nature plants

ARTICLES

PUBLISHED: 13 JUNE 2016 | ARTICLE NUMBER: 16084 | DOI: 10.1038/NPLANTS.2016.84

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] and Jeffrey Ross-Ibarra[1,5]*

Genetic diversity is shaped by the interaction of drift and selection, but the details of this interaction are not well understood. The impact of genetic drift in a population is largely determined by its demographic history, typically summarized by its long-term effective population size (N_e). Rapidly changing population demographics complicate this relationship, however. 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 produce the first whole-genome estimate of the demography of maize domestication, showing that maize was reduced to approximately 5% the population size of teosinte before it experienced rapid expansion post-domestication to population sizes much larger than its ancestor. Evaluation of patterns of nucleotide diversity in and near genes shows little evidence of selection on beneficial amino acid substitutions, and that the domestication bottleneck led to a decline in the efficiency of purifying selection in maize. Young alleles, however, show evidence of much stronger purifying selection in maize, reflecting the much larger effective size of present day populations. Our results demonstrate that recent demographic change—a hall-mark of many species including both humans and crops—can have immediate and wide-ranging impacts on diversity that conflict with expectations based on long-term N_e alone.

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 favouring particular alleles at the expense of others as a result of their effects on fitness. Researchers often study drift by excluding potentially selected sites[2,3], or selection by focusing on site-specific patterns under the assumption that genome-wide diversity reflects primarily the action of drift[4].

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[5,6]. Linked selection, refers to the effects of selection at one site on diversity at linked sites[6]. 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[5], or background selection, where diversity is reduced at loci linked to a site undergoing selection against deleterious alleles[7]. Recent work in *Drosophila*, for example, has shown that virtually the entire genome is impacted by the combined effects of these processes[8–10].

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[6,11–13]. The effective size of a population is not static, and nearly all species, including flies[14], humans[15], domesticates[16] and non-model species[17] have experienced recent or ancient changes in N_e. Although much is known about how the long-term average N_e affects linked selection[11], 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[18], and involved a population bottleneck followed by recent expansion[19–21]. 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, although patterns of overall nucleotide diversity reflect long-term differences in N_e, 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 analysed data from 23 maize and 13 teosinte genomes from the maize HapMap 2 and HapMap 3 projects[22,23]. As a preliminary step, we evaluated levels of diversity inside and outside genes across the genome. We found broad differences in genic and intergenic diversity consistent with earlier results[24] (Fig. 1). In maize, mean pairwise diversity (π) within genes was significantly lower than at sites at least 5 kb away from genes (0.00668 versus 0.00691,

[1]Department of Plant Sciences, University of California, Davis, California 95616, USA. [2]US Department of Agriculture, Agricultural Research Service, Columbia, Missouri 65211, USA. [3]Division of Plant Sciences, University of Missouri, Columbia, Missouri 65211, USA. [4]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011, USA. [5]Genome Center and Center for Population Biology, University of California, Davis, California 95616, USA. *e-mail: rossibarra@ucdavis.edu; beissingert@missouri.edu

