Domestication and the impacts of linked selection on the maize genome

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 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 [**?**]. This is a result of both the anthropomorphic nature of artificial selection on domesticates [**?**] as well as the demographic characteristics of the domestication bottleneck(s) that they tend to have experienced [**?**]. 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 [**?**, **?**, **?**, **?**], 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 [**?**, **?**]. 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 [**?**, **?**]. For example, in *Drosophila* it appears that the majority of the genome is impacted by the effects of selection through linkage [**?**]. 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 [**?**], 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 the potential to be adaptive, and how these affect genome-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 [**?**] and as a model for understanding plant evolution [**?**]. Broadly speaking, archaeological and genetic studies have established that maize domestication is likely to have taken place in Mexico approximately 9,000 years bp [**?**, **?**]. Teosinte, the most recent wild ancestor to maize, remains extant throughout much of the Americas [**?**]. Additionally, several large-effect domestication loci [**?**, **?**, **?**] and putative domestication regions [**?**] 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)

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 cite the Corbett-Detig plos-bio paper here – he 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

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 [**?**], 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 [**?**].

Results

Patterns of diversity differ between genic and non-genic regions of the genome. Previous research of the maize domestication process has relied upon observations drawn from genic DNA [**?**,**?**,**?**]. Our data were generated through whole genome sequencing, which eliminated this constraint. Importantly, we observed substantial differences in patterns of diversity between genic and non-genic regions of the genome for both maize and teosinte. For maize, mean pairwise diversity (π) within genes was significantly lower than at positions at least 5 kb away from genes (0.00668 within, 0.00691 away, p<2e-44). The same pattern of significantly greater diversity outside of genes was observed in teosinte, but to a larger extent. For teosinte we observed π within genes to be 0.0088 and at least 5 kb from genes it was 0.0115 (p≈0). These observations suggest that genes are not evolving neutrally in maize or teosinte. Instead, some form of selection is likely reducing diversity within genes. Additionally, these observations suggest that demographic inference, which relies on the assumption of neutrally evolving DNA, will be more accurate if it is based on observations from non-genic genetic material.

Hard sweeps do not shape maize (or teosinte) diversity. A mutation that is immediately beneficial and positively selected leaves a classical hard sweep signature in the genome, whereby surrounding genetic diversity is reduced as the haplotype where the mutation first arose increases in frequency to fixation. The prevalence of such hard sweeps was evaluated by comparing diversity surrounding non-synonymous (specifically missense) and synonymous substitutions in maize and teosinte since divergence from *Tripsacum*, roughly 1-1.2 million years before present [**?**]. For both taxa, no difference in diversity surrounding these classes of substitutions was identified (Figure 2). This observation suggests that hard sweeps are not a primary form of selection for either maize or teosinte.

To ensure that this analysis is resilient to the intricacies of our sample, we investigated diversity surrounding the maize gene *tga1*, one of the few known instances of a hard sweep corresponding to a causal missense substitution in maize [**?**]. In the case of this known sweep, we observed a reduction

