

Demography and linked selection in wild and domesticated maize

Timothy M. Beissinger * † ‡ Li Wang §, Kate Crosby *, Arun Durvasula *, Matthew Hufford §, and Jeffrey Ross-Ibarra * ¶

*Dept. of Plant Sciences, University of California, Davis, CA, USA, †US Department of Agriculture, Agricultural Research Service, Columbia, MO, USA, ‡Division of Plant Sciences, University of Missouri, Columbia, MO, USA, §Iowa State University, Ames, IA, USA, and ¶Genome Center and Center for population biology, University of California, Davis, CA, USA

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Unique selective and demographic forces operate on domesticated plant species. These forces interact during and after domestication to generate the patterns of DNA variability that are persistent in crop-species today. To quantify the interplay between demography and selection, we investigated genetic diversity in maize, one of the most important crops for food, feed, and fuel world-wide. We utilized Our sample included whole genome sequence data from 23 maize and 13 teosinte individualsto make inferences. We obtained a complete estimate of the population size fluctuations and other demographic parameters experienced by maize as it was domesticated from teosinte. Here, we show that maize went through a domestication bottleneck with a population size of approximately 5% that of teosinte before it experienced rapid population size expansion post-domestication. We observe that hard sweeps, specifically positive selection on new genic mutations, on genic mutations are not the primary force driving maize evolution. We find that a reduced population size during domestication decreased the efficiency of purifying selection to purge deleterious alleles from maize. However, expansion after domestication has since increased the efficiency of purifying selection to levels superior to exceeding those seen in teosinte. Our results demonstrate that particularly in domesticated species or bottlenecked species, demographic and selective history in the ancient and recent past both contribute to genetic variability that is present today, providing substantial implications for the continued improvement of domesticated species.

we need a starting paragraph on linked selection and demography. check out slotte 2014 briefings functional genomics for a great review of ideas. then we do a paragraph on why domesticates are great for this (modified version of your below)

Domesticated plant species evolve in a unique fashion compared to their wild counterparts [1]. This is a result of both the anthropomorphic nature of artificial selection on domesticates [2] as well as the demographic characteristics of the domestication bottleneck(s) that they tend to have experienced [3]. However, the complex interplay between selective pressures and demographic limitations, and the impact that this interplay has on identifying selection and understanding demography, is not fully understood. Although a large body of research that involves searching the genomes of domesticated species for evidence of positive selection exists [4–7], these studies tend to focus on identifying or mapping particular genes or regions that play an important role in phenotypic evolution. In contrast, knowledge regarding the impacts that demography and selection have on whole-genome patterns of genetic variability remains limited.

Maize represents an excellent organism to study these phenomena. Maize is a species of tremendous importance worldwide as both a staple crop [8] and as a model for understanding plant evolution [9]. Broadly speaking, archaeological and genetic studies have established that maize domestication is likely to have taken place in Mexico approximately 9,000 years bp [10,11]. Teosinte, the most recent wild ancestor to maize, remains extant throughout much of the Americas [12]. Additionally, several large-effect domestication loci [13–15] and putative domestication regions [4] have been identified. But despite all that is known about maize domes-

tication, the parameters of the domestication process remain uncertain. Specifically, the size of the maize domestication bottleneck has not been estimated independently of the bottleneck's duration, nor are there sequence-based estimates of the effective population size of modern maize. Sequence information from maize and teosinte plants may therefore be utilized to address these questions.

To that end, the objectives of our study were to 1) investigate the relative importance of different forms of selection on whole-genome variability in both maize and teosinte 2) research the impact that the domestication process has had on genetic variability in maize, and how this compares to the

FigsAndFiles/Pi_and_Tajima.png

Fig. 1. Pairwise diversity π (Pi_{A,B}) and Tajimas D were calculated inside (C,D) in 1kb windows from genic and outside nongenic regions of genes for maize (A,C) and teosinte —. In both taxa (B), a dearth of rare alleles was observed within genes relative to outside of genes, which led to higher values of Tajima's D outside of genes than inside). Shown in A :Mean and standard deviation of pi for 1 kb genic and non-genic windows of maize; B :Mean are means and one standard deviation of pi for 1 kb genic and non-genic windows of teosinte C:Histogram of Tajima's D in 1 kb windows in maize; D:Histogram of Tajima's D in 1 kb windows in teosinte.

Significance

Both selection and demographic change play important roles in shaping diversity across the genome, but clear empirical examples of the interplay of these two forces are lacking. Here we document the combined effects of demography and linked selection on genome-wide diversity in domesticated maize and its wild ancestor teosinte. We estimate that maize underwent a bottleneck of 5% of the size of the ancestral teosinte population, but that recent expansion has resulted in a maize population perhaps orders of magnitude larger than teosinte. We show that positive selection on new mutations has had relatively little effect on genetic diversity, but that selection against deleterious mutations has dramatically reduced diversity in and immediately around genes in both taxa. We find that the relative effect of selection depend qualitatively on the age of the polymorphisms evaluated: while older polymorphisms in maize show more limited effects of linked selection, new mutations instead reflect its larger current size and more efficient purifying selection. Our results demonstrate that a complete understanding of genome-wide patterns of diversity will require careful assessment of both demographic history and the effects of linked selection.

Reserved for Publication Footnotes

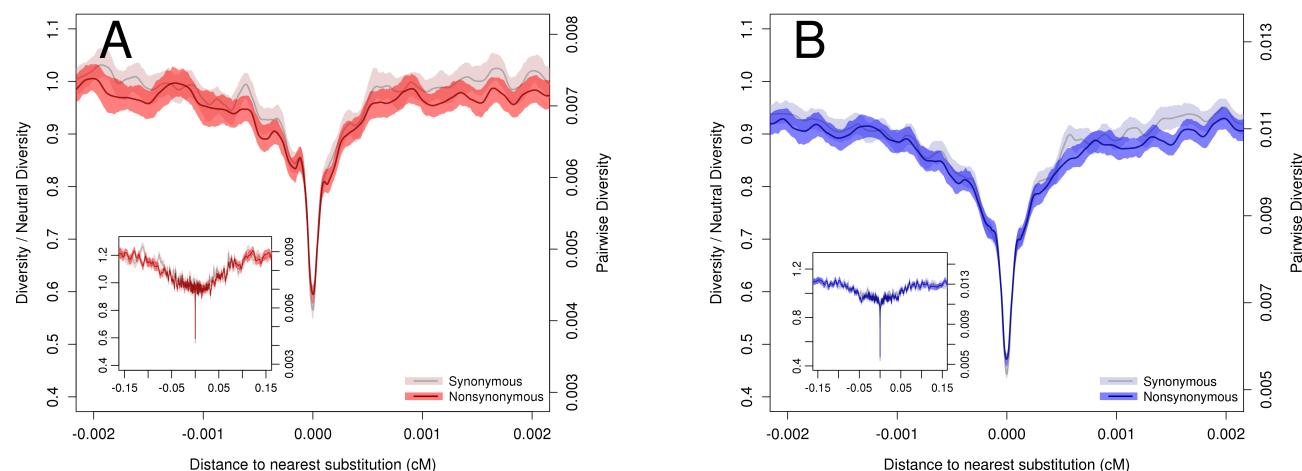


Fig. 2. Pairwise diversity surrounding synonymous and non-synonymous (missense) substitutions in **A** maize and **B** teosinte. Axes show both absolute diversity values (on right) and values relative to mean nucleotide diversity in windows $\geq 0.01\text{cM}$ from a substitution (on left). Lines depict a loess curve (span of 0.01) and shading represents bootstrap-based 95% confidence intervals. Inset plots depict a larger range on the x-axis.

impact of a different demographic history in teosinte; and 3) precisely estimate the parameters of the maize domestication bottleneck. [16].

Results

Patterns of diversity differ between genic and non-genic regions of the genome. To investigate how demography and linked selection have shaped patterns of diversity in maize and teosinte, we reanalyzed data from 23 maize and 13 teosinte genomes from the maize HapMap 2 project [16] and HapMap 3 projects [16]. We find broad differences in genic and non-genic intergenic diversity consistent with earlier results [4] (Figure 1). In maize, mean pairwise diversity (π) within genes was significantly lower than at positions site at least 5 kb away from genes (0.00668 vs. 0.00691 , $p < 2 \times 10^{-44}$). Diversity differences in teosinte are even more pronounced (0.0088 vs. 0.0115 , $p \approx 0$). Differences were also apparent in the site frequency spectrum, with mean Tajima’s D positive in genic regions in both maize (0.4) and teosinte (0.013) but negative outside of genes (-0.087 in maize and -0.25 in teosinte, $p \approx 0$ for both comparisons). These observations suggest that diversity in genes is not evolving neutrally, but instead is reduced by the impacts of selection on linked sites.