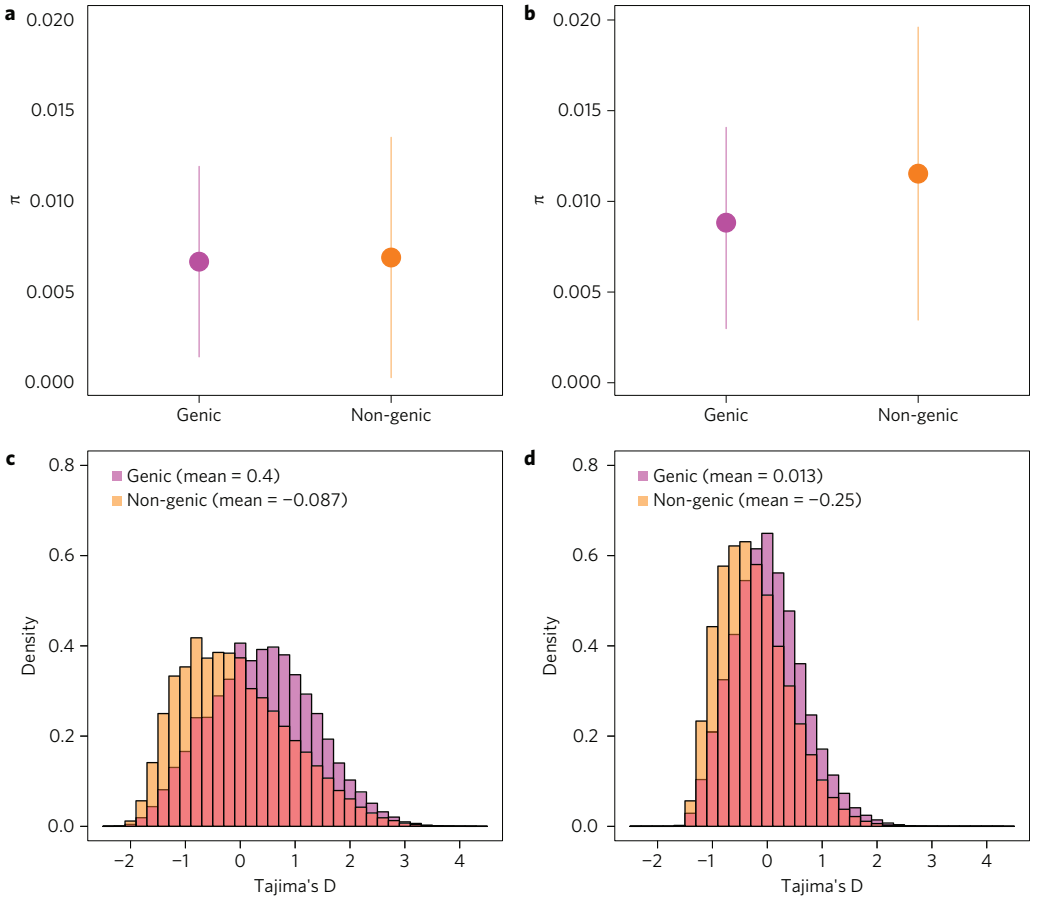

Figure 1 | Genetic diversity in maize and teosinte. a,**b**, Mean pairwise diversity $\pi \pm 1$ s.d. in maize (**a**) and teosinte (**b**). **c**,**d**, Tajima's D in 1 kb windows from genic and non-genic regions of maize (**c**) and teosinte (**d**).

$P < 2 \times 10^{-44}$). Diversity differences in teosinte are even more pronounced (0.0088 versus 0.0115, $P \approx 0$). Differences were also apparent in the site frequency spectrum, with the mean Tajima's D positive in genic regions in both maize (0.4) and teosinte (0.013) but negative outside 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.

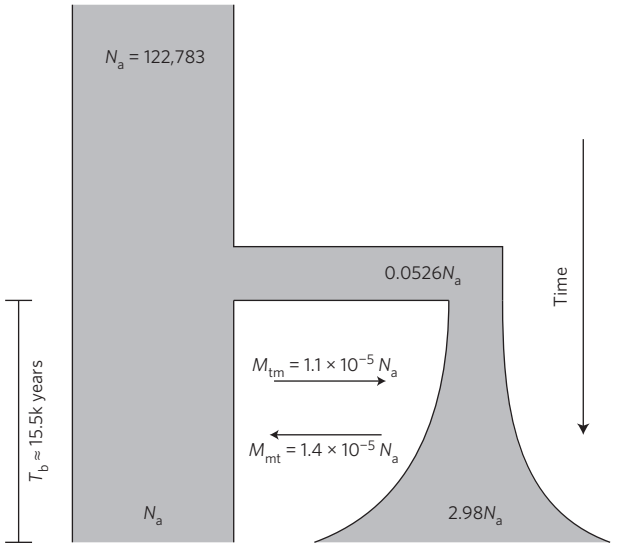

Figure 2 | Estimated demographic history of maize and teosinte. Parameter estimates for a basic bottleneck model of maize domestication. See Methods for details.

Demography of maize domestication. We next estimated a demographic model of maize domestication (Fig. 2). To minimize the impact of selection on our estimates[25], we only included sites >5 kb from genes. The most likely model estimates an ancestral population mutation rate of $\theta = 0.0147$ per base pair (bp), which translates to an ancestral effective population size of $N_a \approx 123{,}000$ teosinte individuals. We estimate that maize split from teosinte ≈15,000 generations in the past, with an initial size of only ≈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 $N_m \approx 370{,}000$. Although our model provides only a rough approximation of migration rates, we included migration parameters during demographic inference because omitting these could bias our population size estimates. We observe that maize and teosinte have continued to exchange migrants after the population split, with gene flow from teosinte to maize estimated to be $M_{tm} = 1.1 \times 10^{-5} \times N_a$ migrants per generation, and from maize to teosinte $M_{mt} = 1.4 \times 10^{-5} \times N_a$ migrants per generation.

