# Recent demography drives changes in linked selection across the maize genome

Timothy M. Beissinger[1,2,3,*], Li Wang[4], Kate Crosby[1], Arun Durvasula[1], Matthew B. Hufford[4], Jeffrey Ross-Ibarra[1,5,*],

**1 Department of Plant Sciences, University of California, Davis, Caliornia, USA**
**2 US Department of Agriculture, Agriculture Research Service, Columbia, Missouri, USA**
**3 Division of Plant Sciences, University of Missouri, Columbia, Missouri, USA**
**4 Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa, USA**
**5 Genome Center and Center for Population Biology, University of California, Davis, California, USA**

**\* rossibarra@ucdavis.edu or beissingert@missouri.edu**

## Abstract

Genetic diversity is shaped by the interaction of drift and selection, but the details of this interaction are not well understood. Although a population's propensity to drift is typically summarized by its long-term effective population size ($N_e$), rapidly changing population demographics complicate this relationship. To better understand how changing demography impacts selection, we used whole-genome sequencing data to investigate patterns of linked selection in domesticated and wild maize (teosinte). We show that maize went through a domestication bottleneck with a population size of approximately 5% that of teosinte before it experienced rapid expansion post-domestication. We observe that hard sweeps on genic mutations are not the primary force driving maize evolution. As expected, a reduced population size during domestication decreased the efficiency of purifying selection to purge deleterious alleles from maize, but rapid expansion after domestication has since increased the efficiency of purifying selection to levels exceeding those seen in teosinte. This final observation demonstrates that rapid demographic change can have wide-ranging impacts on diversity that conflict with would-be expectations based on long-term $N_e$.

## Author Summary

Both selection and demographic change play important roles in shaping genetic diversity, but there are few empirical examples of the interplay between them. Although it is well known that both diversity and the efficacy of natural selection are reduced in smaller populations, the short-term consequences of demographic change on linked selection have not been well studied. We use wild and domesticated maize to demonstrate that rapid changes in population size impact the effect of selection on linked sites in ways not well described by long-term average population size. We also

produce an unbiased estimate of maize demographic history and show that positive selection on new genic mutations plays a lesser role in shaping diversity than purifying selection against deleterious mutations.

## Introduction

The genetic diversity of populations is determined by a constant interplay between genetic drift and natural selection. Drift is a consequence of a finite population size and the random sampling of gametes each generation [1]. In contrast to the stochastic effects of drift, selection systematically alters allele frequencies by favoring particular alleles at the expense of others as a result of their effects on fitness. Researchers often study drift by excluding potentially selected sites [2–4], or selection by focusing on site-specific patterns under the assumption that genome-wide diversity reflects primarily the action of drift [5].

Drift and selection do not operate independently to determine genetic variability, however, in large part because linkage allows the effects of selection to be wide-ranging [6–8]. Linked selection can take the form of hitch-hiking, when the frequency of a neutral allele changes as a result of positive selection at a physically linked site [6], or background selection, where diversity is reduced at loci linked to a site undergoing selection against deleterious alleles [9]. Recent work in *Drosophila*, for example, has shown that virtually the entire genome is impacted by the combined effects of these processes [10–12].

The impact of linked selection, in turn, is heavily influenced by the effective population size ($N_e$), as the efficiency of natural selection is proportional to the product $N_e s$, where $s$ is the strength of selection on a variant [8,13–15]. The effective size of a population is not static, and nearly all species, including flies [16], humans [17], domesticates [18,19], and non-model species [20] have experienced recent or ancient changes in $N_e$. Although much is known about how the long-term average $N_e$ affects linked selection [13], relatively little is understood about the immediate effects of more recent changes in $N_e$ on patterns of linked selection.