Hard sweeps do not explain diversity differences. Selection acting to increase the frequency of a new beneficial mutation will leave a signature of reduced diversity at surrounding linked sites [17]. To evaluate whether patterns of such “hard sweeps” could explain observed differences in diversity between genic and non-genic intergenic regions of the genome, we compared diversity around missense and synonymous substitutions between *Tripsacum* and either maize or teosinte (Figure 2). If a proportion of missense mutations have been fixed due to hard sweeps, diversity around these substitutions should be lower than around synonymous substitutions. We observe this pattern around the causative amino acid substitution in the the domestication locus *tga1* (Figure S1), likely the result of a hard sweep during domestication [15,18]. Genome-wide, however, we observe no differences in diversity between synonymous and missense substitutions in either maize or teosinte (Figure 2).

Previous analyses have suggested that this approach may have limited power because a higher proportion of nonsynonymous relatively high proportion of missense substitutions will be found in genes under weak purifying selection and thus with higher genetic diversity [19]. To address this concern, we took advantage of genome-wide estimates of evolutionary constraint [20], calculated using genomic evolutionary rate profile (GERP) scores [21], a measure of evolutionary constraint, calculated across the maize genome [20]. We re-analyzed substitutions. We then evaluated substitutions only in subsets of genes with in the highest and lowest 10% quantile of mean GERP score, putatively representing genes under the strongest and weakest purifying selection (Figure S3). As expected, we see higher diversity around substitutions in genes under weak purifying selection there a table or figure we can reference?, but we still see find no difference between synonymous and missense substitutions in either subset of the data. Taken together, these data suggest hard sweeps do not play a major role in patterning genic diversity in either maize or teosinte.

Diversity is strongly influenced by purifying selection. Selection can also reduce diversity in functional regions of the genome via removal of deleterious mutations, a process known to as purifying or background selection [22]. We investigated purifying selection in maize and teosinte by evaluating the reduction of diversity within around genes. Pairwise diversity is strongly reduced within genes for both maize and teosinte (Figure 3A) but recovers quickly at sites outside of genes, consistent with the low levels of linkage disequilibrium generally observed in maize [16,23]. The reduction in relative diversity is more pronounced in teosinte, however, reaching lower levels in genes and occurring over a wider region.

Our initial comparison of synonymous and missense substitutions has low power to detect the effects of selection acting on multiple mutations or standing genetic variation, because in such cases diversity is not necessarily reduced [24,25]. Such “soft sweeps”, however, are still expected to occur more frequently in genes [26].

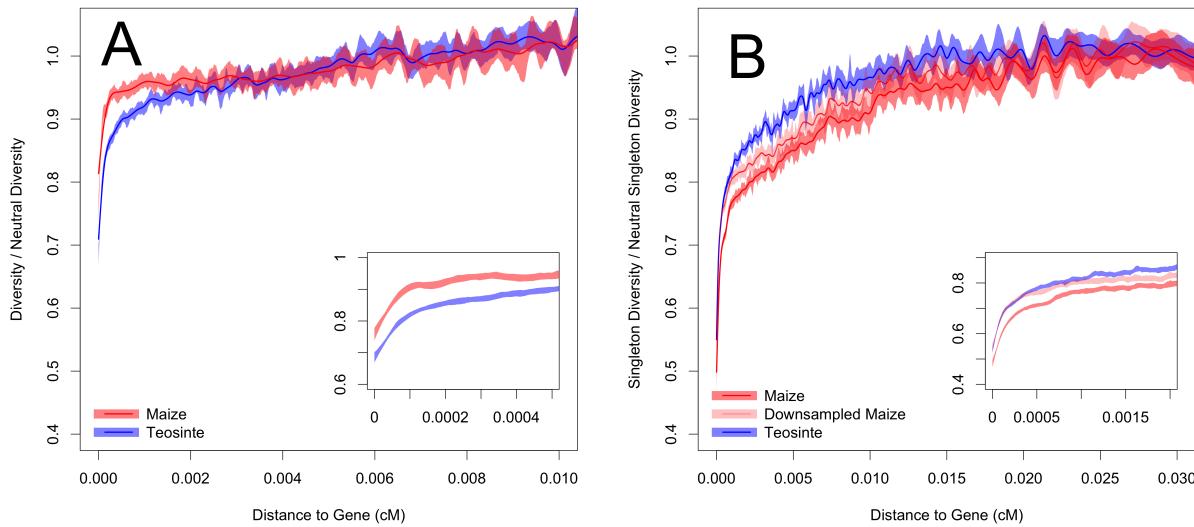


Fig. 3. Relative diversity versus distance to nearest gene in maize and teosinte. Shown are **A** pairwise nucleotide diversity and **B** singleton diversity. Relative diversity is calculated compared to the mean diversity in windows $\geq 0.01\text{cM}$ or $\geq 0.02\text{cM}$ from the nearest gene for pairwise diversity and singletons, respectively. Lines depict cubic smoothing splines with smoothing parameters chosen via generalized cross validation and shading depicts bootstrap-based 95% confidence intervals. Inset plots depict a smaller range on the x-axis.

quently in functional regions of the genome and could provide an alternative explanation for the observed reduction of diversity in genes. To test this possibility, we performed a genome-wide scan for selection using a method the H12 statistic, expected to be reasonably sensitive to both hard and soft sweeps [26]. Even qualitative differences between maize and teosinte remained unchanged even after removing genes in the top 20% of the H12 statistic used to identify targets of selection, however, qualitative differences between maize and teosinte in patterns of diversity in and near genes remained unchanged (Figure S6). We interpret these combined results as suggesting that purifying selection has left a more pronounced signature in the teosinte genome due to the increased efficacy of selection resulting from differences in effective population size.

Demography of maize domestication. To explore whether differences in the efficacy of purifying selection between maize and teosinte can be explained by demographic processes, we estimated the parameters of a simple domestication bottleneck model (Figure 4). The most likely model estimates an ancestral population mutation rate of $\theta = 0.0147$ per bp, which translates to an effective population size of $N_a \approx 123,000$ individuals given the mutation rate [27]. The maize population splits. We estimate that the maize population split from teosinte $\approx 15,000$ generations in the past, with an initial size of only $\approx 5\%$ of the ancestral N_a . Following domestication we estimate considerable gene flow between the populations: $M_{tn} = N_a \times 1.1 \times 10^{-5}$ migrants per generation from teosinte to maize and $M_{mt} = N_a \times 1.4 \times 10^{-5}$ migrants from maize to teosinte. After its split from teosinte, our model posits exponential population growth in maize, estimating a final modern effective population size of $N_m \approx 370,000$. Maize and teosinte continue to exchange migrants after the population split, with gene flow between the populations estimated at $M_{tm} = 1.1 \times 10^{-5} \times N_a$ migrants per generation from teosinte

to maize and $M_{mt} = 1.4 \times 10^{-5} \times N_a$ migrants from maize to teosinte.

Because our relatively small sample size limits our ability to characterize the rare variants that are most informative of population expansion, we investigated two alternative approaches for demographic inference. First, we took advantage of a complementary data set of genotyping data from more than 4,000 maize landraces collected from across the Americas [28] to estimate the modern maize effective population size from using low frequency variants informative of population expansion. This analysis yields a much higher estimate of the modern maize effective population size at $N_m \approx 993,000$. Finally, we reconstructed the demographic history of maize and teosinte using a recently developed coalescent approach CITE that does not require conditioning on a particular model (Figure 4) applied a model-free coalescent approach [29] using a subset of our samples. Though this analysis does highlight a potential weakness with our assumption of a constant size for teosinte suggests non-equilibrium dynamics for teosinte not included in our initial model, it is nonetheless broadly consistent, it nonetheless is broadly consistent with our simple model, identifying a clear domestication bottleneck followed by rapid population expansion (Figure 4) in maize to an extremely large extant size of $\approx 10^9$ (Figure S2).