Because our modest sample size of fully sequenced individuals has limited power to infer recent population expansion, we investigated two alternative approaches for demographic inference. First, we utilized genotyping data from more than 4,000 maize landraces[26] to estimate the modern maize effective population size. Because rare variants provide the best information about recent effective population sizes[27], we estimate N_e using a

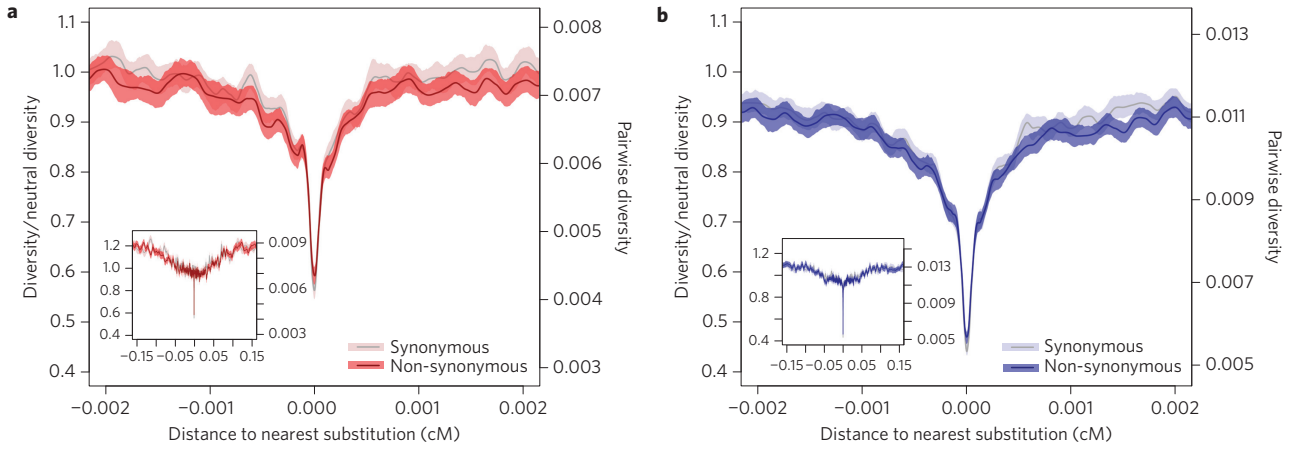

Figure 3 | Relative diversity versus distance to nearest substitution in maize and teosinte. a,**b**, Pairwise diversity surrounding synonymous and missense substitutions in maize (**a**) and teosinte (**b**). Axes show absolute diversity values (right) and values relative to mean nucleotide diversity in windows ≥0.01 cM 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.

singleton-based estimator[28] of the population mutation rate $\theta = 4N_e\mu$ and published values of the mutation rate[29] (see Methods for details). This yields a much higher estimate of the modern maize effective population size at $N_m \approx 993{,}000$. Finally, we employed a model-free coalescent approach[30] to estimate population size change using a subset of six genomes each of maize and teosinte. Though this analysis suggests non-equilibrium dynamics for teosinte not included in our initial model, it is nonetheless broadly consistent with the other approaches, identifying population isolation beginning between 10,000 and 15,000 generations ago, a clear domestication bottleneck, and ultimately rapid population expansion in maize to an extremely large extant size of $\approx 10^9$ (Supplementary Fig. 2). 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[5]. 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 either maize or teosinte and the sister genus *Tripsacum*. If a substantial proportion of missense mutations have been fixed because of 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 maize domestication locus *tga1* (Supplementary Fig. 1), likely to be the result of a hard sweep during domestication[31]. Genome-wide, however, we observe no differences in diversity at sites near synonymous versus missense substitutions in either maize or teosinte (Fig. 3).

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, because of weak purifying selection, have higher genetic diversity[32]. To address this concern, we took advantage of genome-wide estimates of evolutionary constraint[33] calculated using genomic evolutionary rate profile (GERP) scores[34]. 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 found no difference in diversity near synonymous and missense substitutions in either subset of the data (Supplementary Fig. 3). 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[7]. 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 (Fig. 4a) but recovers quickly at sites outside genes, consistent with the low levels of linkage disequilibrium generally observed in these subspecies[22]. The reduction in relative diversity is more pronounced in teosinte, reaching lower levels in genes and occurring a across wider region.

