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, U
Davis, CA, USA

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

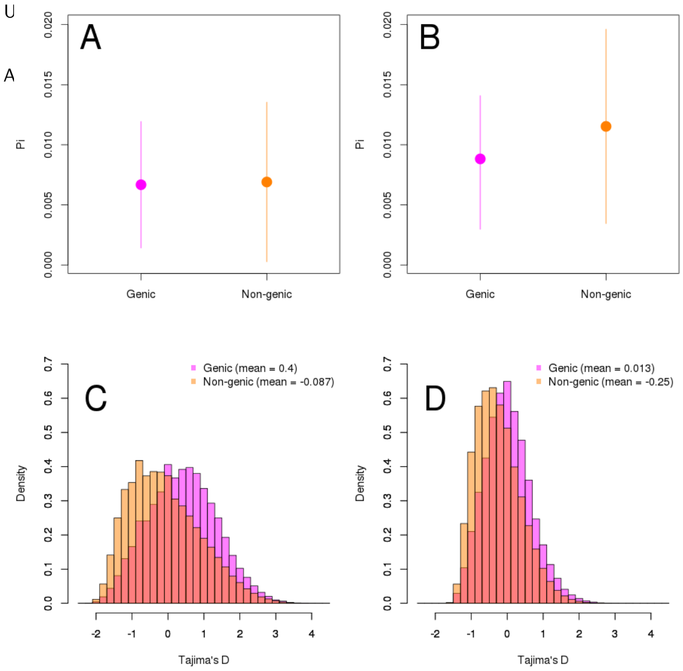

Fig. 1. Pairwise diversity (Pi) and Tajimas D were calculated inside and outside of genes for maize and teosinte. In both taxa, a dearth of rare alleles was observed within genes relative to outside of genes, which led to higher values of Tajima's D outside of genes than inside.**A:** Mean and standard deviation of pi for 1 kb genic and non-genic windows of maize; **B:** Mean and 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.

Unique selective and demographic 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 whole genome sequence data from 23 maize and 13 teosinte individuals to 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, 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 those seen in teosinte. Our results demonstrate that in domesticated 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.

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.

For instance, supposed neutrally-evolving DNA is often used to estimate historical demographic parameters of a population such as effective population size, structure, and expansion history [8, 9]. However, researchers have called into question whether or not there are reasonable approaches for identifying neutral regions of the genome, since the effects of selection can be wide-ranging [10, 11]. For example, in *Drosophila* it appears that the majority of the genome is impacted by the effects of selection through linkage [12]. A natural next question, therefore, is how can demographic parameters be estimated independently of selective parameters? Human researchers have attempted to address this problem by limiting analyses to only sites far from genes [13], but as *Drosophila* demonstrates, it is difficult to be certain that sites even in gene-poor regions of the genome are not influenced by linked selection. Ultimately, an understanding of which sites have

Significance

Patterns of linked selection, or the impact of selection on sites that neighbor a functional variant, have been carefully evaluated in only a few species. not really. we need to rephrase here and cite the Corbett-Detig plos-bio paper somewhere – they looked at a lot of species including maize! **In this work, we demonstrate that selection against deleterious mutations leaves a pronounced signature on the maize genome, reducing diversity in and immediately around genes. We show how demography interacts with selection to impact genome-wide patterns of diversity, including the important observation that rapid population expansion can increase the efficiency of selection as much as a sudden population bottleneck can weaken it. Along the way, we develop the first estimate the demographic parameters of the maize domestication from whole genome sequence data.**

Reserved for Publication Footnotes

wide patterns of variability through linkage, is required for reasonable demographic inference.

Maize represents an excellent organism to study these phenomena. Maize is a species of tremendous importance worldwide as both a staple crop [14] and as a model for understanding plant evolution [15]. Broadly speaking, archaeological and genetic studies have established that maize domestication is likely to have taken place in Mexico approximately 9,000 years bp [16, 17]. Teosinte, the most recent wild ancestor to maize, remains extant throughout much of the Americas [18]. Additionally, several large-effect domestication loci [19–21] and putative domestication regions [4] have been identified. But despite all that is known about maize domestication, 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 impact of a different demographic history in teosinte; and 3) precisely estimate the parameters of the maize domestication bottleneck. We show that our third objective, estimating the parameters of domestication, is not possible without first completing objectives one and two, which demonstrate that as in humans [13], the majority of maize non-genic DNA may reasonably be treated as neutral. We achieve these objectives by utilizing whole-genome-sequence information from 23 maize and 13 teosinte lines sequenced as part of the Maize HapMap 2 panel [22].

Results

Patterns of diversity differ between genic and non-genic regions of the genome. Our reanalysis of the maize HapMap 2 data [22] is consistent with earlier findings that patterns of diversity differ between genic and non-genic regions of the genome in both maize and teosinte [4] (Figure 1). In maize, mean pairwise diversity (π) within genes was significantly lower than at positions at least 5 kb away from genes (0.00668 vs 0.00691, $p < 2\times10^{-44}$). 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, $p < X$ for both comparisons). These observations suggest that diversity in genes is not evolving neutrally, but instead 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 **CITE**. To evaluate whether patterns of such “hard sweeps” could explain observed differences in diversity between genic and non-genic regions of the genome, we compared diversity around missense and synonymous substitutions between *Tripsacum* and either maize or teosinte (Figure 3). 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 [21, 23]. Genome-wide, however, we observe no differences in diversity between synonymous and missense substitutions in either maize or teosinte.

Previous analyses have suggested that this approach may have limited power because a higher proportion of nonsynonymous substitutions will be found in genes under weak purifying selection and thus with higher genetic diversity [24]. To address this concern, we took advantage of genomic evolutionary rate profile (GERP) scores [25], a measure of evolutionary constraint, calculated across the maize genome [26]. We re-analyzed substitutions in subsets of genes with the highest and lowest 10% quantile of mean GERP score, putatively representing genes under the strongest and weakest purifying selection (Figure S2). As expected, we see higher diversity around substitutions in genes under weak purifying selection, but we still see 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 [27]. We investigated purifying selection in maize and teosinte by evaluating the reduction of diversity within genes. Pairwise diversity is strongly reduced within genes for both maize and teosinte (Figure 4A) but recovers quickly at sites outside of genes, consistent with the low levels of linkage disequilibrium generally observed in maize **CITE**. The reduction in relative diversity is more pronounced in teosinte, however, reaching lower levels in genes and occurring over a wider region. make sure we have tense consistent throughout. is/was etc.

