this difference becomes more extreme when considering more severe (*i.e.*, higher GERP scores) mutations. The extreme founder event accompanying colonization of the Andes 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 S8; Figure S6), a comparable result to the differences observed in load between maize and *parviglumis*.

The effect of introgression on prevalence of deleterious alleles. Highly variable rates of *mexicana* introgression were detected across our landrace populations. 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 separate its effect from introgression due to the positive correlations observed between gene density and both introgression ($p = 3.48e - 08$) and deleterious alleles ($p \approx 0$). When identifying deleterious sites under both additive and recessive models, we found a strong negative correlation with introgression ($p \approx 0$ under both models). These findings indicate that introgression from *mexicana* has resulted in a non-random distribution of deleterious alleles across the genome, likely reflecting 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, 41, 42]. It is likely, however, that the demographic history of domesticates is more complex than a simple bottleneck followed by recovery. 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 [43]. By characterizing maize demography from domestication through initial expansion, we provide both a more complete history of maize and a blueprint of how demography has influenced the prevalence of deleterious variants across populations and individual genomes.

3.1 Historical changes in maize population size

Our analysis of maize demography provides two primary insights regarding historical variation in maize effective population size. First, the domestication bottleneck was a protracted process spanning several millennia and lasting until the recent past. Second, like other species with a history of expansion, maize has been subject to serial founder effects with genetic diversity declining in the crop with increasing distance from its center of origin.

Previous maize demography modeling has suggested that domestication bottleneck population size and duration had a ratio between $\approx 2.5 : 1$ and $\approx 5 : 1$, but little statistical support was found for specific estimates of these individual parameters [23, 28]. An alternative model based on full-genome data assumed an instantaneous bottleneck followed by immediate, exponential growth in maize and found N_e at the domestication bottleneck was 5% of the ancestral teosinte population [2]. In comparison to earlier work, our MSMC analyses suggest a much longer $\approx 9,000$-generation bottleneck. Consistent with this estimate, a recently sequenced series of maize samples from various strata in the Tularosa Cave shows a decline in diversity spanning thousands of years and lasting until the recent past, suggesting the potential for a protracted bottleneck [34]. Analysis of the domestication bottleneck in African rice, inferred using an approach similar to MSMC, identified a similarly duration of $\approx 10,000$ generations [44], suggesting demographic bottlenecks during crop evolution may have generally occurred over substantial periods of time.

In striking contrast to the protracted 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* [2, 23, 45], 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 [22, 46]. While our *parviglumis* samples are drawn from a single population in the Balsas, these data collectively suggest *parviglumis* has experienced expansion over the last several millennia.

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 genetic diversity. During maize range expansion, serial founder effects would have occurred if seed carried to each successive colonized region during human migration was limited to a sample of whole infructescences that contained a fraction of the origin's maize diversity [29]. Consistent with serial founder effects, a correlation between geographic and genetic distance has been observed in maize landraces [47, 48], but this was previously attributed to genetic drift due to isolation by distance (IBD). However, genetic drift in combination with the steady decline in heterozygosity we observe with increasing distance from the maize center of origin suggests the framework of serial founder effects may broadly explain patterns of diversity in maize landraces. Deviations we observe in our data from a simple correlation between distance and diversity are potentially due to long-range dispersal events and localized geographic factors [29]. Neutral expectations of allele frequencies across populations under serial founder effects differ substantially from those predicted under equilibrium conditions [16]. Theory incorporating both demography and selection has been developed that predicts the probability that an allele from the origin survives a series of founder effects and reaches high frequency once an expansion is complete [16]. Studies attempting to identify loci underlying crop adaptation during post-domestication expansion may therefore more accurately compare empirical data to theoretical expectations under a serial founder effects demography in order to identify extreme changes in allele frequency driven by selection. Many of the world's crops have experienced such histories of expansion and may be most correctly understood within this theoretical framework.

While a demography of serial founder effects partially explains the variation in diversity across maize landraces, there are deviations from this model. For example, diversity in the Andean population of maize appears lower than would be expected based on its distance from the Balsas alone. Our results modeling changes in effective population size, gauging diversity, and calculating cumulative ROH suggest Andean maize experienced 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 relative 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.

3.2 The prevalence of gene flow during maize diffusion

Increasingly, range-wide analyses of crops and their wild relatives are revealing that crops have received gene flow during post-domestication expansion from newly sympatric populations of their progenitor taxa and closely related species [51, 52, 53]. Consistent with previous results from genotyping data [6], 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. More limited support is also found for *mexicana* introgression in the southwestern US, particularly when looking at specific chromosomal regions such as a putative inversion polymorphism on chromosome 3. These results confirm previous findings suggesting maize from the highlands of Mexico originally colonized the southwestern US [34]. The more limited signal of *mexicana* introgression here may be due to subsequent, extensive gene flow from lowland maize as suggested by [34]. 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 (see [49]).

3.3 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 [11, 13, 19] and introgression [10] account for patterns of deleterious variants. 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 [54] and that complementation of deleterious alleles likely underlies the heterosis observed in hybrid breeding programs [54, 55]. 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 with newer, singleton diversity 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 the prevalence of deleterious alleles (but see [56] 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 that a higher burden of deleterious alleles is found in domesticates compared to their wild progenitors[13, 38, 40, 57, 58]. Consistent with these previous studies, we detect an excess of deleterious alleles in maize relative to *parviglumis*, 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

[13, 57]. While we do find substantial evidence in support of the former in maize, we see very little data that support the latter. A cost is also detected for the diffusion of maize away from its center of origin, particularly in its colonization of the Andes. The increase in deleterious alleles in the Andes appears greater than would be expected due to "expansion load" [59] alone and may be symptomatic of the extreme founder event we propose above.

Much more dramatic differences in the number of deleterious alleles are observed between maize and *parviglumis* and non-Andean and Andean maize under a recessive model than an additive model. This trend may indicate that the bulk of deleterious alleles in maize are recessive and heterozygous sites do not contribute meaningfully to a reduction in individual fitness. Previous work in human populations has shown that the majority of deleterious mutations are recessive or partially recessive [60] and data from knock-out mutations in yeast have revealed that large-effect mutations tend to be more recessive [61]. Likewise, when [62] tested the hypothesis that deleterious mutations generally tend to be recessive, they discovered a dominance coefficient of mildly deleterious mutations of 0.25 (partially recessive) across a number of organisms. In maize, Yang *et al.* [54] find that most deleterious alleles in maize are at least partially recessive and note a negative correlation between the severity of a deleterious variant and its dominance. Indeed, models incorporating very recent bottlenecks, such as those implied by our demographic modeling, show strong differences in load only under a recessive model [8].

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 has increased the overall mutation load in modern humans [10]. We have observed the opposite pattern in maize: introgression from the wild taxon *mexicana* reduced the proportion of deleterious variants in maize. This result may be due to the fact that *mexicana* has had a larger ancestral N_e than maize [27], and introgressed regions have 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. In addition to domesticates, invasive species that have recently undergone founder events followed by expansion and endangered species that have experienced precipitous declines in N_e will likely show clear genetic signs of the interplay between demography and selection. This relationship between demography and selection 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

5.1 Samples, whole genome resequencing, and read mapping

A total of 31 maize landrace accessions were obtained from the US Department of Agriculture (USDA)ś National Plant Germplasm System and through collaborators (Figure S10). 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 [63]. The Illumina HiSeq 2000 was used to generate 100-bp paired-end reads. BWA v. 0.7.5.a [64] was used to map the reads to the maize B73 reference genome v3 [65] 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 [66]. 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.