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[35]. 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[36]. Qualitative differences between maize and teosinte in patterns of diversity within and outside genes remained unchanged even after removing genes in the top 20% quantile of H12 (Supplementary Fig. 7A). 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 because of 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, which is determined primarily by intermediate frequency—therefore comparably older—alleles. Singleton diversity around missense and synonymous substitutions (Supplementary Fig. 4) appears

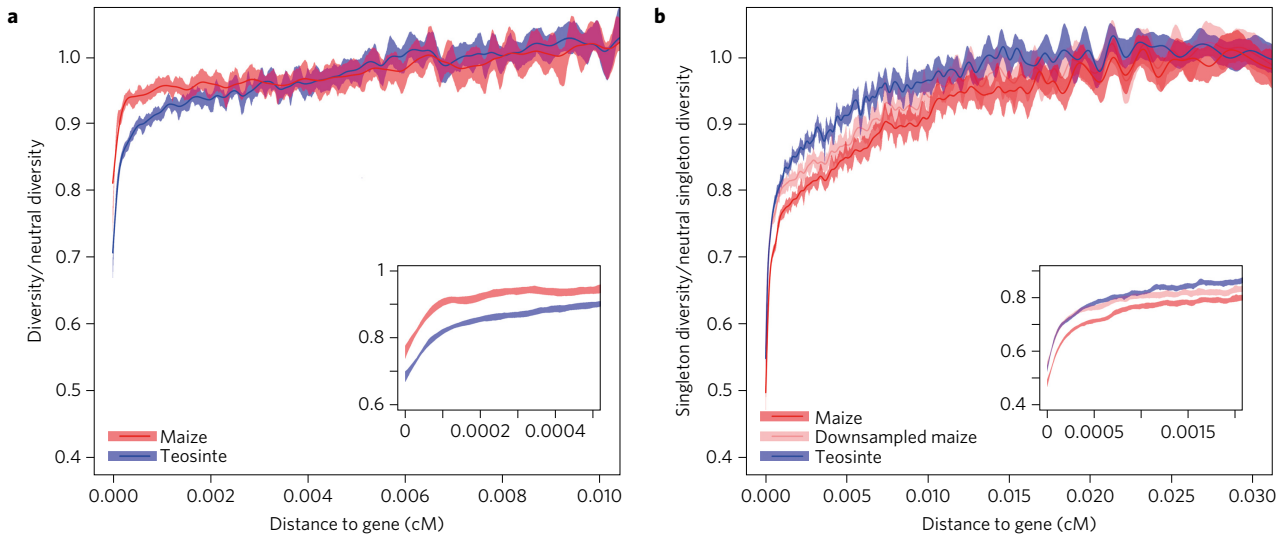

Figure 4 | Relative diversity versus distance to nearest gene in maize and teosinte. a, Pairwise nucleotide diversity. **b**, Singleton diversity. Relative diversity is calculated compared to the mean diversity in windows ≥0.01 cM or ≥0.02 cM from the nearest gene for pairwise diversity and singletons, respectively. Lines depict cubic smoothing splines with smoothing parameters chosen via generalized cross-validation and shading depicts bootstrap-based 95% confidence intervals. Inset plots depict a smaller range on the *x*-axis.

nearly identical to results from pairwise diversity (Fig. 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 (Fig. 4b). This result is the opposite of the pattern observed for π, 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 (Supplementary Fig. 7). Although 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 (Supplementary Fig. 5) and mirror results from simulations of purifying selection (Supplementary Fig. 6).

Discussion

Demography of domestication. Although a number of authors have investigated the demography of maize domestication[19–21], 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 genes demonstrates that sites far from genes can be reasonably used for demographic inference. Furthermore, by utilizing the full joint site frequency spectrum (SFS), we are able to estimate population growth, gene flow and the strength of the domestication bottleneck without making assumptions about its duration. This model paves the way for future work on the demography of domestication, evaluating for example the significance of differences in gene flow estimated here or removing assumptions about demographic history in teosinte.

One surprising result from our model is the estimated divergence time of maize and teosinte approximately 15,000 generations before present. While this appears to conflict with archaeological estimates[37], we emphasize that this estimate reflects the fact that the genetic split between populations is likely to precede anatomical changes that can be identified in the archaeological record. We also note that our result may be inflated owing to population structure, as our geographically diverse sample of teosinte may include populations diverged from those that gave rise to maize.