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. need a citation? Such “soft sweeps”, however, would still be expected to occur more frequently 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 expected to be reasonably sensitive to both hard and soft sweeps [28]. Even after removing genes in the top 20% of the

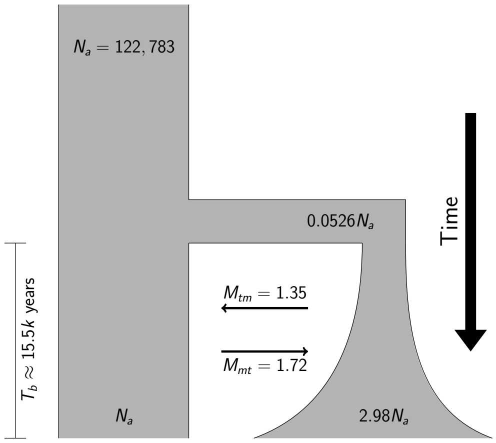

Fig. 2. Parameter estimates for a simple bottleneck model of maize domestication. See methods for details.

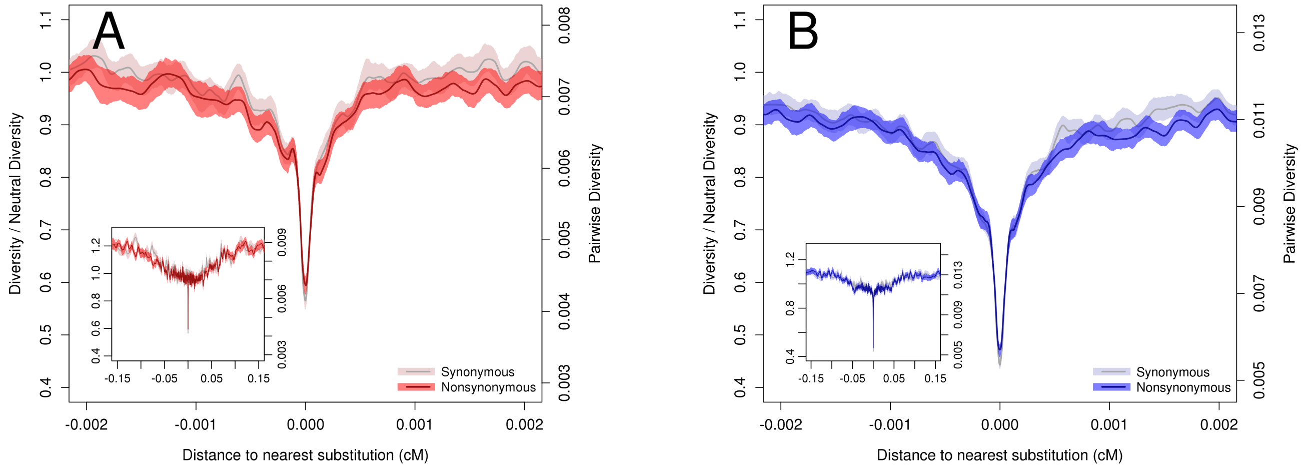

Fig. 3. Pairwise diversity surrounding synonymous and non-synonymous substitutions in **A** maize and **B** teosinte. Axes show both absolute diversity values and values relative to mean nucleotide diversity in windows $\geq 0.01cM$ from a substitution. is this correct or did i confuse things? Lines 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.

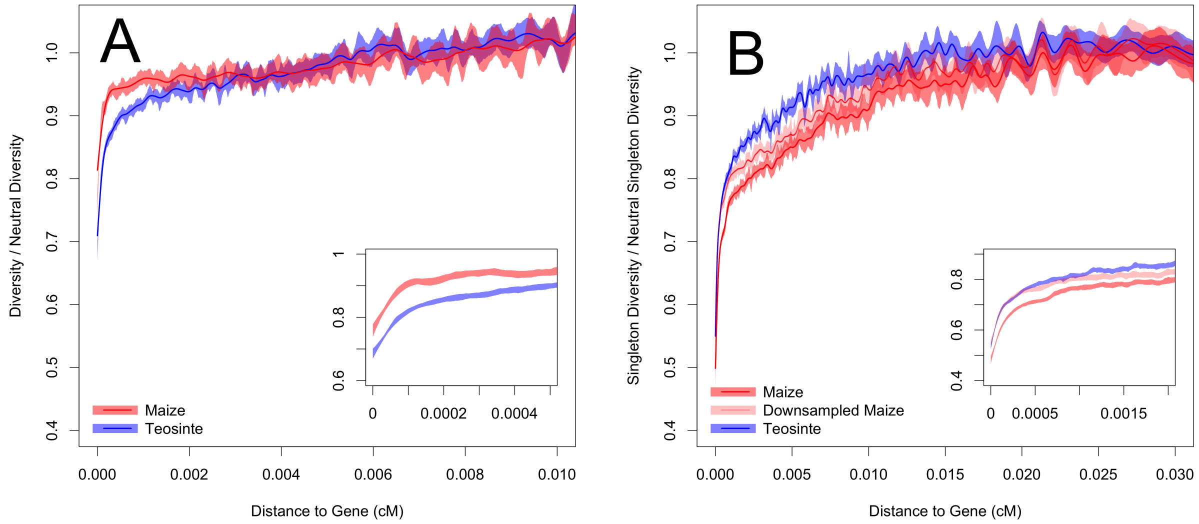

Fig. 4. Relative diversity versus distance to nearest gene in maize and teosinte. Shown are **A** pairwise nucleotide diversity and **B** singleton diversity. Relative diversity is calculated compared to the mean diversity in windows $\geq 0.01cM$ or $\geq 0.02cM$ from the nearest gene for pairwise diversity and singletons, respectively. Lines depict a loess curve (span of 0.01) and shading represents bootstrap-based 95% confidence intervals. Inset plots depict a smaller range on the x-axis.

