# Demography and selection since maize domestication

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

[*]Dept. of Plant Sciences, University of California, Davis, CA, USA,[†]Iowa State University, Ames, IA, USA, and [‡]Genome Center and Center for population biology, University of California, Davis, CA, USA

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

**This is the abstract. It should probably be somewhere around 200 words.**

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

For instance, supposed neutrally-evolving DNA is often used to estimate historical demographic parameters of a population such as effective population size, structure, and expansion history [5, 6]. However, researchers have called into question whether or not there are reasonable approaches for identifying neutral regions of the genome, since the effects of selection can be wide-ranging [7, 8]. For example, in *Drosophila* it appears that the majority of the genome is impacted by the effects of selection through linkage [9]. This observation begs the question, how can demographic parameters be estimated independently of selective parameters? Human researchers have attempted to address this problem by limiting analyses to only sites far from genes [10], but as *Drosophila* demonstrates, it is difficult to be certain that sites even in gene-poor regions of the genome are not influenced by linked selection. Ultimately, an understanding of which sites have the potential to be adaptive, and how these affect genome-wide patterns of variability through linkage, is required for reasonable demographic inference.

Maize represents an excellent organism to study these phenomena. Maize is a species of tremendous importance worldwide as both a staple crop [?] and as a model for understanding crop evolution [?]. Broadly speaking, archaeological and genetic studies have established that maize domestication is likely to have taken place in Central Mexico approximately 9,000 years bp [?, ?]. Teosinte, the most recent wild ancestor to maize, remains extant throughout much of the Americas [?]. Additionally, several large-effect domestication loci [?] and putative domestication regions [3] have been identified. But despite all that is known about maize domestication, the parameters of the domestication process remain uncertain. Specifically, the size of the maize domestication bottleneck has not been estimated independently of the bottleneck’s duration, nor are there sequence-based estimates of the effective population size of modern maize. Sequence information from maize and teosinte plants may therefore be utilized to address these questions.

To that end, the objectives of our study were to 1) investigate the relative importance of different forms of selection on whole-genome variability in both maize and teosinte, as different forms of selection affect DNA variability remarkably differently; 2) research the impact that the domestication process has had on genetic variability in maize, and how this compares to the impact of a different demographic history in teosinte; and 3) precisely estimate the parameters of the maize domestication bottleneck. We show that our third objective, estimating the parameters of domestication, is not possible without first completing objectives one and two, which demonstrate that as in humans [10], the majority of maize non-genic DNA may reasonably be treated as neutral. We achieve these objectives by utilizing whole-genome-sequence information from 23 maize and 13 teosinte lines sequenced as part of the Maize HapMap 2 panel [11].

## Results

**Patterns of variability differ between genic and nongenic regions of the genome.** Previous research of the maize domestication process has relied upon observations drawn from genic DNA (several references here). Our data were generated through whole genome sequencing, which eliminated this constraint. Importantly, we observed substantial differences in patterns of diversity between genic and non-genic regions of the genome for both maize and teosinte. For maize, mean pairwise diversity ($\pi$) within genes was significantly lower than at positions at least 5kb away from genes (0.00668 within, 0.00691 away, p<2e-44). The same pattern of significantly greater diversity outside of genes was observed in teosinte, but to a larger extent. For teosinte we observed $\pi$ within genes to be 0.0088 and at least 5kb from genes it was 0.1153 (p≈0). These obserrvations suggest that genes are not evolving neutrally in maize or teosinte. Instead, some form of selection is likely reducing diversity within genes. Additionally, these observations suggest that demographic inference, which relies on the assumption of neutrally evolving DNA, will be more

**Significance**

**This work is insignificant ;-)**

**Reserved for Publication Footnotes**

accurate if it is based on observations from non-genic genetic material.