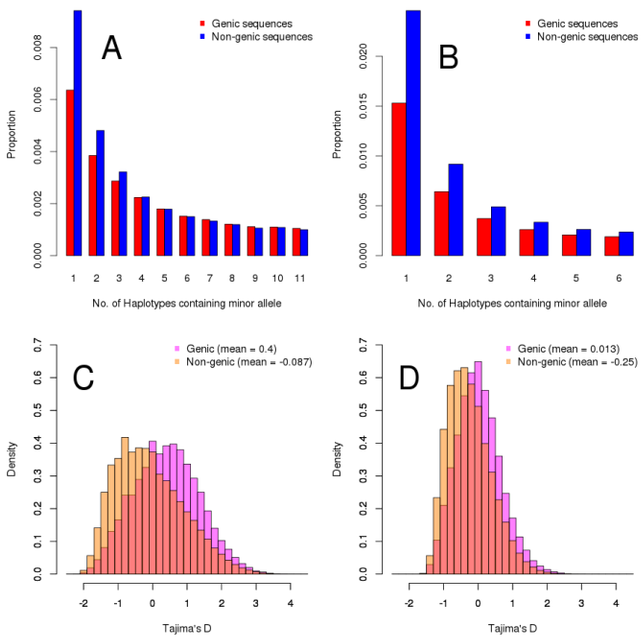

Fig. 1. The folded SFS and Tajimas D were calculated both inside and outside of genes for maize and teosinte. In both taxa, a dearth of rare alleles was observed within genes relative to outside of genes, which led to higher values of Tajima's D outside of genes than inside.**A:** Folded SFS for maize; **B:** Folded SFS for teosinte; **C:** Maize Tajima's D; **D:** Teosinte Tajima's D.

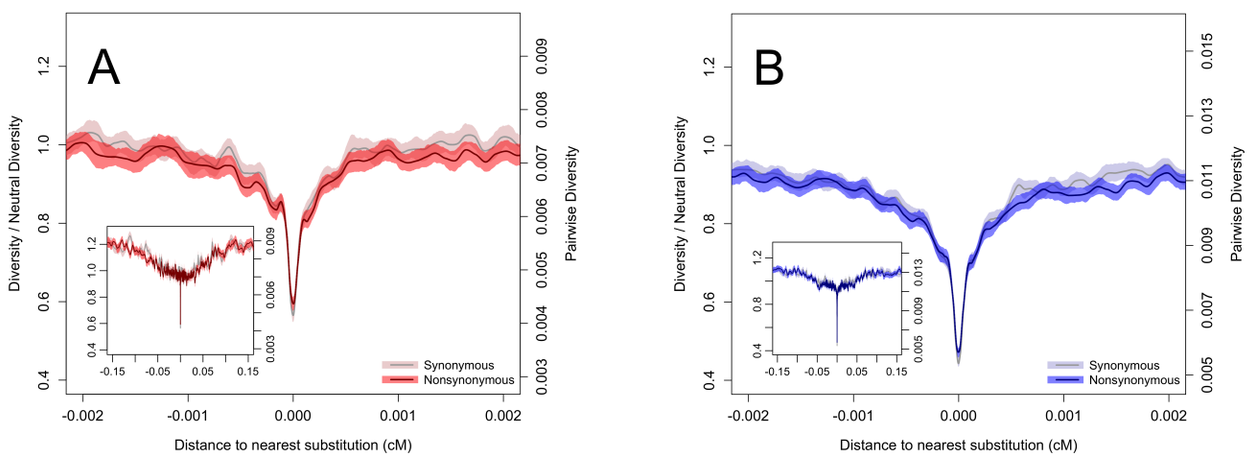

Fig. 2. Pairwise diversity surrounding synonymous and non-synonymous substitutions in maize and teosinte. *a bit too much vertical white space* **A:** Maize diversity; **B:** Teosinte diversity. Bootstrap-based 95% confidence intervals are depicted via shading. Inset plots depict a larger range on the x-axis.

in diversity surrounding the causative substitution to levels much lower than seen for synonymous substitutions (Figure S1) *please add refs to link to supp. figs if possible* , confirming that this method should be informative if hard sweeps were the prevailing mode of selection in maize. Still, to ensure that the approach is robust to potential shortcomings such as limited power in the face of background selection [**?**], we sought to determine if patterns of diversity surrounding synonymous and non-synonymous substitutions differ when we restrict our analysis to the most- or least-conserved genes in maize. When we re-analyzed subsets of genes with the 10% highest and 10% lowest genomic evolutionary rate profile (GERP) score, putatively corresponding to genes undergoing the strongest and weakest purifying selection [**?**, **?**, **?**], we again saw no significant difference in diversity surrounding synonymous and non-synonymous sites (Figure S2).

