REVIEW ARTICLE
https://doi.org/10.1038/s41477-018-0210-1

nature plants

Demography and its effects on genomic variation in crop domestication

Brandon S. Gaut[1], Danelle K. Seymour[1], Qingpo Liu[2] and Yongfeng Zhou[1*]

Over two thousand plant species have been modified morphologically through cultivation and human use. Here, we review three aspects of crop domestication that are currently undergoing marked revisions, due to analytical advancements and their application to whole genome resequencing (WGS) data. We begin by discussing the duration and demographic history of domestication. There has been debate as to whether domestication occurred quickly or slowly. The latter is tentatively supported both by fossil data and application of WGS data to sequentially Markovian coalescent methods that infer the history of effective population size. This history suggests the possibility of extended human impacts on domesticated lineages prior to their purposeful cultivation. We also make the point that demographic history matters, because it shapes patterns and levels of extant genetic diversity. We illustrate this point by discussing the evolutionary processes that contribute to the empirical observation that most crops examined to date have more putatively deleterious alleles than their wild relatives. These deleterious alleles may contribute to genetic load within crops and may be fitting targets for crop improvement. Finally, the same demographic factors are likely to shape the spectrum of structural variants (SVs) within crops. SVs are known to underlie many of the phenotypic changes associated with domestication and crop improvement, but we currently lack sufficient knowledge about the mechanisms that create SVs, their rates of origin, their population frequencies and their phenotypic effects.

Over two thousand plant species have been domesticated; that is, they have been modified morphologically and genetically through human cultivation[1]. The modifications associated with domestication encompass a suite of morphological changes such as enhanced robustness, production of fewer but larger fruits (or grains), and altered seed dormancy and dispersal[2]. Morphological change does not stop after domestication however, because an incipient crop is often dispersed to new locations that present novel biotic and abiotic stresses. These stresses drive local adaptation through additional phenotypic and genetic change.

Geneticists, evolutionary biologists, crop scientists and archaeologists have long been fascinated by the origin and spread of domesticated crops, in part due to the multifaceted nature of the associated questions[3]: Who domesticated a species and how? Did domestication occur once or on multiple occasions? Was domestication rapid or protracted? The answers to these questions enlighten our understanding of human history and also have the potential to affect crop productivity.

In this Review, we focus on three aspects of crop domestication that are currently undergoing marked revisions, due principally to analytical advancements and their application to whole genome resequencing (WGS) data. We begin by considering the demographic history of domesticated populations. Although the demography of domesticates has been investigated for decades, new inferential techniques have the potential to alter our understanding of the time frame of human impacts on domesticated species. Timing matters, because demographic history affects levels of genomic variants in a crop. We illustrate some of these effects by discussing the distribution and accumulation of deleterious single nucleotide polymorphism (SNP) variants within crop genomes, which may contribute to a 'cost of domestication'[4]. Finally, we turn to an under-studied component of crop genomic variation: structural variants (SVs). In theory, demography should impact the diversity of SVs in a manner analogous to SNP variants, but there is currently little information about the rate, pattern and frequency of SVs within crop genomes. The characterization of SVs in populations is the next frontier of plant genomics, both because SVs affect phenotypes and because they will yield unique insights into population history.

Demography and the duration of crop domestication

In this section, we define domestication as a process that occurs in four stages. We then consider critical features of these stages, focusing particularly on their duration and potential demographic effects.

Stages of domestication. Domestication is an expansive process that consists of four stages[5] (Fig. 1a). Stage 1 is 'management', representing the potential for human stewardship and harvesting of wild plant populations prior to their formal cultivation. Stage 2 is purposeful cultivation, a process that winnows genetic diversity through a domestication bottleneck and simultaneously increases the frequency of domestication alleles. Stage 3 is the geographic expansion of the domesticate, which often requires adaptation to new environments. Stage 4 is deliberate breeding, a process that has occurred over the last few hundred years[5]. Not all domesticated plants are equal with respect to the duration or severity of these stages. For example, most cereal crops have experienced strong in situ selection and pronounced genetic bottlenecks (stage 2), but many perennial crops have not[6].

The time to fix favourable variants. To date, most of the domestication literature has focused on the second stage of domestication—that is, the intentional cultivation of crops by incipient farmers. Because early farmers likely used only a limited number of wild individuals for initial cultivation, this stage is often associated with a genetic bottleneck that reduced and repatterned genetic diversity in a crop compared to its wild relative[7]. These domestication

[1]Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA, USA. [2]College of Agriculture and Food Science, Zhejiang A&F University, Lin'an, Hangzhou, China. *e-mail: yongfez1@uci.edu

