bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: Genetic costs of domestication and improvement

Authors: Brook T. Moyers[1], Peter L. Morrell[2], and John K. McKay[1]

Affiliations: [1] Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO, USA 80521. [2] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN, USA 55108.

Address for correspondence:
Brook T. Moyers
Bioagricultural Sciences and Pest Management
Colorado State University
Fort Collins, CO, USA 80521
Phone: 707-235-1237
Email: brook.moyers@gmail.com

Running title: Cost of domestication

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

The 'cost of domestication' hypothesis posits that the process of domesticating wild species increases genetic load by increasing the number, frequency, and/or proportion of deleterious genetic variants in domesticated genomes. This cost may limit the efficacy of selection and thus reduce genetic gains in breeding programs for these species. Understanding when and how genetic load evolves can also provide insight into fundamental questions about the interplay of demographic and evolutionary dynamics. Here we describe the evolutionary forces that may contribute to genetic load during domestication and improvement, and review the available evidence for 'the cost of domestication' in animal and plant genomes. We identify gaps and explore opportunities in this emerging field, and finally offer suggestions for researchers and breeders interested in addressing genetic load in domesticated species.

Keywords: genetic load, deleterious mutations, crops, domesticated animals

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Recently, we have seen a resurgence of evolutionary thinking applied to domesticated plants and animals (e.g. Walsh 2007; Wang *et al*. 2014; Gaut, Díez, and Morrell 2015; Kono *et al*. 2016). One particular wave of this resurgence proposes a general 'cost of domestication': that the evolutionary processes experienced by lineages during domestication are likely to have increased genetic load. This 'cost' was first hypothesized by Lu *et al*. (2006), who found an increase in nonsynonymous substitutions, particularly radical amino acid changes, in domesticated compared to wild lineages of rice. These putatively deleterious mutations were negatively correlated with recombination rate, which the authors interpreted as evidence that these mutations hitchhiked along with the targets of artificial selection (Figure 1B). Lu *et al*. (2006) conclude that: "The reduction in fitness, or the genetic cost of domestication, is a general phenomenon." Here we address this claim by examining the evidence that has emerged in the last decade on genetic load in domesticated species.

The processes of domestication and improvement potentially impose a number of evolutionary forces on populations (Box 1), starting with mutational effects. New mutations can have a range of effects on fitness, from lethal to beneficial. Deleterious variants constitute the mostly directly observable, and likely the most important, source of genetic load in a population. The shape of the distribution of fitness effects of new mutations is difficult to estimate, but theory predicts that a large proportion of new mutations, particularly those that occur in coding portions of the genome, will be deleterious at least in some proportion of the environments that a species occupies (Ohta 1972; 1992; Gillespie 1994). These predictions are supported by experimental responses to artificial selection and mutation accumulation experiments, and molecular genetic studies (discussed in Keightley and Lynch 2003).

The rate of new mutations in eukaryotes varies, but is likely at least 1×10^{-8} / base pair / generation (Baer, Miyamoto, and Denver 2007). For the average

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

eukaryotic genome, individuals should thereby be expected to carry a small number of new mutations not present in the parent genome(s) (Agrawal and Whitlock 2012). The realized distribution of fitness effects for these mutations will be influenced by inbreeding and by effective population size (Figure 2A; Gillespie 1999; Whitlock 2000; Arunkumar *et al*. 2015; Keightley and Eyre-Walker 2007). Generally, for a given distribution of fitness effects for new mutations, we expect to observe relatively fewer strongly deleterious variants and more weakly deleterious variants in smaller populations and in populations with higher rates of inbreeding (Figure 2A).

Domestication and improvement often involve increased inbreeding. Allard (1999) pointed out that many important early cultigens were inbreeding, with inbred lines offering morphologically consistency in the crop. In some cases, domestication came with a switch in mating system from highly outcrossing to highly selfing (e.g. in rice; Kovach, Sweeney, and McCouch 2007). The practice of producing inbred lines, often through single-seed descent, with the specific intent to reduce heterozygosity and create genetically 'stable' varieties, remains a major activity in plant breeding programs. Reduced effective population sizes (N_e) during domestication and improvement and artificial selection on favorable traits also constitute forms of inbreeding (Figure 1). Even in species without the capacity to self-fertilize, inbreeding that results from selection can dramatically change the genomic landscape. For example, the first sequenced dog genome, from a female boxer, is 62% homozygous haplotype blocks, with an N50 length of 6.9 Mb (versus heterozygous region N50 length 1.1 Mb; Lindblad-Tor *et al*. 2005). This level of homozygosity drastically reduces the effective recombination rate, as crossover events will essentially be switching identical chromosomal segments. The likelihood of a beneficial allele moving into a genomic background with fewer linked deleterious alleles is thus reduced.

Linked selection, or interference among mutations, has much the same effect. For a trait influenced by multiple loci, interference between linked loci can limit

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the response to selection (Hill and Robertson 1966). Deleterious variants are more numerous than those that have a positive effect on a trait and thus should constitute a major limitation in responding to selection (Felsenstein 1974). This forces selection to act on the net effect of favorable and unfavorable mutations in linkage (Figure 1B). As inbreeding increases, selection is less effective at purging moderately deleterious mutations and new slightly beneficial mutations more likely to be lost by genetic drift, which will shift the distribution of fitness effects for segregating variants and result in greater accumulation of genetic load over time (Figure 2A–C).

Overall, the cost of domestication hypothesis posits that compared to their wild relatives domesticated lineages will have:

1. Deleterious variants at higher frequency, number, and proportion
2. Enrichment of deleterious variants in linkage disequilibrium with loci subject to strong, positive, artificial selection

With domestications, these effects may differ between lineages that experienced domestication followed by modern improvement (e.g. 'elite' varieties and commercial breeds) and those that experienced domestication only (e.g. landraces and non-commercial populations). Generally speaking, the process of domestication is characterized by a genetic bottleneck followed by a long period of relatively weak and possibly varying selection, while during the process of improvement, intense selection over short time periods is coupled with limited recombination and an additional reduction in N_e, often followed by rapid population expansion (Figure 1A; Yamasaki, Wright, and McMullen 2007). Gaut, Díez, and Morrell (2015) propose that elite crop lines should harbor a lower proportion of deleterious variants relative to landraces due to strong selection for yield during improvement, but the opposing pattern could be driven by increased genetic drift due to limited recombination, lower N_e, and rapid population expansion. At least one study shows little difference in genetics load between landrace and elite lines in sunflower (Renaut and Rieseberg 2015). It is likely that

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relative influence of these factors varies dramatically across domesticated systems.

Box 1. Domesticated lineages may experience:

- **Reduced efficacy of selection:**
  - Deleterious variants that are in linkage disequilibrium with a target of artificial selection can rise to high frequency through genetic hitchhiking, as long as the their fitness effects are smaller than the effect of the targeted variant (Figure 1B).
  - Domesticated lineages can experience reduced effective recombination rate as inbreeding increases and heterozygosity drops (Figure 1B). Increased linkage disequilibrium allows larger portions of the genome to hitchhike, and negative linkage disequilibrium across advantageous alleles on different haplotypes can cause selection interference (Felsenstein 1974).
  - The loss of allelic diversity through artificial selection, genetic drift, and inbreeding (Figure 1B) can reduce trait variance (Eyre-Walker, Woolfit, and Phelps 2006), which reduces the efficacy of selection.
  - **However**: any significant increase in inbreeding, particularly the transition from outcrossing to selfing, **allows selection to more efficiently purge recessive, highly deleterious alleles**, as these alleles are exposed in homozygous genotypes more frequently (Arunkumar *et al*. 2015). This shifts the more deleterious end of the distribution of fitness effects for segregating variants towards neutrality (Figure 2A).

- **Increased genetic drift:**
  - Domestication involves one or more genetic bottlenecks (Figure 1A), where N_e is reduced. In smaller populations, genetic drift is stronger relative to the strength of selection, and deleterious variants can reach higher frequencies.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- N_e varies across the genome (Charlesworth 2009). In particular, when regions surrounding loci under strong selection (e.g. domestication QTL) rapidly rise to high frequency, the N_e for that haplotype rapidly drops, and so drift has a relatively stronger effect in that region.
- The rapid population expansion coupled with long-distance migration typical of many domesticated lineages' demographic histories can result in the accumulation of deleterious genetic variation known as 'expansion load' (Peischl *et al*. 2013; Lohmueller 2014). This occurs when serial bottlenecks are followed by large population expansions (e.g., large local carrying capacities, large selection coefficients, long distance dispersal; Peischl *et al*. 2013). This phenomenon is due to the accumulation of new deleterious mutations at the 'wave front' of expanding populations, which then rise to high frequency via drift regardless of their fitness effects (i.e. 'gene surfing', or 'allelic surfing' Klopfstein, Currat, and Excoffier 2006; Travis *et al*. 2007).

- **Increased mutation load**
  - Domesticated and wild lineages may differ in basal mutation rate, although it is unclear if we would expect this to be biased towards higher rates in domesticated lineages.
  - More likely, deleterious mutations **accumulate** at higher rates in domesticated lineages due to the reduced efficacy of selection. Mutations that would be purged in larger N_e or with higher effective recombination rates are instead retained. This is reflected in a shift in the distribution of fitness effects for segregating variants towards more moderately deleterious alleles (Figure 2A).
  - Smaller population sizes decrease the likelihood of beneficial mutations arising, as these mutations are likely relatively rare and

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

for a given mutation rate fewer individuals mean fewer opportunities for mutation.

Genetic Load in Domesticated Species

The increasing number of genomic datasets has made the study of genome-wide genetic load increasingly feasible. Since Lu *et al*. (2006) first hypothesized genetic load as a 'cost of domestication' in rice, similar studies have examined evidence for such costs in other crops and domesticated animals.

One approach for estimating genetic load is to compare population genetic parameters between two lineages, in this case domesticated species versus their wild relatives. If domestication comes with the cost of genetic load, we might expect domesticated lineages to have accumulated more nonsynonymous mutations or, if the accumulation of genetic load is due to reduced effective population sizes, to show reduced allelic diversity and effective recombination when compared to their wild relatives. While relatively straightforward, this approach only indirectly addresses genetic load by assuming that: (A) nonsynonymous mutations are on average deleterious, and (B) loss of diversity and reduced recombination indicate a decreased efficacy of selection. This approach also assumes that wild relatives of domesticated species have not themselves experienced population bottlenecks, shifts in mating system, or any of the other processes that could affect genetic load relative to the ancestors.