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 S4).

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 2). The most likely model estimates an ancestral θ of 0.0147 per bp, which translates to an effective population size of , which approximately matches a previous independent estimate [29]. Assuming a mutation rate of $N_a \approx 123,000$ individuals given the mutation rate [30]. The maize population splits from teosinte $\approx$ 15,000 generations in the paste with an initial size of only $\approx 5\%$ of ancestral N_a. Following domestication we estimate considerable gene flow between the populations, estimating $M_{tm} = 1.35$ migrants per generation from teosinte to maize and $M_{mt} = 1.72$ migrants from maize to teosinte. Our model posits exponential growth in maize following its split from teosinte, and estimates a final modern effective population size of $N_m \approx 370,00$.

Because our relatively small sample size limits our ability to characterize the rare variants that are most informative of population expansion, we took advantage of a complementary data set of more than 4,000 maize landraces collected from across the Americas [31] to estimate the modern maize effective population size from low frequency variants. This analysis yields a much higher estimate of the modern maize effective population size at $N_m \approx 993,000$.

Post-domestication expansion is apparent in patterns of singleton diversity. Motivated by the rapid post-domestication expansion of maize evident from our demographic analysis, we investigated whether patterns of diversity at low frequency — and thuse younger — polymorphisms would show similar results in comparisons of maize and teosinte. We thus repeated our analysis of diversity an analysis of genome-wide patterns of diversity using alleles present in only a single individual in the sample. These ingleton diversity measures the abundance of alleles that are only observed in one individual among the sample. As a class, singleton alleles depict the most recent patterns of evolution, but they also have the lowest effect on pairwise diversity. Therefore, unlike pairwise diversity, patterns of singleton diversity reflect recent patterns of evolution. Since our sample size for maize (n=23) was larger than teosinte (n=13), maize singletons were analyzed directly as well as down-sampled to reflect the sample size of teosinte. For singletons, our definition of neutral diversity was defined as diversity at positions at least 0.02 cM distal from genes. When evaluating the data in this manner, a very different pattern emerged compared to what we saw with respect to pairwise diversity. Maize singleton diversity was at least as reduced as teosinte singleton diversity near genes, but recovered more slowly (Figure 4), implying that in the recent past maize has been more influenced by purifying selection than has teosinte.

Together, these findings suggest that demographic history has a strong influence on the effect of purifying selection. Historically, teosinte has had a larger population size than maize, and only recently has maize population size overcome that of teosinte. Since the efficacy of purifying selection scales with population size, these results likely reflect changes in N_e more than they reflect underlying changes in selection pressure.

Similarly, to complement our investigation of pairwise diversity surrounding substitutions, we studied patterns of singleton diversity surrounding synonymous and non-synonymous substitutions in maize (Figure S3). A nearly identical pattern to that shown in Figure 3 was observed. This further demonstrates that even the most recent selection patterns in maize are not dominated by hard sweeps.

Discussion

Little positive selection on new genic mutations. mention divergence time from *Tripsacum*, roughly 1-1.2 million years before present [32].Our findings indicate that hard selective sweeps have not contributed substantially to genome-wide patterns of diversity in maize. More specifically, we observe that positive selection on new genic mutations in maize is not common compared to other drivers of diversity. This finding does not exclude the possibility that hard sweeps have taken place at non-genic sites, such as in those in enhancer regions. One known example of positive selection on a non-genic mutation involves the *tb1* locus of maize, which is one of the best characterized examples of positive selection on a "domestication gene" in any crop [33]. However, the maize *tb1* allele was already present in teosinte before domestication [34], so this this example is in agreement with our finding that hard sweeps are rare. The *gt1* locus is another well-characterized case of positive selection operating on standing variation at an enhancer region [20]. Instances of selection on standing variation such as these, often called soft sweeps, may be a major contributor to maize patterns of diversity [35], and this could be part of the explanation as to why hard sweeps appear to be so rare, despite obvious morphological differences between maize and teosinte.

Unfortunately, our ability to accurately identify soft sweeps, and particularly to distinguish them from hard sweeps, remains limited [36,37]. However, we implemented a scan based on the H12 statistic, which is designed to identify both hard and soft sweeps with reasonable power [28]. Although the goal of this study was not to identify specific sweeps, it may be that the outlier sites distinguished by H12 are primarily composed of soft sweeps instead of hard. Similarly, previous studies scanning for evidence of positive selection in maize [4] may have picked up primarily evidence of soft sweeps. In domesticated species, artificial selection beginning at the onset of domestication represents a drastic shift in selective pressure. One can easily imagine a scenario in which previously neutral or slightly deleterious variants that were segregating in the population when this shift took place suddenly became beneficial, leading to soft sweeps. As appealing as this explanation is, however, our data do not speak to the abundance of soft sweeps, beyond demonstrating that they are one possibility for the lack of observed hard sweeps.

We should additionally note that our observations do not exclude the possibility of infrequent hard sweeps having taken place during maize evolution. In fact, the maize locus *tga1* that has been shown to correspond to a hard sweep at an amino-acid changing mutation [21]. Surrounding *tga1*, our data demonstrate a pattern consistent with a hard sweep. But, our data demonstrate that instances such as this are infrequent. This contrasts sharply with *Drosophila* [38] and *Capsella* [39], where differences in diversity surrounding synonymous and non-synonymous substitutions are clear suggesting hard-sweeps are abundant. However, it agrees with what has been seen for humans [40].