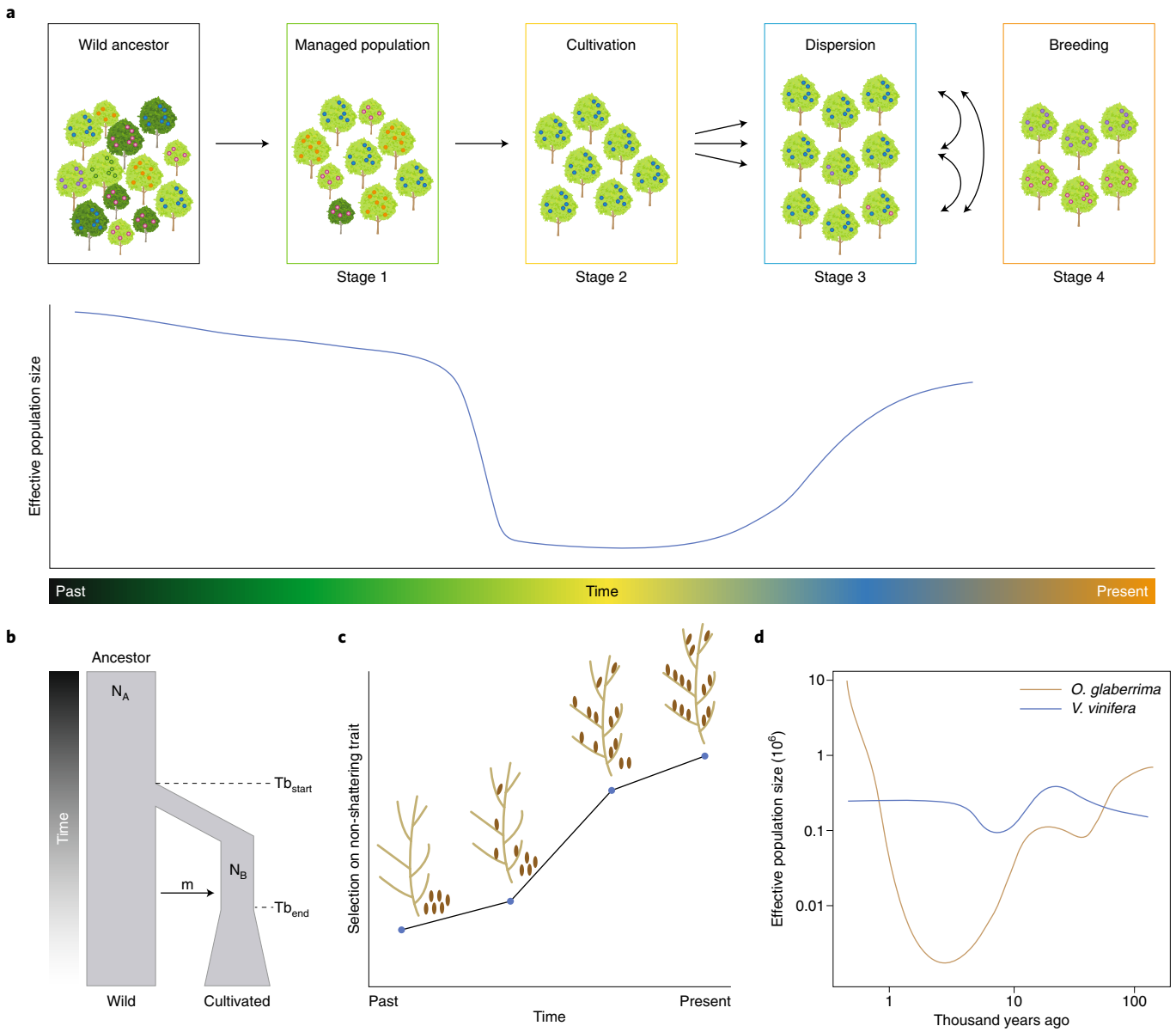

Fig. 1 | Features of demography and selection during plant domestication. a, A schematic representing four stages of domestication. The far left population represents wild populations with substantial genetic diversity. Human management may alter diversity to favour particular phenotypes (stage 1). Eventually, cultivation separates domesticated from wild genotypes (stage 2), which often features genetic bottlenecks and in situ increases in allele frequency. The domesticate is then dispersed to new locales (stage 3), where it must adapt locally, perhaps by introgression or new mutations (represented by differing colours among locations). Finally, deliberate breeding (stage 4) often includes crosses among locally adapted varieties. The curve below the stages provides an example of population size through the four stages, including a long population decline throughout stage 1 and an abrupt bottleneck in stage 2, followed by population expansion. **b**, A typical population genetic model that is used to infer demographic parameters, such as the population size of the ancestral population (N_A) and the bottlenecked population (N_B), the time of the start and end of a population bottleneck (Tb_{start} and Tb_{end}) and migration rates (m). **c**, An idealized representation of artificial selection on the non-shattering phenotype over time. The strength of selection varies with slope. **d**, A cartoon of inferred populations size for African rice (*O. glaberrima*) and grapes (*V. vinifera*) based on SMC methods[30,31]. Note that time on the *x* axis is reversed relative to the other graphs, keeping with the standard for SMC analyses.

bottlenecks have been studied extensively for ~20 years using SNP data[8], but it is surprising to realize that the duration of this stage remains unclear for most plants.

Traditionally, the duration of stage 2 has been studied by two methods—population modelling and fossil analyses—that have led to contrasting viewpoints. The contrast focuses on the time it has taken for favourable domestication alleles to become fixed. Historically, it has been thought that fixation occurred rapidly. Both Zohary[9] and Ladzinsky[10] suggested that domestication alleles predominate less than 100 years after the appearance of a relevant mutant. Similarly, Hillman and Davies, used field experiments to simulate the process of domestication, focusing on phenotypic changes that lead to the non-shattering phenotype, a key feature that separates domesticated cereals from their wild relatives[11]. From their experiment, they estimated that non-shattering could become predominant in less than 200 years and perhaps as rapidly as 20–30 years[9,12]. Population genetic models have generally been consistent with this view. For example, selection in Asian rice (*Oryza sativa*)

may have been strong enough to fix Shattering 4 (*sh4*) within ~100 years[13]. The *sh4* allele is crucial because it accounts for 69% of phenotypic variation in non-shattering between indica rice and its wild annual species *Oryza rufipogon*[14].

These and other modelling efforts typically rely on three assumptions. The first is that the selection coefficient *s* is constant over time, which is unlikely to be true. Second, modelling typically assumes that there are only a few loci of major effect for any trait, but most agronomic traits are polygenic, with hundreds of contributing loci. Finally, even if *s* were constant, the time of fixation of an allele is a function of demographic history, and modelling often relies on assumptions about that history. These assumptions can sometimes introduce an unrecognized circularity, because demographic inferences often (in turn) rely on a pre-determined demographic model that requires simplifying assumptions about the timing, shape and duration of population history. They often assume that a domestication bottleneck begins abruptly (Fig. 1b), at a discrete time in the past, and often without the complications of linkage among variable sites and of continuing genetic introgression from wild populations[8,15,16]. More recent treatments have been able to relax some of these assumptions, particularly by allowing migration between the wild species and the incipient crop[17–19] (Fig. 1b), but some of the basic limitations remain.

In contrast to modelling approaches, fossil analyses examine archaeological trends for phenotypic traits, such as grain/seed size or the shape of the rachis, which contributes to the non-shattering phenotype (Fig. 1c). These analyses suggest that the fixation of domestication traits has taken thousands of years, rather than ~100 years[20–22]. For example, Allaby et al. traced the history of non-shattering in four species—rice, einkorn wheat, barley and emmer wheat—by studying a chronological progression of fossils, with some as old as ~10,000 years[22]. By measuring the rate of change of the rachis over time, they concluded that the *s* on non-shattering had not been constant (Fig. 1c). They also inferred that the maximum selection pressure on the non-shattering phenotype occurred at different times in different species. Their data suggest that the most intense selection on the rice rachis occurred between 7,000 to 6,500 years ago (ya), while that on wheat was between 10,500 and 9,500 ya (ref. [22]).

The use of fossil data for the study of domestication is insightful, but it too has shortcomings. The non-shattering phenotype can be difficult to evaluate based on a fossilized rachis[12], and fossil samples may not represent discrete chronological lineages (or 'selection chains')[22]. Nonetheless, the inference that *s* has not been constant over time may help resolve apparent differences between fossil- and modelling-based inferences about fixation times. Another difference between the two approaches may reflect incomplete correlations between alleles and phenotypes, particularly if polygenic traits evolve more slowly than the fixation of a subset of their component genes. For example, Zhang et al. argued that relatively weak selection may have slowed shifts in phenotype, even if *sh4* itself fixed rapidly[12]. There is some evidence to support this claim, because individuals with homozygous *sh4* alleles in a wild genomic background still shed most of their seeds[23,24], suggesting that the shattering phenotype was shaped by polygenic selection.

Protracted periods of management prior to cultivation. Based on extrapolation from their fossil data, Allaby et al. argue that the onset of domestication for cereals should be extended backwards in time to the Pleistocene era[22]. As this potentially predates the onset of agricultural settlements, they speculate that some selection by humans occurred within wild populations (stage 1, Fig. 1a), prior to direct cultivation (stage 2, Fig. 1a).