5.2 Population structure, genetic diversity and inbreeding coefficients

We first evaluated population structure by principle component analysis (PCA) with ngsCovar [67] in ngsTools [68] based on the matrix of posterior probabilities of SNP genotypes produced in ANGSD v. 0.614 [69], and then utilized NGSadmix v. 32 [70] 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 [6] was constructed with the R package phangorn [71]. The sample RIMMA0677 was not clustered with other highland samples, but embedded within lowland haplotypes (Figure S9). Thus, this sample was removed from further analyses.

The genetic diversity measures Watterson's θ and θ_π were calculated in ANGSD [69] for each population. The neutrality test statistic Tajimaś D was calculated with an empirical Bayes approach [72]. The approach was implemented in ANGSD by first estimating a global site frequency spectrum (SFS) and then calculating posterior sample allele frequencies using the global SFS as a prior. The three statistics were calculated across the genome using a 10-kb non-overlapping sliding window approach.

Inbreeding coefficients of each individual were estimated with ngsF [73], which takes the uncertainty of genotype assignments into account 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.

5.3 Demography of maize domestication and diffusion

The recently-developed method MSMC [30], which explicitly models ancestral relationships under recombination and mutation, was utilized to infer effective population size changes in both *parviglumis* and maize populations. SNPs were called via HaplotypeCaller and filtered via VariantFiltration in GATK [66] across all the 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 [74] was used to further filter SNPs to include only bi-allelic sites. SNPs were phased using BEAGLE v. 4.0 [75] with SNP data from the maize HapMap2 panel [56] used as a reference. Only sites with depth between half and twice of the mean depth were chosen for the analyses. In addition, since the maize genome contains over 80 % repetitive regions [65], we used Heng Li's software SNPable (http://lh3lh3.users.sourceforge.net/snpable.shtml) to mask out genomic regions in which short sequencing reads were not uniquely mapped. The mappability mask file for the MSMC analysis 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 [65]. Sites with the majority of overlapping 100-mers being 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.

Furthermore, the percentage of heterozygous sites was counted for each maize landrace sample. 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 Balsas Valley (latitude: 18.099138; longitude: -100.243303) was examined with linear regression in ggplot2 [76]. Geographical distance in kilometers was calculated based on great circle distance using the haversine [18], which transforms the longitudes and latitudes of two points into kilometers with the radius of the Earth assumed to be 6371 *km*. The correlation between percentage of heterozygous sites and distance away from domestication center was also explored in the SeeDs data set.

5.4 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/) [77] 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 [78]. The total length of ROHs for each genome was summarized for all ROHS, ROHs greater than 1 cM, and ROHs less than 0.5 cM.

5.5 Detection of Introgression

To assess per-genome evidence of population admixture between maize landraces and teosinte, we performed the ABBA-BABA test [**?**], also known as the D statistic. The D statistic was calculated using ANGSD [69]. We counted sites at which *Tripsacum dactyloides* was homozygous and defined the respective allele as ancestral (A). The consensus sequence of *T. dactyloides* was generated in ANGSD [69]. We then required that the individual specified in the third column had a derived homozygous genotype (B). We also required that, of the individuals specified in the first and second columns, one be homozygous for the ancestral allele (A) and the other be homozygous for derived allele (B). This left two possible configurations named ABBA and BABA, respectively. The D statistic is then:

$$D = (n_{ABBA} - n_{BABA})/(n_{ABBA} + n_{BABA}) \tag{1}$$

We used *mexicana* (TIL25 in maize HapMap2) as the third column species. One accession from the Mexican Lowland population was randomly sampled as the second column taxon. We chose all samples from all other populations except the Mexican Lowland individuals as the first column taxon. The D statistic was calculated in a $1kb$-block and then jackknife bootstrapping was conducted to estimate the significance of the D statistic.

In addition, we calculated the statistic $\hat{f}_d$ [35] on a site-by-site basis in order to identify introgressed genomic regions. $\hat{f}_d$ is a further development of the statistic f (the fraction of a genome shared through introgression), which was first proposed by [79], and is related to the ABBA-BABA statistic. Given three populations and an outgroup with the relationship $(((P_1, P_2), P_3), O)$, $\hat{f}_d$ is calculated as follows:

$$\hat{f}_d = \frac{S(P_1, P_2, P_3, O)}{S(P_1, P_D, P_D, O)} \tag{2}$$

In this equation, S is the difference between sums of ABBA and BABA patterns. P_D in the denominator represents the donor population (P_2 or P_3), with the higher frequency of derived alleles for each site independently. $\hat{f}_d$ is estimated by comparing the observed value of S to a value estimated under a scenario of complete introgression between P_2 and P_3. This definition will restrict $\hat{f}_d$ to a range between 0 and 1, as $\hat{f}_d$ is only calculated for sites with positive D.

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 goal here is to identify introgressed chromosomal regions from *mexicana* into five test populations. 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 [80]. 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 minor adjustments with a custom script.

5.6 Estimating burden of deleterious mutations

We estimated the individual burden of deleterious alleles based on GERP scores for each site in the maize genome [81], which reflect the strength of purifying selection. Scores above 0 may be interpreted as subject to the historical action of purifying selection, and thus mutations at such sites are more likely to be 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 [81] 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 not ancestral following [8]. We divided SNPs where the B73 allele is ancestral into bins of 1% derived allele frequency based on maize HapMap3 [82] and used this frequency distribution to estimate the posterior probability of GERP scores for SNPs where B73 allele is derived. Our estimate of burden of deleterious alleles in teosinte and maize was based on the corrected GERP scores.

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. This burden was calculated under a recessive model as the sum of GERP scores multiplied by one for each deleterious homozygous site. In addition, for each individual we also summarized A) the number of deleterious alleles ($1 \times$ HET $+ 2 \times$ HOM), the additive model and B) the number of deleterious homozygous sites, the recessive model. 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 ≤ 0, nearly neutral; $0 <$ GERP ≤ 2, slightly deleterious; $2 <$ GERP ≤ 4, moderately deleterious; GERP > 4, strongly deleterious) and recapitulated the above mentioned statistics.

5.7 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) and will be made available upon acceptance of the manuscript.

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(5866):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(7):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(15):6044–6049, 2011.

[5] H. Jabran Zahid, Erick Robinson, and Robert L. Kelly. Agriculture, population growth, and statistical analysis of the radiocarbon record. *Proceedings of the National Academy of Sciences*, 113(4):931–935, 2016.

[6] 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 genetics*, 2013.

[7] 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(7481):43–49, 01 2014.

[8] 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(3):220–224, 2014.

[9] 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(2):126–131, 2015.

[10] Kelley Harris and Rasmus Nielsen. The genetic cost of neanderthal introgression. *Genetics*, 203(2):881–891, 2016.

[11] 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(4):421–436, 2014.

[12] M. Zhang, L. Zhou, R. Bawa, H. Suren, and J.A. Holliday. Recombination rate variation, hitchhiking, and demographic history shape deleterious load in poplar. *Molecular Biology and Evolution*, 2016.

[13] 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(1):152–157, 2016.

[14] 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.

[15] 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.

[16] Montgomery Slatkin and Laurent Excoffier. Serial founder effects during range expansion: A spatial analog of genetic drift. *Genetics*, 191(1):171–181, 2012.

[17] 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(2):148–164, 1997.

[18] 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(44):15942–15947, 2005.

[19] 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*, 2015.

[20] Ivan Juric, Simon Aeschbacher, and Graham Coop. The strength of selection against neanderthal introgression. *PLOS Genetics*, 12(11):1–25, 11 2016.

[21] 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(9):6080–6084, 2002.

[22] 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(29):11874–11881, 2007.