Pairwise diversity is heavily influenced by purifying selection. Purifying selection refers to the situation where deleterious mutations arising in a population are continuously selected against. When this form of selection is operating, it can serve to reduce genetic variability at linked neutral sites, a phenomenon often called background selection [**?**]. Purifying and background selection lead to lower diversity within genes and other functional sites relative to neutral regions. We investigated purifying selection in maize and teosinte by evaluating the average magnitude of reduced diversity within genes and recovery away from genes in both taxa. To enable comparisons between maize and teosinte, we standardized by neutral levels within each taxa, where neutral levels were defined as mean diversity at positions at least 0.01 cM distal from genes. For both maize and teosinte, pairwise diversity was reduced within genes and a recovery of diversity away from genes was observed. Interestingly, a stronger reduction of diversity and slower recovery was observed for teosinte than for maize, implying that purifying selection has left a more pronounced signature on pairwise diversity in the teosinte genome (Figure 4).

Demography of maize domestication. To explore whether differences in purifying selection between maize and teosinte can be explained by demographic processes, we first estimated the parameters of maize domestication using large-scale sequencing data involving 23 inbred maize landraces and 13 teosinte inbred lines included in the HapMap 2 panel [**?**]. The maize lines were collected from across the Americas and the teosinte lines came from central Mexico. Before estimating demography, we compared the site frequency spectrum (SFS) in genic and non-genic regions, and observed substantial differences in the evolution of these classes of sites. For both maize and teosinte, the SFS within genes showed a dearth of low-frequency alleles and Tajima's D [**?**] was therefore shifted to more positive values, as previously shown (Figure 4). This

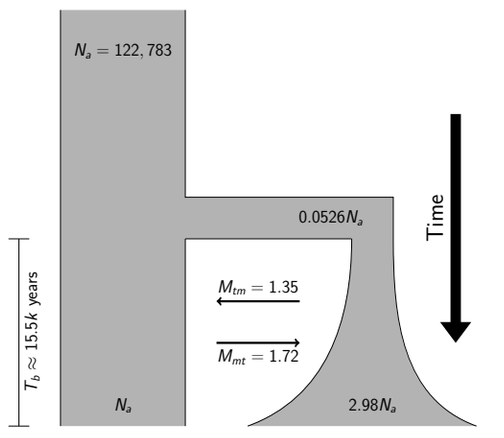

Fig. 3. Parameters of domestication as estimated by dadi. Notably, the maize effective population size (N_e) during the domestication bottleneck appears to have consisted of approximately 5.26% the ancestral N_e, before recovering two at least 2.98 times as large as the ancestral N_e.