Demography of domestication. add note that demography inference will require regions far from genes, compare to [29,41,42]. cite some theory or empirical paper showing this to be the case in humans/drosophila Since we observed that classical hard sweeps do not explain patterns of maize diversity, it is perhaps unsurprising that we instead found evidence of demographic history contributing substantially to genetic diversity. Our estimate of the demographic history of maize, which is the first estimate of maize demographic history to utilize whole-genome-sequence (WGS) data, allowed us to disentangle the confounding parameters of bottleneck strength and bottleneck intensity. This confounding limited previous domestication estimates [41]. By utilizing the two-dimensional site frequency spectrum (SFS) between maize and teosinte for inference [9] we estimate that the maize effective population size during the bottleneck was approximately 5% that of teosinte. A bottleneck of this magnitude may have enabled moderately deleterious alleles to reach intermediate frequency in maize, while also increasing the probability that strongly deleterious alleles segregating at low frequency in teosinte were purged from the maize gene pool, due to elevated drift as is predicted during a domestication bottleneck [43].

add note here that We note that a population split over 15 thousand generations before present precedes estimates from archaeological data, which suggest maize domestication began no more than 10,000 years before present [16]. This could reflect teosinte population structure that was present before domestication. Also, note that the genetic time of the population split must precede physiological changes that could be identified archaelogically.

After the initial bottleneck, our analysis suggests that maize N_e expanded to reach levels much greater than that of teosinte. Our WGS-based estimate shows that the N_e of landrace maize is at least 3X that of ancestral teosinte, while our singleton-based estimate from the the SeeDs project data (citation) implies an N_e of close to 1 million individuals, or $\approx 8X$ that of ancestral teosinte. note ascertainment of GBS biases it downward From the perspective of a new mutation, both of these estimates are biased downward, since they represent past reductions in diversity due to the aforementioned bottleneck. A back-of-the-envelope calculation of modern N_e can be made by assuming there are 47.9 million ha of landrace maize in production [44], 42,000 plants per ha for open pollinated maize varieties [45], and conservatively that no more than one in 1,000 plants contribute gametes to the next generation. This calculation implies a modern $N_e \geq 2$ billion individuals, and incorporates no individuals from hybrid breeding programs. If for some reason this back-of-the-envelope calculation is an overestimate, even by several orders of magnitude, it is clear that the post-domestication expansion in maize N_e is enormous. Patterns of singleton diversity in maize, as discussed in more detail in the next section, demonstrate that this recent expansion has a notable impact on the maize genome.

Demography influences the efficiency of purifying selection.

The observation that maize pairwise diversity 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 [46], 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 patterns of diversity, which are most heavily influenced by intermediate frequency alleles based on the definition of π [47], 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 4.

A consequence of the inefficient purifying selection that maize experienced during its bottleneck is likely that it harbors more 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 [48]; 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 select

Ultimately, we have shown that purifying selection in maize has operated very differently than purifying selection in teosinte. This observation, along with the potential for soft sweeps, appear to explain the phenotypic divergence pretty hard to argue that purifying selection explains phenotypic divergence! between maize and teosinte much more completely than does positive selection generating hard-sweeps at protein-coding mutations. 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.

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 sequence of inbred accessions of teosinte (Z. *mays* ssp. *parviglumis*) and maize landraces from the Maize HapMap2 panel [22, 49]. We removed **X**teosinte and **X**maize individuals that appeared as outliers in an initial principal component analysis (Figure S5), leaving 13 teosinte and 23 maize that were used for all subsequent analyses (Table S6).

Physical and genetic maps. Sequences were mapped to the maize B73 version 3 reference genome [50] (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/) as described by [22].did we use their bams? I thought we used bwa bams, which are not listed and weren’t used for hapmap2. if that’s the case, we need to mention this difference and acknowledge Robert for providing those bams. either way, we need to include a link to the bamfiles too. somewhere here or above mention URL for Tripsacum bams. did we align those too or download someone else’s? 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 [51]. We converted physical coordinates to genetic coordinates via linear interpolation of the 1cM resolution NAM genetic map [52], i thought this should be wallace not glaubitz? need to specify what you do with physical positions before and after last marker on each chrom and markers, if any, that have physical and genetic order that disagree.

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 [53, 54] of R [55] to parse annotations from the **X**annotation include annotation version number and link to ENSEMBL annotation used. We estimated single population and joint SFS with the software ANGSD [56], 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. i removed options details because they don’t make sense to casual reader and the interested reader can find the script. Because ANGSD cannot calculate a folded joint SFS, we first polarized SNPs using the maize reference genome and then folded spectra using $\delta\alpha\delta i$ [9].

Demographic inference. We used the software $\delta\alpha\delta i$ [9] 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 gives rise to a maize population of size N_b which grows exponentially to size N_m in the present (Figure 2) The model includes migration of M_{mt} individuals each generation from maize to teosinte and M_{tm} individuals from teosinte to maize. M_{mt} and tm are the reverse of what are in figure We fixed N_a using $\delta\alpha\delta i$'s estimation of $\theta = 4N_a\mu$ from the data and a mutation rate of 3×10^{-8} [30]. We estimated all other parameters using 1,000 $\delta\alpha\delta i$ optimizations and allowing initial values to be randomly perturbed by a factor of 2. Initial values and upper and lower bounds can be found in table **X**. this looked terrible in text, please make a supp. table. keep track of notation as e.g. we explain N_b as population size of bneck, but in reality parameter we fit is $\frac{N_b}{N_a}$. my stab at notation is $\frac{N_b}{N_a} = 0.02$, $\frac{N_m}{N_a} = 3$, $\frac{T_b}{2N_a} = 0.04$, and $\frac{M_{mt}}{N_c} = \frac{M_{tm}}{N_a} = 10^{-5}$, where the maize population size Nc at time i generations in the past is $N_c = N_a e^{\frac{ln(\frac{N_m}{N_a})(T_b - i)}{T_b}}$ but if you use these make sure they are right (e.g. is it 2Na or 4Na for Tb?). 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 [31] 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.cite or mention how projection was done. We then estimated N_e from the data assuming $\mu = 3 \times 10^{-8}$ [30] and the relation $4N_e\mu = \frac{S}{L}$ [57], where where S is the total number of singleton SNPs and L is the total number of SNPs in the dataset.