[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(5726):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(7):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(50):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(5):1755–1759, 2012.

[27] Jeffrey Ross-Ibarra, Maud Tenaillon, and Brandon S Gaut. Historical divergence and gene flow in the genus zea. *Genetics*, 181(4):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(7):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(8):e12060, 2010.

[30] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 2014.

[31] Daniel R. Schrider, Alexander G. Shanku, and Andrew D. Kern. Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 2016.

[32] Sonia Zarrillo, Deborah M Pearsall, J Scott Raymond, Mary Ann Tisdale, and Dugane J Quon. Directly dated starch residues document early formative maize (zea mays l.) in tropical ecuador. *Proceedings of the National Academy of Sciences*, 105(13):5006–5011, 2008.

[33] Eric Y. Durand, Nick Patterson, David Reich, and Montgomery Slatkin. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.

[34] 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. *bioRxiv*, page 013540, 2015.

[35] 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(1):244–257, 2015.

[36] J Alberto Romero Navarro, Martha Wilcox, 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*, 2017.

[37] 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(7):901–913, 2005.

[38] 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*, page msv106, 2015.

[39] Torsten Günther and Karl J Schmid. Deleterious amino acid polymorphisms in arabidopsis thaliana and rice. *Theoretical and Applied Genetics*, 121(1):157–168, 2010.

[40] 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. *bioRxiv*, page 033175, 2016.

[41] 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(3):875–888, 2007.

[42] 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(12):1053–1059, 2010.

[43] Genomics and the contrasting dynamics of annual and perennial domestication. *Trends in Genetics*, 31(12):709 – 719, 2015.

[44] 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(9):1083–1088, 09 2016.

[45] Jeffrey Ross-Ibarra, Maud Tenaillon, and Brandon S. Gaut. Historical divergence and gene flow in the genus zea. *Genetics*, 181(4):1399–1413, 2009.

[46] A. Correa-Metrio, S. Lozano-Garca, S. Xelhuantzi-Lpez, S. Sosa-Njera, and S. E. Metcalfe. Vegetation in western central mexico during the last 50 000 years: modern analogs and climate in the zacapu basin. *Journal of Quaternary Science*, 27(5):509–518, 2012.

[47] 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(10):1240–1253, 2008.

[48] 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(3):1088–1092, 2011.

[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*, 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] 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):1, 2015.

[52] 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(5):562–570, 2016.

[53] 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*, page msw274, 2016.

[54] 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.

[55] Sofiane Mezmouk and Jeffrey Ross-Ibarra. The pattern and distribution of deleterious mutations in maize. *G3: Genes— Genomes— Genetics*, 4(1):163–171, 2014.

[56] 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(7):803–807, 2012.

[57] 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(3):126–131, 2006.

[58] 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(52):E5661–E5669, 2014.

[59] Stephan Peischl, Isabelle Dupanloup, Mark Kirkpatrick, and Laurent Excoffier. On the accumulation of deleterious mutations during range expansions. *Molecular ecology*, 22(24):5972–5982, 2013.

[60] 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(7):e1002655, 2012.

[61] Aneil F Agrawal and Michael C Whitlock. Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics*, 187(2):553–566, 2011.

[62] Federico Manna, Guillaume Martin, and Thomas Lenormand. Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics*, 189(3):923–937, 2011.

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

[64] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrowswheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[65] 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(5956):1112–1115, 2009.

[66] 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(5):491–498, 2011.

[67] 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(3):979–992, 2013.

[68] Matteo Fumagalli, Filipe G Vieira, Tyler Linderoth, and Rasmus Nielsen. ngstools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, 30(10):1486–1487, 2014.

[69] Thorfinn S Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):356, 2014.

[70] Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, 2013.

[71] Klaus Peter Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

[72] 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(1):289, 2013.

[73] 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(11):1852–1861, 2013.

[74] 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(15):2156–2158, 2011.

[75] 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(5):1084–1097, 2007.

[76] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009.

[77] 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(3):559–575, 2007.

[78] 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(6):552–563, 2015.

[79] Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, et al. A draft sequence of the neandertal genome. *science*, 328(5979):710–722, 2010.

[80] Stéphane De Mita and Mathieu Siol. Egglib: processing, analysis and simulation tools for population genetics and genomics. *BMC genetics*, 13(1):27, 2012.

[81] 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(12):3823–3828, 2015.

[82] 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.

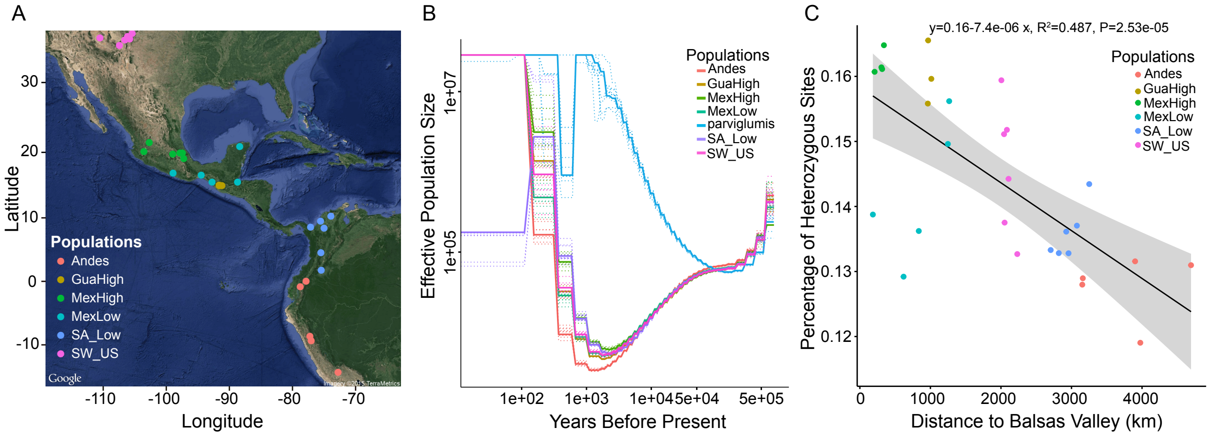

Figure 1: Maize domestication and expansion. A. Sampling locations. B. Estimates of effective population size over time. C. The percentage of polymorphic sites versus distance from the maize domestication center. Abbreviations for populations: GuaHigh, Guatemalan Highlands; MexHigh, Mexican Highlands; MexLow, Mexican Lowlands; SA_-Low, South American Lowlands; SW_US, Southwestern US Highlands.