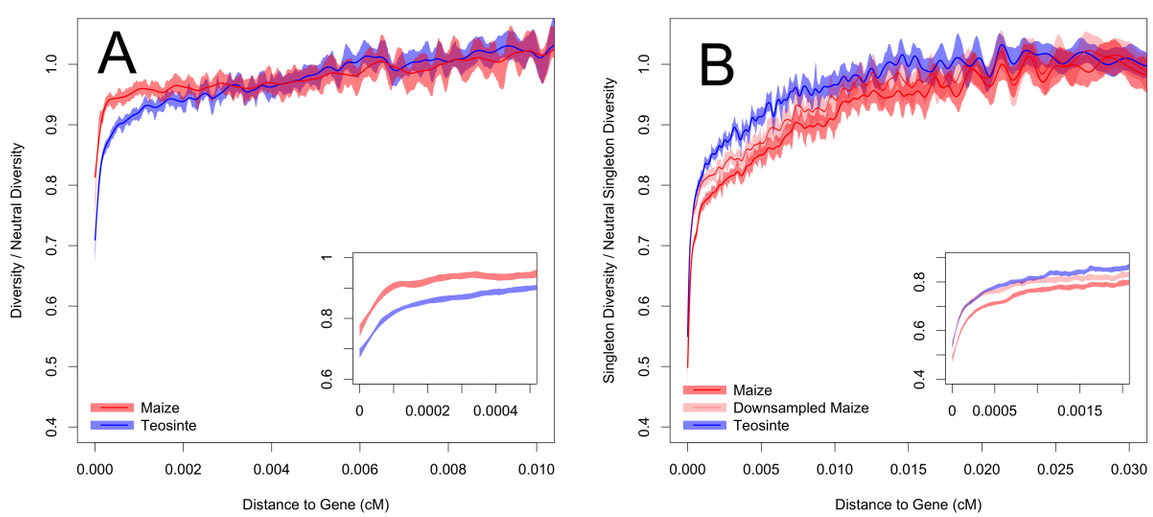

Fig. 4. Relative level of diversity versus distance to the nearest gene, in maize and teosinte. Two measures of diversity were investigated. **A** displays pairwise diversity, which is most influenced by intermediate frequency alleles and therefore depicts more ancient evolutionary patterns, and **B** 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.

is consistent with the aforementioned purifying selection and indicates that genic regions are not evolving neutrally. Therefore, for demographic modeling we restricted analysis to sites at least 5 kb from gene start/stop positions.

We used diffusion approximation as implemented in dadi [**?**] to find the domestication parameters that best explain the joint site frequency spectrum of maize and teosinte. The model we optimized began with an equilibrium ancestral population of size N_a splitting into separate maize and teosinte populations T_B generations in the past. Moving forward in time from the split, teosinte maintained the ancestral population size, while maize experienced an immediate effective population size change to N_b individuals, followed by exponential growth to size N_m. Additionally, since the split M_{tm} individuals migrated from teosinte into maize and M_{mt} individuals migrated from maize to teosinte.

The most likely model suggested an ancestral θ ($4N\mu$) per base pair of 0.014734, which approximately matches a previous independent estimate [**?**]. Assuming a mutation rate of $\mu = 3 \times 10^{-8}$ [**?**], this suggests an effective population size of $N_a = 122,783$ individuals, a population split $T_B = 15,523$ generations in the past, a maize bottleneck effective populations size of $N_b = 6,455$ individuals (5.26% of N_a), a modern maize effective population size of $N_m = 366,973$ individuals, $M_{tm} = 1.35$ migrants per generation from teosinte to maize, and $M_{mt} = 1.72$ migrants per generation from maize to teosinte (Figure 3). 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 [**?**]. 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. Additionally, because recent expansion is most evidenced by rare alleles, and since our limited sample size provides low power to detect rare alleles, we expect that the estimate of $N_m = 366,973$, or $\sim 3N_a$, is likely an underestimate.

Therefore, we utilized a complementary data set of 4,021 maize land-race individuals collected from across the Americas (SeeDs reference here) to obtain a less downward-biased estimate of the current maize effective population size. According to a singleton-based estimate of θ [**?**], which was utilized since rapid post-bottleneck expansion implies that rare alleles will most accurately reflect current maize demography, we obtained the estimate $N_e = 992,713.5$. Although less biased than the previously mentioned estimate, we note that ascertainment bias of the genotyping platform used to generate these data again biases this estimate downward.

Post-domestication expansion is apparent in patterns of singleton diversity

Motivated by the rapid post-domestication expansion in maize that is evident from our demographic analysis, we conducted an analysis of genome-wide patterns of singleton diversity. Singleton 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 2 was observed. This further demonstrates that even the most recent selection patterns in maize are not dominated by hard sweeps.