Because of its relatively simple demographic history and well-developed genomic resources, maize (*Zea mays*) represents an excellent organism to study these effects. Archaeological and genetic studies have established that maize domestication began in Central Mexico at least 9,000 years bp [21,22], and involved a population bottleneck followed by recent expansion [23–25]. Because of this simple but dynamic demographic history, domesticated maize and its wild ancestor teosinte can be used to understand the effects of changing $N_e$ on linked selection. In this study, we leverage the maize-teosinte system to study these effects by first estimating the parameters of the maize domestication bottleneck using whole-genome resequencing data and then investigating the relative importance of different forms of linked selection on diversity in the ancient and more recent past. We show that, while patterns of overall nucleotide diversity reflect long-term differences in $Ne$, recent growth following domestication qualitatively changes these effects, thereby illustrating the importance of a comprehensive understanding of demography when considering the effects of selection genome-wide.

# RESULTS

## Patterns of diversity differ between genic and intergenic regions of the genome

To investigate how demography and linked selection have shaped patterns of diversity in maize and teosinte, we analyzed data from 23 maize and 13 teosinte genomes from the maize HapMap 2 and HapMap 3 projects [26, 27]. As a preliminary step, we evaluated levels of diversity inside and outside of genes across the genome. We find broad differences in genic and intergenic diversity consistent with earlier results [28](Figure 1). In maize, mean pairwise diversity ($\pi$) within genes was significantly lower than at sites 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.

**Figure 1. Diversity and Tajima's D in maize and teosinte. A** and **B** show mean pairwise diversity $\pi$, $\pm$ one standard deviation, while **C** and **D** depict and Tajimas D in 1kb windows from genic and nongenic regions of maize and teosinte.

## Demography of maize domestication

After establishing that genic sites are not evolving neutrally, we used sites > 5kb from genes to estimate the parameters of a simple domestication bottleneck model (Figure 2). The most likely model estimates an ancestral population mutation rate of $\theta = 0.0147$ per bp, which translates to an effective population size of $N_a \approx 123,000$ teosinte individuals. We estimate that maize split from teosinte $\approx 15,000$ generations in the past, with an initial size of only $\approx 5\%$ of the ancestral $N_a$. 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 have continued 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.

**Figure 2. Parameter estimates for a basic bottleneck model of maize domestication.** See methods for details.

In addition to our simple bottleneck model, we investigated two alternative approaches for demographic inference. First, we utilized genotyping data from more than 4,000 maize landraces [29] to estimate the modern maize effective population size 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 applied a model-free coalescent approach [30] using a subset of our samples. Though this analysis suggests non-equilibrium dynamics for teosinte not included in our initial model, it is nonetheless broadly consistent, identifying a clear domestication bottleneck followed by rapid population expansion in maize to an extremely large extant size of $\approx 10^9$ (Figure S2). Our assessment of the historical demography of maize and teosinte provides context for subsequent analyses of linked selection.

## Hard sweeps do not explain diversity differences

When selection increases the frequency of a new beneficial mutation, a signature of reduced diversity is left at surrounding linked sites [6]. To evaluate whether patterns of such "hard sweeps" could explain observed differences in diversity between genic and intergenic regions of the genome, we compared diversity around missense and synonymous substitutions between *Tripsacum* and either maize or teosinte. If a substantial 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 maize domestication locus *tga1* (Figure S1), likely the result of a hard sweep during domestication [31, 32]. Genome-wide, however, we observe no differences in diversity at sites near synonymous versus missense substitutions in either maize or teosinte (Figure 3).

**Figure 3. Pairwise diversity surrounding synonymous and missense substitutions in A: maize and B: teosinte.** Axes show absolute diversity values (right) and values relative to mean nucleotide diversity in windows $\geq 0.01cM$ from a substitution (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.