Population expansion leads to stronger purifying selection in modern maize. Motivated by the rapid post-domestication expansion of maize evident from in our demographic analysis, we investigated whether patterns of diversity at low frequency reasoned that low-frequency and therefore on average thus younger polymorphisms would show similar results in comparisons of maize and teosinte. We thus repeated our analysis of diversity using alleles present in only a single individual in the sample. In direct contrast to patterns observed for pairwise nucleotide might show patterns distinct from pairwise diversity. Singleton diversity around missense and synonymous substitutions (Figure S4) appears

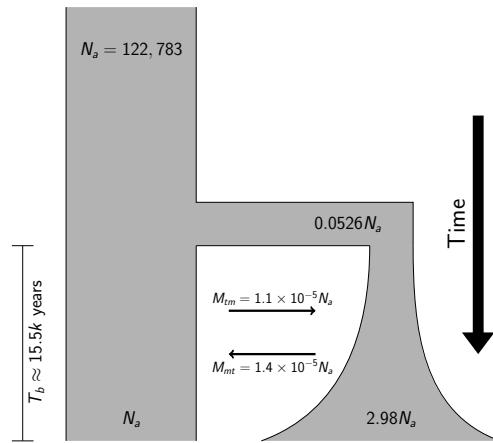


Fig. 4. Parameter estimates for a simple bottleneck model of maize domestication. See methods for details.

nearly identical to results from pairwise diversity (Figure 3A), diversity of singletons was 2, providing little support for a substantial increase in the number or strength of hard sweeps occurring in maize. In contrast, we observe a qualitative shift in the effects of purifying selection: singleton polymorphisms are more strongly reduced in and near genes in maize than in teosinte (Figure 3B), even after downsampling our maize data to account for differences in sample size. This relationship again remained after removing the 20% of genes with the highest H12 values (Figure S6). Finally, while direct comparison of curves for pairwise and singleton diversity within taxa are consistent with at least some non-equilibrium dynamics in teosinte, as indicated by our coalescent-based demographic analysis, these too reveal much stronger differences in maize (Figure S5).

These results suggest that demographic history and population size change have impacted the efficacy of purifying selection during maize evolution. In contrast, analysis of singleton diversity around missense and synonymous substitutions (Figure S4) provides a nearly identical pattern to that shown in Figure 2, providing little support for a substantial increase in the number or strength of hard sweeps occurring in maize.

Discussion

Hard sweeps do not shape genome-wide diversity in maize. Our findings demonstrate that classic hard selective sweeps have not contributed substantially to genome-wide patterns of diversity in maize, and our analysis restricting to sites undergoing the strongest and weakest purifying selection confirm this result we show is robust to previously highlighted limitations [19]. Our approach does not address sweeps on noncoding variants, but it agrees with concerns about power due to the effects of background selection [19]. Although our approach ignores the potential for hard sweeps in noncoding regions of the genome, a growing body of previous evidence arguing evidence argues against hard sweeps as the prevalent mode of selection shaping maize variability. For instance, population genetic studies of Among well-characterized domestication loci, only the gene *tga1* shows evidence of a hard sweep on a missense mutation [15], while several loci are consistent with "soft sweeps" from standing variation [30,31] or multiple mutations [14]. Moreover,

genome-wide studies of domestication [4], local adaptation [32] and modern breeding [33] demonstrate all support the importance of selection on standing variation as opposed to on new mutations. Even during domestication, very few differences between the primary source of adaptive variation. Soft sweeps are expected to be common when $2N_e\mu_b \geq 1$, where μ_b is the mutation rate of beneficial alleles with selection coefficient s_b [34]. Assuming a mutation rate of 3×10^{-8} Clark2005 and that on the order of $\approx 1 - 5\%$ of mutations are beneficial eyre2007 distribution, this implies that soft sweeps should be common in both maize and teosinte were fixed [4]. Moreover, of domestication loci that have been well studied from a population genetics standpoint, only one shows a signal of a hard sweep [15], although there are several consistent with selection on standing variation [14,30] or selection on multiple mutations [14], for mutational targets $>> 10kb$ — a plausible size for quantitative traits or for regulatory evolution targeting genes with large up- or down-stream control regions [e.g.] studer2011. Indeed, many adaptive traits in both maize CITE and teosinte CITE are highly quantitative cite wallace for maize? weber and/or lauter for teosinte? and adaptation in both maize hufford2012 and teosinte pyhajarvi2013 complex has involved selection on regulatory variation. if number is nice, add something like "X of the windows of top H12 score genome-wide occur outside of genes"

The lack of hard sweeps patterning diversity differs from *drosophila* [35] and *capsella* paragraph needs work. maybe eliminate/move? i do think we need some comparison to other stuff, but the broader Ne-selection efficiency argument should come later. The absence of evidence for a genome-wide impact of hard sweeps differs markedly from observations in *Drosophila* [35] and *Capsella* [36], but is consistent with humans [37] data from humans [37,38]. In a related study, [39] showed that selection tends to have an elevated impact on reducing diversity in species with large population size compared to small. This may explain why maize and humans, both bottlenecked species, show similar patterns while *drosophila* and *capsella* are different. However, we also do not observe abundant hard sweeps in teosinte, which calls this hypothesis into question. behave differently. However, we also do not observe abundant hard sweeps in teosinte, which calls th

Another possibility is that maize and teosinte don't show evidence of abundant hard sweeps because instead positive selection tends to operate in the form of soft sweeps. The theory of when soft sweeps are expected was worked out

by [34]. They showed that for $\theta_b = 2N\mu_b$, where μ_b is the mutation rate of beneficial alleles with selection coefficient $s_b > 0$, if $\theta_b > \frac{1}{\log(Ns_b)}$, then soft sweeps in the form of adaptation from multiple mutations should be common. Conversely, for neutral sites that somehow become beneficial (e.g., change in environment), if $\theta_b > \frac{1}{\log(2Ns_b)}$ then soft sweeps in the form of adaptation from standing neutral variation should be common. Alternatively, for previously deleterious sites that become beneficial, if $\theta_b > \frac{1}{\log\left(1 + \frac{2Ns_b}{2Ns_d + 1}\right)}$, where s_d is the previous selection coefficient, soft sweeps again in the form of adaptation from standing variation should be common. The implications of this theory are that soft sweeps are common when $\theta_b \geq 1$, which can be achieved through a large population size or a high mutation rate. If $\mu = 3 \times 10^{-8}$ [27], this implies a population size of ≥ 16.67 million, which our demographic analysis shows is larger than that of teosinte but not out of the question for modern maize.

Demography of domestication. Although many other authors have investigated the demography of maize domestication [40–42], these efforts relied exclusively on data only from genic regions of the genome and made a number of limiting assumptions about the demographic model. We show that diversity within genes has been strongly reduced by the effects of linked selection, making genic data inappropriate for estimating demography. Because levels of diversity alone do not allow independent inference of the strength and duration of a bottleneck [41], previous investigations have instead estimated the ratio of these parameters. We have overcome this limitation through the use of the such that even synonymous polymorphisms in genes are not representative of diversity at unconstrained sites. By utilizing the full joint SFS of maize and teosinte and consideration of a more realistic model of exponential growth, enabling us to estimate for the first time the we are able to estimate population growth, gene flow, and the strength of the domestication bottleneck. Given that modern maize exhibits $\approx 80\%$ of the diversity of teosinte, our estimate that the effective population size of the initial maize population was only $\approx 5\%$ of the ancestral teosinte population seems surprisingly low. The impact of a strong domestication bottleneck is likely ameliorated to some degree by maize-teosinte gene flow and the rapid post-domestication growth estimated in our model, without requiring assumptions about its duration.

Another One surprising result from our model is the estimation of the timing of domestication. Although we estimate the maize population split $\approx 5,000$ years earlier than previous molecular [11] or archaeological at $\approx 15,000$ years before present. While this appears to conflict with archaeological [43] estimates, we do not interpret these results as strongly conflicting. We first note that the timing of the instead argue this estimate reflects the fact that the genetic split between populations likely preceded the anatomical changes that can be identified in the archaeological record. We also caution, however, that our estimate may be inflated by population structure note that our result may also be inflated due to population structure, as our geographically diverse sample of teosinte may include populations diverged from those that gave rise to maize.