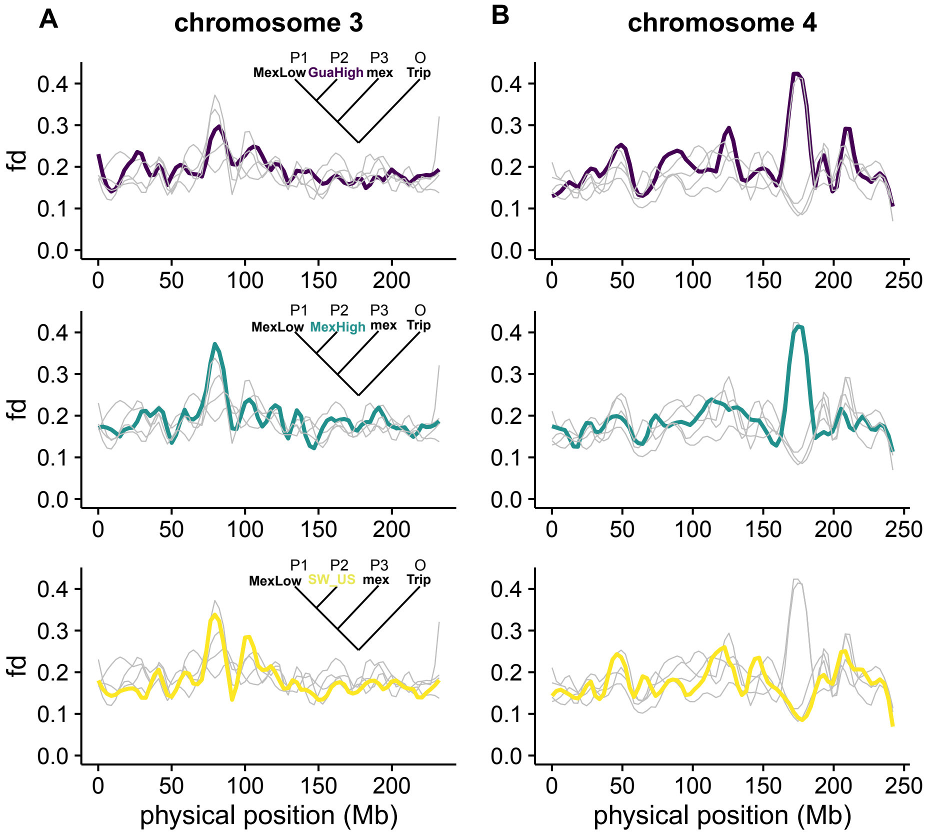

Figure 2: Loess regression of $\hat{f}_d$ in $10kb$ nonoverlapping windows across (A) chromosome 3 and (B) chromosome 4. Shown are results for three populations; other populations are drawn in grey. mex: *mexicana*; Trip: *Tripsacum*.

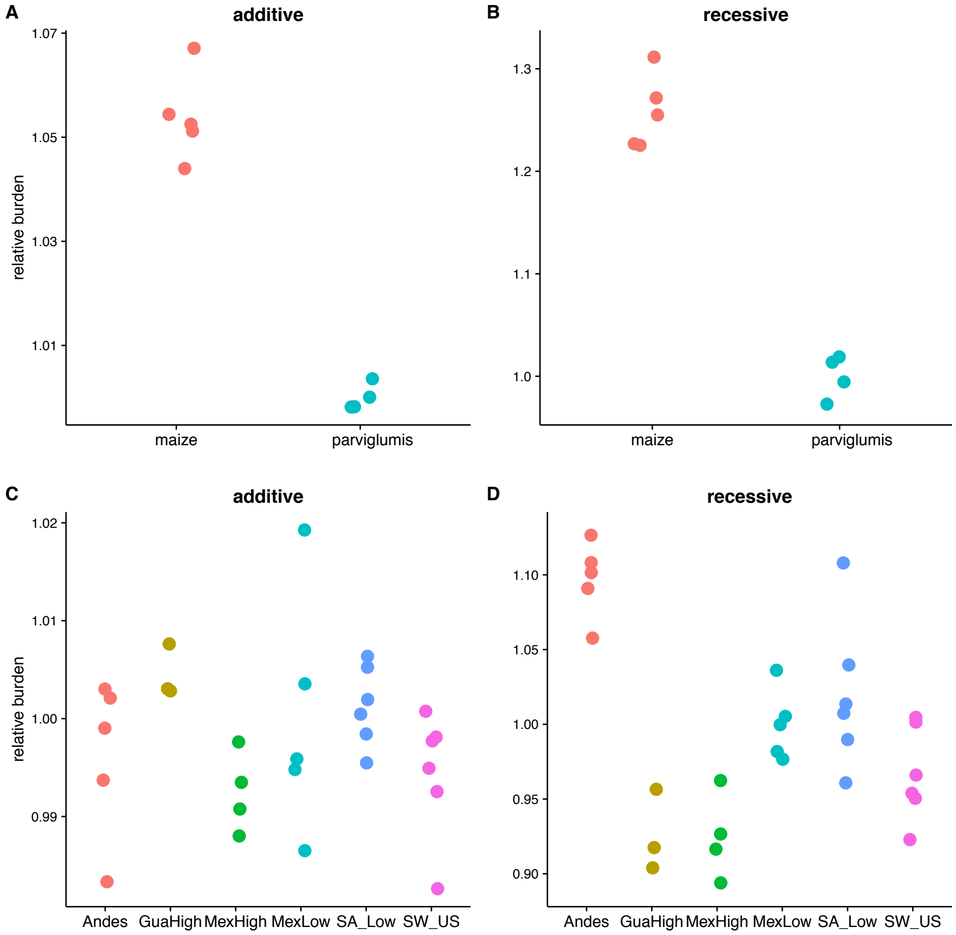

Figure 3: Comparison of counts of deleterious sites (A-B) between *parviglumis* and maize and (C-D) among maize populations. In each case an additive (A,C) and recessive (B,D) model are shown.

Supporting Information

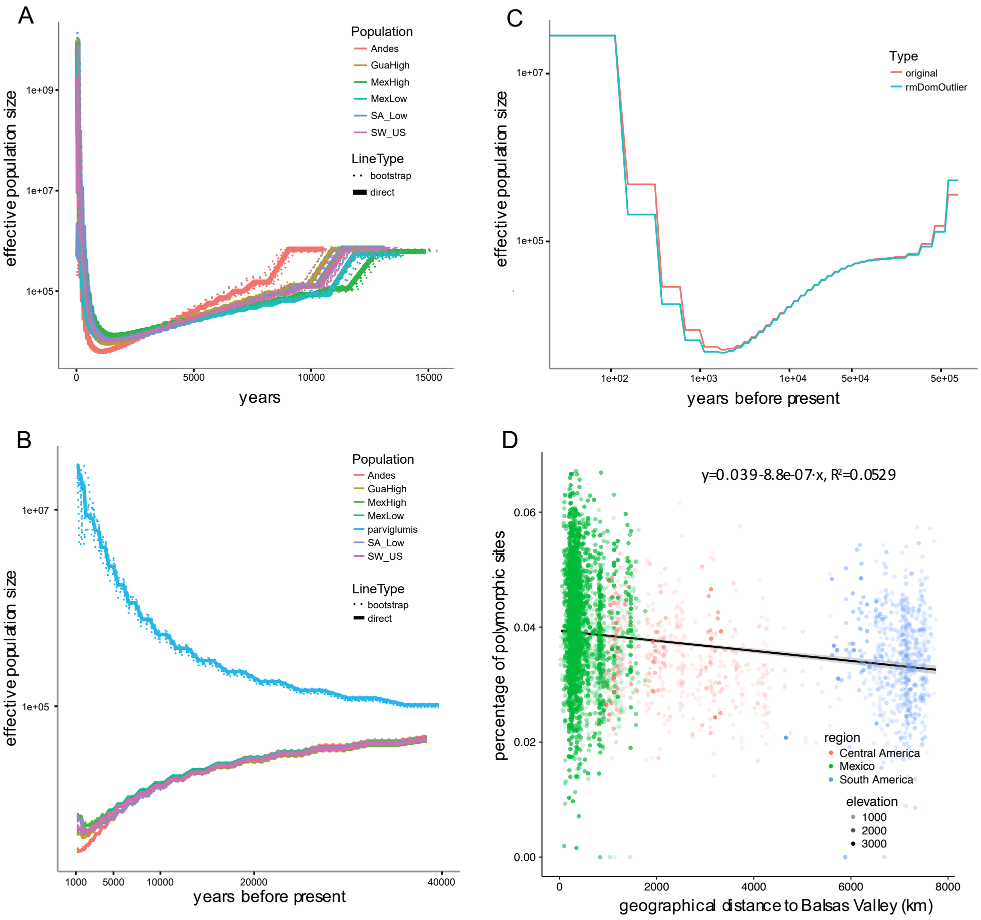

Figure S1: Demography of maize populations. A. Estimates of N_e over time among six maize populations with MSMC version 1. B. Change of N_e over 40,000 - 1,000 years BP with MSMC version 2 (x axis in linear scale). C. MSMC2 results before and after masking candidate regions under selection during domestication. D. Percentage of heterozygous sites versus distance from the Balsas Valley in 3520 samples from the SeeDs data set.