Previous analyses have suggested that this approach may have limited power because a relatively high proportion of missense substitutions will be found in genes that, due to weak purifying selection, have higher genetic diversity [33]. To address this concern, we took advantage of genome-wide estimates of evolutionary constraint [34] calculated using genomic evolutionary rate profile (GERP) scores [35]. We then evaluated substitutions only in subsets of genes in the highest and lowest 10% quantile of mean GERP score, putatively representing genes under the strongest and weakest purifying selection. As expected, we see higher diversity around substitutions in genes under weak purifying selection, but we still find no difference in diversity near synonymous and missense substitutions in either subset of the data (Figure S3). 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

In the case of purifying or background selection, diversity is reduced in functional regions of the genome via removal of deleterious mutations [9]. We investigated purifying selection in maize and teosinte by evaluating the reduction of diversity around 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 these subspecies [26, 36]. The reduction in relative diversity is more pronounced in teosinte, reaching lower levels in genes and occurring over a wider region.

**Figure 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 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.

Our previous comparison of synonymous and missense substitutions has low power to detect the effects of selection acting on multiple beneficial mutations or standing genetic variation, because in such cases diversity around the substitution may be reduced to a lesser degree [37,38]. Nonetheless, such "soft sweeps" are still expected to occur more frequently in functional regions of the genome and could provide an alternative explanation to purifying selection for the observed reduction of diversity at linked sites in genes. To test this possibility, we performed a genome-wide scan for selection using the H12 statistic, a method expected to be sensitive to both hard and soft sweeps [39]. Qualitative differences between maize and teosinte in patterns of diversity within and outside of genes remained unchanged even after removing genes in the top 20% quantile of H12 (Figure S6A). We interpret these combined results as suggesting that purifying selection has predominantly shaped diversity near genes and left a more pronounced signature in the teosinte genome due to the increased efficacy of selection resulting from differences in long-term effective population size.

## Population expansion leads to stronger purifying selection in modern maize

Motivated by the rapid post-domestication expansion of maize evident in our demographic analyses, we reasoned that low-frequency — and thus younger — polymorphisms might show patterns distinct from pairwise diversity. Singleton diversity around missense and synonymous substitutions (Figure S4) appears nearly identical to results from pairwise diversity (Figure 3), providing little support for a substantial recent increase in the number or strength of hard sweeps occurring in maize.

In contrast, we observe a significant shift in the effects of purifying selection: singleton polymorphisms are more strongly reduced in and near genes in maize than in teosinte, even after downsampling our maize data to account for differences in sample size (Figure 4B). This result is the opposite of the pattern observed for $\pi$, where teosinte demonstrated a stronger reduction of diversity in and around genes than did maize. As before, this relationship remained after we removed the 20% of genes with the highest H12 values (Figure S6). Finally, while direct comparison of pairwise and singleton diversity within taxa is consistent with non-equilibrium dynamics in teosinte, these too reveal much stronger differences in maize (Figure S5).

# Discussion

## Demography of domestication

Although a number of authors have investigated the demography of maize domestication [23–25], these efforts relied 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, such that even synonymous polymorphisms in genes are not representative of diversity at unconstrained sites. This implies that genic polymorphism data are unable to tell the complete or accurate demographic history of maize, but the rapid recovery of diversity outside of genes demonstrates that sites far from genes can be reasonably used for demographic inference. Furthermore, by utilizing the full joint SFS, we are able to estimate population growth, gene flow, and the strength of the domestication bottleneck without making assumptions about its duration.

One surprising result from our model is the estimated timing of domestication at $\approx 15,000$ years before present. While this appears to conflict with archaeological

estimates [40], we emphasize that this estimate reflects the fact that the genetic split   161
between populations likely preceded anatomical changes that can be identified in the   162
archaeological record. We also note that our result may be inflated due to population   163
structure, as our geographically diverse sample of teosinte may include populations   164
diverged from those that gave rise to maize.   165