The estimated bottleneck of $\approx 5\%$ of the ancestral teosinte population seems low given that maize landraces exhibit $\approx 80\%$ of the diversity of teosinte hufford2012, but our model suggests that the effects of the bottleneck on diversity are

likely ameliorated to some extent by both gene flow and rapid post-domestication growth (Figure 4). Although we estimate that the modern effective size of maize is three times that of larger than teosinte, the small size of our sample dramatically reduces our power to estimate identify the low frequency alleles most sensitive to rapid population growth since rapid growth is most evidenced by rare alleles [44]. A better estimate of the effective size of modern maize comes from our analysis of singleton diversity in a large sample of maize landraces [28]. These data suggest, which suggests a modern effective size of ≈ 1 million or nearly 1 million, nearly eight times higher than modern teosinte. Nonetheless, Even this estimate, however, is likely downwardly biased due to missing data, sample size, and ascertainment bias in the genotyping data all bias this estimate downward as well. We can, however, get reasonable An alternative approach is to use estimates of the census size of modern maize. There: there are 47.9 million ha of open-pollinated maize in production [45], with a lower limit of 25,000 individuals planter likely planted at a density of $\approx 25,000$ individuals per hectare [46]. Using this lower bound on planting density, ignoring all industrial maize agriculture, and even assuming Assuming the effective size is only 0.1% $\approx 0.4\%$ of the census size (i.e. 1 ear for every 1000 plants), this still implies a modern effective population size of more than one billion. Clearly, the four billion. While these rough estimates are likely poor, all of the evidence points to the fact that the effective size of modern maize is extremely large.

Demography influences the efficiency of purifying selection. The observation that maize pairwise diversity section needs help – i'm not happy with the writing We find a more pronounced decrease in nucleotide diversity around genes in teosinte compared to domesticated maize (Figure 3A). This difference is likely due to the effects of purifying selection. Purifying selection is less impacted by the distance to the nearest gene than is teosinte pairwise diversity is reasonable from a long-term evolutionary standpoint. Theory has established that purifying selection is more efficient in a large population than in a small one larger populations [47], and this observation most likely reflects that prediction. More specifically, if teosinte N_e remained relatively constant while maize bottlenecked and recovered exponentially, this provides that the average N_e of maize over the previous several thousand generations is much smaller than that of teosinte, regardless of how much maize has ballooned in the recent past. Therefore, our observation shows that purifying selection in maize has not purged deleterious alleles, or the neutral alleles they are linked to, as effectively as in teosinte. The reversal of this trend when we analyze only singleton diversity instead of pairwise diversity, however, stands out as a notable observation. Every mutation begins as a singleton (an allele present in only one individual), and therefore singletons are, on average, the youngest class of alleles that can be observed. Therefore unlike pairwise our observation is thus consistent with our demographic model and previous work [40–42, 48] showing a smaller long-term effective population size in maize. To understand how recent population growth has changed the effects of linked selection, we also analyzed diversity of singleton polymorphisms. Under a simple neutral model, lower frequency polymorphisms are expected to be younger, and patterns of diversity, which are most heavily influenced by intermediate frequency alleles based on the definition of π [49], singleton diversity is most influenced by recent patterns of evolution. Hence, because our demographic estimation indicates dramatic expansion of maize N_e in the recent past, we expect for purifying selection to presently operate more efficiently in maize than in teosinte.

This observation is very clearly demonstrated by the fact that singleton diversity in maize is more impacted by the distance to the nearest gene than it is in teosinte, as was shown in Figure 3. At singleton SNPs should thus represent recent evolutionary history. Consistent with its larger current size and thus more efficient purifying selection, we see a stronger decline of singleton diversity around genes in maize (Figure 3B than teosinte.

I would like to see a supplemental figure or maybe just number here on the partial correlation in both teosinte and maize of diversity with recombination after accounting for gene density. do this in 500kb windows and then in 10kb windows. we should do the correlation with both pi and singletons. This let's us compare more explicitly with studies like C-D who evaluate overall patterns of linked selection. will be especially cool if we don't see a big difference in big windows but we do in 10kb windows. A consequence of the inefficient purifying selection that maize experienced during its bottleneck is likely that it harbors more weakly deleterious alleles segregating at intermediate frequency than does teosinte. This could be a part of the explanation of why maize inbreds have continued to improve over the past several decades [50]; if deleterious alleles tend to be recessive and are particularly frequent, they will have ample opportunities to display their phenotypes in inbreds. Our results also demonstrate that recent purifying selection in maize has become much more effective, potentially explaining the ongoing improvement of these inbreds as maize lines are continuously selected. Additionally, the large N_e of modern maize compared to modern teosinte implies that for new mutations, selection will operate much more efficiently in maize.

inbreds stuff OK, but should be shortened. and we need more/broader. Chia, Rogers-Melnick, Mezmouk all looked at load. does this have implications for large effect load (i.e. should bneck have purged those?), for heterosis, architecture of quant. traits, etc.

Importantly, our estimation of the parameters of the maize domestication bottleneck contribute to the understanding of how the demography of crop domestication can impact crop diversity. The bottleneck effects from a sudden collapse in population size have been well studied and are known to impact crops for thousands of generations. Complementing this knowledge, our results demonstrate that the rapid expansion experienced by many crops after domestication can also have a profound influence on patterns of diversity, and the effects of this expansion should be accounted for as important contributors to long-term evolution. These results have implications for lots of studies. recent demographic history often ignored (examples). in crops, bneck is considered and loss of diversity, but expansion also likely ubiquitous and consequences not well studied. in humans XYZ, and we see some of that already here. this argues that a full understanding of diversity in both wild and cultivated plants will require recent demography. mention C-D ignored recent demography – here's where rec/pi windows could help.

Materials and Methods

BASH, R, and Python scripts. All scripts used for analysis are available in an online repository at [REPO ADDRESS HERE](#).

Plant materials. We made use of published sequences from inbred accessions of teosinte (*Z. mays* ssp. *parviglumis*) and maize landraces from the Maize HapMap3 panel as part of the Panzea project (bam files are available at [/iplant/home/shared/panzea/hapmap3/bam_internal/v3.bams.bwamem](#)) [16, 51, 52]. From these data, we removed 4 teosinte individuals that were not ssp. *parviglumis* or appeared as outliers in an initial principal component analysis conducted with the package adegenet [53] (Figure S7), leaving 13 teosinte and 23 maize that were used for all subsequent analyses (Table S8). We also utilized a single *Tripsacum* (*T. dactyloides*) individual as an outgroup (bam file available at [/iplant/home/shared/panzea/hapmap3/bam_internal/v3.bams.bwamem](#)). some

clarification here. did we use Lemmon's higher depth teosinte? If so, why does Li say we only had 7 high-depth bam files? there should be more

Physical and genetic maps. Sequences were mapped to the maize B73 version 3 reference genome [54] ([ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/](#)) as described by [52]. All analyses made use of uniquely mapping reads with mapping quality score ≥ 30 and bases with base quality score ≥ 20 ; quality scores around indels were adjusted following Li *et al.* [55]. We converted physical coordinates to genetic coordinates via linear interpolation of the previously published 1cM resolution NAM genetic map [56].

Estimating the site frequency spectrum. We estimated both the genome-wide site frequency spectrum (SFS) as well as a separate SFS for genic (within annotated transcript) and intergenic ($\geq 5kb$ from a transcript) regions. We used the biomaRt package [57, 58] of R [59] to parse annotations from genebuild version 5b of AGPv3. We estimated single population and joint SFS with the software ANGSD [60], including all positions with at least one aligned read in $\geq 80\%$ of samples in one or both populations. We assumed individuals were fully inbred and treated each line as a single haplotype. Because ANGSD cannot calculate a folded joint SFS, we first polarized SNPs using the maize reference genome and then folded spectra using $\delta\alpha\delta i$ [61].

Demographic inference. We used the software $\delta\alpha\delta i$ [61] to estimate parameters of a domestication bottleneck from the joint maize-teosinte SFS, using only sites $> 5kb$ from a gene to ameliorate the effects of linked selection. We modeled a teosinte population of constant effective size N_a , that at time T_b generations in the past gave rise to a maize population of size N_b which grew exponentially to size N_m in the present (Figure 4). The model includes migration of M_{mt} individuals each generation from maize to teosinte and M_{tm} individuals from teosinte to maize. We estimated N_a using $\delta\alpha\delta i$'s estimation of $\theta = 4N_a\mu$ from the data and a mutation rate of $\mu = 3 \times 10^{-8}$ [27]. We estimated all other parameters using 1,000 $\delta\alpha\delta i$ optimizations and allowing initial values between runs to be randomly perturbed by a factor of 2. Optimized parameters along with their initial values and upper and lower bounds can be found in table S9. We report parameter estimates from the optimization run with the highest log-likelihood.