There is now indirect genetic evidence for sustained periods of pre-domestication for two crops, African rice (*Oryza glaberrima*) and grape (*Vitis vinifera*). The evidence is based on analysis of WGS data using sequentially Markovian coalescent methods (or 'SMC' methods) for demographic inference[25–27]. These methods estimate the chronological history of effective population size (N_e) without a pre-defined a demographic model, potentially circumscribing some of the simplifying assumptions of previous demographic analyses. SMC methods rely on the fact that the time since the most recent common ancestor (TMRCA) can be estimated between two alleles in an individual or in a sample of individuals. The key feature of the TMRCA is that it varies across independent regions of the genome as a function of population size. If many regions of the genome share the same TMRCA, this reflects a period when many allelic lineages coalesced at the same time, indicative of a small population. By studying the distribution of TMRCAs across the genome, SMC methods identify historical patterns of N_e.

To date, SMC methods have been applied to only a few crops[18,28–31], but with interesting and suggestive results. In one study, WGS data from African rice was used to infer that a discrete domestication bottleneck approximately 3,500 years ago (stage 2) was preceded by a long, roughly 14,000-year decline in the N_e of the progenitor population[30] (Fig. 1d). The authors hypothesized that this protracted N_e decline reflects the effects of abiotic and biotic stresses, which may include a history of low-intensity management by humans prior to modern domestication[30]. Under this hypothesis, humans favoured some subset of wild plants, spread the seeds of those favoured individuals and hastened a long N_e decline.

A less severe but protracted N_e decline of over 12,000 years was also inferred for the lineage leading to domesticated grapes[31] prior to a mild domestication bottleneck roughly 8,000 ya (Fig. 1d). Four things are notable about the grape results. First, a sample of wild grapes did not exhibit a similar N_e decline; in fact, the wild grapes show population histories broadly commensurate with expectations based on glacial history. Second, the cultivated and wild lineages were estimated to have diverged over 20,000 ya, roughly the same time as the beginning of the long population decline. Third, population simulations suggest that the observed site-frequency spectrum (SFS) from the grape sample, better fits a model of protracted N_e decline than a rapid bottleneck. Finally, some sites near the purported region of grape domestication in the southern Caucasus mountains contain evidence of human habitation for more than 20,000 years[32]. Thus, the lineage that led to domesticated grapes had a distinct demographic history from at least one sample of wild grapes. That history included a long N_e decline in a region inhabited by humans.

These studies complement fossil data in suggesting the possibility of protracted periods of N_e decline prior to purposeful cultivation. If true, these are important findings as such a decline would exacerbate many of the genetic effects typically associated with a domestication bottleneck. This conclusion is tentative however, partly because it is based on only a few studies, but also because SMC methods have at least five limitations. One important limitation is phasing. If SNPs are phased incorrectly, the inference of recent demographic history can be wildly incorrect[27]. Fortunately, the most recent method (SMC++)[27] does not require phasing, but it has not yet been applied widely to domesticates (see ref. [31]). A second limitation is that SMC inference may spread-out an abrupt bottleneck over time, making the process seem more protracted than reality[22–24] and raising the possibility that protracted N_e declines are methodological artefacts. Third, these methods can also confuse ancient population structure with changes in N_e. If domesticated germplasm is derived from a single deme of a larger metapopulation, then this will be inferred as a long-term signal of N_e decline, even if that single deme had not experienced an N_e decline[33,34]. Fourth, selective sweeps and linked selection complicate the inference of demography[35], but SMC methods can also be applied with putatively selected regions masked[29,31]. A final limitation rests on interpretation rather than inference; it is a logical leap to infer that human management led to a protracted

population decline, because many other factors (such as climate, biotic and abiotic stresses) could also drive such declines.

Is there genetic evidence to suggest that long-term management of wild populations preceded purposeful cultivation? If so, how many crops experienced long-term management prior to purposeful cultivation? We have summarized the evidence and consider these to be open questions of tremendous importance. Crucial next steps will include further application of SMC-based approaches to genomic data from more crop species, simulation studies that test the inferential robustness of SMC methods under reasonable demographic scenarios, and population simulations that assess whether the empirical data fit SMC-based inferences of population decline[31]. Although there is much work to do, the results thus far are nonetheless consistent with a growing appreciation that humans altered plant species and ecosystems long before the onset of agriculture[22,36,37].

Demography shapes the spectrum of deleterious variants

Why should anyone care about the duration of human effects and particularly the demographic history of domestication? One reason is that more accurate demographic inference may inform our understanding of human history and specifically humans' role in ecological niche construction. Furthermore, demographic history has shaped extant genetic diversity, which may, in turn, affect crop productivity. Here, we illustrate the effect of demographic history on one aspect of crop genetic variation—the accumulation of deleterious SNP variants.

Demography and the cost of domestication. The 'cost of domestication' refers to an increased genetic load within cultivars, which can in theory be studied by identifying deleterious variants within crop genomes[4,38]. This hypothesized cost originates primarily from at least two sources: the decreased N_e during stages 1 and 2 of domestication, which in turn both increase genetic drift[39] and reduce the efficacy of genome-wide selection[40]; and linked selection, which can drag deleterious mutations to high frequency[31,41,42].

To better understand how a reduction in population size may contribute to the accumulation of deleterious alleles, it is helpful to first consider the distribution of fitness effects (DFE) of mutations in a wild progenitor population. The effects of mutations range from strongly deleterious to highly advantageous. The former (strongly deleterious mutations) will typically be lost rapidly from the gene pool. The latter (highly advantageous mutations) are expected to represent only a small proportion of total mutations[43]. Thus, most mutations in the wild progenitors of crops have fitness effects ranging from zero (that is, selectively neutral) to slightly deleterious[43–45].

Slightly deleterious mutations are typically found at a low frequency within a population with large N_e, because selection is efficacious in large populations[46]. When N_e decreases during a domestication bottleneck, two things happen. First, the bottleneck reduces genetic variation. By reducing genetic variation, the bottleneck is likely to remove many (and perhaps most) slightly deleterious genetic variants. Second, selection is less efficacious[40], so that deleterious variants that are not lost tend to rise to higher frequency due to genetic drift[47–49]. Remarkably, these two forces balance out for outcrossing organisms when mutations are strictly additive[50]—that is, with a dominance coefficient of $h = 0.5$. As a result, the genome of a domesticated individual is expected to contain the same number of derived deleterious variants as the genome of a wild individual after a genetic bottleneck[50]. In other words, the bottleneck decreases the total number of slightly deleterious variants while the frequencies of surviving variants increase. These cancel out so that each individual contains roughly the same number of deleterious variants (Fig. 2).

When outcrossing and additivity do not hold however, the number of slightly deleterious variants within a single individual can actually increase after a population bottleneck, depending on dominance coefficients and other factors[29,51,52]. For example, genetic diversity within domesticated dogs is lower than in wolf populations, but the number of deleterious alleles per dog is actually higher[53].

Putatively deleterious variants in crops. Although crops typically contain less genetic diversity than their wild relatives, the question remains as to whether they have increased numbers of deleterious variants. Several recent studies have addressed this question using WGS data. All of these studies have relied on bioinformatic methods to predict whether an SNP variant is a potentially deleterious (as opposed to neutral) genetic modification. There are several methods to perform these predictions; most of them identify deleterious variants as nucleotide changes within evolutionarily conserved genomic regions[53–55]. The various methods appear to perform roughly equivalently in *Arabidopsis thaliana*[56], although their accuracy has yet to be tested explicitly in a crop.