**Hard sweeps do not shape maize (or teosinte) diversity.** A mutation that is immediately beneficial and positively selected leaves a classical hard sweep signature in the genome, whereby genetic diversity surrounding that mutation is reduced as the haplotype where the mutation first arose increases in frequency to fixation. The prevalence of such hard sweeps was evaluated by comparing diversity surrounding non-synonymous and synonymous substitutions in maize and teosinte since divergence from *Tripsacum*, roughly ZZZ years before present (ref here). For both taxa, no difference in diversity surrounding these classes of substitutions was identified (Figure 2). Additionally, diversity around maize substitutions not seen in teosinte was investigated. These sites have the potential to correspond to recent maize sweeps, occuring after the split from teosinte. Again, no difference was observed between diversity around synonymous and nonsynonymous substitutions (Supplemental figure here). Together, these observations suggest that hard sweeps are not a primary form of selection for either maize or teosinte.

**Patterns of purifying and background selection are influenced by demographic history.** Purifying selection refers to the situation where deleterious mutations arising in a population are continuously selected against. When this form of selection is operating, it can serve to reduce genetic variability at linked neutral sites, a phenomenon called background selection [12]. Purifying and background selection lead to lower diversity within genes and other functional sites relative to neutral regions (ref here). We investigated purifying selection in maize and teosinte by evaluating the average magnitude of reduced diversity within genes and recovery away from genes in both taxa. When standardized by neutral levels, which were defined as mean diversity at postions at least 0.01 cM distal from genes, a stronger reduction of diversity and slower recovery was observed for teosinte than for maize, implying that purifying selection has left a more pronounced signature in the teosinte genome (Figure 3). This conflicted with our *a priori* hypothesis; we expected that strong artificial selection since domestication may have elevated the intensity of purifying selection for maize. We therefore conducted a parallel analysis based on singleton diversity. As a class, singleton alleles depict the most recent patterns of evolution, but also have the lowest effect on pairwise diversity. Therefore, unlike pairwise diversity, patterns of singleton diversity reflect recent patterns of evolution. Since our sample size for maize (n=23) was larger than teosite (n=13), maize singletons were analyzed directly as well as downsampled to reflect the sample size of teosinte. For singletons, our definition of neutral diversity was modified to be mean diversity at positions at least 0.02 cM distal from genes. When evaluating the data in this manner, a very different pattern emerged. Maize singleton diversity was just at least as low as teosinte singleton diversity near genes, but recovered more slowly (Figure 3), implying that in the recent past maize has been at least as influenced by purifying selection as teosinte. Together, these findings suggest that demographic history has a strong influence on the effect of purifying selection. Historically, teosinte has had a larger population size than maize, and only recently has maize population size overcome that of teosinte. Since the efficacy of purifying selection scales with population size,

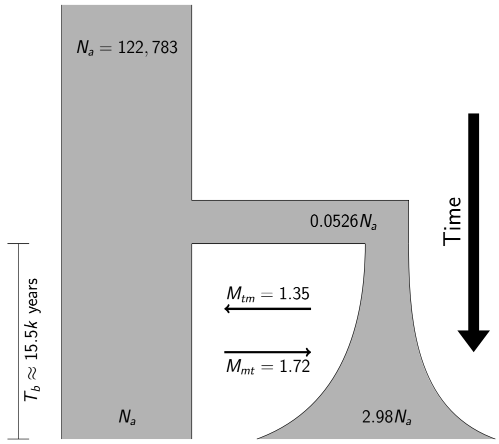

**Fig. 1.** Parameters of domestication as estimated by dadi. Notably, the maize effective population size ($N_e$) during the domestication bottleneck appears to have consisted of approximately 5.26% the ancestral $N_e$, before recovering two at least 2.98 times as large as the ancestral $N_e$.