We further made use of a large genotyping data set of more than 4,000 maize landraces [28] to estimate the modern maize N_e from singleton counts. We filtered these data to include only SNPs with data in $\geq 1,500$ individuals, and then projected the SFS down to a sample of 500 individuals by sampling each marker without replacement 1,000 times according to the observed allele frequencies. We then estimated N_e from the data assuming $\mu = 3 \times 10^{-8}$ [27] and the relation $4N_e\mu = \frac{S}{L}$ [62], where where S is the total number of singleton SNPs and L is the total number of SNPs in the dataset.

Finally, As a final estimate of demography, we employed MSMC [29] to complement our model-based demographic inference. We used six each of maize and teosinte (BKN022, BKN025, BKN029, BKN030, BKN031, BKN033, TIL01, TIL03, TIL09, TIL10, TIL11 and TIL14), treating each inbred genome as a single haplotype. We called SNPs in ANGSD [60] using a SNP p-value of $1e-6$ against a reference genome masked using SNPable (). We then removed heterozygous genotypes and filtered sites with a mapping quality ≤ 30 , a base quality ≤ 20 , or a $\log_2(depth) < 1$. We ran MSMC with pattern parameters $20 \times 2 + 20 \times 4 + 10 \times 2$.

Diversity. We made use of the software ANGSD [60] for diversity calculations and genotype calling. We calculated diversity statistics in maize and teosinte in 1 kb non-overlapping windows using filters as described above for the SFS. We used allele counts to estimate the number of singleton polymorphisms in each window, and used binomial sampling to create a second maize data set down-sampled to have the same number of samples as teosinte. We called genotypes in maize, teosinte, and *Tripsacum* at sites with a SNP p-value $< 10^{-6}$ and when the genotype posterior probability > 0.95 . We identified substitutions in maize and teosinte as all sites with a fixed difference with *Tripsacum* and $\leq 20\%$ missing data. Substitutions were classified as synonymous, missense, or noncoding using the ensemble variant effects predictor [63]. For each window with $\geq 100bp$ of data we computed the genetic distance between the window center and the nearest synonymous and missense substitution as well as the genetic distance to the center of the nearest gene transcript.

Selection scan. We scanned the genome to identify sites that have experienced recent positive selection using the H12 statistic [26] in sliding windows of 200 SNPs with a step of 25 SNPs.

ACKNOWLEDGMENTS. We are indebted to Graham Coop and Simon Aeshbacher for their constructive input during this study. We thank Robert Bukowski and Qi Sun for providing early-access P

We thank Robert Bukowski and Qi Sun for providing early-access data from maize HapMap3. Funding was provided by NSF Plant Genome Research Project 1238014.

1. Doebley, J. F., Gaut, B. S., & Smith, B. D. (2006) *Cell* 127, 1309–1321.
2. Purugganan, M. D. & Fuller, D. Q. (2009) *Nature* 457, 843–848.
3. Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007) *Proceedings of the National Academy of Sciences* 104, 8641–8648.
4. Hufford, M. B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E., Kaeplinger, S. M., et al. (2012) *Nature genetics* 44, 808–811.
5. He, Z., Zhai, W., Wen, H., Tang, T., Wang, Y., Lu, X., Greenberg, A. J., Hudson, R. R., Wu, C.-I., & Shi, S. (2011) *PLoS genetics* 7, e1002100.
6. Vigouroux, Y., McMullen, M., Hittinger, C., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y., & Doebley, J. (2002) *Proceedings of the National Academy of Sciences* 99, 9650–9655.
7. Chapman, M. A., Pashley, C. H., Wenzler, J., Hvala, J., Tang, S., Knapp, S. J., & Burke, J. M. (2008) *The Plant Cell* 20, 2931–2945.
8. Shiferaw, B., Prasanna, B. M., Hellin, J., & Bänziger, M. (2011) *Food Security* 3, 307–327.
9. Strable, J. & Scanlon, M. J. (2009) *Cold Spring Harbor Protocols* 2009, pdbemo132.
10. Smith, B. D. (1995) The emergence of agriculture. (*Scientific American Library New York*).
11. Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez, J., Buckler, E., & Doebley, J. (2002) *Proceedings of the National Academy of Sciences* 99, 6080–6084.
12. Wilkes, H. C. et al. (1967) Teosinte: the closest relative of maize.
13. Doebley, J., Stec, A., & Gustus, C. (1995) *Genetics* 141, 333.
14. Wills, D. M., Whipple, C. J., Takuno, S., Kursel, L. E., Shannon, L. M., Ross-Ibarra, J., & Doebley, J. F. (2013) *PLoS Genet* 9, e1003604.
15. Wang, H., Studer, A. J., Zhao, Q., Meeley, R., & Doebley, J. F. (2015) *Genetics pp. genetics*–115.
16. Chia, J.-M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., et al. (2012) *Nature genetics* 44, 803–807.
17. Smith, J. M. & Haigh, J. (1974) *Genetical research* 23, 23–35.
18. Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Fallar, M., Bomblies, K., Lukens, L., & Doebley, J. F. (2005) *Nature* 436, 714–719.
19. Enard, D., Messer, P. W., & Petrov, D. A. (2014) *Genome research* 24, 885–895.
20. Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., Li, C., Li, Y., & Buckler, E. S. (2015) *Proceedings of the National Academy of Sciences* 112, 3823–3828.
21. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010) *PLoS Comput Biol* 6, e1001025.
22. Charlesworth, B., Morgan, M., & Charlesworth, D. (1993) *Genetics* 134, 1289–1303.
23. Tenaiillon, M. I., Sawkins, M. C., Anderson, L. K., Stack, S. M., Doebley, J., & Gaut, B. S. (2002) *Genetics* 162, 1401–1413.
24. Innan, H. & Kim, Y. (2004) *Proceedings of the National Academy of Sciences of the United States of America* 101, 10667–10672.
25. Messer, P. W. & Petrov, D. A. (2013) *Trends in ecology & evolution* 28, 659–669.
26. Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015) *PLoS genetics* 11, e1005004.
27. Clark, R. M., Tavaré, S., & Doebley, J. (2005) *Molecular biology and evolution* 22, 2304–2312.
28. Hearne, S., Chen, C., Buckler, E., & Mitchell, S. (2015) Unimputed gbs derived snps for maize landrace accessions represented in the seed-maize gwaa panel (.). Accessed: 2015-02-16.
29. Schiffels, S. & Durbin, R. (2014) *Nature genetics*.
30. Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011) *Nature genetics* 43, 1160–1163.
31. Gallavotti, A., Zhao, Q., Kyozuka, J., Meeley, R. B., Ritter, M. K., Doebley, J. F., Pè, M. E., & Schmidt, R. J. (2004) *Nature* 432, 630–635.
32. Takuno, S., Ralph, P., Swarts, K., Elshire, R. J., Glaubitz, J. C., Buckler, E. S., Hufford, M. B., & Ross-Ibarra, J. (2015) *Genetics*.
33. Beissinger, T. M., Hirsch, C. N., Vaillancourt, B., Deshpande, S., Barry, K., Buell, C. R., Kaeplinger, S. M., Gianola, D., & de Leon, N. (2014) *Genetics* 196, 829–840.
34. Messer, P. W. & Petrov, D. A. (2013) *Trends in ecology & evolution* 28, 659–669.
35. Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y., & Sella, G. (2011) *PLoS genetics* 7, e1001302.
36. Williamson, R., Josephs, E., Platts, A., Hazzouri, K., Haudry, A., Blanchette, M., & Wright, S. (2014) *PLoS genetics* 10, e1004622–e1004622.
37. Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., Sella, G., Przeworski, M., et al. (2011) *science* 331, 920–924.
38. Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010) *Current Biology* 20, R208–R215.
39. Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015) *PLoS Biol* 13, e1002112.
40. Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L., & Gaut, B. S. (1998) *Proceedings of the National Academy of Sciences* 95, 4441–4446.
41. Tenaiillon, M. I., U'Ren, J., Tenaiillon, O., & Gaut, B. S. (2004) *Molecular Biology and Evolution* 21, 1214–1225.
42. Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., & Gaut, B. S. (2005) *Science* 308, 1310–1314.
43. Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J., & Dickau, R. (2009) *Proceedings of the National Academy of Sciences* 106, 5019–5024.
44. Keinan, A. & Clark, A. G. (2012) *science* 336, 740–743.
45. Program, T. M. (1999) Development, maintenance, and seed multiplication of open-pollinated maize varieties. (*CIMMYT, Mexico, D.F.*), 2 edition.
46. Baden, W. W. & Beekman, C. S. (2001) *American Antiquity* pp. 505–515.
47. Kimura, M. (1984) The neutral theory of molecular evolution. (*Cambridge University Press*).
48. Ross-Ibarra, J., Tenaiillon, M., & Gaut, B. S. (2009) *Genetics* 181, 1399–1413.
49. Nei, M. & Li, W.-H. (1979) *Proceedings of the National Academy of Sciences* 76, 5269–5273.
50. Meghji, M., Dudley, J., Lambert, R., & Sprague, G. (1984) *Crop Science* 24, 545–549.
51. Lemmon, Z. H., Bukowski, R., Sun, Q., & Doebley, J. F. (2014) *PLoS Genet* 10, e1004745.
52. Panzea. (In prep) TBD.
53. Jombart, T. & Ahmed, I. (2011) *Bioinformatics* 27, 3070–3071.
54. Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., et al. (2009) *science* 326, 1112–1115.
55. Li, H. (2011) *Bioinformatics* 27, 2987–2993.
56. Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014) *PLoS One* 9, e90346.
57. Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009) *Nature protocols* 4, 1184–1191.
58. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005) *Bioinformatics* 21, 3439–3440.
59. R Core Team. (2014) R: A Language and Environment for Statistical Computing (*R Foundation for Statistical Computing, Vienna, Austria*).
60. Corneliusen, T. S., Albrechtsen, A., & Nielsen, R. (2014) *BMC bioinformatics* 15, 356.
61. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009) *PLoS genetics* 5, e1000695.
62. Fu, Y.-X. & Li, W.-H. (1993) *Genetics* 133, 693–709.
63. McLaren, W., Pritchard, B., Ríos, D., Chen, Y., Flieck, P., & Cunningham, F. (2010) *Bioinformatics* 26, 2069–2070.