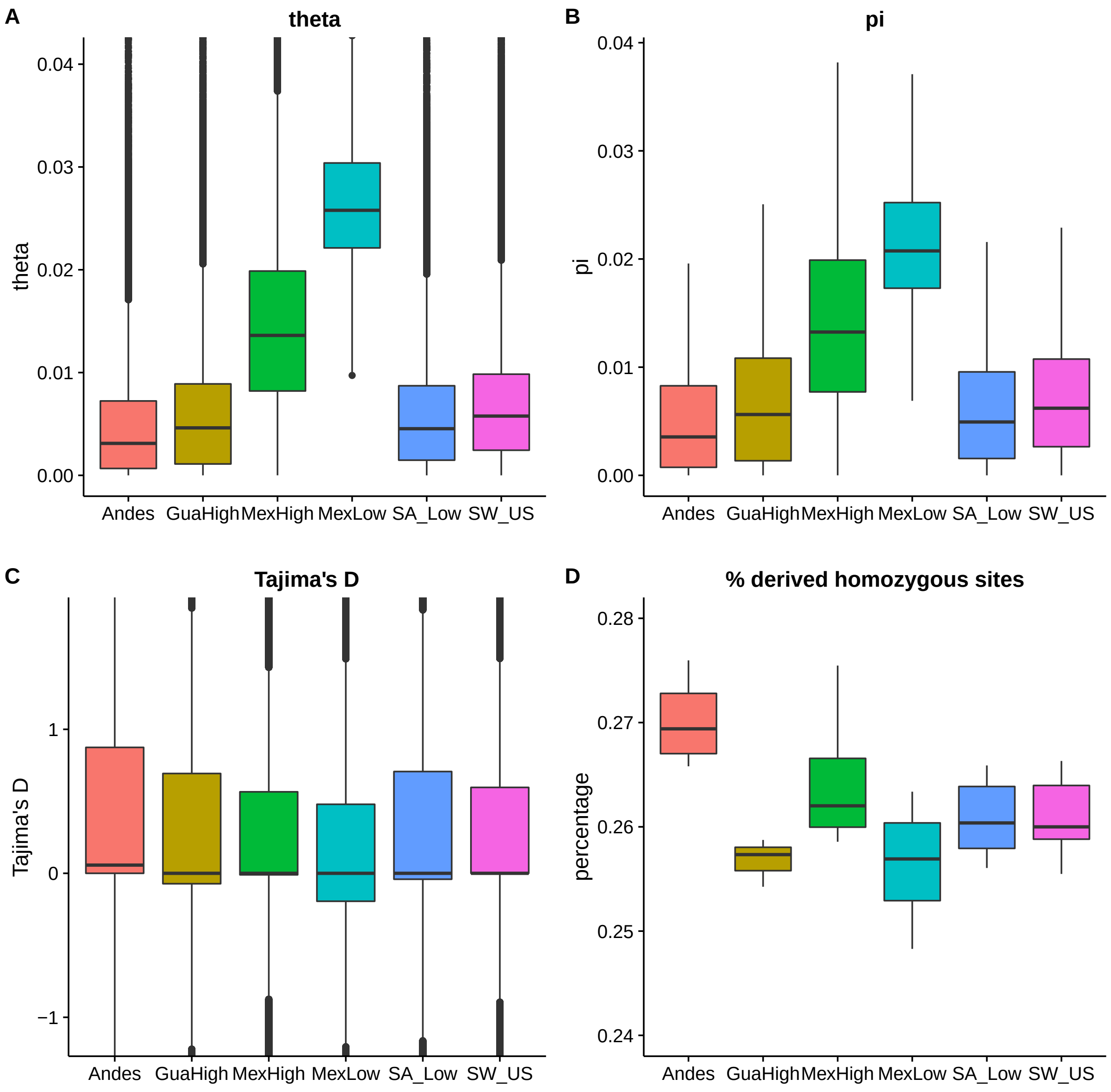

Figure S2: Boxplot of multiple population genetic statistics. Watterson's *theta* (A), θ_π)(B) and Tajima's D (C) are based on values in 10-kb non-overlapping windows across the genome. Percentage of derived homozygous sites was calculated for each individual and reported per population

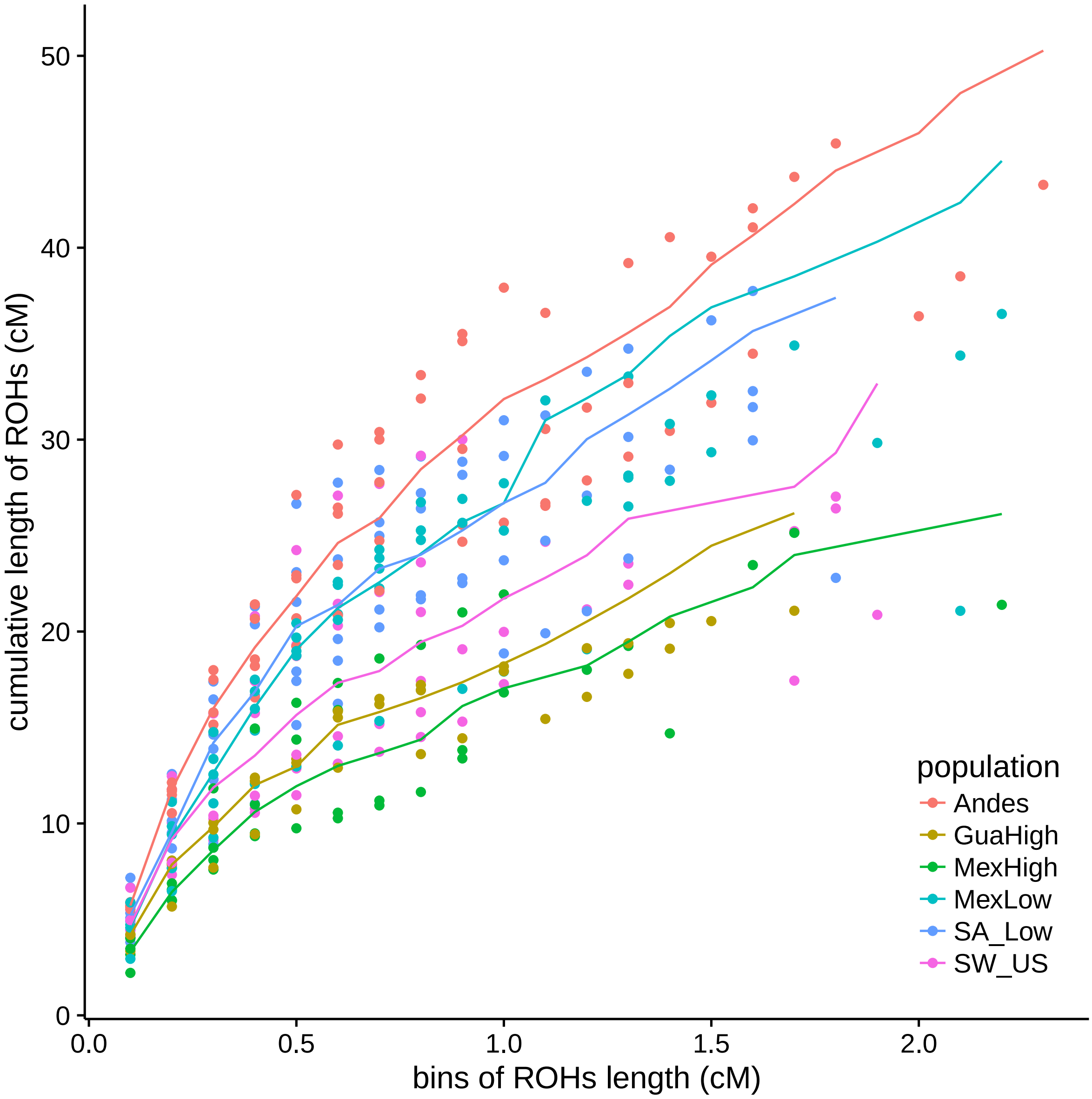

Figure S3: Cumulative length of ROHs in cM among populations. The lines indicate median level in each population. ROH: runs of homozygosity.