The estimated bottleneck of $\approx 5\%$ of the ancestral teosinte population seems low   166
given that maize landraces exhibit $\approx 80\%$ of the diversity of teosinte [28], but our   167
model suggests that the effects of the bottleneck on diversity are likely ameliorated by   168
both gene flow and rapid population growth (Figure 2). Although we estimate that   169
the modern effective size of maize is larger than teosinte, the small size of our sample   170
reduces our power to identify the low frequency alleles most sensitive to rapid   171
population growth [41], and our model is unable to incorporate growth faster than   172
exponential. Both alternative approaches we employ estimate a much larger modern   173
effective size of maize in the range of $\approx 10^6 - 10^9$, an order of magnitude or more than   174
the current size of teosinte. Census data suggest these estimates are plausible: there   175
are 47.9 million ha of open-pollinated maize in production [42], likely planted at a   176
density of $\approx 25,000$ individuals per hectare [43]. Assuming the effective size is only   177
$\approx 0.4\%$ of the census size (i.e. 1 ear for every 1000 male plants), this still implies a   178
modern effective population size of more than four billion. While these genetic and   179
census estimates are likely inaccurate, all of the evidence points to the fact that the   180
effective size of modern maize is extremely large.   181

## Hard sweeps do not shape genome-wide diversity in maize   182

Irrespective of ancient or current population size, our findings demonstrate that classic   183
hard selective sweeps have not contributed substantially to genome-wide patterns of   184
diversity in maize, a result we show is robust to concerns about power due to the   185
effects of purifying selection [33]. Although our approach ignores the potential for   186
hard sweeps in noncoding regions of the genome, a growing body of evidence argues   187
against hard sweeps as the prevalent mode of selection shaping maize variability.   188
Among well-characterized domestication loci, only the gene *tga1* shows evidence of a   189
hard sweep on a missense mutation [32], while several loci are consistent with "soft   190
sweeps" from standing variation [44, 45] or multiple mutations [46]. Moreover,   191
genome-wide studies of domestication [28], local adaptation [47] and modern   192
breeding [48, 49] all support the importance of standing variation as primary sources of   193
adaptive variation. Soft sweeps are expected to be common when $2N_e\mu_b \geq 1$, where $\mu_b$   194
is the mutation rate of beneficial alleles with selection coefficient $s_b$ [38]. Assuming a   195
mutation rate of $3 \times 10^{-8}$ [50] and that on the order of $\approx 1 - 5\%$ of mutations are   196
beneficial [51], this implies that soft sweeps should be common in both maize and   197
teosinte for mutational targets $>> 10kb$ — a plausible size for quantitative traits or   198
for regulatory evolution targeting genes with large up- or down-stream control   199
regions [44, e.g.]. Indeed, many adaptive traits in both maize [52] and teosinte [53] are   200
highly quantitative, and adaptation in both maize [28] and teosinte [54] has involved   201
selection on regulatory variation.   202

The absence of evidence for a genome-wide impact of hard sweeps in coding regions   203
differs markedly from observations in *Drosophila* [55] and *Capsella* [56], but is   204
consistent with data from humans [57, 58]. Comparisons of the estimated percentages   205
of nonsynonmyous substitutions fixed by natural selection [10, 56, 59, 60] give similar   206
results. While differences in long-term $N_e$ likely explains some of the observed   207
variation across species, we see little change in the importance of hard sweeps in genes   208
in singleton diversity in modern maize (Figure S4), perhaps suggesting other factors   209
may contribute to these differences as well. One possibility, for example, is that, if   210
mutational target size scales with genome size, the larger genomes of human and   211

maize may offer more opportunities for noncoding loci to contribute to adaptation, with hard sweeps on nonsynonymous variants then playing a relatively smaller role. Support for this idea comes from numerous cases of adaptive transposable element insertion modifying gene regulation in maize [44,61–63] and studies of local adaptation that show enrichment for SNPs in regulatory regions in teosinte [54] and humans [64] but for nonsynonymous variants in the smaller *Arabidopsis* genome. Our results, for example, are not dissimilar to findings in the comparably-sized mouse genome, where no differences are seen in diversity around nonsynonymous and synonymous substitutions in spite of a large $N_e$ and as many as 80% of adaptive substitutions occuring outside of genes [65]. Future comparative analyses using a common statistical framework (e.g. [14] )and considering additional ecological and life history factors (c.f. [15]) should allow explicit testing of this idea.