Finally, MSMC text here

Diversity. We made use of the software ANGSD [56] for diversity calculations and genotype calling. We calculated diversity statistics in maize and teosinte in 100bp non-overlapping windows using filters as described above for the SFS. this said 1,000 kb which is 1Mb, but should say 100bp right? 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 ensembl variant effects predictor [58]. 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 [28] in sliding windows of 200 SNPs with a step of 25 SNPs.

Notes:

- should also add the singleton vs pi comparison within species as a supplemental figure. if π/singleton curves **don't** match, will need to explain this suggests teosinte not in equilibrium. good justification i think for msmc model, which we then use to say "gee whiz, not perfect equilibrium but who cares because the big interesting demography is in maize."

ACKNOWLEDGMENTS. We are indebted to Graham Coop and Simon Aeshbacher for their constructive input during this study. 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, Kaeppler, 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. Luikart, G, England, P. R, Tallmon, D, Jordan, S, & Taberlet, P. (2003) Nature Reviews Genetics 4, 981–994.
9. Gutenkunst, R. N, Hernandez, R. D, Williamson, S. H, & Bustamante, C. D. (2009) PLoS genetics 5, e1000695.
10. Li, J, Li, H, Jakobsson, M, Li, S, SjÖDin, P, & Lascoux, M. (2012) Molecular Ecology 21, 28–44.
11. Slotte, T. (2014) Briefings in functional genomics 13, 268–275.
12. Sella, G, Petrov, D. A, Przeworski, M, & Andolfatto, P. (2009) PLoS genetics 5, e1000495.
13. Gazave, E, Ma, L, Chang, D, Coventry, A, Gao, F, Muzny, D, Boerwinkle, E, Gibbs, R. A, Sing, C. F, Clark, A. G, et al. (2014) Proceedings of the National Academy of Sciences 111, 757–762.
14. Shiferaw, B, Prasanna, B. M, Hellin, J, & Bänziger, M. (2011) Food Security 3, 307–327.
15. Strable, J & Scanlon, M. J. (2009) Cold Spring Harbor Protocols 2009, pdb–emo132.
16. Smith, B. D. (1995) The emergence of agriculture. (Scientific American Library New York).
17. Matsuoka, Y, Vigouroux, Y, Goodman, M. M, Sanchez, J, Buckler, E, & Doebley, J. (2002) Proceedings of the National Academy of Sciences 99, 6080–6084.
18. Wilkes, H. G et al. (1967) Teosinte: the closest relative of maize.
19. Doebley, J, Stec, A, & Gustus, C. (1995) Genetics 141, 333.
20. 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.
21. Wang, H, Studer, A. J, Zhao, Q, Meeley, R, & Doebley, J. F. (2015) Genetics pp. genetics–115.
22. 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.
23. Wang, H, Nussbaum-Wagler, T, Li, B, Zhao, Q, Vigouroux, Y, Faller, M, Bomblies, K, Lukens, L, & Doebley, J. F. (2005) Nature 436, 714–719.
24. Enard, D, Messer, P. W, & Petrov, D. A. (2014) Genome research 24, 885–895.
25. Davydov, E. V, Goode, D. L, Sirota, M, Cooper, G. M, Sidow, A, & Batzoglou, S. (2010) PLoS Comput Biol 6, e1001025.
26. 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.
27. Charlesworth, B, Morgan, M, & Charlesworth, D. (1993) Genetics 134, 1289–1303.
28. Garud, N. R, Messer, P. W, Buzbas, E. O, & Petrov, D. A. (2015) PLoS genetics 11, e1005004.
29. 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.
30. Clark, R. M, Tavaré, S, & Doebley, J. (2005) Molecular biology and evolution 22, 2304–2312.
31. Hearne, S, Chen, C, Buckler, E, & Mitchell, S. (2015) Unimputed gbs derived snps for maize landrace accessions represented in the seed-maize gwas panel (). Accessed: 2015-02-16.
32. Ross-Ibarra, J, Tenaillon, M, & Gaut, B. S. (2009) Genetics 181, 1399–1413.
33. Clark, R. M, Wagler, T. N, Quijada, P, & Doebley, J. (2006) Nature genetics 38, 594–597.
34. Studer, A, Zhao, Q, Ross-Ibarra, J, & Doebley, J. (2011) Nature genetics 43, 1160–1163.
35. Beissinger, T. M, Hirsch, C. N, Vaillancourt, B, Deshpande, S, Barry, K, Buell, C. R, Kaeppler, S. M, Gianola, D, & de Leon, N. (2014) Genetics 196, 829–840.
36. Innan, H & Kim, Y. (2004) Proceedings of the National Academy of Sciences of the United States of America 101, 10667–10672.
37. Messer, P. W & Petrov, D. A. (2013) Trends in ecology & evolution 28, 659–669.
38. Sattath, S, Elyashiv, E, Kolodny, O, Rinott, Y, & Sella, G. (2011) PLoS genetics 7, e1001302.
39. Williamson, R, Josephs, E, Platts, A, Hazzouri, K, Haudry, A, Blanchette, M, & Wright, S. (2014) PLoS genetics 10, e1004622–e1004622.
40. 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.
41. 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.
42. Wang, R.-L, Stec, A, Hey, J, Lukens, L, & Doebley, J. (1999) Nature 398, 236–239.
43. Tajima, F. (1989) Genetics 123, 585–595.
44. Program, T. M. (1999) Development, maintenance, and seed multiplication of open-pollinated maize varieties. (CIMMYT, Mexico, D.F.), 2 edition.
45. Van Heerwaarden, J, Van Eeuwijk, F, & Ross-Ibarra, J. (2010) Heredity 104, 28–39.
46. Kimura, M. (1984) The neutral theory of molecular evolution. (Cambridge University Press).
47. Nei, M & Li, W.-H. (1979) Proceedings of the National Academy of Sciences 76, 5269–5273.
48. Meghji, M, Dudley, J, Lambert, R, & Sprague, G. (1984) Crop Science 24, 545–549.
49. Lemmon, Z. H, Bukowski, R, Sun, Q, & Doebley, J. F. (2014) PLoS Genet 10, e1004745.
50. 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.