Another approach for estimating genetic load is to assay the number and proportions of putatively deleterious variants present in populations of domesticated species, using various bioinformatic approaches. This approach directly addresses the question of genetic load, but does not tell us whether the load is the result of domestication or other evolutionary processes (unless wild relatives are also assayed, again assuming that their own evolutionary trajectory

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

has not simultaneously been affected). It also requires accurate, unbiased algorithms for identifying deleterious variants from neutral variants (see below).

Studies taking these approaches in domesticated species are presented Table 1. We searched the literature using Google Scholar with the terms “genetic load” and “domesticat*”, and in many cases followed references from one study to the next. To the best of our knowledge, the studies in Table 1 represent the majority of the extant literature on this topic. We excluded studies examining only the mitochondrial or other non-recombining portions of the genome. In a few cases where numeric values were not reported, we extracted values from published figures using relative distances as measured in the image analysis program ImageJ (Schneider, Rasband, and Eliceiri 2012), or otherwise extrapolated values from the provided data. Where exact values were not available, we provide estimates. Please note that no one study (or set of genotypes) contributed values to all columns for a particular domestication event, and that in many cases methods used or statistics examined varied across species.

Recombination and linkage disequilibrium

The process of domestication may increase recombination rate, as measured in the number of chiasmata per bivalent. Theory predicts that recombination rate should increase during periods of rapid evolutionary change (Otto and Barton 1997), and in domesticated species this may be driven by strong selection and limited genetic variation under linkage disequilibrium (Otto and Barton 2001). This is supported by the observation that recombination rates are higher in many domesticated species compared to wild relatives (e.g., chicken, Groenen *et al*. 2009; honey bee, Wilfert *et al.* 2007; and a number of cultivated plant species, Ross-Ibarra 2004). In contrast, recombination in some domesticated species is limited by genomic structure (e.g., in barley, where 50% of physical length and many functional genes are in (peri-)centromeric regions with extremely low recombination rates; The International Barley Genome Sequencing Consortium 2012), although this structure may be shared with wild relatives. Even when

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

actual recombination rate (chiasmata per bivalent) increases, effective recombination may be reduced in many domesticated species due to an increase in the physical length of fixed or high frequency haplotypes, or linkage disequilibrium (LD). In other words, chromosomes may physically recombine, but if the homologous chromosomes contain identical sequences then no "effective recombination" occurs and the outcome is no different than no recombination. In a striking example, the parametric estimate of recombination rate in maize populations, has decreased by an average of 83% compared to the wild relative teosinte (Wright *et al*. 2005). We note that Mezmouk and Ross-Ibarra (2014) found that in maize deleterious variants are not enriched in areas of low recombination (although see McMullen *et al*. 2009; Rodgers-Melnick *et al*. 2015).

Strong directional selection like that imposed during domestication and improvement can reduce genetic diversity in long genetic tracts linked to the selected locus (Maynard Smith and Haigh 1974). These regions of extended LD are among the signals used to identify targets of selection and reconstruct the evolutionary history of domesticated species (e.g., Tian, Stevens, and Buckler 2009). A shift in mating system from outcrossing to selfing, or an increase in inbreeding, will have the same effect across the entire recombining portion of the genome, as heterozygosity decreases with each generation (Charlesworth 2003). Extended LD has consequences for the efficacy of selection: deleterious variants linked to larger-effect beneficial alleles can no longer recombine away, and beneficial variants that are not in LD with larger-effect beneficial alleles may be lost. We see evidence for this pattern in our review of the literature: LD decays most rapidly in the outcrossing plant species in Table 1 (maize and sunflower), and extends much further in plant species propagated by selfing and in domesticated animals. In all cases where we have data for wild relatives, LD decays more rapidly in wild lineages than in domesticated lineages (Table 1).

While we primarily report mean LD in Table 1, linkage disequilibrium also varies among varieties and breeds of the same species. For example, in domesticated

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pig breeds the length of the genome covered by runs of homozygosity ranges from 13.4 to 173.3 Mb, or 0.5–6.5% (Traspov *et al*. 2016). Similarly, among dog breeds average LD decay (to $r^2 \leq 0.2$) ranges from 20 kb to 4.2 Mb (Gray *et al*. 2009). In both species, the decay of LD for wild individuals occurs over even shorter distances. This suggests that patterns of LD in these species have been strongly impacted by breed-specific demographic history (i.e., the process of improvement), in addition to the shared process of domestication.

Genetic diversity and dN/dS

In the species we present in Table 1, we see consistent loss of genetic diversity when 'improved' or 'breed' genomes are compared to domesticated 'landrace' or 'non-commercial' genomes, and again when domesticated genomes are compared to the genomes of wild relatives. This ranges from ~5% nucleotide diversity lost between wolf populations and domesticated dogs (Gray *et al*. 2009) to 77% lost between wild and improved tomato populations (Lin *et al*. 2014). The only case where we see a gain in genetic diversity is in the Andean domestication of the common bean, where gene flow with the more genetically diverse Mesoamerican common bean is likely an explanatory factor (Schmutz *et al*. 2013). This pattern is consistent with a broader review of genetic diversity in crop species: Miller and Gross (2011) found that annual crops had lost an average of ~40% of the diversity found in their wild relatives. This same study found that perennial fruit trees had lost an average of ~5% genetic diversity, suggesting that the impact of domestication on genetic diversity is strongly influenced by life history (see also Gaut, Díez, and Morrell 2015). Given a relatively steady evolutionary trajectory for wild populations, loss of genetic diversity in domesticated populations can be attributed to: genetic bottlenecks, increased inbreeding, artificial selection. On an evolutionary timescales, even the oldest domestications occurred quite recently relative to the rate at which new mutations can recover the loss. This is especially true for non-recombining portions of the genome like the mitochondrial genome or non-recombining sex chromosomes. Most modern animal breeding programs are strongly sex-biased,

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with few male individuals contributing to each generation. In horses, for example, this has likely led to almost complete loss of polymorphism on the Y chromosome via genetic drift (Lippold *et al*. 2011). Loss of allelic diversity can reduce the efficacy of selection by reducing trait variation within species (Eyre-Walker, Woolfit, and Phelps 2006). However, a loss of genetic diversity alone does not necessarily signal a corresponding increase in the frequency or proportion of deleterious variants, and so is not sufficient evidence of genetic load.

In six of seven species, the domesticated lineage shows an increase in genome-wide nonsynonymous to synonymous substitution rate or number compared to the wild relative lineage (Table 1). The exception is in soybean, where the domesticated *Glycine max* and wild *G. soja* genomes contain approximately the same proportion of nonsynonymous to synonymous single nucleotide polymorphisms (Lam *et al*. 2010). These differences in nonsynonymous substitution rate are likely driven by differences in N_e (Eyre-Walker and Keightley 2007; Woolfit 2009). If most nonsynonymous mutations are deleterious, as theory and empirical data suggest, then an increase in nonsynonymous substitutions will reduce mean fitness (Figure 2B-C). The comparisons in Table 1 suggest that this has occurred in domesticated species. This result differs from Moray, Lanfear, and Bromham (2014), who examined rates of mitochondrial genome sequence evolution in domesticated animals and their wild relatives and found no such consistent pattern. This difference may be attributable to the focus of each review (genome-wide versus mitochondria) or, as the authors speculate, to genetic bottlenecks in some of the wild relatives included in their study.

The ratio of nonsynonymous to synonymous substitutions may not be a good estimate for mutational load. For one, nonsynonymous substitutions are particularly likely to be phenotype changing (Kono *et al*. *bioRxiv*) and contribute to agronomically important phenotypes (Kono *et al*. 2016). Thus artificial selection during domestication and improvement is likely to drive these variants to higher frequencies (or fixation) in domesticated lineages. Thus variants that

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

annotate as deleterious based on sequence conservation (see below), can in some case, contribute to agronomically important phenotypes. In addition, estimates of the proportion of nonsynonymous sites with deleterious effects range from 0.03 (in bacteria; Hughes 2005) to 0.80 (in humans; Fay *et al*. 2001; The Chimpanzee Sequencing and Analysis Consortium 2005). It follows that nonsynonymous substitution rate may have a poor correlation with genetic load, at least at larger taxonomic scales. However, the consensus proportion of deleterious variants in Table 1 is between 0.05 and 0.25, which spans a smaller range. Finally, dN/dS may in itself be a flawed estimate of functional divergence because it relies on the assumptions that dS is neutral and mean dS can control for substitution rate variation, which may not hold true, especially for closely related taxa (Wolf *et al*. 2009; see also Kryazhimskiy and Plotkin 2008).

Deleterious variants

An increase in genetic load can come from: (1) increased frequency of deleterious variants, (2) increased number of deleterious variants, and (3) enrichment of deleterious variants relative to total variants. The first can be assessed by examining shared deleterious variants between wild and domesticated lineages, and all three by comparing both shared and private deleterious variants across lineages. Looking at deleterious variants in high frequency across all domesticated varieties may provide insight into the early processes of domestication, while looking at deleterious variants with varying frequencies among domesticated varieties may provide insight into the processes of improvement. We recommend that researchers studying deleterious variants present their results per genome, and then compare across genomes and among lineages. Many, but not all, of the studies in Table 1 take this approach. The total number, frequency, or proportion of deleterious variants within a population will necessarily depend on the size of that population, and so sufficient sampling is important before values can be compared across populations. Values per genome are therefore easier to compare across most available genomic datasets.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Renaut and Rieseberg (2015) found a significant increase in both shared and private deleterious mutations in domesticated relative to wild lines of sunflower, and similar patterns in two additional closely-related species: cardoon and globe artichoke (Table 1). This pattern also holds true for *japonica* and *indica* domesticated rice (Liu *et al*. 2017) and for the domesticated dog (Marsden *et al*. 2016) compared to their wild relatives (Table 1). In horses, deleterious mutation load as estimated using Genomic Evolutionary Rate Profiling (GERP) appears to be higher in both domesticated genomes and the extant wild relative (Przewalski's horse, which went through a severe genetic bottleneck in the last century) compared to an ancient wild horse genome (Schubert *et al*. 2014). These four studies, spanning a wide taxonomic range, suggest that an increase in the number and proportion of deleterious variants may be a general consequence of domestication. However, the other studies we present that examined deleterious variants in domesticated species did not include any sampling of wild lineages. Without sufficient sampling of a parallel lineage that did not undergo the process of domestication, it is difficult to assess whether the 'cost of domestication' is indeed general.