The estimated bottleneck of ≈5% of the ancestral teosinte population seems low given that maize landraces exhibit ≈80% of the diversity of teosinte[24], but our model suggests that the effects of the bottleneck on diversity are likely to be ameliorated by both gene flow and rapid population growth (Fig. 2). Although we estimate that the modern effective size of maize is larger than teosinte, the small size of our sample reduces our power to identify the low frequency alleles most sensitive to rapid population growth[27], and our model is unable to incorporate growth faster than exponential. Both alternative approaches we employ estimate a much larger modern effective size of maize in the range of $\approx 10^6$–10^9, an order of magnitude or more than the current size of teosinte. Census data suggest these estimates are plausible: there are 47.9 million hectares of open-pollinated maize in production[38], likely to be planted at a density of ≈25,000 individuals per hectare[39]. Assuming the effective size is only ≈0.4% of the census size (i.e. 1 ear for every 1,000 male plants), this still implies a modern effective population size of more than four billion. Although these genetic and census estimates are likely to be inaccurate, all of the evidence points to the fact that the modern effective size of maize is extremely large.

Hard sweeps do not shape genome-wide diversity in maize. Our findings demonstrate that classic hard selective sweeps have not contributed substantially to genome-wide patterns of diversity in maize, a result we show is robust to concerns about power due to the effects of purifying selection[32]. Although our approach ignores the potential for hard sweeps in non-coding regions of the genome, a growing body of evidence argues against hard sweeps as the prevalent mode of selection shaping maize variability. Among well-characterized domestication loci, only the *tga1* gene shows evidence of a hard sweep on a missense mutation[31], whereas published data for several loci are consistent with soft sweeps from standing variation[40] or multiple mutations[41]. Moreover, genome-wide studies of domestication[24], local adaptation[42] and modern breeding[43,44] all support the importance of standing variation as

primary sources of adaptive variation. Soft sweeps are expected to be common when $2N_e\mu_b \geq 1$, where μ_b is the mutation rate of beneficial alleles with selection coefficient s_b[35]. Assuming a mutation rate of 3×10^{-8}[29] and that on the order of ≈1–5% of mutations are beneficial[45], this implies that soft sweeps should be common in both maize and teosinte for mutational targets ≫10 kb—a plausible size for quantitative traits or for regulatory evolution targeting genes with, for example, large up- or downstream control regions[40]. Indeed, many adaptive traits in both maize[46] and teosinte[47] are highly quantitative, and adaptation in both maize[24] and teosinte[48] has involved selection on regulatory variation.

The absence of evidence for a genome-wide impact of hard sweeps in coding regions differs markedly from observations in *Drosophila*[49] and *Capsella*[50], but is consistent with data from humans[51]. Comparisons of the estimated percentages of non-synonymous substitutions fixed by natural selection[8,50,52,53] give similar results. Although differences in long-term N_e are likely to explain some of the observed variation across species, we see little change in the importance of hard sweeps in genes in singleton diversity in modern maize (Supplementary Fig. 4), perhaps suggesting other factors may contribute to these differences as well. One possibility, for example, is that, if mutational target size scales with genome size, the larger genomes of human and maize may offer more opportunities for non-coding loci to contribute to adaptation, with hard sweeps on non-synonymous 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[40,54,55] and studies of local adaptation that show enrichment for single nucleotide polymorphism (SNPs) in regulatory regions in teosinte[48] and humans[56] but for non-synonymous variants in the smaller *Arabidopsis* genome[57]. Our results, for example, are not dissimilar to findings in the similarly sized mouse genome, where no differences are seen in diversity around non-synonymous and synonymous substitutions in spite of a large N_e and as many as 80% of adaptive substitutions occurring outside genes[58]. Future comparative analyses using a common statistical framework[12] and considering additional ecological and life history factors (see ref. 13) 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 π around genes in teosinte than maize (Fig. 4a), but the opposite trend when we evaluate diversity using singleton polymorphisms (Fig. 4b). The efficiency of purifying selection is proportional to effective population size[59], and these results are thus consistent with our demographic analyses which show a domestication bottleneck and smaller long-term N_e in maize[19–21,52] followed by recent rapid expansion and a much larger modern N_e. Simple foward-in-time population genetic simulations qualitatively confirm these results, and further suggest that the observed patterns are likely cause by sites under relatively weak purifying selection (Supplementary Fig. 6).

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[60,61] and human populations out of Africa[27,62], and such variants may play an important role in phenotypic variation[62–64]. Nonetheless, demographic history may have little impact on the overall genetic load of populations[65,66], 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[66,67].

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. Although a number of authors have tested for selection using methods that explicitly incorporate or are robust to demographic change[53,68] and others have compared estimates of the efficiency of adaptive and purifying selection across species[69] or populations[70], previous analyses of the impact of linked selection on genome-wide diversity have relied on single estimates of the effective population size[12,13]. 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 to be informative about current processes affecting allele frequency trajectories.