Based on the application of these methods, most crops examined to date contain an enriched proportion of deleterious SNP variants compared to their wild relatives[29,31,44,57–59]. The early papers on this topic tended to measure the ratio of putatively deleterious to synonymous variants. More recently, it has been recognized that a better measure of potential genetic load is the total number of deleterious alleles per genome[38,50,60]. Of the handful of crops that have been studied thus far, cassava (*Manihot esculenta*) is the current champion, because a single cassava individual has approximately 26% more predicted deleterious alleles, on average, than a corresponding wild cassava genome[59].

Why does a crop like cassava have some many putatively deleterious variants? Generally speaking, the accumulation of deleterious mutations will vary among crops due to several factors. One is population history, where protracted N_e decline (stage 1) exacerbates the accumulation of deleterious variants[31]. Accumulation will also vary as a function of the strength and duration of a domestication bottleneck[61]. This accumulation is also affected by the DFE, which may differ among species and between wild and cultivated forms[38]. A further factor is the mating system, as it impacts the possible retention or purging of deleterious variants. Clonal propagation can hide recessive deleterious variants in a heterozygous state, but self-fertilization cannot. Finally, dominance coefficients[52,62] interact with the mating system. For example, Zhou et al.[31] showed that recessive deleterious alleles ($h = 0.0$) accumulate without a fitness detriment in clonally propagated crops. In fact, fitness increases dramatically after a shift to clonal propagation (Fig. 3), because clonal propagation hides new recessive deleterious mutations in a heterozygous state and no homozygotes are generated to increase genetic load[31]. In contrast, inbreeding exposes deleterious alleles, so that recessive deleterious alleles are purged after a bottleneck event, eventually returning load to pre-bottleneck levels[61]. Given these considerations, the remarkable 26% increase of deleterious variants in cassava is probably due to its clonal propagation as well as a severe population bottleneck during domestication[59].

The phenotypic impacts of deleterious variants. A further question is whether these predicted deleterious variants affect phenotype and fitness. If they do, then the 'cost of domestication' is more than an esoteric outcome of demographic history, and a potential path toward crop improvement is through breeding to remove deleterious variants[63,64]. The phenotypic effects of deleterious variants have been explored to only a limited extent thus far[65], but putatively deleterious SNPs in maize (*Zea mays* ssp. *mays*) explain more phenotypic variance than random sets of SNPs and consistently have larger estimated effect sizes[66]. The inclusion of deleterious SNPs also improved the predictive performance of genomic selection models by as much as 4.3% for grain yield[66]. Finally, rare, putatively deleterious variants have an outsized effect on dysregulation of maize gene expression, with correlative effects on fitness measures[64]. When deleterious variants do affect phenotype, as they appear to do in maize, a chal-

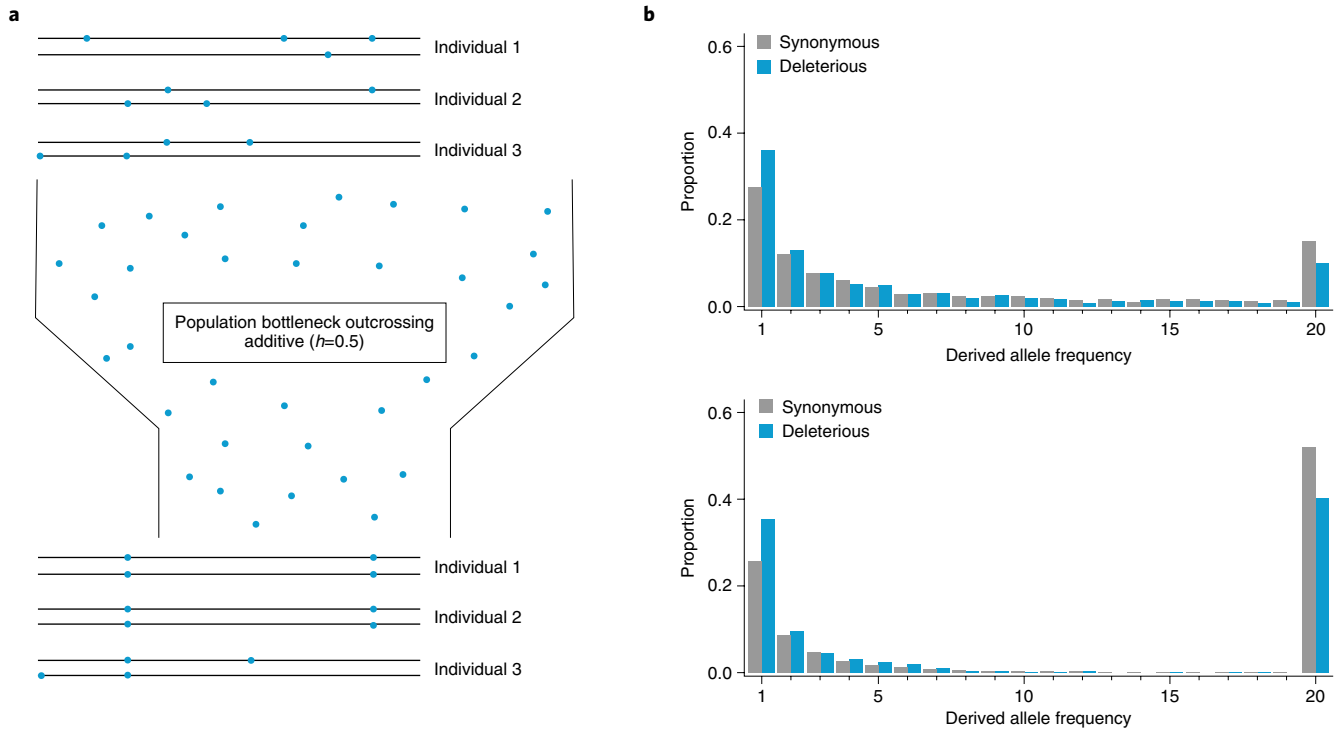

Fig. 2 | The effects of demography on deleterious variants. a, The figure illustrates individual chromosomes before and after a bottleneck. As shown, the number of derived deleterious variants per individual is expected to be equivalent before and after the bottleneck under an outcrossing, additive model[50]. **b**, The site-frequency-spectrum (SFS) of neutral and deleterious derived variants in an outcrossing population with additive (h=0.5) selection. The top SFS represents a population before a population bottleneck; the bottom SFS illustrates an increase in fixed derived variants after a bottleneck, due to enhanced genetic drift. The SFS are based on forward simulations that assume $N_B/N_A = 0.05$, with other parameters identical to reference[44].