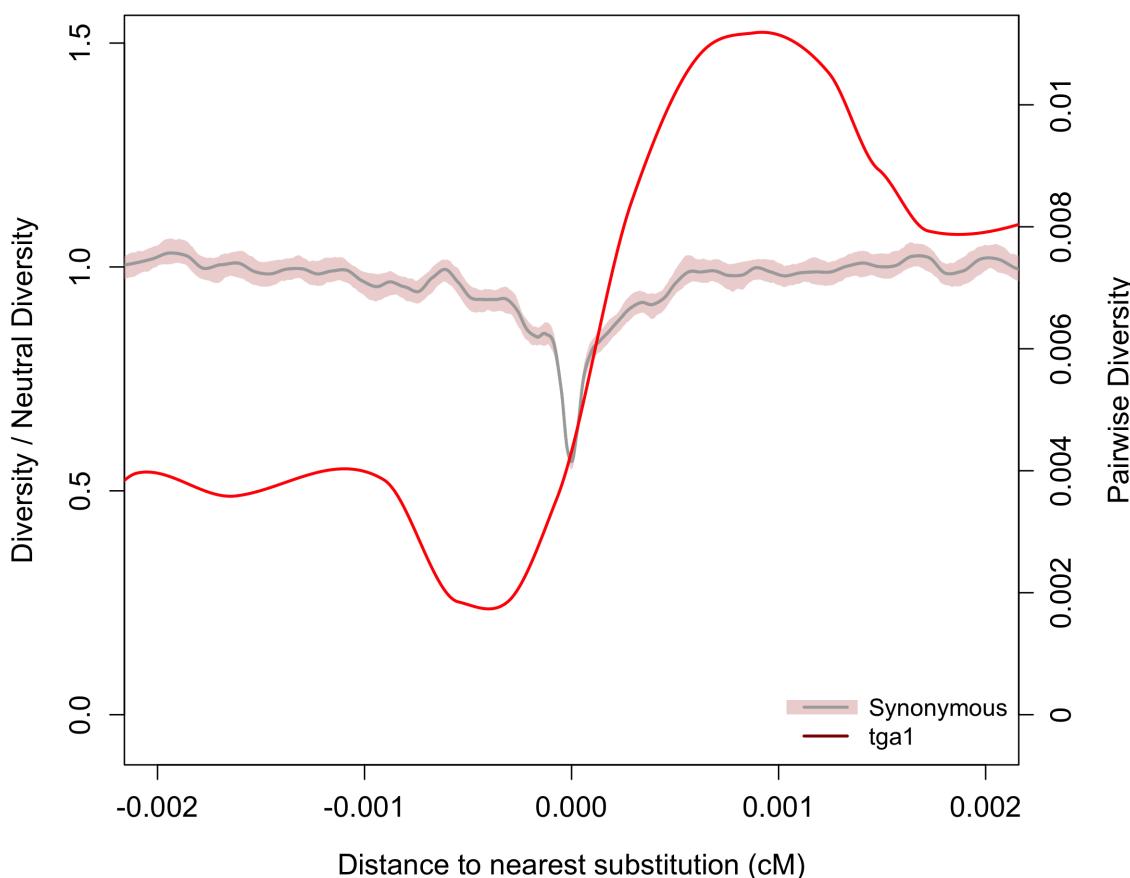


Fig. S1. Diversity surrounding the causitive polymorphism at the *tga1* locus is plotted. Since this is only one gene, the large amount of noise compared to our average plots is expected. However, notice that diversity precisely at the causitive polymorphism is reduced and a recovery of diversity is observed away from that site.

Supporting Information

Fig. S2. need MSMC caption

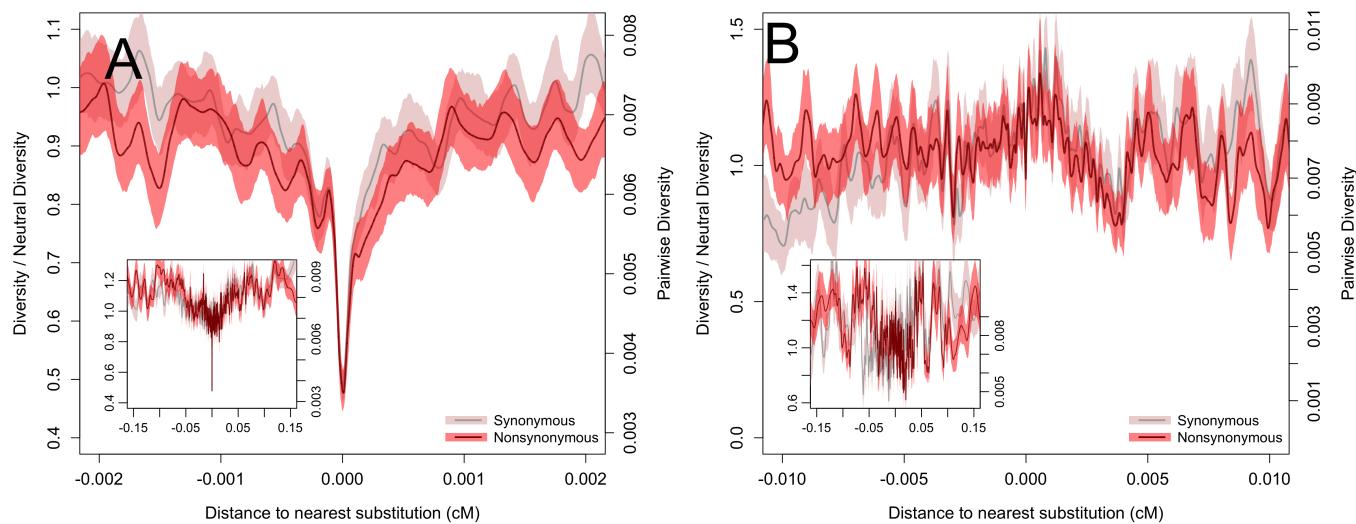


Fig. S3. Pairwise diversity surrounding synonymous and nonsynonymous substitutions in maize at highly conserved (A) or unconserved (B) sites. Bootstrap-based 95% confidence intervals are depicted via shading. Inset plots depict a larger range on the x-axis.