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[22,23,71]. From these data, we removed four teosinte individuals that were not ssp. *parviglumis* or appeared as outliers in an initial principal component analysis conducted with the package adegenet[72] (Supplementary Fig. 8), leaving 13 teosinte and 23 maize that were used for all subsequent analyses (Supplementary Table 1). We also utilized a single individual of (*Tripsacum dactyloides*) as an outgroup. All BAM files are available from CyVerse at /iplant/home/shared/commons_repo/curated/Beissinger_MaizeTeo_2016 (http://dx.doi.org/10.7946/P2QP4N).

Physical and genetic maps. Sequences were mapped to the maize B73 version 3 reference genome[73] (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/) as described previously[23]. All analyses made use of uniquely mapping reads with mapping quality score ≥30 and bases with base quality score ≥20; quality scores around indels were adjusted following ref. 74. We converted physical coordinates to genetic coordinates using linear interpolation of the previously published 1 cM resolution NAM genetic map[75].

Estimating the site frequency spectrum. We estimated both the genome-wide SFS as well as a separate SFS for genic (within annotated transcript) and intergenic (≥5 kb from a transcript) regions. We used the biomaRt package[76,77] of R (ref. 78) to parse annotations from genebuild version 5b of AGPv3. We estimated single population and joint SFS with the software ANGSD[79], including all positions with at least one aligned read in ≥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 $\delta a \delta i^3$.

Demographic inference. We used the software $\delta a \delta i^3$ to estimate parameters of a domestication bottleneck from the joint maize-teosinte SFS, using only sites >5 kb from a gene to ameliorate the effects of linked selection. To minimize the number of parameters estimated, we employed a simple demographic model which posits a teosinte population of constant effective size N_a. At time T_b generations in the past, this population gave rise to a maize population of size N_b that grew exponentially to size N_m in the present (Fig. 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 $\mu = 3 \times 10^{-8}$ (ref. 29). 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 Supplementary Table 2. 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[26] to estimate the modern maize N_e from singleton counts. We filtered these data to include only SNPs with data in ≥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 $\mu = 3 \times 10^{-8}$ (ref. 29) and the relation $4N_e\mu = S/L$ (ref. 28), 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 (ref. 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 (ref. 79) using a SNP P value of 1×10^{-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 parameter $20 \times 2 + 20 \times 4 + 10 \times 2$ (Supplementary Fig. 2A) for population size inference. To estimate the rate of cross-coalescence we used four maize and four teosinte haplotypes with pattern parameter $20 \times 1 + 20 \times 2$ (Supplementary Fig. 2B).

Diversity. We made use of the software ANGSD (ref. 79) 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 downsampled 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 ≤20% missing data. Substitutions were classified as synonymous or missense using the ensembl variant effects predictor[80]. For each window with ≥100 bp 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 centre of the nearest gene transcript.

Selection scan. We scanned the genome to identify sites that have experienced recent positive selection using the H12 statistic[36] in sliding windows of 200 SNPs with a step of 25 SNPs.

Simulations. We used the program *bneck_selection_ind* included in version 0.4.4 of the forward-in-time population genetic simulation library *fwdpp*[81] (https://github.com/molpopgen/fwdpp[]thornton2014genetics). All simulations used a population mutation rate of $\theta = 20$, a population recombination rate of $\rho = 20$ and simulated 150,000 burn-in generations at an ancestral population size of $N_1 = 15{,}000$ to establish equilibrium, after which the population instantly changed to size N_2 and then grew exponentially for 1,000 generations to size N_3. To simulate a constant size population emulating teosinte, we set $N_2 = N_3 = 15{,}000$. For maize we simulated a bottleneck similar to that estimated in Fig. 2 by setting $N_2 = 750$, followed by exponential growth to a large modern population size of $N_3 = 150{,}000$. For each taxon, we performed 1,000 simulations for each of five values of the strength of purifying selection: $s = \{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. All mutations were assumed to be codominant. To mimic nonsynonymous changes at a coding locus, we assumed that three out of four mutations were selected. We calculated summary statistics across all sites using version 0.3.4 of *msstats* (https://github.com/molpopgen/msstats/releases).