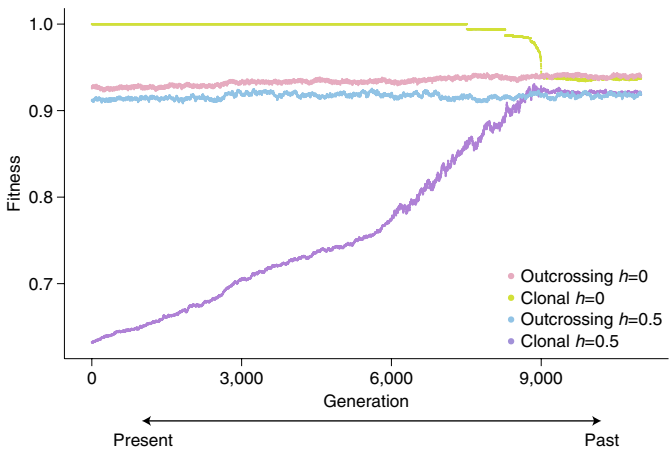

Fig. 3 | The fitness effects of domestication and clonal propagation. Forward simulations show the effects of domestication on fitness for a clonally propagated crop under additive (h=0.5) and recessive (h=0) selection. These simulations follow those in reference[31]. For the clonal lineage, the shift in mating system and a genetic bottleneck both occur 9,000 generations in the past. The bottleneck was of severity $N_B/N_A = 0.1$ but recovered 5,000 generations in the past. The outcrossing populations are for comparison and had a constant effective population size.

lenging question will be to decide which among the thousands of deleterious variants are appropriate targets for crop improvement. As most deleterious mutations accumulate in recombination-limited regions[44,58,67] (also see ref. [65]), they may be difficult to purge with standard breeding techniques[58].

Interestingly, crops may be able to escape some of their deleterious burden through introgression, which is a particularly important evolutionary force when crops disperse from their domestication centres to new geographic locales (stage 3, Fig. 1a). Dispersion requires adaptation to new environments, which can be achieved either by selection within the crop or by introgression of alleles from locally adapted wild populations. Introgression can change genetic load when two populations differ in N_e, and therefore in their deleterious burden. There is at least one known instance where a genomic region with fewer deleterious variants introgressed into a crop from wild populations[29]. This observation raises the intriguing possibility that adaptation by introgression may in some cases result from a reduction of genetic burden rather than from better-adapted alleles. Introgression has shaped genetic diversity in many crops, including rice[68], olives[69], apples[70,71] and maize[29,72,73], but it is an open question whether it has played a role in ameliorating the cost of domestication.

Structural variants remain a great unknown

The demographic shifts associated with domestication also shape the number and frequency of SVs in a crop. Yet, there is currently little information about the population and evolutionary dynamics of SVs in crops or their wild relatives. We begin this section by considering potential effects of demography on SVs, and we close by considering the phenotypic effect of SVs. Throughout we argue that better characterization of SVs is necessary for understanding crop history and the genetic variants that affect phenotypes.

Structural variants in populations. Like SNPs, most new SV mutations are expected to be either neutral or slightly deleterious in their effects. If most SVs are slightly deleterious, they should be affected by demography similarly to deleterious SNPs. That is, as

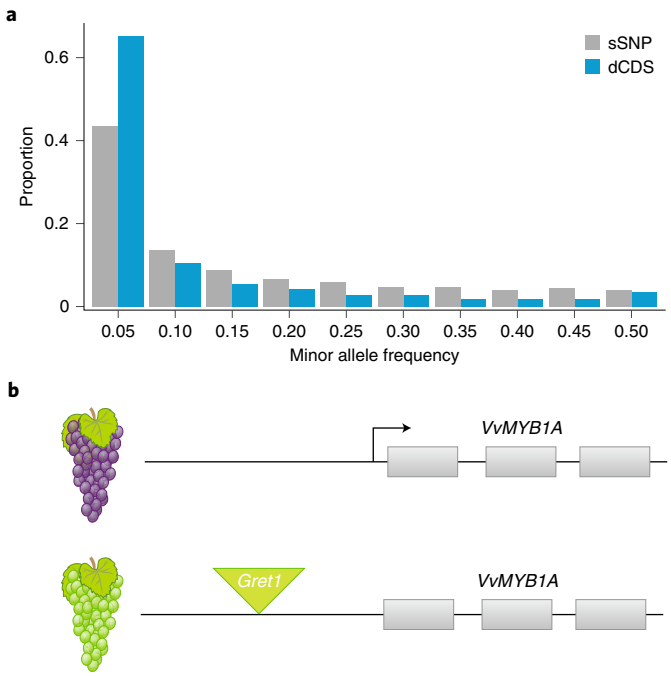

Fig. 4 | SVs in populations and an example of an affected phenotype. **a**, The folded SFS of inferred duplicated coding DNA sequence (cCDS) relative to the SFS of presumed neutral synonymous SNPs (sSNPs) for a sample of 116 indica rice accessions. The SFS of cCDS shows an excess of rare alleles compared to sSNPs, suggesting they are predominantly deleterious and under purifying selection. The cCDS were called using the pipeline from reference[117] (see Supplementary Information). **b**, Expression of the grape *VvMYB1A* gene is modified by the insertion of the *Gret1* retrotransposon, altering grape colour[114].

a wild progenitor is domesticated through a protracted N_e decline (stage 1) and/or a bottleneck (stage 2), most SV loci will lose genetic variation and no longer segregate within the crop. However, just as a subset of slightly deleterious SNPs drift to high frequency or fixation in small populations, so might a subset of slightly deleterious SVs. As the crop expands (stage 3), selection will act against these high frequency variants and mutation will introduce new SVs.

This model predicts that crops will have fewer polymorphic SV loci than their wild relatives after stages 1 and 2, and that the remaining derived and newly introduced SVs will contribute to the cost of domestication. To address the latter point, we examined the minor allele frequency of duplicated coding DNA sequences (dCDSs) in cultivated Asian rice (*O. sativa* ssp. *indica*). The results are consistent with the notion that derived dCDS events in rice are slightly deleterious and under purifying selection (Fig. 4a). Beyond this simple example, there are few empirical data about the population dynamics of SVs during and after domestication. Most population genomic surveys of crops report the number of SVs detected in a crop sample[74–77] (see also Supplementary Table 1), but not the population frequencies of individual SVs between crops and their wild ancestor. Notable exceptions include the observations that the SFS of maize CNVs illustrate a slight uptick in high frequency relative to teosintes[78] (consistent with the fixation of CNVs through the domestication bottleneck), and that small indels between African rice and its wild progenitor (*Oryza barthii*) exhibit a similar pattern[79]. Some studies have also contrasted population frequencies between crops to identify positively selected SVs (refs [80,81]). Nonetheless, our understanding of the prevalence of SVs, their evolutionary dynamics and their effects on domestication is in its infancy.