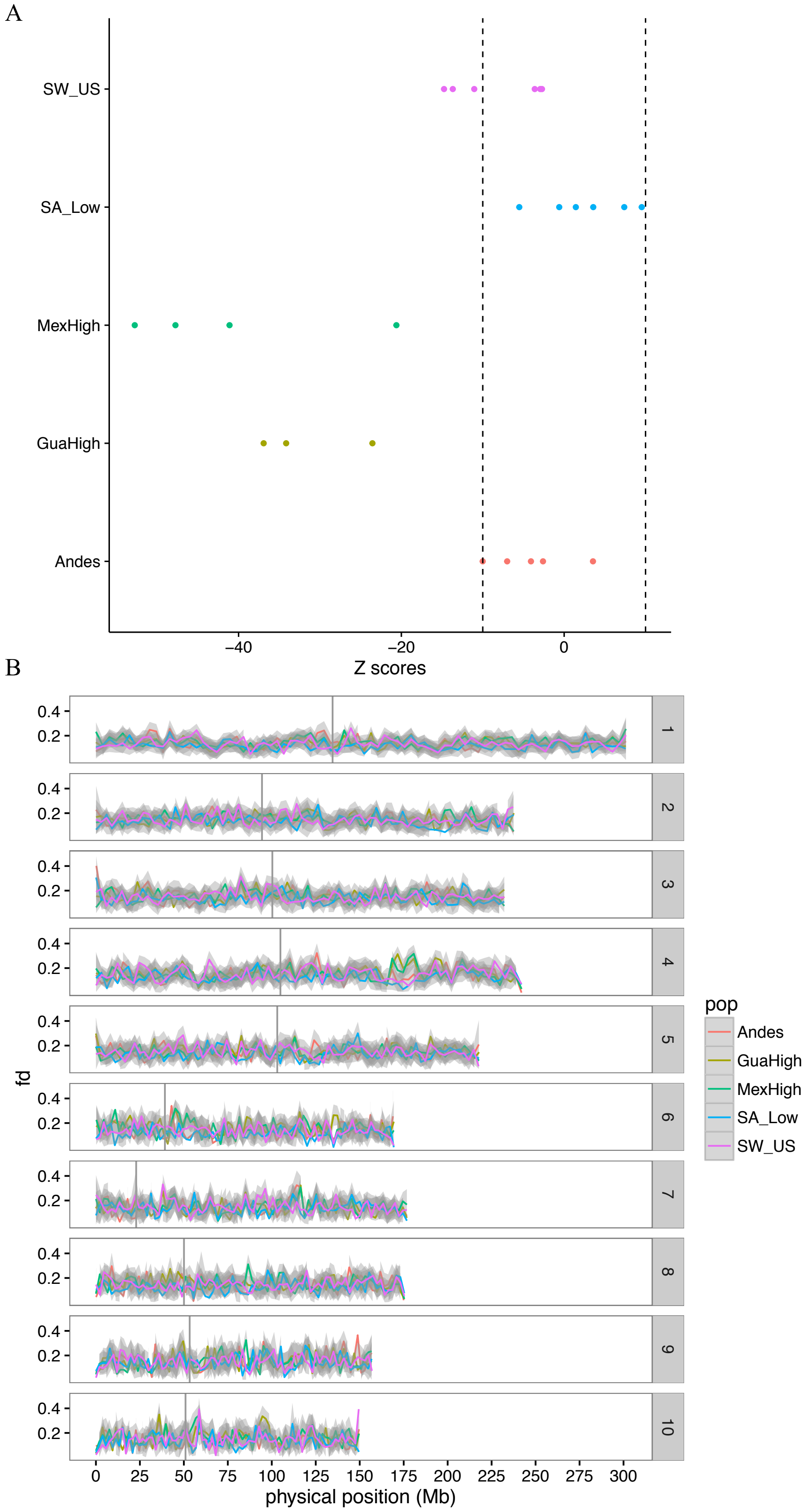

Figure S4: Evidence of Introgression from *mexicana* into maize populations. A. D statistic results. The dashed lines correspond to Z scores equal to -10 and 10. B. Loess regression of $\hat{f}_d$ in 10-kb nonoverlapping windows across all the chromosomes.