Purifying selection results are robust to soft-sweeps. As a category, instances of soft-sweeps, or cases of selection on alleles that have already reached intermediate frequency before becoming advantageous, are difficult to identify [**?**]. However, this form of selection is likely to be frequent during crop domestication [**?**]. We were concerned that the disparate patterns of pairwise and singleton diversity in maize and teosinte, which seem to reflect differences in the efficiency of purifying selection due to demographic history, may instead reflect differences in the location and abundance of soft-sweeps. Therefore, we re-analyzed patterns of pairwise and singleton diversity in maize and teosinte including only a subset of sites that are least likely to correspond to soft-sweeps. To achieve this, we performed a genome-wide scan for soft or hard sweeps in maize or teosinte based on the H12 statistic [**?**]. H12 is sensitive to both hard and soft sweeps. Sites nearest to genes that displayed among the top 20% of H12 values in either maize or teosinte were removed from our data set, and both pairwise and singleton diversity at remaining sites, as a function of the distance to the nearest gene, was calculated (Figure S4). This subset of the data is no longer subject to the potentially confounding effects of soft-sweeps, and nearly identical patterns of pairwise and singleton diversity surrounding genes are observed. This provides further evidence that the patterns of diversity we observe in maize as compared to teosinte are the result of rapid population expansion as opposed to differing selective patterns in the domesticated and wild taxa.

Discussion

Little positive selection on new genic mutations. 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 [**?**]. However, the maize *tb1* allele was already present in teosinte before domestication [**?**], 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 [**?**]. Instances of selection on standing variation such as these, often called soft sweeps, may be a major contributor to maize patterns of diversity [**?**], 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 [**?**, **?**]. However, we implemented a scan based on the H12 statistic, which is designed to identify both hard and soft sweeps with reasonable power [**?**]. 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 [**?**] 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 [**?**]. 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* [**?**] and *Capsella* [**?**], 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 [**?**].

Demography of domestication. 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 [**?**]. By utilizing the two-dimensional site frequency spectrum (SFS) between maize and teosinte for inference [**?**] 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 [**?**].

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 ∼8X that of ancestral teosinte. 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 [**?**], 42,000 plants per ha for open pollinated maize varieties [**?**], 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 strongly impacts 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 [**?**], 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 [**?**], 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 [**?**]; 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 analysis scripts are available in an online repository at REPO ADDRESS HERE.

Plant materials. Accessions studied were selected from the Maize HapMap2 panel [**?**, **?**]. *this needs slight clarification. cite chia for hapmap, but that we used higher-coverage of the teosinte from lemmon. if possible, include link to fastq or bamfiles used* Principal component analysis was employed to ensure that closely related individuals were not included due to their potential to bias results (Figure S5). Ultimately, 23 maize inbreds derived from a diverse assortment of landraces were selected for inclusion. Thirteen teosinte inbred lines, all members of subspecies Z. *mays* ssp. *parviglumis*, were utilized. Included lines are listed in (Table S1). Sequences were mapped to the maize B73 version 3 reference genome [**?**] (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/) as described by [**?**].*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.*

Interpolating genetic position.

Physical positions were converted to genetic positions by interpolating from the NAM genetic map [**?**], *does jeff's paper really have the NAM map? I didn't think so and quick look didn't find it. i thought this came from wallace?* which provides a 1 cM resolution for physical to genetic conversion. Within R [**?**], physical positions with corresponding genetic positions in the NAM map were used as anchors. Physical positions in our dataset without corresponding genetic position were assigned genetic positions by scaling the anchored genetic positions according to the physical distance between the unlabeled position and the flanking anchors.