51. Li, H. (2011) Bioinformatics 27, 2987–2993.
52. Glaubitz, J. C, Casstevens, T. M, Lu, F, Harriman, J, Elshire, R. J, Sun, Q, & Buckler, E. S. (2014) PLoS One 9, E90346.
53. Durinck, S, Spellman, P. T, Birney, E, & Huber, W. (2009) Nature protocols 4, 1184–1191.
54. Durinck, S, Moreau, Y, Kasprzyk, A, Davis, S, De Moor, B, Brazma, A, & Huber, W. (2005) Bioinformatics 21, 3439–3440.
55. R Core Team. (2014) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
56. Korneliussen, T. S, Albrechtsen, A, & Nielsen, R. (2014) BMC bioinformatics 15, 356.
57. Fu, Y.-X & Li, W.-H. (1993) Genetics 133, 693–709.
58. McLaren, W, Pritchard, B, Rios, D, Chen, Y, Flicek, P, & Cunningham, F. (2010) Bioinformatics 26, 2069–2070.

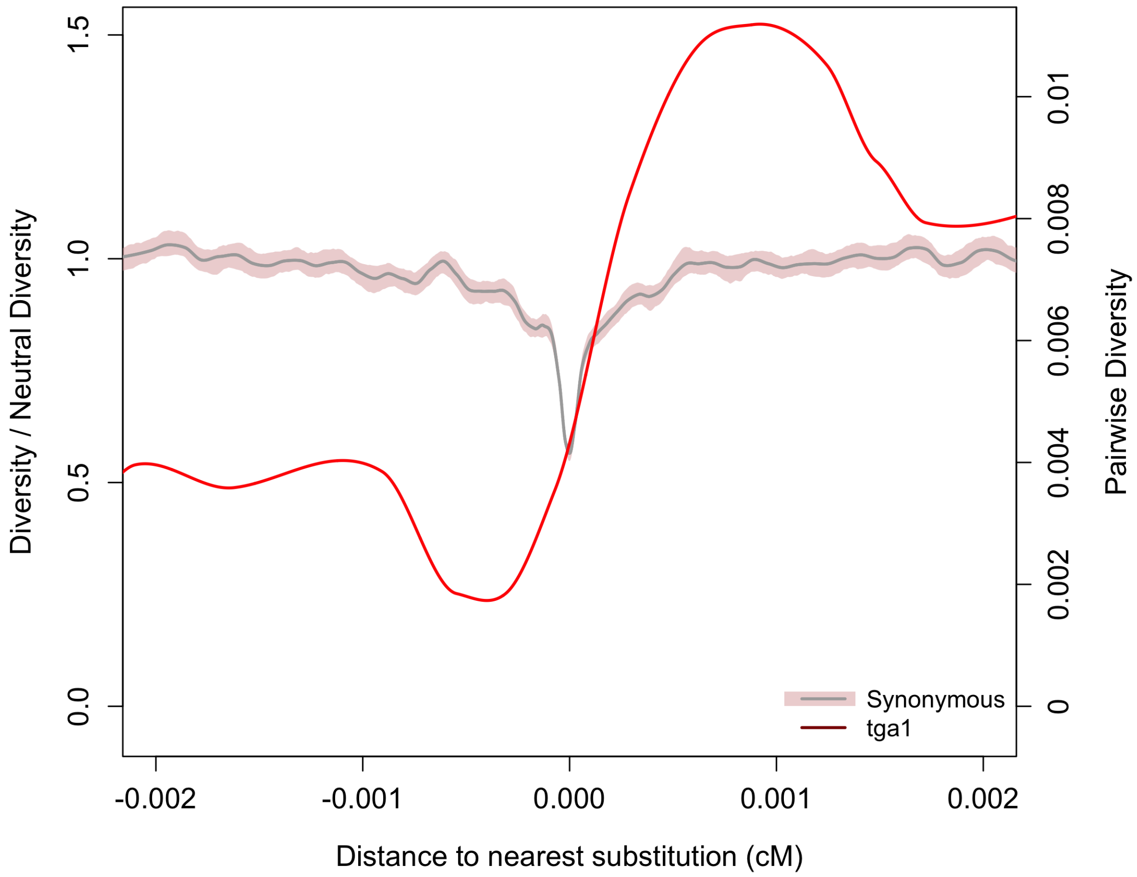

Fig. S1. Diversity surrounding the causitive polymorphism at the *tga1* locus is plotted. Since this is only one gene, the large amount of noise compared to our average plots is expected. However, notice that diversity precisely at the causitive polymorphism is reduced and a recovery of diversity is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