Received 5 January 2016; accepted 12 May 2016; published 13 June 2016

References

1. Dobzhansky, T. & Pavlovsky, O. An experimental study of interaction between genetic drift and natural selection. *Evolution* **11,** 311–319 (1957).
2. Voight, B. F. *et al.* Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. USA* **102,** 18508–18513 (2005).
3. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5,** e1000695 (2009).
4. Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* **19,** 711–722 (2009).
5. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23,** 23–35 (1974).
6. Slotte, T. The impact of linked selection on plant genomic variation. *Brief. Funct. Genomics* **13,** 268–275 (2014).
7. Charlesworth, B., Morgan, M. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134,** 1289–1303 (1993).
8. Sella, G., Petrov, D. A., Przeworski, M. & Andolfatto, P. Pervasive natural selection in the drosophila genome? *PLoS Genet.* **5,** e1000495 (2009).
9. Elyashiv, E. *et al.* A genomic map of the effects of linked selection in drosophila. Preprint at arXiv:1408.5461 (2014).
10. Andolfatto, P. Adaptive evolution of non-coding DNA in drosophila. *Nature* **437,** 1149–1152 (2005).
11. Cutter, A. D. & Payseur, B. A. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Rev. Genet.* **14,** 262–274 (2013).
12. Corbett-Detig, R. B., Hartl, D. L. & Sackton, T. B. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.* **13,** e1002112 (2015).
13. Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species. *PLoS Biol.* **10,** e1001388 (2012).
14. 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* **193,** 291–301 (2013).
15. Reich, D. E. & Goldstein, D. B. Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl Acad. Sci. USA* **95,** 8119–8123 (1998).
16. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nature Rev. Genet.* **14,** 840–852 (2013).
17. Ellegren, H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* **29,** 51–63 (2014).
18. Matsuoka, Y. *et al.* A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl Acad. Sci. USA* **99,** 6080–6084 (2002).
19. Wright, S. I. *et al.* The effects of artificial selection on the maize genome. *Science* **308,** 1310–1314 (2005).
20. Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L. & Gaut, B. S. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl Acad. Sci. USA* **95,** 4441–4446 (1998).
21. Tenaillon, M. I., U'Ren, J., Tenaillon, O. & Gaut, B. S. Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21,** 1214–1225 (2004).
22. Chia, J.-M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genet.* **44,** 803–807 (2012).
23. Bukowski, R. *et al.* Construction of the third generation *Zea mays* haplotype map. Preprint at http://biorxiv.org/content/early/2015/09/16/026963 (2015).
24. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nature Genet.* **44,** 808–811 (2012).
25. Ewing, G. B. & Jensen, J. D. The consequences of not accounting for background selection in demographic inference. *Mol. Ecol.* **25,** 135–141 (2016).
26. Hearne, S., Chen, C., Buckler, E. & Mitchell, S. *Unimputed GBS Derived SNPs for Maize Landrace Accessions Represented in the Seed-Maize GWAS Panel* (CIMMYT Dataverse Network, 2015); http://hdl.handle.net/11529/10034
27. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336,** 740–743 (2012).
28. Fu, Y.-X. & Li, W.-H. Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709 (1993).
29. Clark, R. M., Tavaré, S. & Doebley, J. Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Mol. Biol. Evol.* **22,** 2304–2312 (2005).
30. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature Genet.* **46,** 919–925 (2014).
31. Wang, H., Studer, A. J., Zhao, Q., Meeley, R. & Doebley, J. F. Evidence that the origin of naked kernels during maize domestication was caused by a single amino acid substitution in tga1. *Genetics* **200,** 965–974 (2015).
32. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Res.* **24,** 885–895 (2014).
33. Rodgers-Melnick, E. *et al.* Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl Acad. Sci. USA* **112,** 3823–3828 (2015).
34. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).
35. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28,** 659–669 (2013).
36. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* **11,** e1005004 (2015).
37. Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J. & Dickau, R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proc. Natl Acad. Sci. USA* **106,** 5019–5024 (2009).
38. Program, T. M. *Development, Maintenance, and Seed Multiplication of Open-Pollinated Maize Varieties* 2nd edn (CIMMYT, 1999).
39. Baden, W. W. & Beekman, C. S. Culture and agriculture: a comment on Sissel Schroeder, maize productivity in the eastern woodlands and great plains of North America. *Am. Antiq.* **66,** 505–515 (2001).
40. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genet.* **43,** 1160–1163 (2011).
41. Wills, D. M. *et al.* From many, one: genetic control of prolificacy during maize domestication. *PLoS Genet.* **9,** e1003604 (2013).
42. Takuno, S. *et al.* Independent molecular basis of convergent highland adaptation in maize. *Genetics* **200,** 1297–1312 (2015).
43. van Heerwaarden, J., Hufford, M. B. & Ross-Ibarra, J. Historical genomics of North American maize. *Proc. Natl Acad. Sci. USA* **109,** 12420–12425 (2012).
44. Beissinger, T. M. *et al.* A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number. *Genetics* **196,** 829–840 (2014).
45. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* **8,** 610–618 (2007).
46. Wallace, J., Larsson, S. & Buckler, E. Entering the second century of maize quantitative genetics. *Heredity* **112,** 30–38 (2014).