Estimating the site frequency spectrum. To estimate individual and joint site frequency spectra (SFS) for maize and teosinte from inbred lines, each inbred individual was treated as representing a single haplotype from its population. SFS were separately computed for genic and intergenic regions, as well as for the whole genome together. First, genic and intergenic regions were isolated using the biomaRt package [**?**, **?**] of R [**?**]. Genic regions were defined as DNA between the start and stop position of a *you meant transcription start/stop not translation, right?* gene, while intergenic regions were required to be at least 5kb up- or down-stream from a gene start/stop. With regions defined, SFS were estimated with ANGSD [**?**]. Individual population SFS were estimated using all positions observed in at least 80% of the individuals in the population, and joint SFS were estimated using all positions observed in at least 80% of individuals in both populations. Individuals were assumed to be fully inbred (-doSaf 2). Quality filters were employed such that reads with quality score below 30 and bases with quality score below 20 were discarded (-minMapq 30 and -minQ 20), as were reads that didn't map uniquely (-uniqueOnly 0). Quality scores around indels were adjusted as in Samtools (-baq 0). Genotype likelihoods were estimated using the samtools method (-GL 1). Major and minor alleles were inferred from the data (-doMaf 1). Because ANGSD cannot calculate a folded joint SFS, the maize reference genome was used for polarization and then unfolded spectra were folded using $\delta\alpha\delta i$ [**?**].

Demographic inference. We used dadi [**?**] to estimate the parameters of maize domestication based on diffusion approximation of the 2-dimensional SFS for maize and teosinte. Our first step was to specify a model, as shown in (Figure 3). In words, our model began with an ancestral teosinte population with effective population size N_a individuals. The ancestral population split into two distinct populations, one of maize and the other of teosinte, T_b generations in the past. The maize population immediately shrank to size N_b, and then grew exponentially to size N_{maize} today. From the split until today, the teosinte effective population size remained constant. Each year, M_{mt} individuals migrated from maize to teosinte, and M_{tm} individuals migrated from teosinte to maize.The free parameters of our model were $N_a, N_b, N_{maize}, T_b, M_{mt}$, and M_{tm}. We implemented 1,000 dadi optimizations of this model using the "dadi.Inference.optimize_log_fmin" optimizer. In dadi optimizations, initial values for $N_b, N_{maize}, T_b, M_{mt}$, and M_{tm}, respectively, were randomly perturbed up to a factor of two from 0.02, 3, 0.04, 1e-5, and 1e-5. Lower bounds were set at 1e-7, 1e-7, 0, 1e-10, and 1e-10, respecively, while upper bounds were respectively set at 2, 200, 1, .001, and .001. N_a did not have an initial value, upper, or lower bounds because it is estimated from equilibrium ancestral theta ($\theta = 4N_a\mu$), which dadi determines automatically regardless of model. Population sizes were converted from dadi's units using a mutation rate estimate of 3e-8 [**?**]. Among the 1,000 dadi runs, parameters from the run with the lowest log-likelihood were taken as estimates of the parameters of the domestication process.

Recent population expansion, as is observed for maize, most influences the abundance of rare alleles. However, our demographic parameter estimation using dadi suffered from a relatively small sample size of 23 maize individuals. Therefore, our approach likely underestimated current maize effective population size. To address this, we utilized a complimentary dataset of 4,021 maize landrace individuals collected from across the Americas (SeeDs reference here) to generate an independent estimate of N_{maize}. These individuals were genotyped using GBS [**?**], a protocol that suffers from a high rate of missing data [**?**]. To minimize biases imposed from missing data, we used R to isolate only those SNPs that were observed in at least 1,500 individuals and subsequently projected the SFS down to a sample of 500 individuals. Then, we utilized the singleton-based estimate of theta ($\theta = 4N_e\mu = \frac{S}{L}$) to estimate modern maize N_e [**?**], where S is the total number of singletons and L is the total length of DNA sequenced. Based on the GBS protocol implemented [**?**], L is equal to the total number of SNPs in the dataset.

Evaluating diversity around substitutions. To investigate diversity around substitutions, maize and teosinte pairwise diversity was first calculated in 1,000 kb non-overlapping windows using ANGSD [**?**]. This was performed separately for both maize and teosinte, using the same filters as employed for estimating the SFS. Next, SNPs and genotypes among maize, teosinte, and *Tripsacum* were called. *Tripsacum* bam files were downloaded from (TRIPSICUM FILES), and then all SNPs with a p-value less than 1e-6 were called using ANGSD. Quality filters were as the same as before, and genotypes were only called when the posterior probability was above 0.95. From the set of called SNPs and genotypes, substitutions between maize and tripsicum, as well as between teosinte and *Tripsacum* were identified using R [**?**] as all positions with no more than 20% missing data for which every maize or teosinte allele differed from the observed *Tripsacum* allele. At each substitution, effects were estimated using the ensembl variant effects predictor [**?**].