Identifying dangerous hitchhikers

The effect size of a deleterious mutation is negatively correlated with its likelihood of increasing in frequency through any of the mechanisms we discuss here. That is, variants with a strongly deleterious effect are more likely to be purged by selection than mildly deleterious variants, regardless of the efficacy of selection. Similarly, mutations that have a consistent, environmentally-independent deleterious effect are more likely to be purged than mutations with environmentally-plastic effects. In the extreme case, we would never expect mutations that have a consistently lethal effect in a heterozygous state to contribute to genetic load, as these would be lost in the first generation of their appearance in a population. However, mutations with consistent, highly deleterious effects are likely rare relative to those with smaller or environment-

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

412 dependent effects, especially in inbred populations (Figure 2A; Arunkumar *et al*.
413 2015). When thinking about genetic load, it is therefore important to recognize
414 that the effect of any particular mutation can depend on context, including
415 genomic background and developmental environment. This complexity likely
416 makes these classes of deleterious mutations more difficult to identify, and we
417 might not expect these variants to show up in our bioinformatic screens
418 (discussed below). Furthermore, all of these expectations are modified by
419 linkage: when deleterious variants are in LD with targets of artificial selection,
420 they are more likely to evade purging even with consistent, large, deleterious
421 effects (Figure 1B).
422
423 We searched the literature for examples of specific deleterious variants that
424 hitchhiked along with targets of selection during domestication and improvement.
425 We did not find many such cases, so we describe each in detail here. The best
426 characterized example comes from rice, where an allele that negatively affects
427 yield under drought (*qDTY1.1*) is tightly linked to the major green revolution
428 dwarfing allele *sd1* (Vikram *et al*. 2015). Vikram *et al*. (2015) found that the
429 *qDTY1.1* allele explained up to 31% of the variance in yield under drought across
430 three RIL populations and two growing seasons. Almost all modern elite rice
431 varieties carry the *sd1* allele (which increases plant investment in grain yield),
432 and as a consequence are drought sensitive. The discovery of the *qDTY1.1*
433 allele has enabled rice breeders to finally break the linkage and create drought
434 tolerant, dwarfed lines.
435
436 In sunflower, the *B* locus affects branching and was a likely target of selection
437 during domestication (Bachlava *et al*. 2010). This locus has pleiotropic effects on
438 plant and seed morphology that, in branched male restorer lines, mask the effect
439 of linked loci with both ‘positive’ (increased seed weight) and ‘negative’ (reduced
440 seed oil content) effects (Bachlava *et al*. 2010). To properly understand these
441 effects required a complex experimental design, where these linked loci were
442 segregated in unbranched (*b*) and branched (*B*) backgrounds. Managing these

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

443 effects in the heterotic sunflower breeding groups has likely also been
444 challenging.
445
446 A similarly complex narrative has emerged in maize. The gene *TGA1* was key to
447 evolution of 'naked kernels' in domesticated maize from the encased kernels of
448 teosinte (Wang *et al*. 2015). This locus has pleiotropic effects on kernel features
449 and plant architecture, and is in linkage disequilibrium with the gene *SU1*, which
450 encodes a starch debranching enzyme (Brandenburg *et al*. 2017). *SU1* was
451 targeted by artificial selection during domestication (Whitt *et al*. 2002), but also
452 appears to be under divergent selection between Northern Flints and Corn Belt
453 Dents, two maize populations (Brandenburg *et al*. 2017). This is likely because
454 breeders of these groups are targeting different starch qualities, and this work
455 may have been made more difficult by the genetic linkage of *SU1* with *TGA1*.
456
457 In the above cases, the linked allele(s) with negative agronomic effects are
458 unlikely to be picked up in a genome-wide screen for deleterious variation, as
459 they are segregating in wild or landrace populations and are not necessarily
460 disadvantageous in other contexts. One putative 'truly' deleterious case is in
461 domesticated chickens, where a missense mutation in the thyroid stimulating
462 hormone receptor (*TSHR*) locus sits within a shared selective sweep haplotype
463 (Rubin *et al*. 2010). However, Rubin *et al*. (2010) argue this is more likely a case
464 where a 'deleterious' (i.e. non-conserved) allele was actually the target of artificial
465 selection and potentially contributed to the loss of seasonal reproduction in
466 chickens.
467
468 In the Roundup Ready (event 40-3-2) soybean varieties released in 1996, tight
469 linkage between the transgene insertion event and another allele (or possibly an
470 allele created by the insertion event itself) reduced yield by 5-10% (Elmore *et al*.
471 2001). This is not quite genetic hitchhiking in the traditional sense, but the yield
472 drag effect persisted through backcrossing of the transgene into hundreds of
473 varieties (Benbrook 1999). This effect likely explains, at least in part, why

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

transgenic soybean has failed to increase realized yields (Xu *et al*. 2013). A second, independent insertion event in Roundup Ready 2 Yield® does not suffer from the same yield drag effect (Horak *et al*. 2015).

It is likely that 'dangerous hitchhiker' examples exist that have either gone undetected by previous studies (possibly due to low genomic resolution, limited phenotyping, or limited screening environments), have been detected but not publicized, or are buried among other results in, for example, large QTL studies. It is also possible that the role of genetic hitchhiking has not been as important in shaping genome-wide patterns of genetic load as previously assumed.

Gaps and Opportunities

For major improvement alleles, how much of the genome is in LD?

Artificial selection during domestication targets a clear change in the optimal multivariate phenotype. This likely affects a significant portions of the genome: available estimates include 2–4% of genes in maize (Wright *et al*. 2005) and 16% of the genome in common bean (Papa *et al*. 2007) targeted by selection during domestication. For crops, traits such as seed dormancy, branching, indeterminate flowering, stress tolerance, and shattering are known to be selected for different optima under artificial versus natural selection (Takeda and Matsuoka 2008; Gross and Olsen 2010). In some cases, we know the loci that underlie these domestication traits. One well studied example is the green revolution dwarfing gene, *sd1* in rice. *sd1* is surrounded by a 500kb region (~13 genes) with reduced allelic diversity in *japonica* rice (Asano *et al*. 2011). Another example in rice is the *waxy* locus, where a 250 kb tract shows reduced diversity consistent with a selective sweep in temperate *japonica* glutinous varieties (Olsen *et al*. 2006). The difference in the size of the region affected by these two selective sweeps may be because the strength of selection on these two traits varied, with weaker selection at *waxy* then *sd1*. Unfortunately, the relative strength of selection on domestication traits is largely unknown, and other factors can also influence the size of the genomic region affected by artificial selection.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

505 The physical position of selected mutation can have large effect on this via gene
506 density and local recombination rate (e.g. in rice, Flowers *et al*. 2012). This
507 explanation has been invoked in maize (Wright *et al*. 2005), where the extent of
508 LD surrounding domestication loci is highly variable. For example a 1.1 Mb
509 region (~15 genes) lost diversity during a selective sweep on chromosome 10 in
510 maize (Tian, Stevens, and Buckler 2009), but only a 60–90 kb extended
511 haplotype came with the *tb1* domestication allele (Clark *et al*. 2004). While these
512 case studies provide examples of sweeps resulting from domestication, they also
513 show that the size of the affected region is highly variable, and we don't yet know
514 how this might impact genetic load. This is compounded by the fact that even in
515 highly researched species we don't always know what were the genomic targets
516 of selection during domestication (e.g., in maize; Hufford *et al*. 2012), especially if
517 any extended LD driven by artificial selection has eroded or the intensity or mode
518 of artificial selection has changed over time.
519
520 *What's "worse" in domesticated species: hitchhikers, drifters, or inbreds?*
521 We do not have a clear sense of which evolutionary processes contribute most to
522 genetic load, and sometimes see contrasting patterns across species. In maize,
523 relatively few putatively deleterious alleles are shared across all domesticated
524 lines (hundreds vs. thousands; Mezmouk and Ross-Ibarra 2014), which points to
525 a larger role for the process of improvement than for domestication in driving
526 genetic load. In contrast, the increase in dog dN/dS relative to wolf populations
527 appears to not be driven by recent inbreeding (i.e., improvement) but by the
528 ancient domestication bottleneck common to all dogs (Marsden *et al*. 2016).
529
530 Of the deleterious alleles segregating in more than 80% of maize lines, only 9.4%
531 show any signal of positive selection (Mezmouk and Ross-Ibarra 2014). This
532 suggests that hitchhiking during domestication played a relatively small role in
533 the evolution of deleterious variation in maize. The same study found little
534 support for enrichment of deleterious SNPs in areas of reduced recombination
535 (Mezmouk and Ross-Ibarra 2014). However, Rodgers-Melnick *et al*. (2015)

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present contrasting evidence supporting enrichment of deleterious variants in regions of low recombination, and the authors argue that this difference is due to the use of a tool that does not rely on genome annotation (Genomic Evolutionary Rate Profiling, or GERP; Cooper *et al*. 2005). This complex narrative has emerged from just one domesticated system, and it is likely that each species will present new and different complexities.

Specific differences

Examining general differences between wild and domesticated lineages ignores species-specific demographic histories and changes in life history, which may be important contributors to genetic load. Although we found some general patterns (e.g., loss of genetic diversity, Table 1), we also see clear exceptions (e.g., the Andean common bean). We can attribute these exceptions to particular demographic scenarios (e.g., gene flow with Mesoamerican common bean populations), assuming we have sufficient archeological, historical, or genetic data. One clear problem is our inability to sample ancestral (pre-domestication) lineages, and our subsequent reliance on sampling of current wild relative lineages that have their own unique evolutionary trajectories. Sequencing ancient DNA can provide some insight into the history of these lineages and their ancestral states (e.g., in horses; Schubert *et al*. 2014). Currently, we know very little about most domesticated species' histories. We are still working towards understanding dynamics since domestication: even in highly-researched species like rice, the number of domestication events and subsequent demographic dynamics are hotly contested (Kovach, Sweeney, and McCouch 2007; Gross and Zhao 2014; Chen, Huang, and Han 2016; Choi *et al*. 2017).