Supporting Information

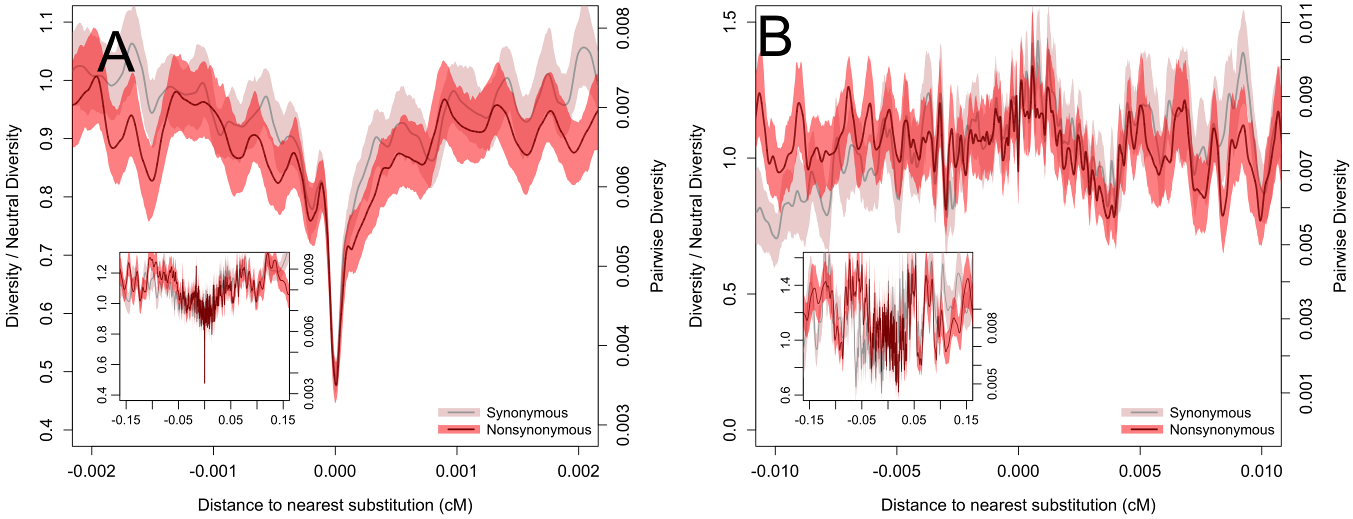

Fig. S2. Pairwise diversity surrounding synonymous and nonsynonymous substitutions in maize at highly conserved (A) or unconserved (B) sites. Bootstrap-based 95% confidence intervals are depicted via shading. Inset plots depict a larger range on the x-axis.

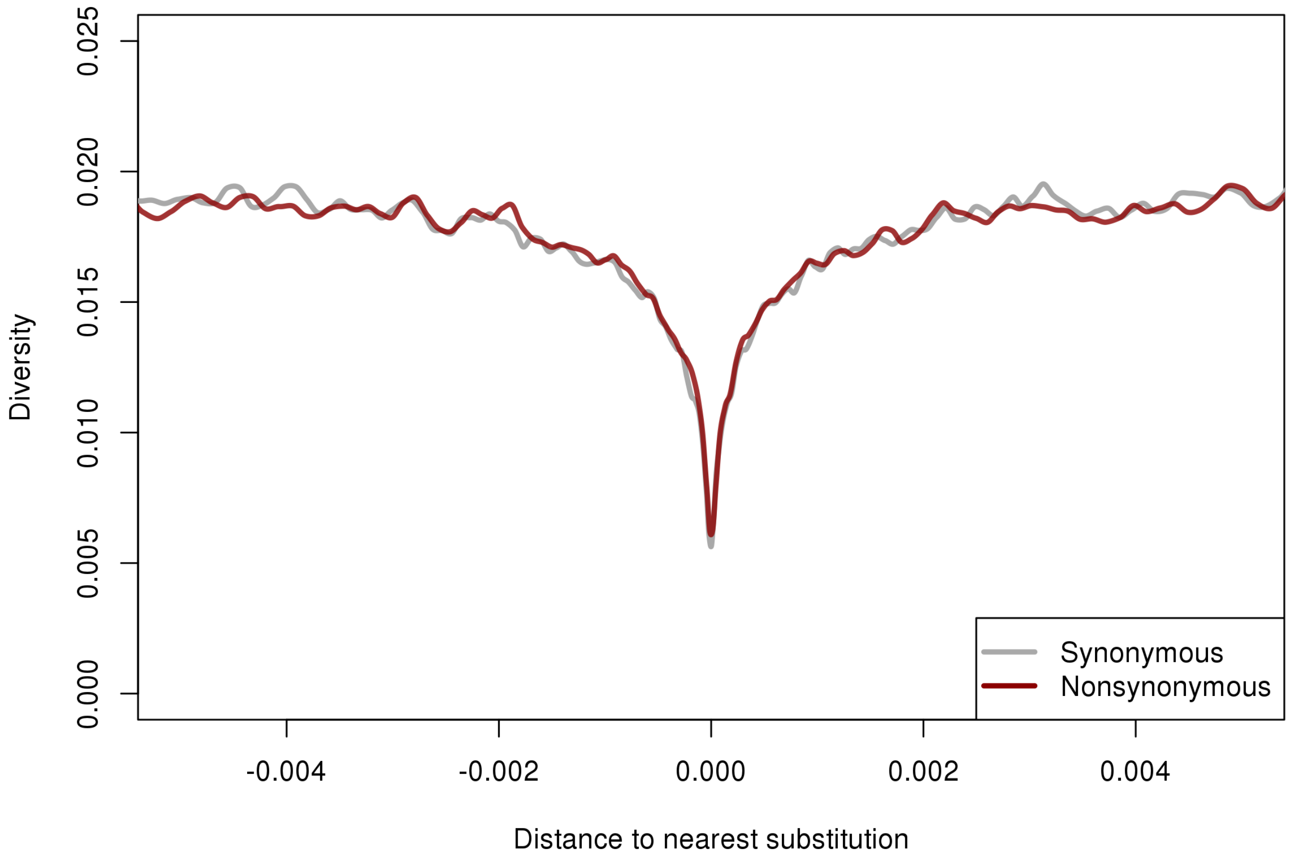

Fig. S3. Singleton diversity surrounding synonymous and nonsynonymous substitutions in maize.