For each diversity window with at least 100 bps observed, the distance from the window center to the nearest synonymous and missense substitution was computed. Then, following the methods of [**?**], a loess curve was plotted for diversity values against the distance to the nearest synonymous or nonsynonymous substitution. A span of 0.01 was utilized.

To verify that this approach is robust to the criticisms of [**?**], which involve reduced power at highly conserved sites, we applied the method to subsets of sites that included only positions nearest to the most- or least-conserved 10% of genes data. This was achieved by assigning every gene a GERP score corresponding to the average score observed across it [**?**]. Next, genes ing the top and bottom 10% quantiles of GERP score were isolated. Two datasets were created, each including all diversity windows who's nearest gene was in the high or low GERP score sets. Finally, Loess fits of diversity vs. distance to gene were computed from sites within each GERP score.

Evaluating diversity around genes. Two types of diversity surrounding genes were investigated. The first was pairwise diversity in 1kb windows, as described previously. The second was singleton diversity in 1kb windows. Singletons represent the rarest class of alleles that this dataset can identify, and collectively demonstrate the most recent patterns of evolution. Minor allele frequencies were estimated with ANGSD [**?**] using the same quality filters previously described. Then, the number of singletons in each non-overlapping 1kb window was calculated with R for both maize and teosinte [**?**]. For maize, we also generated a parallel set of downsampled singleton data, for which binomial sampling within R was conducted based on allele frequencies at each site, to generate a singleton psuedo dataset of sample size 13, equivalent to our sample size for teosinte. BiomaRt [**?**, **?**] was then used to identify the center of each gene. Next, the distance from each diversity window center to the nearest gene center was computed.

Teosinte diversity is generally higher then maize diversity. Therefore, to enable comparisons between the reduction of diversity around genes in maize and teosinte, a neutral measure of pairwise and singleton diversity for each taxa was estimated. For pairwise diversity, this nuetral measure was defined according to mean pairwise diversity at windows greater than 0.01 cM from the nearest gene. For singletons, both maize and teosinte still showed reduced diversity at a distance of 1 cM, so neutral diversity in this case was defined as mean singleton diversity at positions 0.02 cM or greater from genes. Then, pairwise and singleton diversity at each window was standardized by dividing by the corresponding neutral measure. Separately for pairwise and singleton diversity in maize and teosinte, cubic smoothing splines were fit to describe diversity levels according to the distance to the nearest gene. Significant differences were assessed by taking 100 bootstrap samples from each set of diversity windows and re-fitting the cubic smoothing spline to each. Then, the 2.5% and 97.5% quantiles of values along the bootstrapped splines were identified.

Additionally, we performed a parallel implementation of the above analysis after excluding sites potentially experiencing positive selection in the form of either hard- or soft-sweeps. To isolate these sites, we performed genome-scans separately for maize and teosinte based on the H12 statistic [**?**]. SNPs within each taxa were called in ANGSD [**?**] using the previously specified SNP calling parameters. H12 was computed with a window size of 200 SNPs and a jump of 25 SNPs. H12 windows in the highest 20% for maize or teosinte were identified and considered as regions most likely to be under positive selection. Finally, levels of pairwise and singleton diversity vs. distance to gene was computed using only sites that were not nearest to selected sites in either maize or teosinte.

ACKNOWLEDGMENTS. We are indebted to Graham Coop and Simon Aeshbacher for their constructive input during this study. Funding was provided by NSF Grant number ZZZZZZZZZ.