A second challenge, briefly mentioned above, is understanding the relative importance of each of the factors in any particular domestication event. For example, in rice the shift to selfing from outcrossing during domestication appears to have played a larger role than the domestication bottleneck in shaping genetic load (Liu *et al*. 2017). This is useful in understanding rice

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

domestication and its impact, but similar studies would need to be conducted across domesticated species to understand the generality of this dynamic. It is possible that general patterns may be drawn from subsets of domesticated species (e.g., vertebrates versus vascular plants, short-lived versus long-lived, outcrossing versus selfing versus clonally propagated etc.). For example, there may be general differences between annual and perennial crops, including less severe domestication bottlenecks and higher levels of gene flow from wild populations in perennials (Miller and Gross 2011; Gaut, Díez, and Morrell 2015).

Predictive algorithms

The identification of individual deleterious variants typically relies on sequence conservation. That is, if a particular nucleotide site or encoded amino acid is invariant across a phylogenetic comparison of related species, a variant is considered more likely to be deleterious. More advanced approaches use estimates of synonymous substitution rates at a locus to improve estimates of constraint on a nucleotide site (Chun and Fay 2009). The majority of 'SNP annotation' approaches are intended for the annotation of amino acid changing variants, although at least one (GERP++; Davydov *et al*. 2010) can be applied to noncoding sequences in cases where nucleotide sequences can be aligned across species. This estimation of phylogenetic constraint is heavily dependent on the sequence alignment used, and so new annotation approaches have sought to use more consistent sets of alignments across loci. Both GERP++ and MAPP permit users to provide alignments for SNP annotation (Davydov *et al*. 2010; Stone and Sidow 2005). The recently reported tool BAD_Mutations (Kono *et al*. 2016; Kono *et al*. bioRxiv) permits the use of a consistent set of alignments for the annotation of deleterious variants by automating the download and alignment of the coding portion of plant genomes from Phytozome and Ensembl Plants, which include 50+ sequenced angiosperm genomes (Goodstein *et al*. 2010; Kersey *et al*. 2016). Currently, the tool is configured for use with angiosperms, but could be applied to other organisms.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

598 Using phylogenetic conservation may be problematic in domesticated species, as
599 relaxed, balancing, or diversifying selection in agricultural environments could lift
600 constraint on sites under purifying or stabilizing selection in wild environments
601 (assuming some commonalities within these environment types). In one such
602 case, two AGPase small subunit paralogs in maize appear to be under
603 diversifying and balancing selection, respectively, even though these subunits
604 are likely under selective constraint across flowering plants (Georgelis, Shaw,
605 and Hannah 2009; Corbi *et al*. 2010). Similarly, broadly 'deleterious' traits may
606 have been under positive artificial selection in domesticated species. The loci
607 underlying these traits could flag as deleterious in bioinformatic screens despite
608 increasing fitness in the domesticated environment. For example, the *fgf4*
609 retrogene insertion that causes chrondrodysplasia (short-leggedness) in dogs
610 would likely have a strongly deleterious effect in wolves, but has been positively
611 selected in some breeds of dog (Parker *et al*. 2009). Finally, bioinformatic
612 approaches that rely on phylogenetic conservation are likely to miss variants with
613 effects that are only deleterious in specific environmental or genomic contexts
614 (plastic or epistatic effects), or which reduce fitness specifically in agronomic or
615 breeding contexts. Specific knowledge of the phenotypic effects of putatively
616 deleterious mutations is necessary to address these issues, but as we discuss
617 below these data are challenging to obtain.
618
619 Information from the site frequency spectrum, or the number of times individual
620 variants are observed in a sample, can provide additional information about
621 which variants are most likely to be deleterious. In resequencing data from many
622 species, nonsynonymous variants typically occur at lower average frequencies
623 than synonymous variants (cf. Nordborg *et al*. 2005; Ross-Ibarra *et al*. 2009;
624 Günther and Schmid 2010). Mutations that annotate as deleterious are
625 particularly likely to occur at lower frequencies than other classes of variants
626 (Marth *et al*. 2011; Kono *et al*. 2016; Liu *et al*. 2017) and may be less likely to be
627 shared among populations (Marth *et al*. 2011).
628

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tools for the annotation of potentially deleterious variants continue to be developed rapidly (see Grimm *et al*. 2015 for a recent comparison). This includes many tools that attempt to make use of information beyond sequence conservation, including potential effects of variant on protein structure or function (Adzhubei *et al*. 2010) or a diversity of genomic information intended to improve prediction of pathogenicity in humans (Kircher *et al*. 2014). This contributes to the issue that the majority of SNP annotation tools are designed to work on human data and may not be applicable to other organisms (Kono *et al*. *bioRxiv*). There is the potential for circularity when an annotation tool is trained on the basis of pathogenic variants in humans and then evaluated on the basis of a potentially overlapping set of variants (Grimm *et al*. 2015). Even given these limitations, validation outside of humans is more challenging because of a paucity of phenotype-changing variants. To address this issue, Kono et al. (*bioRxiv*) report a comparison of seven annotation tools applied to a set of 2,910 phenotype changing variants in the model plant species *Arabidopsis thaliana*. The authors find that all seven tools more accurately identify phenotyping changing variants likely to be deleterious in *Arabidopsis* than in humans, potentially because the larger N_e in *Arabidopsis* relative to human populations may allow more effective purifying selection.

No one is perfect, not even the reference

Bioinformatic approaches can suffer from reference bias (Simons *et al*. 2014; Kono *et al*. 2016). One particular challenge in the identification of deleterious variants is that with reference-based read mapping, variants are typically identified as differences from reference, then passed through a series of filters to identify putatively deleterious changes. For example, most annotation approaches single out nonsynonymous differences from references. Because a reference genome, particularly when based on an inbred, has no differences from itself, the reference has no nonsynonymous variants to annotate and thus appears free of deleterious variants. A more concerning type of reference bias is that individuals that are genetically more similar to the reference genome will

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

have fewer differences from reference and thus fewer variants that annotate as deleterious. This pattern is observed by Mezmouk and Ross-Ibarra (2014), who find fewer deleterious variants in the stiff-stalk population of maize (to which the reference genome B73 belongs) than in other elite maize populations. Along similar lines, because gene models are derived from the reference genome, more closely related lines with more similar coding portions of genes will appear to have fewer disruptions of coding sequence. Finally reference bias can contribute to under-calling of deleterious variants, either when different alleles fail to properly align to the reference or if the reference is included in a phylogenetic comparison. Variants that are detected as a difference from reference may then be compared against the reference in an alignment testing for conservation at a site. This conflates diversity within a species with the phylogenetic divergence that is being tested in the alignment. In cases where the reference genome contains the novel (or derived) variant relative to the state in related species, the presence of the reference variant in the alignment will cause the site to appear less constrained. This last problem is easily resolved by leaving the species being tested out of the phylogenetic alignment used to annotate deleterious variants.

Effect size

The evidence we present supports the cost of domestication hypothesis, namely that domesticated lineages carry heavier genetic loads than their wild relatives (Table 1). However, the distribution of fitness effects of variants is important with the regard to the total load within an individual or population (Henn *et al*. 2015). In other words, it is the cumulative effect of the variants carried by an individual that make up its genetic load, not simply what proportion of those variants is 'deleterious'. As we describe above, current bioinformatic approaches that rely on phylogenetic conservation may identify a number of false positives (driven by new fitness optima under artificial selection) or false negatives (with specific epistatic, dominance, or environmentally-dependent effects). Functional and quantitative genetics approaches provide means of assessing the phenotypic

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

691 effect of genetic variants, but there are practical limitations that limit the quantity
692 of evaluations that can be conducted.
693
694 One potential issue is that the phenotypic effect of a genetic variant often
695 depends on genomic background, through both dominance (interaction between
696 alleles at the same locus) and epistasis (interaction among alleles at different
697 loci). Evaluating the effect of these variants is consequently a complex task,
698 requiring the creation and evaluation of multiple classes of recombinant
699 genomes. Similarly, the environment is an important consideration in thinking
700 about effect size. As we saw with the yield under drought *qDTY1.1* allele in rice
701 (*Identifying dangerous hitchhikers*, above; Vikram *et al*. 2015), the effect of a
702 genetic variant can depend on developmental or assessment environment.
703 These kinds of variants might be expected to be involved in local adaptation in
704 wild populations (e.g., Monroe *et al*. 2016), and so would not show up in a screen
705 based on phylogenetic conservation. Nevertheless, these variants may be
706 important in breeding programs that target general-purpose genotypes with, for
707 example, high mean and low variance yields. Identifying these alleles requires
708 assessment in the appropriate environment(s) and large assessment populations
709 (as the power for detecting genotype-phenotype correlations generally scales
710 with the number of genotypes assessed). Addressing effect size in the
711 appropriate context(s) therefore involves challenges of scale. Fortunately, new
712 types of mapping populations (e.g., MAGIC) and high-throughput phenotyping
713 platforms that can enable this work are increasingly available across
714 domesticated systems. Given these tools and the resultant data, we should soon
715 be able to parameterize genomic selection and similar models with putatively
716 deleterious variants and test their cumulative effects.
717
718 *Heterosis*
719 In systems with hybrid production, complementation of deleterious variants
720 between heterotic breeding pools may contribute substantially to heterosis (e.g.,
721 in maize; Hufford *et al*. 2012; Mezmouk and Ross-Ibarra 2014; Yang *et al*.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv). If this is broadly true, hybrid production may be an interesting solution to the cost of domestication, as long as deleterious variants are segregating rather than fixed in domesticated species. Theory suggests that the deleterious variants that contribute to heterosis between populations with low levels of gene flow are likely to be of intermediate effect, and may not play a large role in inbreeding depression (Whitlock, Ingvarsson, and Hatfield 2000). Since fitness in these contexts is evaluated in hybrid individuals rather than inbred parents, parental populations are likely to retain a higher proportion of slightly and moderately deleterious variants than even selfing populations (Figure 2A). As long as hybrid crosses are the primary mode of seed production, these alleles may not be of high importance to breeders even though they contribute to genetic load more broadly. Evaluating effect size in this context requires assessing both parental and hybrid populations, and is therefore that much more difficult. However, assuming that complementation of deleterious variants is a substantial component of heterosis, quantifying these variants should improve the ability of breeding programs to predict trait values in hybrid crosses. Further, the question of whether genetic load limits genetic gains in hybrid production systems remains open.