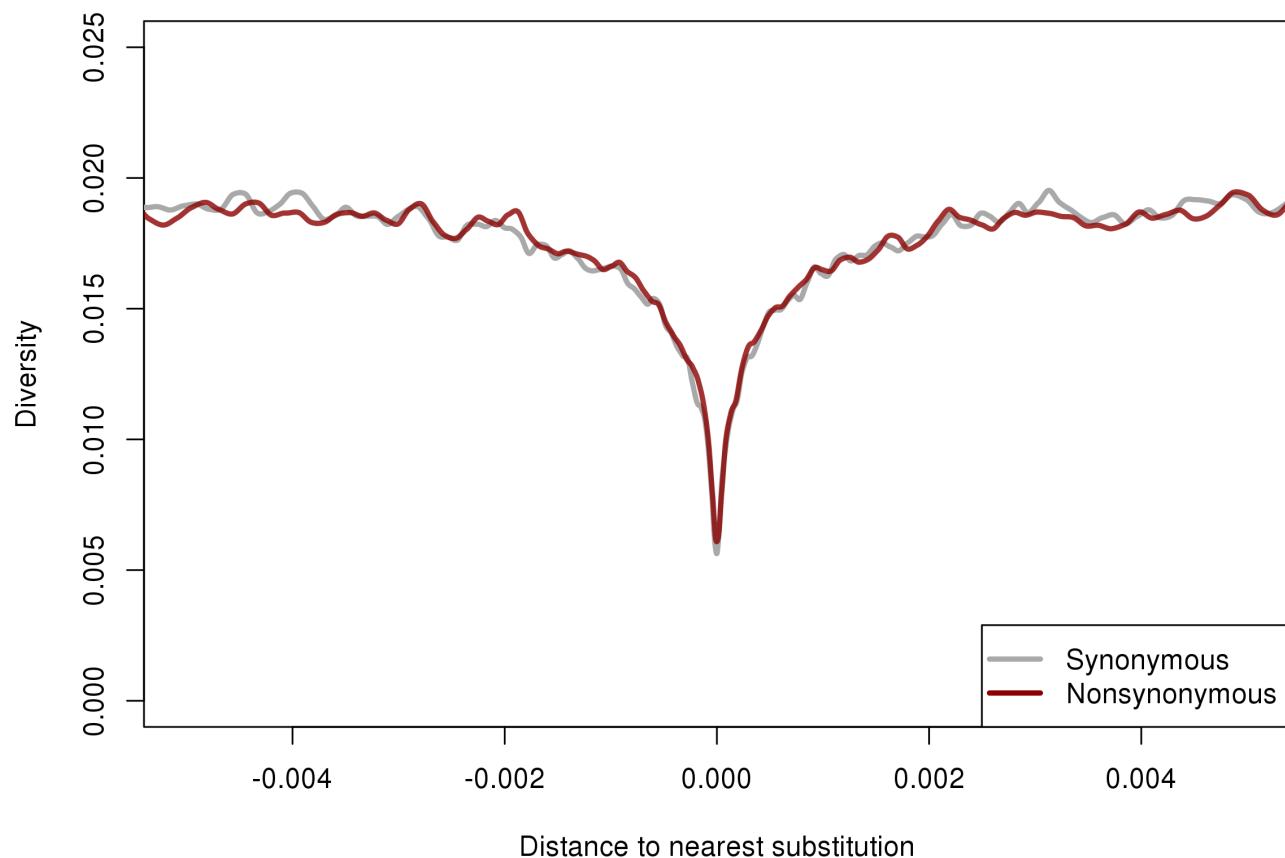


Fig. S4. Singleton diversity surrounding synonymous and nonsynonymous substitutions in maize.

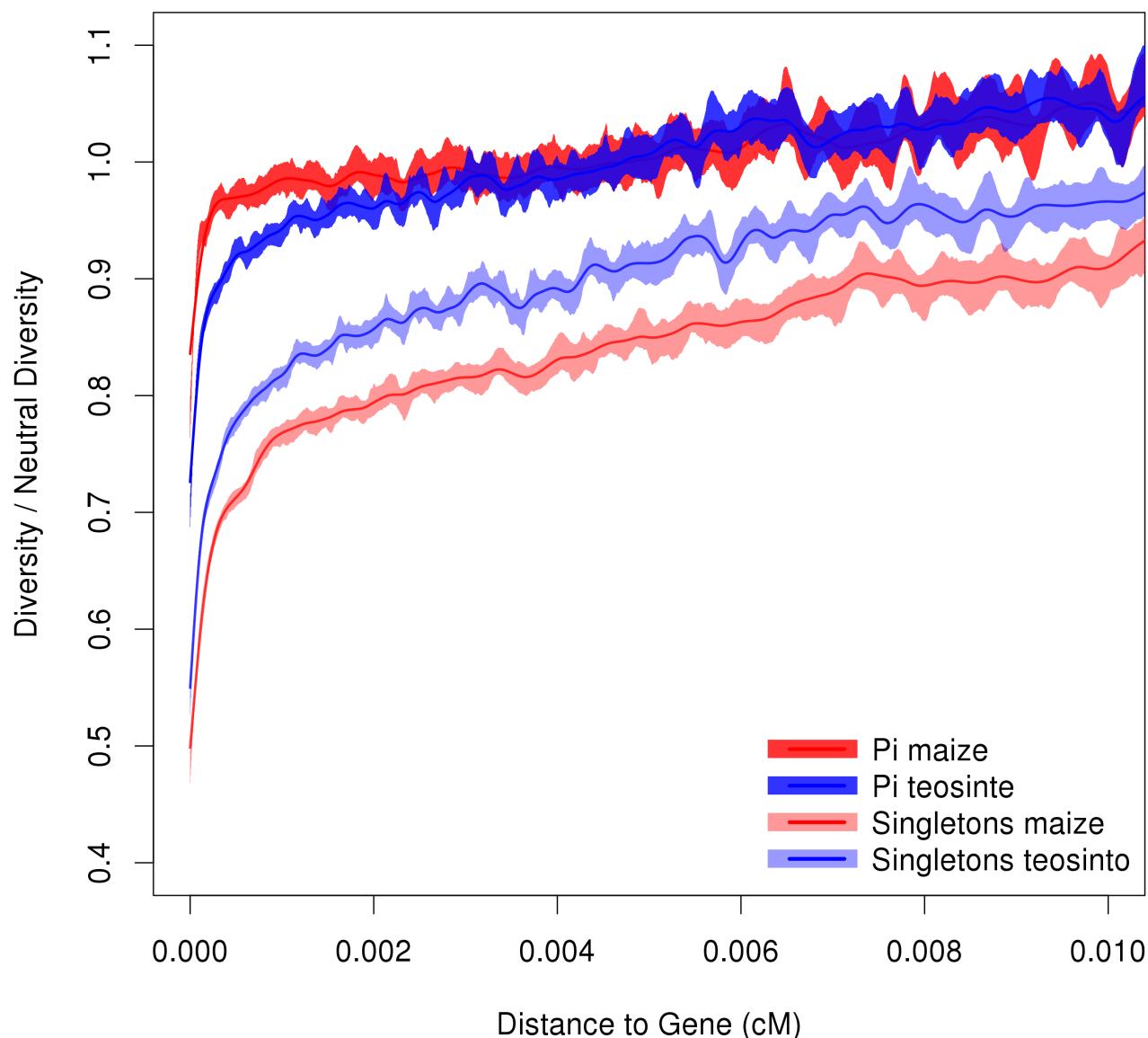


Fig. S5. Relative diversity versus distance to nearest gene in maize and teosinte. Relative diversity is calculated by comparing to the mean diversity in all windows $\geq 0.02\text{cM}$ from the nearest gene. Lines depict cubic smoothing splines with smoothing parameters chosen via generalized cross validation and shading depicts bootstrap-based 95% confidence intervals.