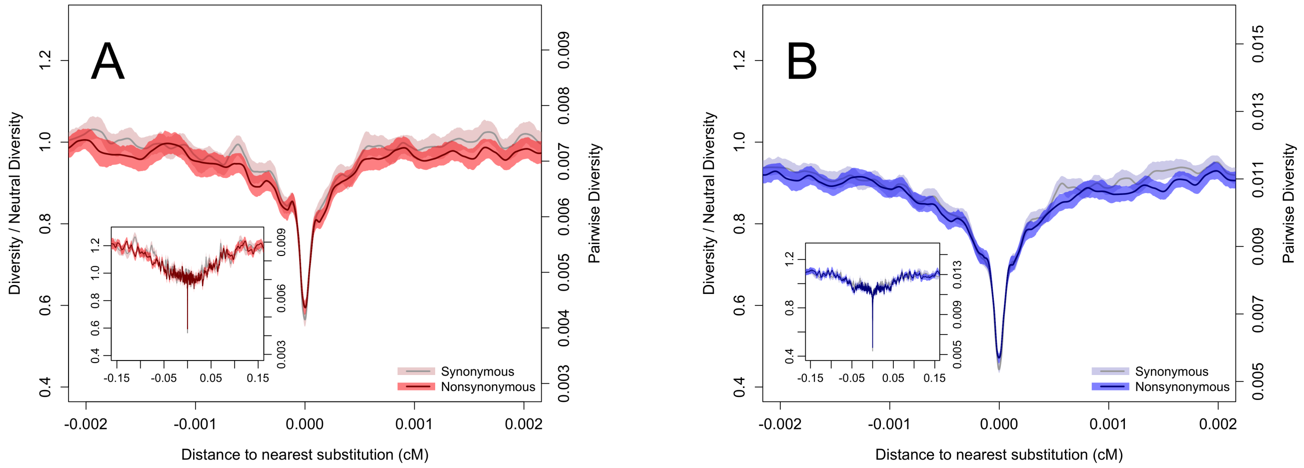

**Fig. 2.** Pairwise diversity surrounding synonymous and nonsynonymous substitutions in maize and teosinte. A: Maize diversity; B: Teosinte diversity.

these results likely reflect changes in $N_e$ more than they reflect underlying changes in selection pressure.

This might be a good place to stick in a paragraph about curve fitting B to estimate s, mu.

**Demography of maize domestication.** To explore whether differences in purifying selection between maize and teosinte can be explained purely by demographic processes, we first estimated the parameters of maize domestication using large-scale sequencing data involving 23 inbred maize landraces and 13 teosinte inbred lines included in the HapMap 2 panel [11] (supplemental table here). The maize lines were collected from across the Americas and the teosinte lines came from central Mexico. Before estimating demography, we compared the site frequency spectrum (SFS) in genic and non-genic regions, and observed substantial differences in the evolution of these classes of sites. For both maize and teosinte, the SFS within genes showed a dearth of low-frequency alleles and Tajima's D (reference here) was therefore shifted to more positive values (figure here). This is consistent with the aforementioned purifying selection and indicates that genic regions are not evolving neutrally. Therefore, for demographic modeling we restricted analysis to non-genic sites.

We used diffusion approximation as implemented in dadi [6] to find the domestication parameters that best explain the joint site frequency spectrum of maize and teosinte. The model we optimized began with an equilibrium ancestral population of size $N_a$ splitting into separate maize and teosinte populations $T_B$ generations in the past. Moving forward in time from the split, teosinte maintained the ancestral population size, while maize experienced an immediate effective population size change to $N_b$ individuals, followed by exponential growth to size $N_m$. Additionally, since the split a $M_{tm}$ individuals migrated from teosinte into maize and $M_{mt}$ individuals migrated from maize to teosinte.