Types and rates of SVs. One of the complicating features of studying SVs is that they encompass many types of genetic events—including sequence insertions, deletions, duplications, inversions, translocations and complex events. SVs are generated by heterogeneous mechanisms and occur at different rates. Estimates of these rates are important, as rapidly evolving SVs are likely to have weak linkage to nearby SNPs, thus making traditional SNP-based approaches (such as genome-wide association studies (GWAS)) unlikely to anchor a causative SV. Indeed, 20% of CNVs in a survey of maize could not be anchored by SNPs[82]; a slightly higher proportion (~27%) could not be anchored by SNPs in humans[83]; and two-thirds of transposable elements (TEs) in a population survey of *Arabidopsis thaliana* were not in linkage disequilibrium with a nearby SNP (ref. [84]). Furthermore, SVs that evolve at different rates from SNPs may prove useful for inferring demographic history, because they provide additional snapshots into the timing and pace of demographic events. For example, dinucleotide microsatellite loci mutate at a rate of 10^{-3} per locus per generation[85], which is six magnitudes faster than most accepted rates of SNP mutation[86]. Given this rate, microsatellites may be able to provide unique insights into very recent demographic events (stages 3 and 4) but fewer insights into more ancient events such as N_e decline (stage 1).

TEs are probably the source of most SVs within crops and their wild relatives, because they compose the majority of plant genomes and facilitate genomic rearrangement. The gain and loss of TEs can be so rapid as to be detectable as differences in genome size within species. For example, *Brachypodium distachon* genomes vary up to ~1.7-fold in size, due largely to variation in the TE-rich, non-genic portion of genomes[87]. However, we have surprisingly few estimates of the rates of TE transposition. It has been measured to occur at 10^{-6} per locus for the *Associator–Dissociator* (*Ac–Ds*) system[88], but the rate may be much higher for other element types[89], particularly during bouts of stress[90]. Interestingly, element loss may be quite slow in some cases, as orthologous TEs are estimated to have been lost at a rate of 3.62 kb per million years per element across *Oryza* species[79]. This slow rate is difficult to reconcile with dramatic differences in genome size and TE content within and across species[79,87,91].

Another potent source of SVs is gene duplication. Plant genomes contain a higher fraction of duplicated genes relative to other eukaryotes[92]; on average 64.5% of plant genes have paralogues[93]. Gene duplications are caused by a variety of mechanisms, including whole genome duplications, tandem duplication, dispersed duplications and retrotransposition[94,95]. We suspect that tandem duplicates are the most frequent source of CNVs, because they represent 10% of plant genes on average[96] and can amass rapidly within a species. As a concrete example, two maize accessions differ by eight tandem-duplicated genes (22 versus 14 genes) in the 22 kDa zein cluster[97]. They are probably generated at a high rate; an experimental tandem array in *A. thaliana* generates CNVs at the rate of approximately 10^{-6} per array per F1 meiosis[98]. Given ~1,500 tandem arrays in the *A. thaliana* reference genome[99], this rate translates to a conservative expectation of one CNV in 1 in 700 seeds, due to expansion or contraction of tandem arrays[100].

These examples are meant to briefly illustrate gaping holes in our knowledge about the rates at which most types of SVs occur across plant species. The potential for SVs to provide information about plant domestication history and the genetic contributions of phenotypic variation, is immense.

SV detection. Given the potential of SVs, it follows that their categorization is a crucial undertaking. To date, SVs have been inferred most often from short-read WGS data based on read coverage, split-read (SR) information and/or misaligned paired-end (PE) reads[101] (Supplementary Fig. 1, Supplementary Table 1). Typically, SVs are defined in such a way that small events are ignored. The human genomics community defines SVs as more than 50 bp in length[101,102],

but to our knowledge there is no community standard in plants. It is also worth noting an interesting dilemma that applies to the use of short read data to detect SVs relative to a reference genome. Reference genomes are commonly only available for cultivars that are expected to have fewer SV loci that are at high population frequencies. If this is true, it creates a potential bias, in that it will be more difficult to identify novel SVs in wild samples than in the (mostly) shared SVs of a cultivated sample.

Fortunately, SVs are now starting to be reported from the comparison of de novo assembled genomes, both between individuals[87,103] and across species[79,104,105]. Long-read sequencing technologies are also substantially enhancing SV detection[106]. One particularly interesting example comes from *Drosophila melanogaster*, where a genome based on long-reads was compared to the existing reference. The comparison identified numerous CNVs that likely contribute to quantitative traits[105], and also revealed that one 'essential gene' was missing from several strains. To date, the most extensive comparison of plant reference genomes is among *Oryza* species[79,104], but within-species comparisons are still rare within a single crop. Ongoing efforts to define crop 'pangenomes' will provide more comprehensive catalogues of SVs, particularly when they are based on long-read sequencing. We fervently hope that these pangenome efforts will include wild accessions, as they may yield important information about domestication history.

SVs affect phenotypes. Another reason to study SVs is because they are implicated, time and again, as genetic variants that affect phenotypes. One line of evidence is that SVs tend to have an outsized effect when they are included in GWAS. For example, read-depth variants (as a proxy for CNVs) were consistently over-represented relative to SNPs in maize GWAS results, even after accounting for their relative abundance in the genome[82]. Similarly, 42% of SNPs associated with traits in rice could not be identified in the reference genome[107], suggesting that they reside in SVs that are missing from the reference. The phenotypic effects of SVs may even extend to microsatellite loci, which are traditionally considered to be evolutionarily neutral, because segregating variation in microsatellites accounts for 12% of gene expression variation in humans[108].

While most new SVs are likely to be deleterious, it is also clear that a subset contribute to adaptive phenotypic change during domestication. As an example, consider 'domestication genes', that is, genes that contribute to morphological changes associated with stage 2 (Fig. 1a). We reviewed the primary literature corresponding to 34 of these genes[109]. In one-third of those genes, the domestication variant was caused by CNVs, large (>1 Kb) indels or similarly large rearrangements (Supplementary Table 2). The remaining two-thirds of causal variants were enriched for point mutations and small indels, especially premature stop codons. However, we expect that our estimate of one-third is overly conservative, as SVs may be linked to nearby SNPs that appear causal but are not, and some 'domestication SVs' likely cause regulatory changes but are difficult to identify because they do not flank the affected gene. A classic example is a TE residing 60–70 Kb away from the maize domestication locus *TEOSINTE BRANCHED 1* (*TB1*), where it acts as an enhancer of expression and confers regulatory variation between the domestication and non-domestication alleles[110].

The prevalence of SVs that contribute to stages 3 and 4 of domestication may be even higher than for stage 2. Genes that are known to alter phenotypes across cultivars include a tandem array duplication that confers flooding resistance in rice[111], a gene deletion that affects rice grain size[112], and TE-induced null mutations such as the sticky gene in foxtail millet[113], berry colour in grapes[114] (Fig. 4c) and fruit colour in oranges[115]. Given these observations, our overriding hypothesis is that most crop adaptation is driven by SVs. If true, most studies of crop adaptation and domestication have ignored the critical genomic variants.

Conclusions

Extant genetic variation within a crop has been shaped by demography, selection, linked selection and mating systems. This Review aims to provide insight into how these factors interact and to emphasize that demographic history dictates many aspects of extant genetic diversity. Yet, one could argue that our understanding of demographic history is at its infancy, because the duration and severity of population bottlenecks during stage 2 have not been investigated for most crops. Moreover, the possibility of a protracted pre-domestication stage exacerbates demographic effects and mimics some of the population genetic features of a bottleneck[31]. SMC methods may prove very useful for demographic inference, but they are not a panacea. To fully understand the protracted process of domestication, we must continue to utilize every tool available, including traditional inference based on population genetic models, fossil analyses[116] and ancient DNA (ref. [73]).