## Demography influences the efficiency of purifying selection

One of our more striking findings is that the impact of purifying selection on maize and teosinte qualitatively changed over time. We observe a more pronounced decrease in $\pi$ around genes in teosinte than maize (Figure 4A), but the opposite trend when we evaluate diversity using singleton polymorphisms (Figure 4B). The efficiency of purifying selection is proportional to effective population size [66], and these results are thus consistent with our demographic analyses which show a domestication bottleneck and smaller long-term $N_e$ in maize [23–25,59] followed by recent rapid expansion and a much larger modern $N_e$.

Although demographic change affects the efficiency of purifying selection, it may have limited implications for genetic load. Recent population bottlenecks and expansions have increased the relative abundance of rare and deleterious variants in domesticated plants [67,68] and human populations out of Africa [41,69], and such variants may play an important role in phenotypic variation [69–71]. Nonetheless, demographic history may have little impact on the overall genetic load of populations [72,73], as decreases in $N_e$ that allow weakly deleterious variants to escape selection also help purge strongly deleterious ones, and the increase of new deleterious mutations in expanding populations is mitigated by their lower initial frequency and the increasing efficiency of purifying selection [73–75].

## Rapid changes in linked selection

Our results demonstrate that consideration of long-term differences in $N_e$ cannot fully capture the dynamic relationship between demography and selection. While a number of authors have tested for selection using methods that explicitly incorporate or are robust to demographic change [60,76,77] and others have compared estimates of the efficiency of adaptive and purifying selection across species [78] or populations [79], previous analyses of the impact of linked selection on genome-wide diversity have relied on single estimates of the effective population size [14,15]. Our results show that demographic change over short periods of time can quickly change the dynamics of linked selection: mutations arising in extant maize populations are much more strongly impacted by the effects of selection on linked sites than would be suggested by analyses using long-term effective population size. As many natural and domesticated populations have undergone considerable demographic change in their recent past, long-term comparisons of $N_e$ are likely not informative about current processes affecting allele frequency trajectories.

# Materials and Methods

## BASH, R, and Python scripts

All scripts used for analysis are available in an online repository at
https://github.com/timbeissinger/Maize-Teo-Scripts.

## 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 [26, 27, 80]. 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 [81] (Figure S7), leaving 13 teosinte and
23 maize that were used for all subsequent analyses (Table S1). We also utilized a
single individual of (*Tripsacum dactyloides*) as an outgroup. All bam files are available
at /iplant/home/shared/panzea/hapmap3/bam_internal/v3_bams_bwamem.

## Physical and genetic maps

Sequences were mapped to the maize B73 version 3 reference genome [82]
(ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/) as
described by [27]. All analyses made use of uniquely mapping reads with mapping
quality score $\geq 30$ and bases with base quality score $\geq 20$; quality scores around indels
were adjusted following [83]. We converted physical coordinates to genetic coordinates
via linear interpolation of the previously published 1cM resolution NAM genetic
map [84].

## 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 [85, 86] of R [87] to parse
annotations from genebuild version 5b of AGPv3. We estimated single population and
joint SFS with the software ANGSD [88], 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 $\delta a \delta i$ [4].

## Demographic inference

We used the software $\delta a \delta i$ [4] 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 2). 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 $\delta a \delta i$'s estimation of
$\theta = 4N_a\mu$ from the data and a mutation rate of $\mu = 3 \times 10^{-8}$ [50]. We estimated all
other parameters using 1,000 $\delta a \delta i$ 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 S2. 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 partially imputed maize landraces [29] 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 $\mu = 3 \times 10^{-8}$ [50] and the relation $4N_e\mu = \frac{S}{L}$ [89], where where $S$ is the total number of singleton SNPs and $L$ is the total number of SNPs in the dataset.