The most likely model suggested an ancestral for $\theta$ ($4N\mu$) of 0.014734, which is similar to previous estimates (references here). Assuming a mutation rate of $\mu = 3 \times 10^{-8}$ (reference here), this suggests an effective population size of $N_a = 122,783$ individuals, a population split $T_B = 15,523$ generations in the past, a maize bottleneck effective populations size of $N_b = 6,455$ individuals (5.26% of $N_a$), a modern maize effective population size of $N_m = 366,973$ individuals, $M_{tm} = 1.35$ migrants per generation from teosinte to maize, and $M_{mt} = 1.72$ migrants per generation from maize to teosinte (Figure 1). We note that a population split over 15 thousand generations before present precedes estimates from archaeological data which suggests maize domestication began approximately 9,000 years before present (reference here). This could result from multiple generations per year, or in may reflect teosinte population structure that was present before domestication. Also, note that the genetic time of the population split must precede morphological changes that could be identified morphologically. Additionally, because recent expansion is most evidenced by rare alleles, and since these data provide low power to detect rare alleles, we expect that the estimate of $N_m = 366,973$, or $\sim 3N_a$, is likely an underestimate.

We utilized a complementary dataset of 4,021 maize landrace individuals collected from across the Americas (SeeDs reference here) to obtain a less downward-biased estimate of the current maize effective population size. According to a singleton-based estimate of $\theta$ [13], which was chosen since rapid post-bottleneck expansion implies that rare alleles will most accurately reflect current maize demography, we obtained the estimate $N_e = 992,713.5$. Although less biased than the previously mentioned esitmate, we note that ascertainment bias of the genotyping platform used to generate these data again biases this estimate downward.

**Simulations confirm the relationship between demographic processes and background selection.** We explored whether or not the demographic patterns that correspond to maize domestication are capable of generating the interesting pattern of background selection that is observed. That is, does de-

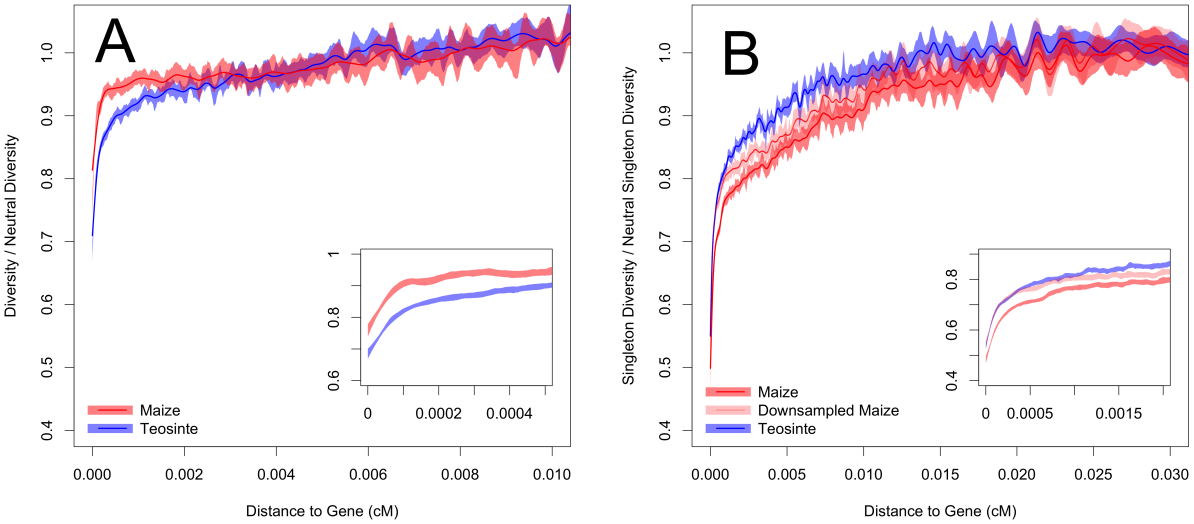

**Fig. 3.** Relative level of diversity versus distance to the nearest gene, in maize and teosinte. Two measures of diversity were investigated. **A** displays pairwise diversity, which is most influenced by intermediate frequency alleles and therefore depicts more ancient evolutionary patterns, and **B** depicts singleton diversity, influenced by rare alleles and thus depicting evolutionary patterns in the recent past. *are these done with latest values we used for curve fitting? should also add the singleton vs pi comparison within species as a supplemental figure*

mography alone explain the weak effect of background selection on pairwise diversity in maize? We begain by estimating the distribution of fitness effects and mutation rate in both maize and teosinte from (formulas)... These were observed to be ZZZZZZ. Then, we simulated mating according to our estimated demographic model and evaluated the resulting levels of pairwise and singleton diversity. A LOT MORE GOES HERE, BUT WHAT IS WRITTEN NOW CAN FRAME THINGS.