A major gap in our knowledge of the genetic process of domestication centres on the numbers, frequencies, rates and types of SVs; how SVs have been affected by demographic history; and whether they are a predominant source of genetic variation affecting phenotypes. An appreciable fraction of SVs may not be detectable by their linkage to SNPs, and so their contributions to phenotypic diversity have been consistently underrepresented in SNP-based analyses of crop diversity. Fortunately, the publication of chromosome-level genomes from multiple species has begun to provide more complete characterizations of SVs. As a consequence of improved long-read sequencing platforms and assembly methods, we expect to learn more about SV diversity (and the processes that shape that diversity) over the next decade than we had in the previous two.

Received: 11 February 2018; Accepted: 27 June 2018;
Published online: 30 July 2018

References

1. Gaut, B. S., Díez, C. M. & Morrell, P. L. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends Genet.* **31**, 709–719 (2015).
2. Hammer, K. Das domestikationssyndrom. *Kulturpflanze* **32**, 11–34 (1984).
3. Gerbault, P. et al. Storytelling and story testing in domestication. *Proc. Natl. Acad. Sci. USA* **111**, 6159–6164 (2014).
4. Lu, J. et al. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* **22**, 126–131 (2006).
5. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
6. Miller, A. J. & Gross, B. L. From forest to field: perennial fruit crop domestication. *Am. J. Bot.* **98**, 1389–1414 (2011).
7. Doebley, J. in *Isozymes in plant biology* (eds Soltis, D. E. & Soltis, P. S.) 165–191 (Chapman and Hall, London, 1989).
8. Eyre-Walker, A., Gaut, R. L., Hilton, H., Feldman, D. L. & Gaut, B. S. Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* **95**, 4441–4446 (1998).
9. Hillman, G. C. & Davies, M. S. Domestication rates in wild-type wheats and barley under primitive cultivation. *Biol. J. Linn. Soc.* **39**, 39–78 (1990).
10. Ladizinsky, G. Pulse domestication before cultivation. *Econ. Bot.* **41**, 60–65 (1987).
11. Harlan, J. R., de Wet, J. M. J. & Price, E. G. Comparative evolution of cereals. *Evolution* **27**, 311–325 (1973).
12. Hillman, G. C. & Davies, M. S. Measured domestication rates in wild wheats and barley under primative cultivation and their archaeological implications. *J. World Prehist.* **4**, 157–222 (1990).
13. Zhang, L. B. et al. Selection on grain shattering genes and rates of rice domestication. *New Phytol.* **184**, 708–720 (2009).
14. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
15. Wright, S. I. et al. The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).
16. Zhu, Q., Zheng, X., Luo, J., Gaut, B. S. & Ge, S. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* **24**, 875–888 (2007).
17. Molina, J. et al. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc. Natl. Acad. Sci. USA* **108**, 8351–8356 (2011).

18. Beissinger, T. M. et al. Recent demography drives changes in linked selection across the maize genome. *Nat. Plants* **2**, 16084 (2016).
19. Caicedo, A. L. et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet.* **3**, 1745–1756 (2007).
20. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
21. Purugganan, M. D. & Fuller, D. Q. Archaeological data reveal slow rates of evolution during plant domestication. *Evolution* **65**, 171–183 (2011).
22. Allaby, R. G., Stevens, C., Lucas, L., Maeda, O. & Fuller, D. Q. Geographic mosaics and changing rates of cereal domestication. *Philos. T. Roy. Soc. B* **372**, 20160429 (2017).
23. Ishikawa, R. et al. Allelic interaction at seed-shattering loci in the genetic backgrounds of wild and cultivated rice species. *Genes Genet. Syst.* **85**, 265–271 (2010).
24. Ishii, T. et al. OsLG1 regulates a closed panicle trait in domesticated rice. *Nat. Genet.* **45**, 462–465 (2013).
25. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
26. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
27. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
28. Velasco, D., Hough, J., Aradhya, M. & Ross-Ibarra, J. Evolutionary genomics of peach and almond domestication. *G3-Genes Genom. Genet.* **6**, 3985–3993 (2016).
29. Wang, L. et al. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).
30. Meyer, R. S. et al. Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* **48**, 1083–1088 (2016).
31. Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D. & Gaut, B. S. Evolutionary genomics of grape (*Vitis vinifera* ssp. *vinifera*) domestication. *Proc. Natl. Acad. Sci. USA* **114**, 11715–11720 (2017).
32. Adler, D. S. & Tushabramishvili, N. in *Settlement Dynamics of the Middle Palaeolithic and Middle Stone Age* Vol. 2 (ed. Conard, N. J.) Ch. 5 (Kerns Verlag, Tubingen, 2004).
33. Nielsen, R. & Beaumont, M. A. Statistical inferences in phylogeography. *Mol. Ecol.* **18**, 1034–1047 (2009).
34. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference. *Heredity* **116**, 362–371 (2016).
35. Schrider, D. R., Shanku, A. G. & Kern, A. D. Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**, 1207–1223 (2016).
36. Roberts, P., Hunt, C., Arroyo-Kalin, M., Evans, D. & Boivin, N. The deep human prehistory of global tropical forests and its relevance for modern conservation. *Nat. Plants* **3**, 17093 (2017).
37. Boivin, N. L. et al. Ecological consequences of human niche construction: Examining long-term anthropogenic shaping of global species distributions. *Proc. Natl. Acad. Sci. USA* **113**, 6388–6396 (2016).
38. Moyers, B. T., Morrell, P. L. & McKay, J. K. Genetic costs of domestication and improvement. *J. Hered.* **109**, 103–116 (2017).
39. Hedrick, P. W. & Garcia-Dorado, A. Understanding inbreeding depression, purging, and genetic rescue. *Trends Ecol. Evol.* **31**, 940–952 (2016).
40. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796 (2009).
41. Chun, S. & Fay, J. C. Evidence for hitchhiking of deleterious mutations within the human genome. *PLoS Genet.* **7**, e1002240 (2011).
42. Charlesworth, B. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. *Genetics* **191**, 233–246 (2012).
43. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**, 610–618 (2007).
44. Liu, Q., Zhou, Y., Morrell, P. L. & Gaut, B. S. Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **34**, 908–924 (2017).
45. Arunkumar, R., Ness, R. W., Wright, S. I. & Barrett, S. C. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* **199**, 817–829 (2015).
46. Ohta, T. The nearly neutral model of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**, 263–286 (1992).
47. Casals, F. et al. Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. *PLoS Genet* **9**, e1003815 (2013).
48. Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. *Curr. Opin. Genet. Dev.* **29**, 139–146 (2014).
49. Lohmueller, K. E. et al. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–997 (2008).
50. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
51. Marsden, C. D. et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proc. Natl. Acad. Sci. USA* **113**, 152–157 (2016).
52. Balick, D. J., Do, R., Cassa, C. A., Reich, D. & Sunyaev, S. R. Dominance of deleterious alleles controls the response to a population bottleneck. *PLoS Genet.* **11**, e1005436 (2015).
53. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688 (2012).
54. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
55. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP^{++}. *PLoS Comput. Biol.* **6**, e1001025 (2010).
56. Kono, T. J. Y. et al. Comparative genomics approaches accurately predict deleterious variants in plants. *bioRxiv* https://doi.org/10.1101/112318 (2018).
57. Kono, T. J. et al. The role of deleterious substitutions in crop genomes. *Mol. Biol. Evol.* **33**, 2307–2317 (2016).
58. Renaut, S. & Rieseberg, L. H. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other composite crops. *Mol. Biol. Evol.* **32**, 2273–2283 (2015).
59. Ramu, P. et al. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
60. Brandvain, Y. & Wright, S. I. The limits of natural selection in a nonequilibrium world. *Trends Genet.* **32**, 201–210 (2016).
61. Kirkpatrick, M. & Jarne, P. The effects of a bottleneck on inbreeding depression and the genetic load. *Am. Nat.* **155**, 154–167 (2000).
62. Henn, B. M. et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl. Acad. Sci. USA* **113**, 440–449 (2016).
63. Morrell, P. L., Buckler, E. S. & Ross-Ibarra, J. Crop genomics: advances and applications. *Nat. Rev. Genet.* **13**, 85–96 (2011).
64. Kremling, K. A. G. et al. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520–523 (2018).
65. Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3-Genes Genom. Genet.* **4**, 163–171 (2014).
66. Yang, J. et al. Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLoS Genet.* **13**, e1007019 (2017).
67. Rodgers-Melnick, E. et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. USA* **112**, 3823–3828 (2015).
68. Wang, H., Vieira, F. G., Crawford, J. E., Chu, C. & Nielsen, R. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. *Genome Res.* **27**, 1029–1038 (2017).
69. Diez, C. M. et al. Olive domestication and diversification in the Mediterranean Basin. *New Phytol.* **206**, 436–447 (2015).
70. Cornille, A. et al. New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genet.* **8**, e1002703 (2012).
71. Duan, N. et al. Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat. Commun.* **8**, 249 (2017).
72. Hufford, M. B. et al. The genomic signature of crop-wild introgression in maize. *PLoS Genet.* **9**, e1003477 (2013).
73. Swarts, K. et al. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* **357**, 512–515 (2017).
74. Wang, M. et al. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* **49**, 579–587 (2017).
75. Cheng, F. et al. Genome resequencing and comparative variome analysis in a *Brassica rapa* and *Brassica oleracea* collection. *Sci. Data* **3**, 160119 (2016).
76. Cao, K. et al. Comparative population genomics reveals the domestication history of the peach, *Prunus persica*, and human influences on perennial fruit crops. *Genome Biol.* **15**, 415 (2014).
77. Hazzouri, K. M. et al. Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. *Nat. Commun.* **6**, 8824 (2015).
78. Swanson-Wagner, R. A. et al. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20**, 1689–1699 (2010).
79. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
80. Zhou, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).