CONCLUSIONS

Research on the putative costs of domestication is still relatively new, and there remain many open questions. However, our review of the literature suggests that genetic loads are generally heavier in domesticated species compared to their wild relatives. This pattern is likely driven by a number of processes that collectively act to reduce the efficacy of selection relative to drift in domesticated populations, resulting in increased frequency of deleterious variants linked to selected loci and greater accumulation of deleterious variants genome-wide. We encourage further research across domesticated species on these processes, and recommend that researchers: (1) Sample domesticated and wild lineages sufficiently to assess diversity within as well as between these groups and (2) present deleterious variant data per genome and as proportional as well as

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

753 absolute values. We also strongly encourage researchers and breeders to think
754 about deleterious variation in context, both genomic and environmental.
755
756 FUNDING
757 This work was supported by the US National Science Foundation (award
758 numbers 1523752 to B.T.M.; and 1339393 to P.L.M.).
759
760 ACKNOWLEDGEMENTS
761 We thank Greg Baute and Kathryn Turner for comments on an earlier version of
762 this manuscript, and Brandon Gaut, Loren Rieseberg, and Greg Owens for
763 helpful discussion.
764
765 REFERENCES
766
767 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al.*
768 (2010). A method and server for predicting damaging missense mutations.
769 *Nat Methods* **7**: 248–249.
770
771 Agrawal AF, Whitlock MC (2012). Mutation load: the fitness of individuals in
772 populations where deleterious alleles are abundant. *Annu Rev Ecol Evol Syst*
773 **43**: 115–135.
774
775 Allard RW (1999). History of plant population genetics. *Annu Rev Genet* **33**: 1–
776 27.
777
778 Amaral AJ, Megens HJ, Crooijmans RPMA, Heuven HCM, Groenen MAM
779 (2008). Linkage disequilibrium decay and haplotype block structure in the pig.
780 *Genetics* **179**: 569–579.
781
782 Arunkumar R, Ness RW, Wright SI, Barrett SCH (2015). The evolution of selfing
783 is accompanied by reduced efficacy of selection and purging of deleterious
784 mutations. *Genetics* **199**: 817–829.
785
786 Asano K, Yamasaki M, Takuno S, Miura K, Katagiri S, Ito T, *et al.* (2011).
787 Artificial selection for a green revolution gene during japonica rice
788 domestication. *PNAS* **108**: 11034–11039.
789
790 Bachlava E, Tang S, Pizarro G, Schuppert GF, Brunick RK, Draeger D, *et al.*
791 (2009). Pleiotropy of the branching locus (B) masks linked and unlinked
792 quantitative trait loci affecting seed traits in sunflower. *Theor Appl Genet* **120**:
793 829–842.
794
795 Baer CF, Miyamoto MM, Denver DR (2007). Mutation rate variation in
796 multicellular eukaryotes: causes and consequences. *Nat Rev Genet* **8**: 619–
797 631.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Benbrook C (1999). *Evidence of the magnitude and consequences of the Roundup Ready soybean yield drag from university-based varietal trials in 1998*. Ag BioTech InfoNet.

Bosse M, Megens H-J, Madsen O, Crooijmans RPMA, Ryder OA, Austerlitz F, *et al.* (2015). Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research* **25**: 970–981.

Bosse M, Megens H-J, Madsen O, Paudel Y, Frantz LAF, Schook LB, *et al.* (2012). Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS Genet* **8**: e1003100.

Brandenburg J-T, Mary-Huard T, Rigaill G, Hearne SJ, Corti H, Joets J, *et al.* (2017). Independent introductions and admixtures have contributed to adaptation of European maize and its American counterparts. *PLoS Genet* **13**: e1006666.

Charlesworth B (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**: 195–205.

Charlesworth D (2003). Effects of inbreeding on the genetic diversity of populations. *Phil Trans Roy Soc B* **358**: 1051–1070.

Chen E, Huang X, Han B (2016). How can rice genetics benefit from rice-domestication study? *National Science Review* **3**: 278–280.

Choi JY, Platts AE, Fuller DQ, Hsing Y-I, Wing RA, Purugganan MD (2017). The rice paradox: Multiple origins but single domestication in Asian rice. *Mol Biol Evol*: msx049.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **7**: e46688.

Chun S, Fay JC (2009). Identification of deleterious mutations within three human genomes. *Genome Research* **19**: 1553–1561.

Clark RM, Linton E, Messing J, Doebley JF (2004). Pattern of diversity in the genomic region near the maize domestication gene *tb1*. *PNAS* **101**: 700–707.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The International Barley Genome Sequencing Consortium (2013). A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**: 711–716.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**: 901–913.

Corbi J, Debieu M, Rousselet A, Montalent P, Le Guilloux M, Manicacci D, *et al.* (2010). Contrasted patterns of selection since maize domestication on duplicated genes encoding a starch pathway enzyme. *Theor Appl Genet* **122**: 705–722.

Cruz F, Vila C, Webster MT (2008). The legacy of domestication: accumulation of deleterious mutations in the dog genome. *Mol Biol Evol* **25**: 2331–2336.

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

Elmore RW, Roeth FW, Nelson LA (2001). Glyphosate-resistant soybean cultivar yields compared with sister lines. *Agron J* **93**: 408–412.

Eyre-Walker A, Keightley PD (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**: 610–618.

Eyre-Walker A, Woolfit M, Phelps T (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**: 891–900.

Fay JC, Wyckoff GJ, Wu C-I (2001). Positive and Negative Selection on the Human Genome. *Genetics* **158**: 1227–1234.

Felsenstein J (1974). The evolutionary advantage of recombination. *Genetics* **78**: 737–756.

Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD (2012). Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Mol Biol Evol* **29**: 675–687.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, *et al.* (2002). The structure of haplotype blocks in the human genome. *Science* **296**: 2225–2229.

Gaut BS, Díez CM, Morrell PL (2015). Genomics and the contrasting dynamics of annual and perennial domestication. *TIG* **31**: 709–719.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Georgelis N, Shaw JR, Hannah LC (2009). Phylogenetic analysis of ADP-glucose pyrophosphorylase subunits reveals a role of subunit interfaces in the allosteric properties of the enzyme. *Plant Physiol* **151**: 67–77.

Gheyas AA, Boschiero C, Eory L, Ralph H, Kuo R, Woolliams JA, *et al.* (2015). Functional classification of 15 million SNPs detected from diverse chicken populations. *DNA Research* **22**: 205–217.

Gillespie JH (1994). Substitution processes in molecular evolution. II. Exchangeable models from population genetics. *Genetics* **138**: 943–952.

Gillespie JH (1999). The role of population size in molecular evolution. *Theor Pop Biol* **55**: 145–156.

Giorgi D, Pandozy G, Farina A, Grosso V, Lucretti S, Gennaro A, *et al.* (2016). First detailed karyo-morphological analysis and molecular cytological study of leafy cardoon and globe artichoke, two multi-use Asteraceae crops. *CCG* **10**: 447–463.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, *et al.* (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–D1186.

Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, *et al.* (2009). A First-Generation Haplotype Map of Maize. *Science* **326**: 1115–1117.

Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, *et al.* (2009). Linkage Disequilibrium and Demographic History of Wild and Domestic Canids. *Genetics* **181**: 1493–1505.

Gregory TR (2017). Animal Genome Size Database. http://www.genomesize.com.

Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, *et al.* (2015). The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation* **36**: 513–523.

Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RPMA, *et al.* (2008). A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research* **19**: 510–519.

Gross BL, Olsen KM (2010). Genetic perspectives on crop domestication. *Trends Plant Sci* **15**: 529–537.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

936 Gross BL, Zhao Z (2014). Archaeological and genetic insights into the origins of
937 domesticated rice. *PNAS* **111**: 6190–6197.
938
939 Günther T, Schmid KJ (2010). Deleterious amino acid polymorphisms in
940 *Arabidopsis thaliana* and rice. *Theor Appl Genet* **121**: 157–168.
941
942 Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S (2015). Estimating
943 the mutation load in human genomes. *Nat Rev Genet* **16**: 333–343.
944
945 Hill WG, Robertson A (1966). The effect of linkage on limits to artificial selection.
946 *Genetics Research* **8**: 269–294.
947
948 Horak MJ, Rosenbaum EW, Kendrick DL, Sammons B, Phillips SL, Nickson TE,
949 *et al.* (2015). Plant characterization of Roundup Ready 2 Yield® soybean,
950 MON 89788, for use in ecological risk assessment. *Transgenic Res* **24**: 213–
951 225.
952
953 Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, *et al.* (2012). A map of
954 rice genome variation reveals the origin of cultivated rice. *Nature* **490**: 497–
955 501.
956
957 Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, *et*
958 *al.* (2012). Comparative population genomics of maize domestication and
959 improvement. *Nat Genet* **44**: 808–811.
960
961 Hughes AL (2005). Evidence for Abundant Slightly Deleterious Polymorphisms in
962 Bacterial Populations. *Genetics* **169**: 533–538.
963
964 Keightley PD, Eyre-Walker A (2007). Joint Inference of the Distribution of Fitness
965 Effects of Deleterious Mutations and Population Demography Based on
966 Nucleotide Polymorphism Frequencies. *Genetics* **177**: 2251–2261.
967
968 Keightley PD, Lynch M (2003). Toward a realistic model of mutations affecting
969 fitness. *Evolution* **57**: 683–685.
970
971 Kersey PJ, Allen JE, Armean I, Boddu S, Bolt BJ, Carvalho-Silva D, *et al.* (2016).
972 Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids*
973 *Res* **44**: D574–D580.
974
975 Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, *et al.* (2007).
976 Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat Genet*
977 **39**: 1151–1155.
978
979 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014). A
980 general framework for estimating the relative pathogenicity of human genetic
981 variants. *Nat Genet* **46**: 310–315.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Klopfstein S, Currat M, Excoffier L (2005). The Fate of Mutations Surfing on the Wave of a Range Expansion. *Mol Biol Evol* **23**: 482–490.

Koenig D, Jiménez-Gómez JM, Kimura S, Fulop D, Chitwood DH, Headland LR, *et al.* (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *PNAS* **110**: E2655–E2662.

Kono TJY, Fu F, Mohammadi M, Hoffman PJ, Liu C, Stupar RM, *et al.* (2016). The Role of Deleterious Substitutions in Crop Genomes. *Mol Biol Evol* **33**: 2307–2317.

Kono TJY, Lei L, Shih C-H, Hoffman PJ, Morrell PL, Fay JC Comparative genomics approaches accurately predict deleterious variants in plants. *bioRxiv*.

Kovach MJ, Sweeney MT, McCouch SR (2007). New insights into the history of rice domestication. *TIG* **23**: 578–587.