## Discussion

**Discussion 1.** Text goes here

**Discussion 2.** Text goes here

## Materials and Methods

**Plant materials.** Accessions studied were selected from the Maize HapMap2 panel [11] . Principal component analysis was employed to ensure that closely related individuals were not included due to their potential to bias results (maybe a supplemental figure here). Ultimately, 23 maize inbreds derived from a diverse assortment of landraces were selected for inclusion. Thirteen teosinte inbred lines, all members of the subspecies Z. *mays* ssp. *parviglumis*, were utilized. Sequences were mapped to the maize B73 version 3 reference genome [14] (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/).

**Interpolating genetic position.** For many of the following analyses, physical position along a chromosome was a less relevant measure of map location than was genetic position. Therefore, physical positions were converted to genetic positions by interpolating from the NAM genetic map (REFERENCE), which provides a 1 cM resolution for physical to genetic conversion. Within R [15], physical positions with corresponding genetic positions in the NAM map were used as anchors. Physical positions in our dataset wihout corresponding genetic position were assigned genetic positions by scaling the anchored genetic positions according to the physical distance between the unlabeled position and the flanking anchors.

**Estimating the site frequency spectrum.** To estimate individual and joint site frequency spectra (SFS) for maize and teosinte from inbred lines, each inbred individual was treated as representing a single haplotype from its population. SFS were separately computed for genic and intergenic regions, as well as for the whole genome together. First, genic and intergenic regions were isolated using the biomaRt package [16,17] of R [15]. Genic regions were defined as DNA between the start and stop position of a gene, while intergenic regions were required to be at least 5kb up- or down-stream from a gene start/stop. With regions defined, SFS were estimated with ANGSD [18]. Individual population SFS were estimated using all positions observed in at least 80% of the individuals in the population, and joint SFS were estimated using all positions observed in at least 80% of individuals in both populations. Individuals were assumed to be fully inbred (-doSaf 2). Quality filters were employed such that reads with quality score below 30 and bases with quality score below 20 were discarded (-minMapq 30 and -minQ 20), as were reads that didn't map uniquely (-uniqueOnly 0). Quality scores around indels were adjusted as in Samtools (-baq 0). Genotype likelihoods were estimated using the samtools method (-GL 1). Major and minor alleles were inferred from the data (-doMaf 1). Because ANGSD cannot calculate a folded joint SFS, the maize reference genome was used for polarization and then unfolded spectra were folded using dadi [6].

**Demographic inference.** We used dadi [6] to estimate the parameters of maize domestication based on diffusion approximation of the 2-dimensional SFS for maize and teosinte. Our first step was to specify a model, as shown in (Figure 1). In words, our model began with an ancestral teosinte population with effective population size $N_a$ individuals. The ancestral population split into two distinct populations, one of maize and the other of teosinte, $T_b$ generations in the past. The maize population immediately shrank to size $N_b$, and then grew exponentially to size $N_{maize}$ today. From the split until today, the teosinte effective population size remained constant. Each year, $M_{mt}$ individuals migrated from maize to teosinte, and $M_{tm}$ individuals migrated from teosinte to maize.The free parameters of our model were $N_a, N_b, N_{maize}, T_b, M_{mt},$ and $M_{tm}$. Python code specifying this model is provided in (SUPPLEMENT). We implimented 1,000 dadi optimizations of this model using the "dadi.Inference.optimize_log_fmin" optimizer. In dadi optimizations, initial values for $N_b, N_{maize}, T_b, M_{mt},$ and $M_{tm}$, respectively, were randomly perturbed up to a factor of two from 0.02, 3, 0.04, 1e-5, and 1e-5. Lower bounds were set at 1e-7, 1e-7, 0, 1e-10, and 1e-10, respecively, while upper bounds were respectively set at 2, 200, 1, .001, and .001. $N_a$ did not have an initial value, upper, or lower bounds because it is estimated from equilibrium ancestral theta ($\theta = 4N_a\mu$), which dadi determines automatically regardless of model. Population sizes were converted from dadi's units using a mutation rate estimate of 3e-8 (REFERENCE). Among the 1,000 dadi runs, parameters from that with the lowest log-likelihood were taken as estimates of the parameters of the domestication process.