47. Weber, A. L. *et al.* The genetic architecture of complex traits in teosinte (*Zea mays* ssp. *parviglumis*): new evidence from association mapping. *Genetics* **180,** 1221–1232 (2008).
48. Pyhäjärvi, T., Hufford, M. B., Mezmouk, S. & Ross-Ibarra, J. Complex patterns of local adaptation in teosinte. *Genome Biol. Evol.* **5,** 1594–1609 (2013).
49. 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 Genet.* **7,** e1001302 (2011).
50. Williamson, R. *et al.* Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. *PLoS Genet.* **10,** e1004622 (2014).
51. Hernandez, R. D. *et al.* Classic selective sweeps were rare in recent human evolution. *Science* **331,** 920–924 (2011).
52. Ross-Ibarra, J., Tenaillon, M. & Gaut, B. S. Historical divergence and gene flow in the genus *Zea*. *Genetics* **181,** 1399–1413 (2009).
53. Eyre-Walker, A. & Keightley, P. D. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26,** 2097–2108 (2009).
54. Mao, H. *et al.* A transposable element in a *nac* gene is associated with drought tolerance in maize seedlings. *Nature Commun.* **6,** 8326 (2015).
55. Yang, Q. *et al.* CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proc. Natl Acad. Sci. USA* **110,** 16969–16974 (2013).
56. Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Res.* **23,** 1089–1096 (2013).
57. Hancock, A. M. *et al.* Adaptation to climate across the *Arabidopsis thaliana* genome. *Science* **334,** 83–86 (2011).
58. Halligan, D. L. *et al.* Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet.* **9,** e1003995 (2013).
59. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1984).
60. Günther, T. & Schmid, K. J. Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. *Theor. Appl. Genet.* **121,** 157–168 (2010).
61. Renaut, S. & Rieseberg, L. H. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other Compositae crops. *Mol. Biol. Evol.* **32,** 2273–2283 (2015).
62. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Commun.* **1,** 131 (2010).
63. Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3 (Bethesda)* **4,** 163–171 (2014).
64. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl Acad. Sci. USA* **107,** 1752–1756 (2010).
65. Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genet.* **47,** 126–131 (2015).
66. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genet.* **46,** 220–224 (2014).
67. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10,** e1004379 (2014).
68. Zeng, K. & Charlesworth, B. The effects of demography and linkage on the estimation of selection and mutation parameters. *Genetics* **186,** 1411–1424 (2010).
69. Popadin, K. Y., Nikolaev, S. I., Junier, T., Baranova, M. & Antonarakis, S. E. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. *Mol. Biol. Evol.* **2,** 347–355 (2013).
70. Elyashiv, E. *et al.* Shifts in the intensity of purifying selection: an analysis of genome-wide polymorphism data from two closely related yeast species. *Genome Res.* **20,** 1558–1573 (2010).
71. Lemmon, Z. H., Bukowski, R., Sun, Q. & Doebley, J. F. The role of *cis* regulatory evolution in maize domestication. *PLoS Genet.* **10,** e1004745 (2014).
72. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27,** 3070–3071 (2011).
73. Schnable, P. S. *et al.* The b73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115 (2009).
74. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
75. Glaubitz, J. C. *et al.* Tassel-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **9,** e90346 (2014).
76. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protoc.* **4,** 1184–1191 (2009).
77. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).
78. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014); http://www.R-project.org/
79. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15,** 356 (2014).
80. McLaren, W. *et al.* Deriving the consequences of genomic variants with the nsemble API and SNP effect predictor. *Bioinformatics* **26,** 2069–2070 (2010).
81. Thornton, K. R. A c++ template library for efficient forward-time population genetic simulation of large populations. *Genetics* **198,** 157–166 (2014).

Acknowledgements

We are indebted to G. Coop and S. Aeschbacher for their constructive input during this study. We thank R. Bukowski and Q. Sun for providing early-access data from maize HapMap3. Funding was provided by National Science Foundation Plant Genome Research Project 1238014, the US Department of Agriculture (USDA) Agricultural Research Service, and USDA Hatch project CA-D-PLS-2066-H.

Author contributions

T.M.B. and J.R.I. devised this study. T.M.B., L.W., J.R.-I. and K.C. analysed the data. A.D. performed early-stage simulations. T.M.B., J.R.-I. and M.B.H. wrote the manuscript.

Additional information

Supplementary information is available online. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to T.M.B. and J.R.I.

Competing interests

The authors declare no competing financial interests.