Kryazhimskiy S, Plotkin JB (2008). The Population Genetics of dN/dS. *PLoS Genet* **4**: e1000304.

Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, *et al.* (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* **42**: 1053–1059.

Lau AN, Peng L, Goto H, Chemnick L, Ryder OA, Makova KD (2009). Horse Domestication and Conservation Genetics of Przewalski's Horse Inferred from Sex Chromosomal and Autosomal Sequences. *Mol Biol Evol* **26**: 199–208.

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, *et al.* (2014). Genomic analyses provide insights into the history of tomato breeding. *Nat Genet* **46**: 1220–1226.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, *et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**: 803–819.

Lippold S, Knapp M, Kuznetsova T, Leonard JA, Benecke N, Ludwig A, *et al.* (2011). Discovery of lost diversity of paternal horse lineages using ancient DNA. *Nat Commun* **2**: 450–6.

Liu A, Burke JM (2006). Patterns of Nucleotide Diversity in Wild and Cultivated Sunflower. *Genetics* **173**: 321–330.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1028 Liu Q, Zhou Y, Morrell PL, Gaut BS (2017). Deleterious variants in Asian rice and
1029 the potential cost of domestication. *Mol Biol Evol*: msw296.
1030
1031 Lohmueller KE (2014). The Impact of Population Demography and Selection on
1032 the Genetic Architecture of Complex Traits. *PLoS Genet* **10**: e1004379.
1033
1034 Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I (2006). The accumulation of
1035 deleterious mutations in rice genomes: a hypothesis on the cost of
1036 domestication. *TIG* **22**: 126–131.
1037
1038 Mandel JR, Dechaine JM, Marek LF, Burke JM (2011). Genetic diversity and
1039 population structure in cultivated sunflower and a comparison to its wild
1040 progenitor, *Helianthus annuus* L. *Theor Appl Genet* **123**: 693–704.
1041
1042 Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C,
1043 *et al.* (2016). Bottlenecks and selective sweeps during domestication have
1044 increased deleterious genetic variation in dogs. *PNAS* **113**: 152–157.
1045
1046 Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, *et al.* (2011). The
1047 functional spectrum of low-frequency coding variation. *Genome Biology* **12**:
1048 R84.
1049
1050 Maynard Smith J, Haigh J (1974). The hitch-hiking effect of a favourable gene.
1051 *Genetical research* **23**: 1–13.
1052
1053 Megens H-J, Crooijmans RP, Bastiaansen JW, Kerstens HH, Coster A, Jalving
1054 R, *et al.* (2009). Comparison of linkage disequilibrium and haplotype diversity
1055 on macro- and microchromosomes in chicken. *BMC Genet* **10**: 86.
1056
1057 Mezmouk S, Ross-Ibarra J (2014). The Pattern and Distribution of Deleterious
1058 Mutations in Maize. *G3* **4**: 163–171.
1059
1060 Michaelson MJ, Price HJ, Ellison JR, Johnston JS (1991). Comparison of plant
1061 DNA contents determined by Feulgen microspectrophotometry and laser flow
1062 cytometry. *Am J Bot* **78**: 183–188.
1063
1064 Miller AJ, Gross BL (2011). From forest to field: Perennial fruit crop
1065 domestication. *Am J Bot* **98**: 1389–1414.
1066
1067 Miyata T, Miyazawa S, Yasunaga T (1979). Two types of amino acid
1068 substitutions in protein evolution. *J Molec Evol* **12**: 219–236.
1069
1070 Monroe JG, McGovern C, Lasky JR, Grogan K, Beck J, McKay JK (2016).
1071 Adaptation to warmer climates by parallel functional evolution of *CBF* genes
1072 in *Arabidopsis thaliana*. *Mol Ecol* **25**: 3632–3644.
1073

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Moray C, Lanfear R, Bromham L (2014). Domestication and the mitochondrial genome: comparing patterns and rates of molecular evolution in domesticated mammals and birds and their wild relatives. *Genome Biol Evol* **6**: 161–169.

Morrell PL, Buckler ES, Ross-Ibarra J (2011). Crop genomics: advances and applications. *Nat Rev Genet* **13**: 85–96.

Muir WM, Wong G, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, *et al.* (2008). Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *PNAS* **105**: 17312–17317.

Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, *et al.* (2014). Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*). *Mol Ecol* **23**: 2210–2227.

Ng PC, Henikoff S (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**: 3812–3814.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, *et al.* (2005). The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biol* **3**: e196.

Ohta T (1972). Population size and rate of evolution. *J Molec Evol* **1**: 305–314.

Ohta T (1992). The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst* **23**.

Olsen KM (2006). Selection Under Domestication: Evidence for a Sweep in the Rice *Waxy* Genomic Region. *Genetics* **173**: 975–983.

Otto SP, Barton NH (2001). Selection for recombination in small populations. *Evolution* **55**: 1921–1931.

Otto SP, Whitlock MC (1997). The probability of fixation in populations of changing size. *Genetics* **146**: 723–733.

Papa R, Bellucci E, Rossi M, Leonardi S, Rau D, Gepts P, *et al.* (2007). Tagging the Signatures of Domestication in Common Bean (*Phaseolus vulgaris*) by Means of Pooled DNA Samples. *Ann Bot–London* **100**: 1039–1051.

Parker HG, vonHoldt BM, Quignon P, Margulies EH, Shao S, Mosher DS, *et al.* (2009). An expressed *Fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. *Science* **325**: 995–998.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Peischl S, Dupanloup I, Kirkpatrick M, Excoffier L (2013). On the accumulation of deleterious mutations during range expansions. *Mol Ecol* **22**: 5972–5982.

Renaut S, Rieseberg LH (2015). The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops. *Mol Biol Evol* **32**: 2273–2283.

Rice A, Glick L, Abadi S, Einhorn M (2015). The Chromosome Counts Database (CCDB)–a community resource of plant chromosome numbers. *New Phytol* **206**: 19–26.

Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, Mitchell SE, *et al.* (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. *PNAS* **112**: 3823–3828.

Ross-Ibarra J (2004). The evolution of recombination under domestication: a test of two hypotheses. *Am Nat* **163**: 105–112.

Ross-Ibarra J, Tenaillon M, Gaut BS (2009). Historical Divergence and Gene Flow in the Genus *Zea*. *Genetics* **181**: 1399–1413.

Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, *et al.* (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587–591.

Scaglione D, Reyes-Chin-Wo S, Acquadro A, Froenicke L, Portis E, Beitel C, *et al.* (2015). The genome sequence of the outbreeding globe artichoke constructed. *Sci Rep*: 1–17.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, *et al.* (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169–175.

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, *et al.* (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet* **46**: 707–713.

Schneider CA, Rasband WS, Eliceiri KW (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**: 671–675.

Schubert M, Jónsson H, Chang D, Sarkissian Der C, Ermini L, Ginolhac A, *et al.* (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *PNAS* **111**: E5661–E5669.

Semon M, Nielsen R, Jones MP, McCouch SR (2005). The Population Structure of African Cultivated Rice *Oryza glaberrima* (Steud.): Evidence for Elevated

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Levels of Linkage Disequilibrium Caused by Admixture with *O. sativa* and Ecological Adaptation. *Genetics* **169**: 1639–1647.

Shi T, Dimitrov I, Zhang Y, Tax FE, Yi J, Gou X, *et al.* (2015). Accelerated rates of protein evolution in barley grain and pistil biased genes might be legacy of domestication. *Plant Molecular Biology* **89**: 253–261.

Simons YB, Turchin MC, Pritchard JK, Sella G (2014). The deleterious mutation load is insensitive to recent population history. *Nat Genet* **46**: 220–224.

Stone EA, Sidow A (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research* **15**: 978–986.

Takeda S, Matsuoka M (2008). Genetic approaches to crop improvement: responding to environmental and population changes. *Nat Rev Genet* **9**: 444–457.

Tian F, Stevens NM, Buckler ES (2009). Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *PNAS* **106**: 9979–9986.

Traspov A, Deng W, Kostyunina O, Ji J, Shatokhin K, Lugovoy S, *et al.* (2016). Population structure and genome characterization of local pig breeds in Russia, Belorussia, Kazakhstan and Ukraine. *Genetics Selection Evolution* **48**: 1–9.

Travis JMJ, Munkemuller T, Burton OJ, Best A, Dytham C, Johst K (2007). Deleterious Mutations Can Surf to High Densities on the Wave Front of an Expanding Population. *Mol Biol Evol* **24**: 2334–2343.

Vikram P, Swamy BPM, Dixit S, Singh R, Singh BP, Miro B, *et al.* (2015). Drought susceptibility of modern rice varieties: an effect of linkage of drought tolerance with undesirable traits. *Sci Rep* **5**.

Wade CM, Giulotto E, Sigurdsson S, Zoli M (2009). Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. *Science* **326**: 865–867.

Walsh B (2007). Using molecular markers for detecting domestication, improvement, and adaptation genes. *Euphytica* **161**: 1–17.

Wang G-D, Xie H-B, Peng M-S, Irwin D, Zhang Y-P (2014). Domestication Genomics: Evidence from Animals. *Annu Rev Anim Biosci* **2**: 65–84.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Wang H, Studer AJ, Zhao Q, Meeley R (2015). Evidence that the origin of naked kernels during maize domestication was caused by a single amino acid substitution in *tga1*. *Genetics* **200**: 965–974.

Whitlock MC (2000). Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution* **54**: 1855–1861.

Whitlock MC, Ingvarsson PK, Hatfield T (2000). Local drift load and the heterosis of interconnected populations. *Heredity* **84**: 452–457.

Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES (2002). Genetic diversity and selection in the maize starch pathway. *PNAS* **99**: 12959–12962.

Wilfert L, Gadau J, Schmid-Hempel P (2007). Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98**: 189–197.

Wolf JBW, Kunstner A, Nam K, Jakobsson M, Ellegren H (2009). Nonlinear Dynamics of Nonsynonymous (dN) and Synonymous (dS) Substitution Rates Affects Inference of Selection. *Genome Biol Evol* **1**: 308–319.

Woolfit M (2009). Effective population size and the rate and pattern of nucleotide substitutions. *Biology Letters* **5**: 417–420.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, *et al.* (2005). The Effects of Artificial Selection on the Maize Genome. *Science* **308**: 1310–1314.

Xu Z, Hennessy DA, Sardana K, Moschini G (2013). The Realized Yield Effect of Genetically Engineered Crops: U.S. Maize and Soybean. *Crop Sci* **53**: 735–745.