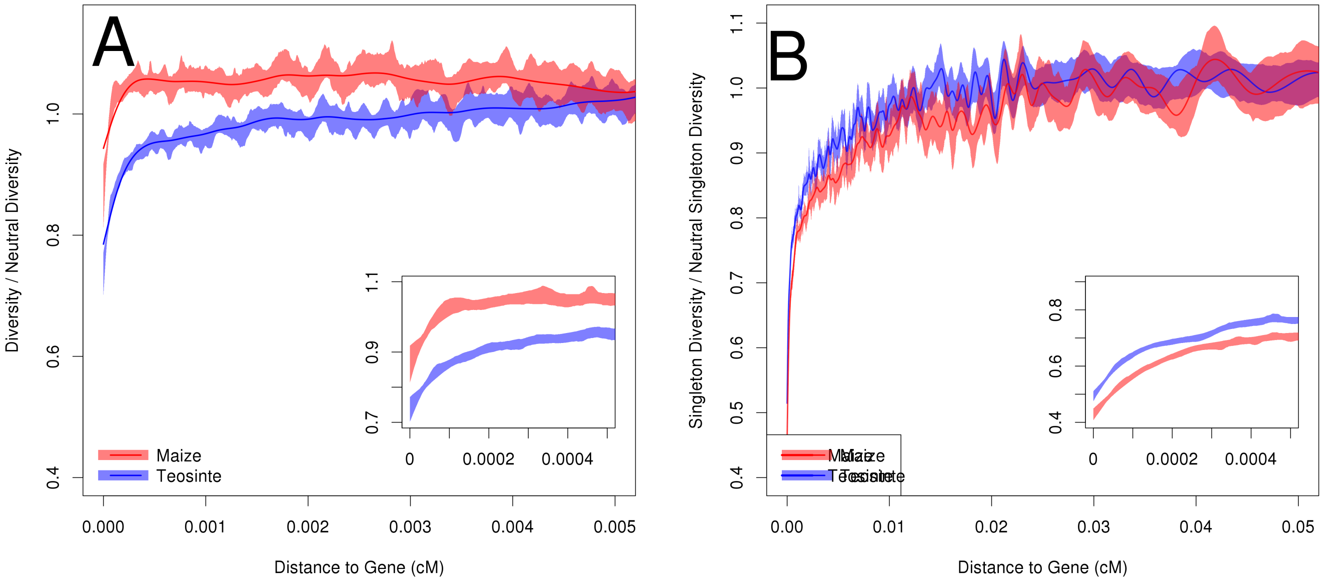

Fig. S4. Relative level of diversity versus distance to the nearest gene, in maize and teosinte, based on only sites that do not show evidence of hard or soft sweeps according to H12. Two measures of diversity were investigated. **A** displays pairwise diversity, which is most influenced by intermediate 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.

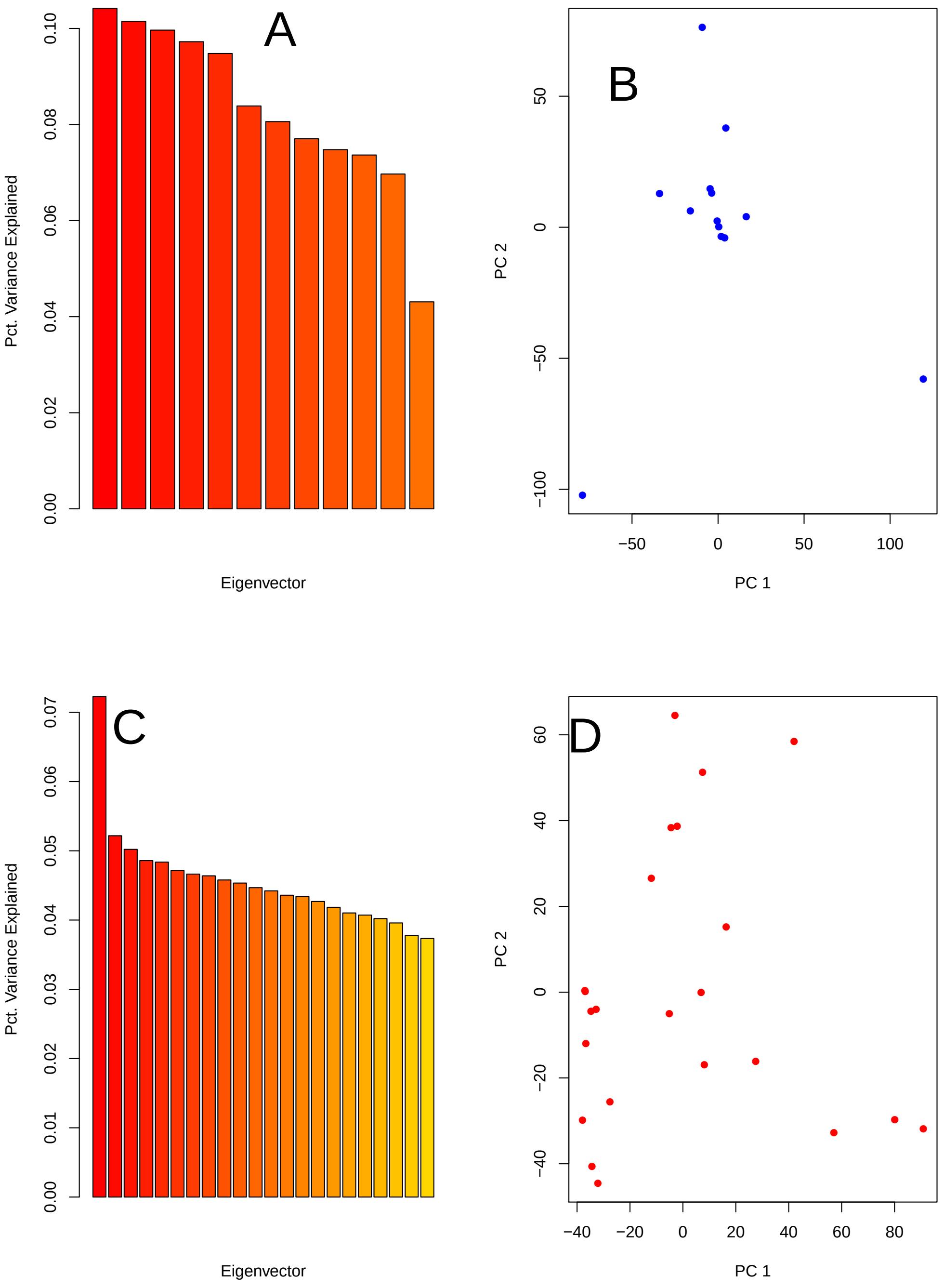

Fig. S5. Principal component analysis of teosinte and maize individuals to ensure that no close relatives were inadvertantly included in our study. Plots are based on a random sample of 10,000 SNPs. **A:** Percentage of total variance explained by each principal component for teosinte. **B:** PC1 vs 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. need to fix letter label positions

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. S6. A list of maize and teosinte individuals included in this study. Sequencing and details were previously described by cite chia and lemmon