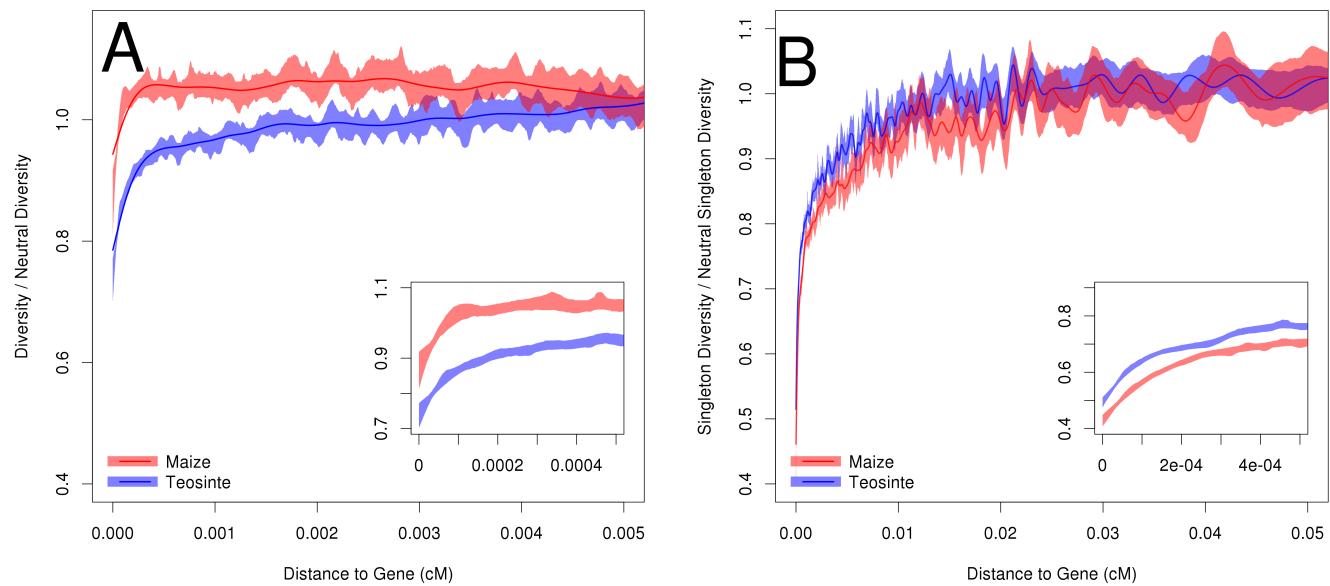


Fig. S6. Relative level of diversity versus distance to the nearest gene, in maize and teosinte, based on only sites that do not show evidence of hard or soft sweeps according to H12. Two measures of diversity were investigated. **A** displays pairwise diversity, which is most influenced by intermediate frequency alleles and therefore depicts more ancient evolutionary patterns, and **B** depicts singleton diversity, influenced by rare alleles and thus depicting evolutionary patterns in the recent past. Bootstrap-based 95% confidence intervals are depicted via shading. Inset plots depict a smaller range on the x-axis.

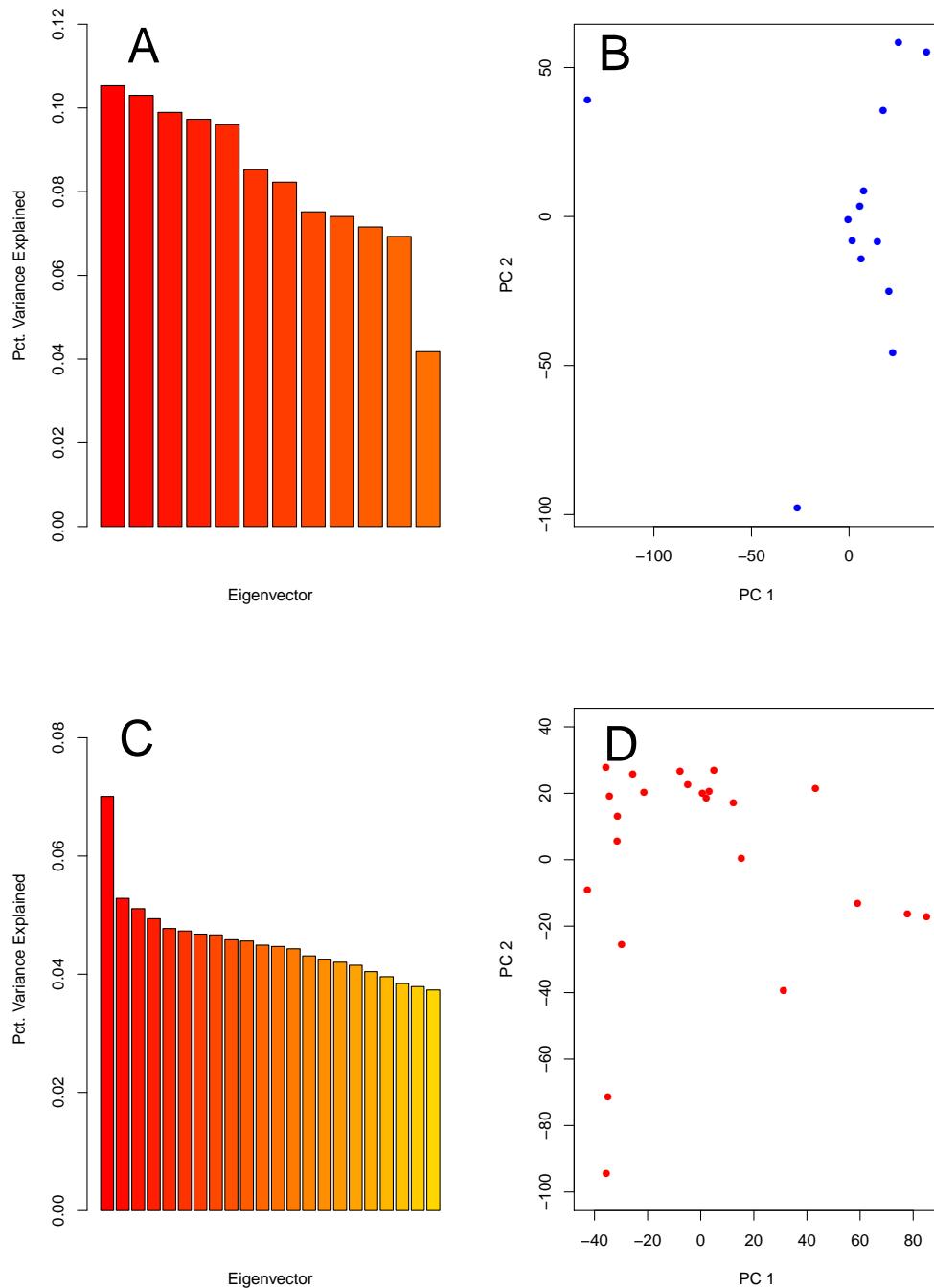


Fig. S7. Principal component analysis of teosinte and maize individuals to ensure that no close relatives were inadvertently included in our study. Plots are based on a random sample of 10,000 SNPs. **A:** Percentage of total variance explained by each principal component for teosinte. **B:** PC1 vs PC2 for all 13 teosinte individuals. **C:** Percentage of total variance explained by each principal component for maize. **D:** PC1 vs PC2 for all 23 maize individuals.

Maize	Teosinte
BKN009	TIL01
BKN010	TIL02
BKN011	TIL03
BKN014	TIL04-TIP454
BKN015	TIL07
BKN016	TIL09
BKN017	TIL10
BKN018	TIL11
BKN019	TIL12
BKN020	TIL14-TIP498
BKN022	TIL15
BKN023	TIL16
BKN025	TIL17
BKN026	
BKN027	
BKN029	
BKN030	
BKN031	
BKN032	
BKN033	
BKN034	
BKN035	
BKN040	

Fig. S8. A list of maize and teosinte individuals included in this study. Sequencing and details were previously described by [cite chia and lemmmon](#)

Parameter	Initial value	Upper bound	Lower bound
$\frac{N_b}{N_a}$	0.02	1×10^{-7}	2
$\frac{N_m}{N_a}$	3	1×10^{-7}	200
$\frac{T_b}{2N_a}$	0.04	0	1
$\frac{M_{mt}}{N_a}$	1×10^{-10}	1×10^{-7}	0.001
$\frac{M_{tm}}{N_a}$	1×10^{-10}	1×10^{-7}	0.001

Fig. S9. Parameters, initial values, and boundaries used for model-fitting with $\delta\alpha\delta i$. Parameters are shown in the units utilized by $\delta\alpha\delta i$, although in the text simplified units are reported.