Recent population expansion, as is observed for maize, most influences the abundance of rare alleles. However, our demographic parameter estimation using dadi suffered from a relatively small sample size of 23 maize individuals. Therefore, our approach likely underestimated current maize effective population size. To address this, we utilized a complimentary dataset of 4,021 maize landrace individuals collected from across the Americas (SeeDs reference here) to generate an independent estimate of $N_{maize}$. These individuals were genotyped using GBS [19], a protocol that suffers from a high rate of missing data [20]. To minimize biases imposed from missing data, we used R to isolate only those SNPs that were observed in at least 1,500 individuals and subsequently projected the SFS down to a sample of 500 individuals. Then, we utilized the singleton-based estimate of theta ($\theta = 4N_e\mu = \frac{S}{L}$) to estimate modern maize $N_e$ [13], where $S$ is the total number of singletons and $L$ is the total length of DNA sequenced. Based on the GBS protocol implemented [19], $L$ is equal to the total number of SNPs in the dataset.

**Evaluating diversity around substitutions.** To investigate diversity around substitutions, maize and teosinte pairwise diversity was first calculated in 1,000 kb non-overlapping windows using ANGSD [18]. This was performed separately for both maize and teosinte, using the same filters as employed for estimating the SFS. Next, SNPs and genotypes among maize, teosinte, and *Tripsacum* were called. *Tripsacum* bam files were downloaded from (TRIPSICUM FILES), and then all SNPs with a p-value less than 1e-6 were called using ANGSD. Quality filters were as the same as before, and genotypes were only called when the posterior probability was above 0.95. From the set of called SNPs and genotypes, substitutions between maize and tripsicum, as well as between teosinte and *Tripsacum* were identified using R [15] as all positions with no more than 20% missing data for which every maize or teosinte allele differed from the observed *Tripsacum* allele. At each class of substitution, effects were estimated using the ensembl variant effects predictor [21].

For each diversity window with at least 100 bps observed, the distance from the window center to the nearest synonymous and nonsynonymous (missense) substitution was computed. Then, following the methods of [22], a loess curve was plotted for diversity values against the distance to the nearest synonymous or nonsynonymous substitution. A span of 0.01 was utilized. Unlike [23], we did not fit separate loess curves in the up- and down-stream directions, but instead fit single curves encompassing both directions.

**Evaluating diversity around genes.** Two types of diversity surrounding genes were investigated. The first was pairwise diversity in 1kb windows, as described previously. The second was singleton diversity in 1kb windows. Singletons represent the rarest class of alleles that this dataset can identify, and collectively demonstrate the most recent patterns of evolution. Minor allele frequencies were estimated with ANGSD [18] using the same quality filters previously described. Then, the number of singletons in each non-overlapping 1kb window was calculated with R for both maize and teosinte [15]. For maize, we also generated a parallel set of downsampled singleton data, for which binomial sampling within R was conducted based on allele frequencies at each site, to generate a singleton psuedo dataset of sample size 13, equivalent to our sample size for teosinte. BiomaRt [16,17] was then used to identify the center of each gene. Next, the distance from each diversity window center to the nearest gene center was computed.