Yamasaki M, Wright SI, McMullen MD (2007). Genomic Screening for Artificial Selection during Domestication and Improvement in Maize. *Ann Bot–London* **100**: 967–973.

Yang J, Mezmouk S, Baumgarten A, Buckler ES, Guill KE, McMullen MD, *et al.* Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *bioRxiv*.

Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, *et al.* (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* **33**: 408–414.

bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Zonneveld BJM, Leitch IJ, Bennett MD (2005). First Nuclear DNA Amounts in more than 300 Angiosperms. *Ann Bot–London* **96**: 229–244.

TABLES AND FIGURES

Table 1. Evidence for genetic load across domesticated plants and animals. *Arabidopsis thaliana* and *Homo sapiens* included for comparison. Plant 1C genome sizes from the RBG Kew database (http://data.kew.org/cvalues/), except tomato (Michaelson *et al*. 1991) and *Cynara* spp. (Giorgi *et al*. 2016). Plant chromosome counts from the Chromosome Counts Database (http://ccdb.tau.ac.il/). Animal 1C genome sizes and chromosome counts from the Animal Genome Size Database (http://www.genomesize.com/, mean value when multiple records were available). Gene numbers are high-confidence (if available) estimates from the vertebrate and plant Ensembl databases (http://uswest.ensembl.org/; http://plants.ensembl.org/), except common bean (Schmutz et al. 2013), sunflower (Compositate Genome Project, unpublished), and *Cynara* spp. (Scaglione et al. 2016). LD N50 is the approximate distance over which LD decays to half of maximum value. Loss of genetic diversity is calculated as 1 – (ratio of $p_{domesticated}$ to p_{wild}), or q if p not available.

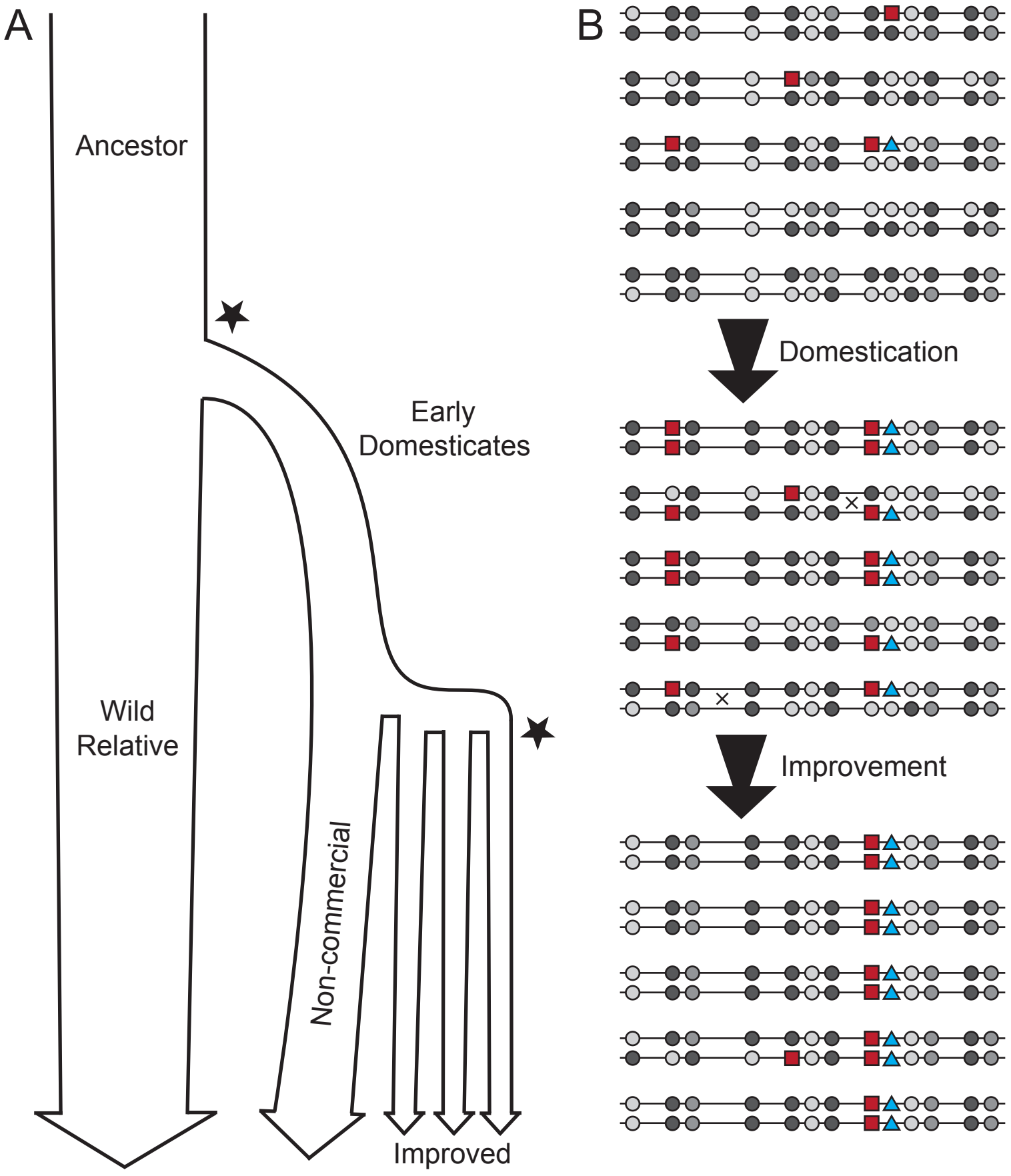
Figure 1. Processes of domestication and improvement. (A) Typical changes in effective population size through domestication and improvement. Stars indicate genetic bottlenecks. These dynamics can be reconstructed by examining patterns of genetic diversity in contemporary wild relative, domesticated non-commercial, and improved populations. (B) Effects of artificial selection (targeting the blue triangle variant) and linkage disequilibrium on deleterious (red squares) and neutral variants (grey circles, shades represent different alleles). In the ancestral wild population, four deleterious alleles are at relatively low frequency (mean = 0.10) and heterozygosity is high (H_O = 0.51). After domestication, the selected blue triangle and linked variants increase in frequency (three remaining deleterious alleles, mean frequency = 0.46), heterozygosity decreases (H_O = 0.35), and allelic diversity is lost at two sites. Recombination may change haplotypes, especially at sites less closely linked to the selected allele (Xs). After improvement, further selection for the blue triangle allele has: lowered heterozygosity (H_O = 0.08), increased deleterious variant frequency (two remaining deleterious alleles, mean = 0.55), and lost allelic diversity at six additional sites.

Figure 2. Effects of inbreeding on mutations and fitness. (A) Theoretical density plot of fitness effects for segregating variants. In outcrossing, non-inbred populations (solid navy line), more variants with highly deleterious, recessive effects persist that are rapidly exposed to selection and purged in inbred populations (dotted red line), shifting the left side of the distribution towards more neutral effects. At the same time, the reduced effective population size created through inbreeding causes on average higher loss of slightly advantageous mutations and retention of slightly deleterious mutations, shifting the right side of

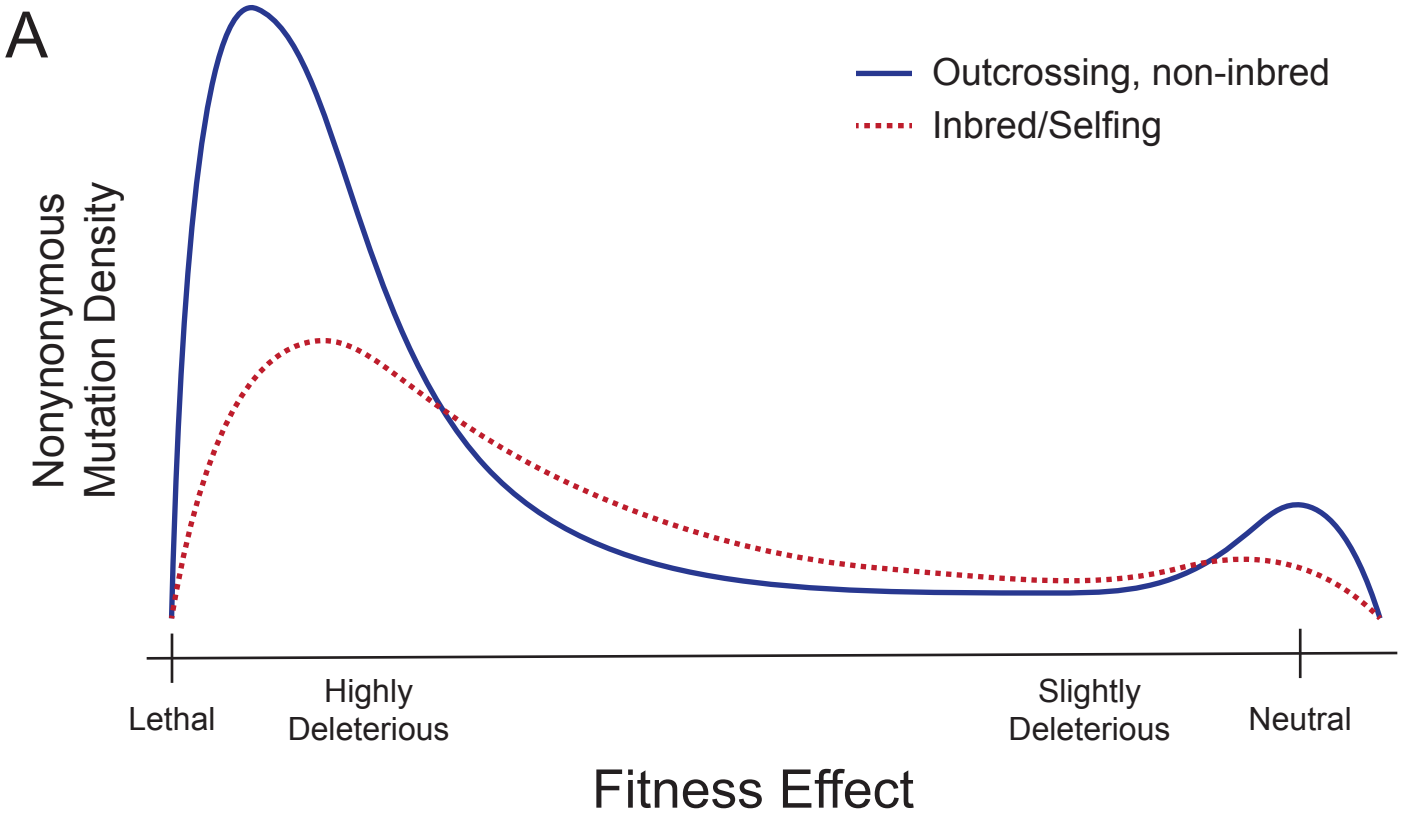
bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