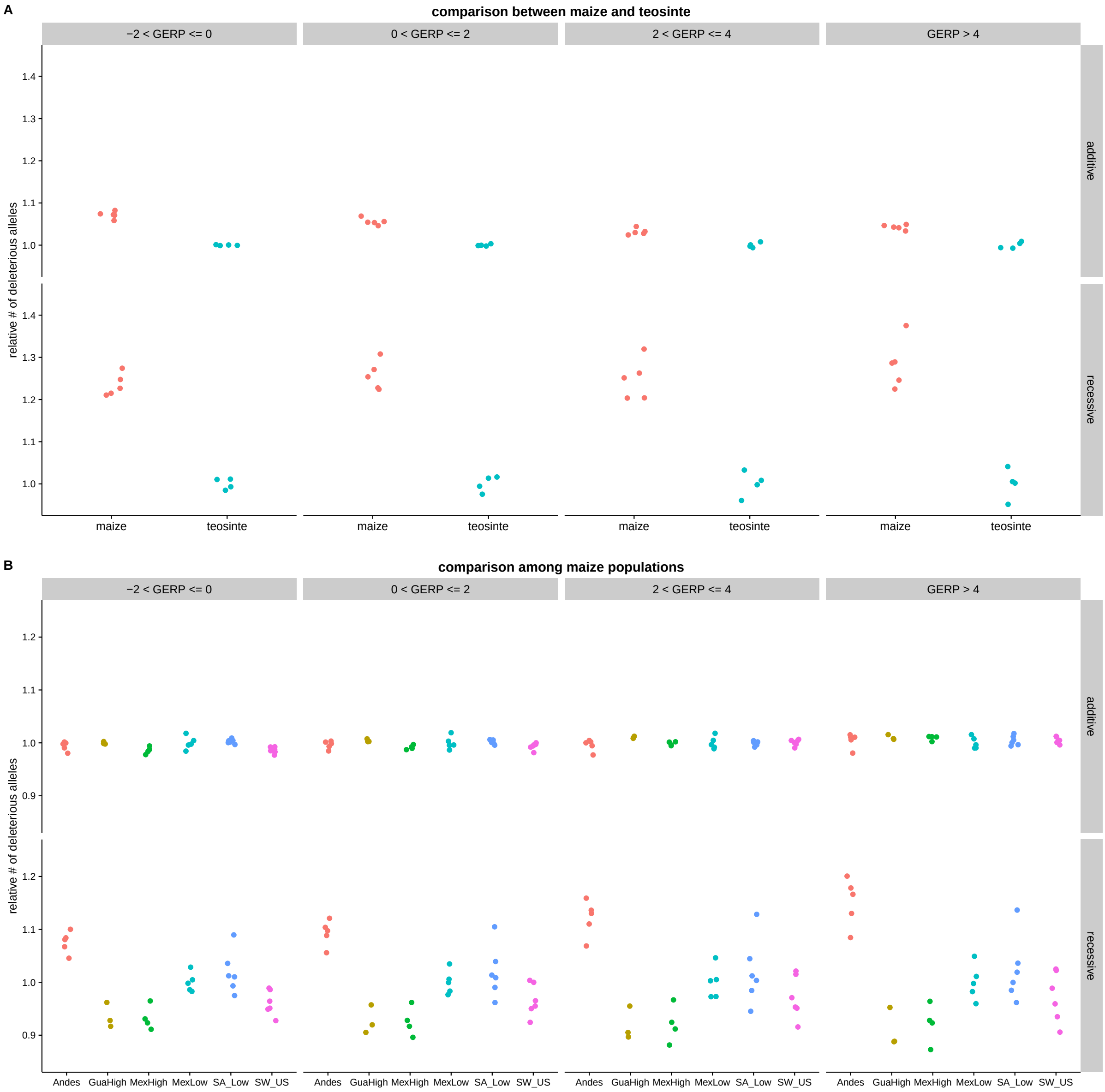

Figure S5: Relative burden of deleterious alleles under both additive and recessive models with different GERP partitions between maize and teosinte (A) and among maize populations (B).

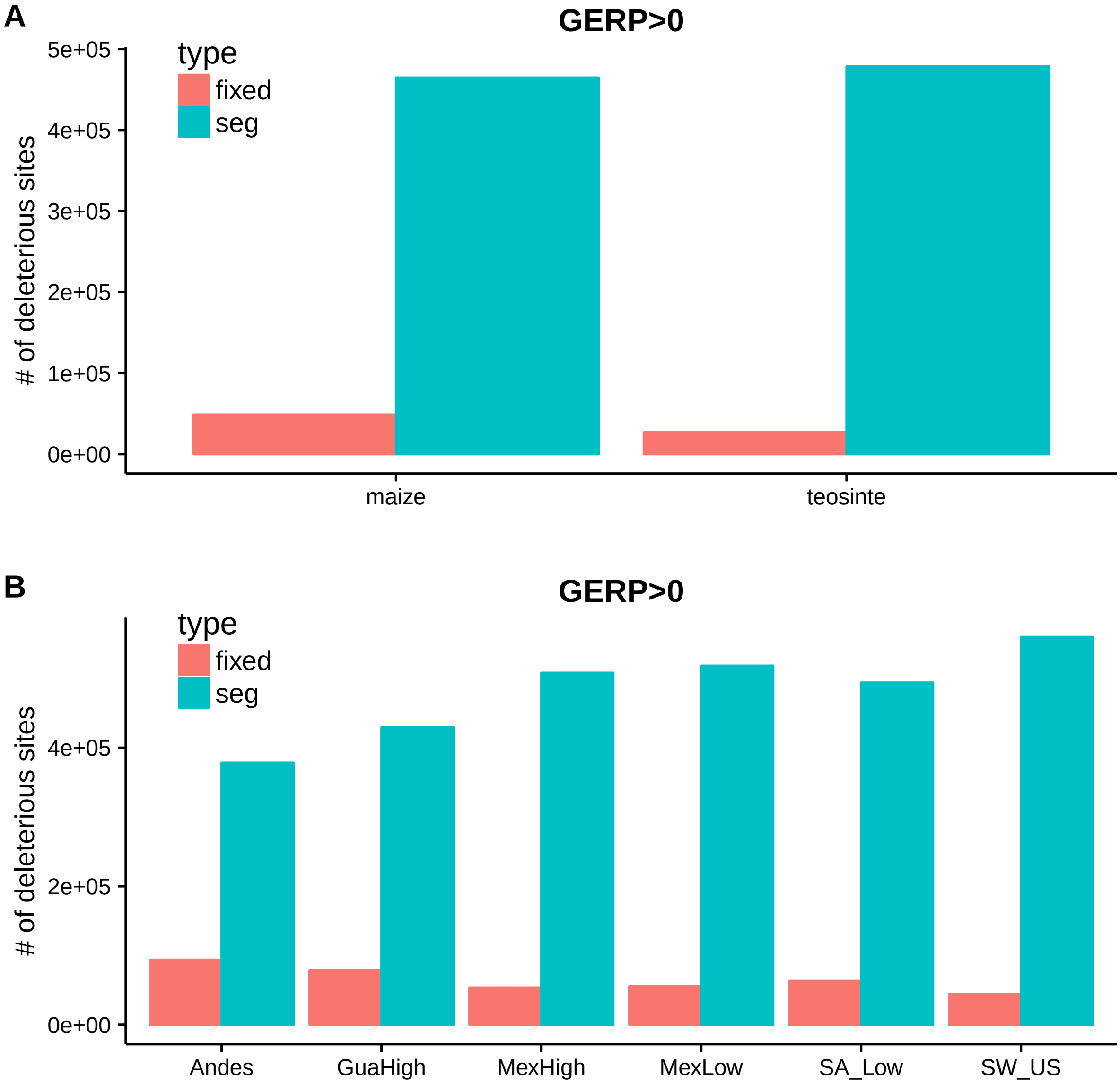

Figure S6: Histogram of counts of fixed and segregating deleterious SNPs between teosinte and maize (A) and among maize populations (B).

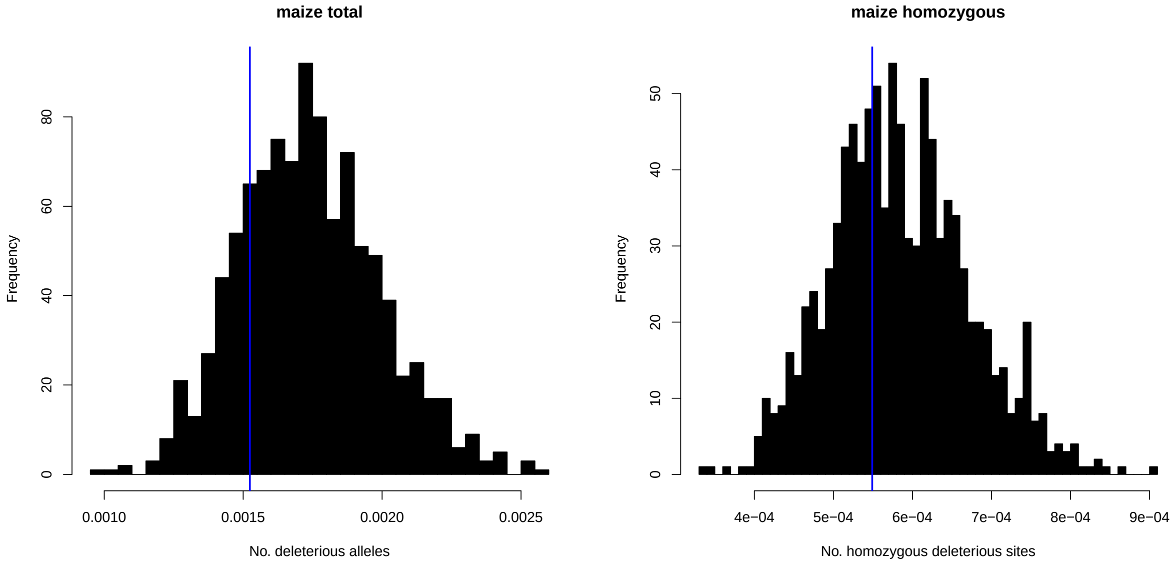

Figure S7: Distribution of number of deleterious sites per bp in 420 domestication candidate genes (indicated with blue line) against genome-wide random samples under an (A) additive model (B) recessive model.