81. Zhang, Z. et al. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* **27**, 1595–1604 (2015).
82. Chia, J. M. et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
83. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
84. Stuart, T. et al. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife* **5**, e20777 (2016).
85. Vigouroux, Y. et al. Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* **19**, 1251–1260 (2002).
86. Gaut, B., Yang, L., Takuno, S. & Eguiarte, L. E. The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. S.* **42**, 245–266 (2011).
87. Gordon, S. P. et al. Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).
88. Walbot, V. Saturation mutagenesis using maize transposons. *Curr. Opin. Plant Biol.* **3**, 103–107 (2000).
89. Naito, K. et al. Dramatic amplification of a rice transposable element during recent domestication. *Proc. Natl. Acad. Sci. USA* **103**, 17620–17625 (2006).
90. Lisch, D. Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.* **60**, 43–66 (2009).
91. Diez, C. M. et al. Genome size variation in wild and cultivated maize along altitudinal gradients. *New Phytol.* **199**, 264–276 (2013).
92. Szathmáry, E., Jordán, F. & Pál, C. Molecular biology and evolution. Can genes explain biological complexity. *Science* **292**, 1315–1316 (2001).
93. Panchy, N., Lehti-Shiu, M. & Shiu, S. H. Evolution of gene duplication in plants. *Plant Physiol.* **171**, 2294–2316 (2016).
94. Lockton, S. & Gaut, B. S. Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* **21**, 60–65 (2005).
95. Wang, Y., Wang, X. & Paterson, A. H. Genome and gene duplications and gene expression divergence: a view from plants. *Ann. NY Acad. Sci.* **1256**, 1–14 (2012).
96. Yu, J. et al. PTGBase: an integrated database to study tandem duplicated genes in plants. *Database* **2015**, bav017 (2015).
97. Dong, J. et al. Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *Proc. Natl. Acad. Sci. USA* **113**, 7949–7956 (2016).
98. Jelesko, J. G., Carter, K., Thompson, W., Kinoshita, Y. & Gruissem, W. Meiotic recombination between paralogous RBCSB genes on sister chromatids of *Arabidopsis thaliana*. *Genetics* **166**, 947–957 (2004).
99. Zhang, L. & Gaut, B. S. Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome. *Genome Res.* **13**, 2533–2540 (2003).
100. Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J. & Anderson, L. K. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.* **8**, 77–84 (2007).
101. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
102. Tattini, L., D'Aurizio, R. & Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng. Biotechnol.* **3**, 92 (2015).
103. Maretty, L. et al. Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, 87–91 (2017).
104. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
105. Chakraborty, M. et al. Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat. Genet.* **50**, 20–25 (2018).
106. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
107. Yao, W. et al. Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.* **16**, 187 (2015).
108. Gymrek, M. et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
109. Olsen, K. M. & Wendel, J. F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
110. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).
111. Xu, K. et al. Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442**, 705–708 (2006).
112. Shomura, A. et al. Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028 (2008).
113. Kawase, M., Fukunaga, K. & Kato, K. Diverse origins of waxy foxtail millet crops in East and Southeast Asia mediated by multiple transposable element insertions. *Mol. Genet. Genom.* **274**, 131–140 (2005).
114. Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
115. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012).
116. Fuller, D. Q. Long and attenuated: comparative trends in the domestication of tree fruits. *Veg. Hist. Archaeobot.* **27**, 165–176 (2017).
117. Rogers, R. L. et al. Landscape of standing variation for tandem duplications in *Drosophila yakuba* and *Drosophila simulans*. *Mol. Biol. Evol.* **31**, 1750–1766 (2014).

Acknowledgements

The authors thank two anonymous reviewers, A. Muyle, E. Solares and T. Batarseh for comments. BSG is supported by NSF grants 1741627 and 1655808. DKS is supported by an NSF Postdoctoral Research Fellowship in Biology (1609024). QL is supported by the National Natural Science Foundation of China (grant no. 31471431) and the Training Program for Outstanding Young Talents of Zhejiang A&F University. YZ is supported by Y.Z. is supported by the International Postdoctoral Exchange Fellowship Program, China.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-018-0210-1.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to Y.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.