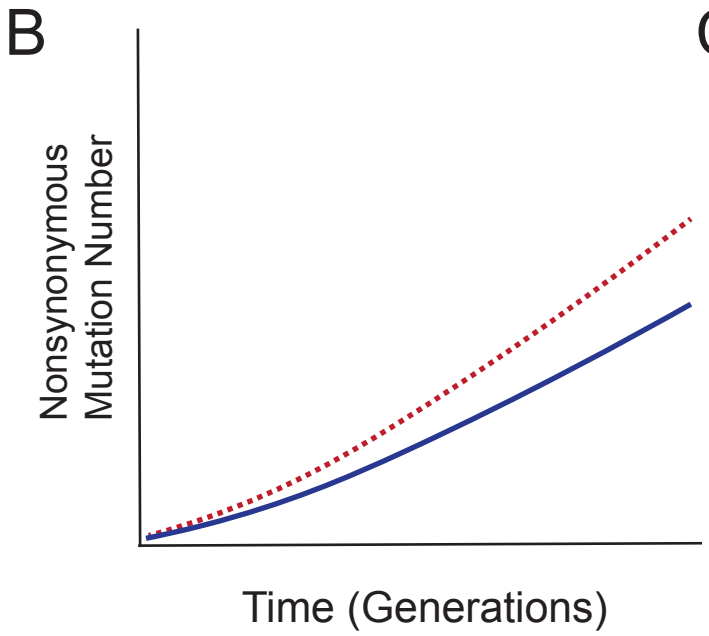
1302 the distribution towards more deleterious effects. (B) Accumulation of
1303 nonsynonymous mutations is accelerated in selfing individuals (dotted red line)
1304 relative to outcrossing individuals (solid navy line). Synonymous mutations
1305 accumulate at similar rates in both populations. Adapted from simulation results
1306 in Arunkumar *et al*. 2015. (C) As a consequence of (B), mean individual fitness
1307 drops in selfing relative to outcrossing populations. Adapted from Arunkumar *et*
1308 *al*. 2015.

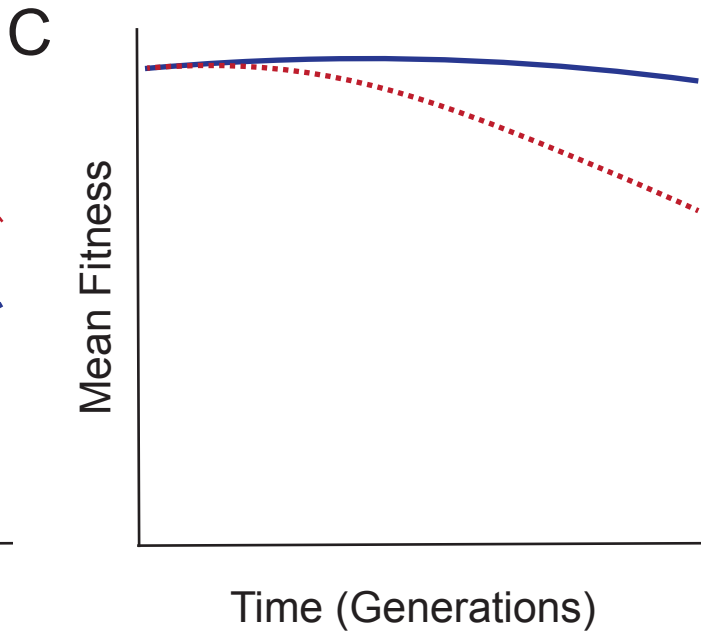
bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/122093; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.







Species	Primary reproduction	Genome size (1C, pg)	Chromosome number (N)	Protein coding genes	Non-coding genes	LD N50	LD N50 in wild	Loss of genetic diversity	dN/dS domesticated to dN/dS wild (or Ka/Ks)*	GERP score load increase relative to wild	Genome-wide deleterious variant number	Deleterious variant proportion (of nonsynonymous variants)	Deleterious variants in wild, # (proportion)	Method(s) for deleterious variant detection	Studies
japonica rice (Oryza sativa var japonica)	selfing	0.50	12	35,679	55,401	167 kb	20 kb	0.67	1.70	NA	~6295*	0.19*	6099 (0.15)*	SIFT, PROVEAN	Lu et al. 2006; Huang et al. 2012; Liu et al. 2017
indica rice (Oryza sativa var indica)	selfing	0.50	12	40,745	45,577	123 kb	20 kb	0.25	1.75	NA	~6351*	0.18*	6099 (0.15)*	SIFT, PROVEAN	Lu et al. 2006; Huang et al. 2012; Liu et al. 2017
African rice (Oryza glaberrima)	selfing	0.83	12	33,164	38,882	~25 Mb*	NA	0.60	1.58	NA	NA	NA	NA		Semon et al. 2005; Nabholz et al. 2014
maize (Zea mays)	outcrossing	2.73	10	39,498	4,976	<1 kb	NA	0.17	NA	NA	~1025–4056*	0.04–0.16*	NA	SIFT, MAPP, Join of these	Gore et al. 2009; Hufford et al. 2012; Mezmouk and Ross-Ibarra 2014
barley (Hordeum vulgare)	selfing	5.55	7	24,287	1,512	~2.2 Mb	~110 kb	0.13 (domestication) and 0.27 (improvement)	2.43–4.60	NA	1,006–3,400	0.06–0.19	NA	SIFT, PolyPhen2, LRT, Intersect of these	The International Barley Genome Sequencing Consortium 2012; Morrell et al. 2014; Shi et al. 2015; Kono et al. 2016
soybean (Glycine max)	selfing	1.13	20	54,174	NA	83 kb (domesticated) and 123 kb (improved)	27 kb	0.52 (domestication) and 0.64 (improvement)	1.01**	NA	784–3,881	0.03–0.13	NA	SIFT, PolyPhen2, LRT, Intersect of these	Lam et al. 2010; Zhou et al. 2015; Kono et al. 2016
common bean (Phaseolus vulgaris, Mesoamerican domestication)	selfing	0.60	11	27,197	NA	NA	NA	0.20	NA	NA	NA	NA	NA		Schmutz et al. 2013
common bean (Phaseolus vulgaris, Andean domestication)	selfing	0.60	11	27,197	NA	NA	NA	-0.50	NA	NA	NA	NA	NA		Schmutz et al. 2013
tomato (Solanum lycopersicum)	selfing	1.00	12	33,837	4,849	257 kb (domesticated) and 866 kb (improved)	9 kb	0.46 (domestication) and 0.77 (improvement)	1.37***	NA	NA	NA	NA		Koenig et al. 2013; Lin et al. 2014
sunflower (Helianthus annuus)	outcrossing	2.43	17	52,232	NA	~1 kb	200 bp	0.33	NA	NA	5,626–20,273*	0.14–0.17*	7,882–24,479 (0.12–0.14)*	PROVEAN	Liu and Burke 2006; Mandel et al. 2011; Renaut and Rieseberg 2015; L. Rieseberg unpublished
globe artichoke (Cynara cardunculus var. scolymus)	clonal propogation, outcrossing	1.20	17	26,889	NA	NA	NA	NA	NA	NA	1,537–1,889*	0.19–0.20*	1,465 (0.17)*	PROVEAN	Renaut and Rieseberg 2015
cardoon (Cynara cardunculus var. altilis)	outcrossing	1.10	17	26,889	NA	NA	NA	NA	NA	NA	1,239–1,458*	0.22–0.23*	1,900 (0.19)*	PROVEAN	Renaut and Rieseberg 2015
thale cress (Arabidopsis thaliana)	*selfing*	*0.30*	*5*	*27,655*	*7,398*	*~3.4 kb*	*NA*	*NA*	*NA*	*NA*	*8,831–9,584*	*0.19–0.21*	*NA*	*SIFT, MAPP*	*Kim et al. 2007; Günther and Schmid 2010*
chicken (Gallus gallus domesticus)	outcrossing	1.25	39	18,346	6,492	~50–250 kb	< 50 kb	~0.50 (from domesticated to 'commercial' breeds*)	NA	NA	35,889–88,655	0.20–0.49	NA	SIFT, PROVEAN, Join and Intersect of these	Muir et al. 2008; Megens et al. 2009; Gheyas et al. 2015
dog (Canis familiaris)	outcrossing	3.12	39	19,856	11,898	~100 kb	< 10 kb	0.05 (domestication) / 0.35 (breed formation)	1.54	2.1%	~5,115*	~0.14	~4,990 (~0.138)*	Mivata distance; GERP	Lindblad-Tor *et al. 2005*; Cruz, Vila & Webster 2008; Gray et al. 2009; Marsden et al. 2016
horse (Equus caballus)	outcrossing	3.22	32	20,449	2,142	~200 kb	NA	0.40**	NA	1.1–11%	NA	NA	NA	GERP	Lau et al. 2009; Wade et al. 2009; Schubert et al. 2014
pig (Sus scrofa)	outcrossing	3.17	19	21,630	3,124	5 kb–1 Mb	< 50 kb	0.19	NA	NA	656*	NA	NA	SIFT	Amaral et al. 2008; Bosse et al. 2012; Bosse et al. 2015
human (Homo sapiens)	*outcrossing*	*3.50*	*23*	*20,441*	*22,219*	*22–44 kb*	*NA*	*NA*	*1.03*	*NA*	*796–838*	*0.08–0.15*	*NA*	*LRT*	*Gabriel et al. 2002; Chun and Fay 2009; Scally et al. 2012*
						* likely due to population structure in sampled accessions (Semon et al. 2005)		*estimated as portion of hypothetical ancestral allele frequency distribution lost	* or versus sister species for human		* per genome	* per genome	*significantly lower in wild		
								** autosomal nucleotide diversity only	** data are ratio of nonsynonymous to synonymous SNP counts rather than substitutions						
									*** comparison of domesticated lineage versus whole tree rates (five wild lineages and one domesticated)						