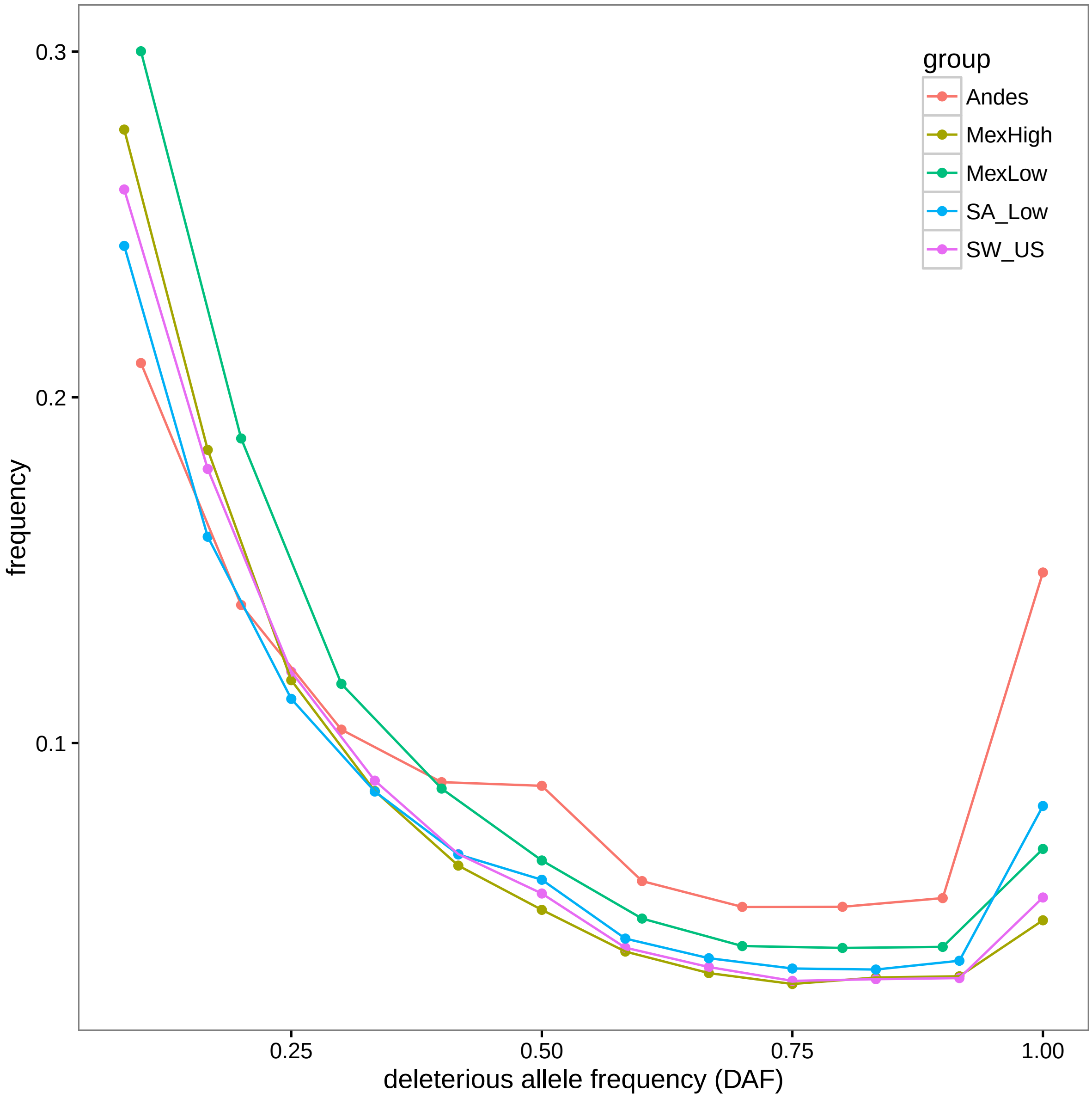

Figure S8: Site frequency spectrum of deleterious SNPs in five populations; GuaHigh is not included since the small sampling limited power for the SFS.

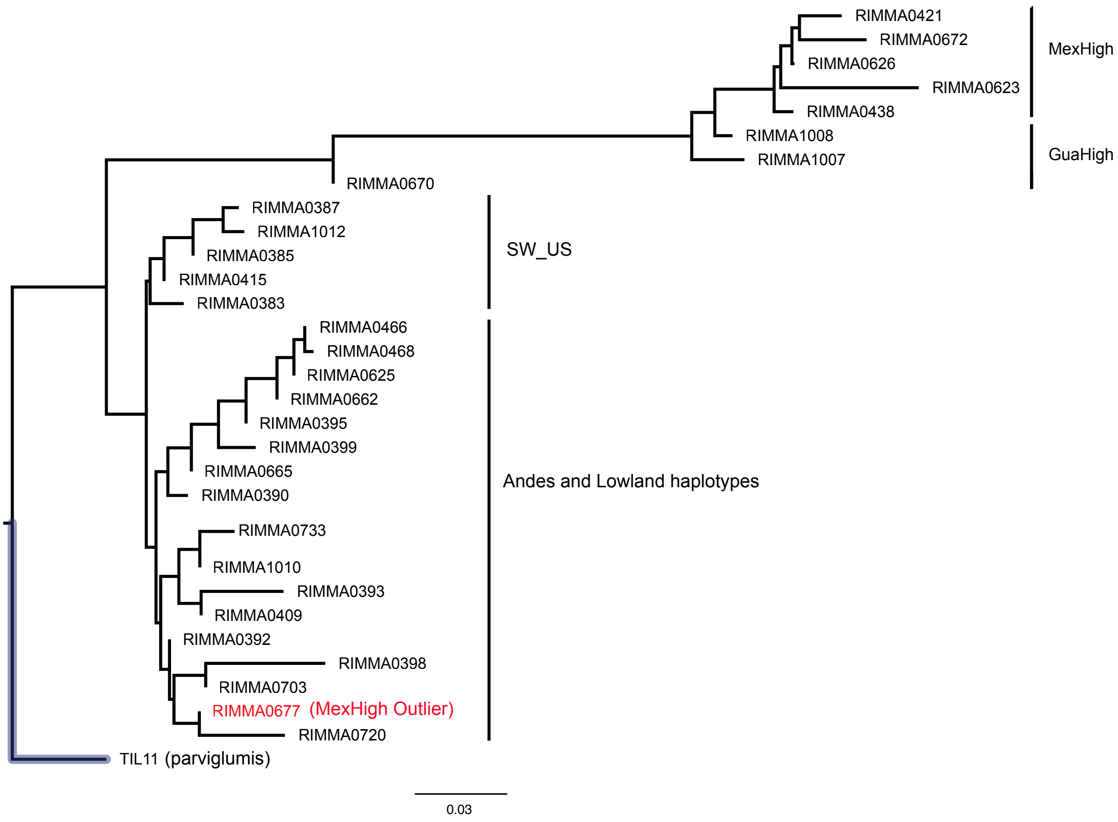

Figure S9: Neighbor Joining tree of SNPs from an inversion on chromosome 4 with a diagnostic haplotype for highland Mexican material.

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 S10: Basic information regarding the sampled maize landrace accessions. NM: New Mexico.