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

**Simulations.**

**ACKNOWLEDGMENTS.** Various thankyous will be in order. *Coop and his lab. funding from NSF Biology of Rare alleles – need grant number.*

1. Purugganan, M. D & Fuller, D. Q. (2009) Nature 457, 843–848.
2. Ross-Ibarra, J, Morrell, P. L, & Gaut, B. S. (2007) Proceedings of the National Academy of Sciences 104, 8641–8648.
3. Hufford, M. B, Xu, X, Van Heerwaarden, J, Pyhäjärvi, T, Chia, J.-M, Cartwright, R. A, Elshire, R. J, Glaubitz, J. C, Guill, K. E, Kaeppler, S. M, et al. (2012) Nature genetics 44, 808–811.
4. He, Z, Zhai, W, Wen, H, Tang, T, Wang, Y, Lu, X, Greenberg, A. J, Hudson, R. R, Wu, C.-I, & Shi, S. (2011) PLoS genetics 7, e1002100.
5. Luikart, G, England, P. R, Tallmon, D, Jordan, S, & Taberlet, P. (2003) Nature Reviews Genetics 4, 981–994.
6. Gutenkunst, R. N, Hernandez, R. D, Williamson, S. H, & Bustamante, C. D. (2009) PLoS genetics 5, e1000695.
7. Li, J, Li, H, Jakobsson, M, Li, S, SjÖDin, P, & Lascoux, M. (2012) Molecular Ecology 21, 28–44.
8. Slotte, T. (2014) Briefings in functional genomics 13, 268–275.
9. Sella, G, Petrov, D. A, Przeworski, M, & Andolfatto, P. (2009) PLoS genetics 5, e1000495.
10. Gazave, E, Ma, L, Chang, D, Coventry, A, Gao, F, Muzny, D, Boerwinkle, E, Gibbs, R. A, Sing, C. F, Clark, A. G, et al. (2014) Proceedings of the National Academy of Sciences 111, 757–762.
11. Chia, J.-M, Song, C, Bradbury, P. J, Costich, D, de Leon, N, Doebley, J, Elshire, R. J, Gaut, B, Geller, L, Glaubitz, J. C, et al. (2012) Nature genetics 44, 803–807.
12. Charlesworth, B, Morgan, M, & Charlesworth, D. (1993) Genetics 134, 1289–1303.
13. Fu, Y.-X & Li, W.-H. (1993) Genetics 133, 693–709.
14. Schnable, P. S, Ware, D, Fulton, R. S, Stein, J. C, Wei, F, Pasternak, S, Liang, C, Zhang, J, Fulton, L, Graves, T. A, et al. (2009) science 326, 1112–1115.
15. R Core Team. (2014) R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria).
16. Durinck, S, Spellman, P. T, Birney, E, & Huber, W. (2009) Nature protocols 4, 1184–1191.
17. Durinck, S, Moreau, Y, Kasprzyk, A, Davis, S, De Moor, B, Brazma, A, & Huber, W. (2005) Bioinformatics 21, 3439–3440.
18. Korneliussen, T. S, Albrechtsen, A, & Nielsen, R. (2014) BMC bioinformatics 15, 356.
19. Elshire, R. J, Glaubitz, J. C, Sun, Q, Poland, J. A, Kawamoto, K, Buckler, E. S, & Mitchell, S. E. (2011) PloS one 6, e19379.
20. Beissinger, T. M, Hirsch, C. N, Sekhon, R. S, Foerster, J. M, Johnson, J. M, Muttoni, G, Vaillancourt, B, Buell, C. R, Kaeppler, S. M, & de Leon, N. (2013) Genetics 193, 1073–1081.
21. McLaren, W, Pritchard, B, Rios, D, Chen, Y, Flicek, P, & Cunningham, F. (2010) Bioinformatics 26, 2069–2070.
22. Hernandez, R. D, Kelley, J. L, Elyashiv, E, Melton, S. C, Auton, A, McVean, G, Sella, G, Przeworski, M, et al. (2011) science 331, 920–924.
23. Sattath, S, Elyashiv, E, Kolodny, O, Rinott, Y, & Sella, G. (2011) PLoS genetics 7, e1001302.