As a final estimate of demography, we employed MSMC [30] 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 [88] using a SNP p-value of $1e - 6$ against a reference genome masked using SNPable (`http://lh3lh3.users.sourceforge.net/snpable.shtml`). We then removed heterozygous genotypes and filtered sites with a mapping quality $< 30$, a base quality $< 20$, or a $|log_2(\text{depth})| < 1$. We ran MSMC with pattern parameters $20 \times 2 + 20 \times 4 + 10 \times 2$.

## Diversity

We made use of the software ANGSD [88] 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, or missense using the ensembl variant effects predictor [90]. 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 [39] in sliding windows of 200 SNPs with a step of 25 SNPs.

# Supporting Information

**Figure S1. Diversity surrounding the causitive substitution at the *tga1* locus.**

**Figure S2. Effective population size estimated over time using MSMC.** Shown are estimates (solid lines) and boostrap resampling (dotted lines) for both maize (red) and teosinte (blue). Time is estimated assuming an annual generation time and a mutation rate of $\mu = 3 \times 10^{-8}$

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

**Figure S4. Singleton diversity surrounding synonymous and nonsynonymous substitutions in maize.**

**Figure 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 $\geq 0.02cM$ 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.

**Figure 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.

**Figure S7. 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** displays the percentage of total variance explained by each principal component for teosinte, while **B** shows PC1 vs PC2 for all 13 teosinte individuals. Simlarly, **C** depicts the percentage of total variance explained by each principal component for maize, and **D** shows PC1 vs PC2 for all 23 maize individuals.

**Table S1. A list of maize and teosinte individuals included in this study.** Sequencing and details were previously described by [26]

**Table S2. Parameters, initial values, and boundaries used for model-fitting with $\delta a \delta i$.** Parameters are shown in the units utilized by $\delta a \delta i$, although in the text simplified units are reported.

# Acknowledgments <span style="float:right">337</span>

<span style="float:right">338<br>339<br>340<br>341</span>

# References

1. Dobzhansky T, Pavlovsky O. An experimental study of interaction between genetic drift and natural selection. Evolution. 1957;p. 311–319.

2. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(51):18508–18513.

3. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genetics. 2003;4(12):981–994.

4. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS genetics. 2009;5(10):e1000695.

5. Akey JM. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome research. 2009;19(5):711–722.

6. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetical research. 1974;23(01):23–35.

7. Li J, Li H, Jakobsson M, Li S, SjÖDin P, Lascoux M. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Molecular Ecology. 2012;21(1):28–44.

8. Slotte T. The impact of linked selection on plant genomic variation. Briefings in functional genomics. 2014;13(4):268–275.

9. Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289–1303.

10. Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? PLoS genetics. 2009;5(6):e1000495.

11. Elyashiv E, Sattath S, Hu TT, Strustovsky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in Drosophila. arXiv preprint arXiv:14085461. 2014;.

12. Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. Nature. 2005;437(7062):1149–1152.

13. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics. 2013;14(4):262–274.

14. Corbett-Detig RB, Hartl DL, Sackton TB. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLoS Biol. 2015 04;13(4):e1002112. Available from: `http://dx.doi.org/10.1371%2Fjournal.pbio.1002112`.

15. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species. PLoS Biol. 2012;10(9):e1001388.

16. Duchen P, Živković D, Hutter S, Stephan W, Laurent S. Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. Genetics. 2013;193(1):291–301.

17. Reich DE, Goldstein DB. Genetic evidence for a Paleolithic human population expansion in Africa. Proceedings of the National Academy of Sciences. 1998;95(14):8119–8123.

18. Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, et al. Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences. 2006;103(45):16666–16671.

19. Consortium BH, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324(5926):528–532.

20. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends in ecology & evolution. 2014;29(1):51–63.

21. Smith BD. The emergence of agriculture. Scientific American Library New York; 1995.

22. Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez J, Buckler E, Doebley J. A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences. 2002;99(9):6080–6084.

23. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, et al. The effects of artificial selection on the maize genome. Science. 2005;308(5726):1310–1314.

24. Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS. Investigation of the bottleneck leading to the domestication of maize. Proceedings of the National Academy of Sciences. 1998;95(8):4441–4446.

25. Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS. Selection versus demography: a multilocus investigation of the domestication process in maize. Molecular Biology and Evolution. 2004;21(7):1214–1225.

26. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nature genetics. 2012;44(7):803–807.

27. Bukowski R, Guo X, Lu Y, Zou C, He B, Rong Z, et al. Construction of the third generation Zea mays haplotype map. bioRxiv. 2015;p. 026963.

28. Hufford MB, Xu X, Van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nature genetics. 2012;44(7):808–811.

29. Hearne S, Chen C, Buckler E, Mitchell S. Unimputed GbS derived SNPs for maize landrace accessions represented in the SeeD-maize GWAS panel; 2015. Accessed: 2015-02-16. `http://hdl.handle.net/11529/10034`.

30. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature genetics. 2014;.

31. Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, et al. The origin of the naked grains of maize. Nature. 2005;436(7051):714–719.

32. Wang H, Studer AJ, Zhao Q, Meeley R, Doebley JF. Evidence that the Origin of Naked Kernels During Maize Domestication Was Caused by a Single Amino Acid Substitution in tga1. Genetics. 2015;p. genetics–115.

33. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome research. 2014;24(6):885–895.

34. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceedings of the National Academy of Sciences. 2015;112(12):3823–3828.

35. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025.

36. Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS. Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.). Genetics. 2002;162(3):1401–1413.

37. Innan H, Kim Y. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(29):10667–10672.

38. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends in ecology & evolution. 2013;28(11):659–669.

39. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. PLoS genetics. 2015;11(2):e1005004.

40. Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R. Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. Proceedings of the National Academy of Sciences. 2009;106(13):5019–5024.

41. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. science. 2012;336(6082):740–743.

42. Program TM. Development, maintenance, and seed multiplication of open-pollinated maize varieties. 2nd ed. Mexico, D.F.: CIMMYT; 1999.

43. Baden WW, Beekman CS. Culture and agriculture: A comment on Sissel Schroeder, maize productivity in the Eastern Woodlands and Great Plains of North America. American Antiquity. 2001;p. 505–515.

44. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nature genetics. 2011;43(11):1160–1163.

45. Gallavotti A, Zhao Q, Kyozuka J, Meeley RB, Ritter MK, Doebley JF, et al. The role of barren stalk1 in the architecture of maize. Nature. 2004;432(7017):630–635.

46. Wills DM, Whipple CJ, Takuno S, Kursel LE, Shannon LM, Ross-Ibarra J, et al. From many, one: genetic control of prolificacy during maize domestication. PLoS Genet. 2013;9(6):e1003604.

47. Takuno S, Ralph P, Swarts K, Elshire RJ, Glaubitz JC, Buckler ES, et al. Independent Molecular Basis of Convergent Highland Adaptation in Maize. Genetics. 2015;Available from: http://www.genetics.org/content/early/2015/06/15/genetics.115.178327.abstr

48. van Heerwaarden J, Hufford MB, Ross-Ibarra J. Historical genomics of North American maize. Proceedings of the National Academy of Sciences. 2012;109(31):12420–12425.

49. Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, Buell CR, et al. A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. Genetics. 2014;196(3):829–840.

50. Clark RM, Tavaré S, Doebley J. Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. Molecular biology and evolution. 2005;22(11):2304–2312.

51. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nature Reviews Genetics. 2007;8(8):610–618.

52. Wallace J, Larsson S, Buckler E. Entering the second century of maize quantitative genetics. Heredity. 2014;112(1):30–38.

53. Weber AL, Briggs WH, Rucker J, Baltazar BM, de Jesus Sánchez-Gonzalez J, Feng P, et al. The genetic architecture of complex traits in teosinte (Zea mays ssp. parviglumis): new evidence from association mapping. Genetics. 2008;180(2):1221–1232.

54. Pyhäjärvi T, Hufford MB, Mezmouk S, Ross-Ibarra J. Complex patterns of local adaptation in teosinte. Genome biology and evolution. 2013;5(9):1594–1609.

55. Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS genetics. 2011;7(2):e1001302.

56. Williamson R, Josephs E, Platts A, Hazzouri K, Haudry A, Blanchette M, et al. Evidence for Widespread Positive and Negative Selection in Coding and Conserved Noncoding Regions of Capsella grandiflora. PLoS genetics. 2014;10(9):e1004622–e1004622.

57. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. science. 2011;331(6019):920–924.

58. Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology. 2010;20(4):R208–R215.

59. Ross-Ibarra J, Tenaillon M, Gaut BS. Historical divergence and gene flow in the genus Zea. Genetics. 2009;181(4):1399–1413.

60. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009;26(9):2097–2108.

61. Castelletti S, Tuberosa R, Pindo M, Salvi S. A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1. G3: Genes— Genomes— Genetics. 2014;4(5):805–812.

62. Mao H, Wang H, Liu S, Li Z, Yang X, Yan J, et al. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nature Communications. 2015;6.

63. Yang Q, Li Z, Li W, Ku L, Wang C, Ye J, et al. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proceedings of the National Academy of Sciences. 2013;110(42):16969–16974.

64. Fraser HB. Gene expression drives local adaptation in humans. Genome research. 2013;23(7):1089–1096.

65. Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, et al. Contributions of Protein-Coding and Regulatory Change to Adaptive Molecular Evolution in Murid Rodents. PLoS Genetics. 2013 Dec;9(12):e1003995–14.

66. Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1984.

67. Günther T, Schmid KJ. Deleterious amino acid polymorphisms in Arabidopsis thaliana and rice. Theoretical and Applied Genetics. 2010 Mar;121(1):157–168.

68. Renaut S, Rieseberg LH. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other Compositae crops. Molecular biology and evolution. 2015;p. msv106.

69. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature communications. 2010;1:131.

70. Mezmouk S, Ross-Ibarra J. The pattern and distribution of deleterious mutations in maize. G3 (Bethesda, Md). 2014 Jan;4(1):163–171.

71. Eyre-Walker A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proceedings of the National Academy of Sciences. 2010;107(suppl 1):1752–1756.

72. Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nature genetics. 2015;47(2):126–131.

73. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nature genetics. 2014;46(3):220–224.

74. Gazave E, Chang D, Clark AG, Keinan A. Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect. Genetics. 2013;195(3):969–978.

75. Lohmueller KE. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. PLoS Genetics. 2014;10(5).

76. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome research. 2010;20(3):393–402.

77. Zeng K, Charlesworth B. The effects of demography and linkage on the estimation of selection and mutation parameters. Genetics. 2010;186(4):1411–1424.

78. Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. Molecular biology and evolution. 2012;p. mss219.

79. Elyashiv E, Bullaughey K, Sattath S, Rinott Y, Przeworski M, Sella G. Shifts in the intensity of purifying selection: An analysis of genome-wide polymorphism data from two closely related yeast species. Genome Research. 2010 Nov;20(11):1558–1573.

80. Lemmon ZH, Bukowski R, Sun Q, Doebley JF. The Role of ¡italic¿cis¡/italic¿ Regulatory Evolution in Maize Domestication. PLoS Genet. 2014 11;10(11):e1004745. Available from: `http://dx.doi.org/10.1371%2Fjournal.pgen.1004745`.

81. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070–3071.

82. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science. 2009;326(5956):1112–1115.

83. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–2993.

84. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One. 2014;9(2):E90346.

85. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols. 2009;4(8):1184–1191.

86. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–3440.

87. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available from: `http://www.R-project.org/`.

88. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. BMC bioinformatics. 2014;15(1):356.

89. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133(3):693–709.

90. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069–2070.