The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops

Sebastien Renaut*,†,[1] and Loren H. Rieseberg[1,2]

[1]Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, BC, Canada

[2]Department of Biology, Indiana University – Bloomington

†Present address: Biodiversity Centre, Institut de recherche en biologie végétale, Université de Montréal, Montréal, QC, Canada

***Corresponding author:** E-mail: sebastien.renaut@gmail.com.

Associate editor: Brandon Gaut

Abstract

For populations to maintain optimal fitness, harmful mutations must be efficiently purged from the genome. Yet, under circumstances that diminish the effectiveness of natural selection, such as the process of plant and animal domestication, deleterious mutations are predicted to accumulate. Here, we compared the load of deleterious mutations in 21 accessions from natural populations and 19 domesticated accessions of the common sunflower using whole-transcriptome single nucleotide polymorphism data. Although we find that genetic diversity has been greatly reduced during domestication, the remaining mutations were disproportionally biased toward nonsynonymous substitutions. Bioinformatically predicted deleterious mutations affecting protein function were especially strongly over-represented. We also identify similar patterns in two other domesticated species of the sunflower family (globe artichoke and cardoon), indicating that this phenomenon is not due to idiosyncrasies of sunflower domestication or the sunflower genome. Finally, we provide unequivocal evidence that deleterious mutations accumulate in low recombining regions of the genome, due to the reduced efficacy of purifying selection. These results represent a conundrum for crop improvement efforts. Although the elimination of harmful mutations should be a long-term goal of plant and animal breeding programs, it will be difficult to weed them out because of limited recombination.

Key words: **deleterious mutation, adaptation, genetic load, recombination rate, crop improvement.**

Introduction

Populations of organisms harbor harmful mutations that prevent them from achieving optimal fitness. These mutations may arise as a consequence of replication errors during cell division. In addition, individuals are constantly exposed to different mutagenic environmental factors. Consequently, new variants, untested by selection, are introduced into all populations, of all life forms, at all times and the majority of these mutations are likely to be deleterious or neutral (Ohta 1992; Eyre-Walker and Keightley 2007).

The existence of deleterious mutations in natural populations is best understood as a balance among mutation, selection, and drift. The demographic conditions and genomic features that permit such mutations to rise in frequency in sexual populations, despite their deleterious effects, are of great interest, both for theoretical and applied reasons (Felsenstein 1974; Charlesworth et al. 1993; Hartl and Clark 1997). Under mutation selection balance, the accumulation of deleterious mutations in sexually reproducing species is infrequent because sex and genetic recombination during meiosis can bring together currently deleterious mutations to create unfit genotypes that are then eliminated from the population. However, under certain circumstances, the beneficial effects of sexual reproduction and recombination may be reduced and the accumulation of deleterious mutations can be substantial (Kondrashov 1988). For example, reduction in population size and inbreeding will lower effective rates of recombination and may allow nonadaptive, putatively deleterious mutations to rise in frequency. In particular, as species expand into new environments (either natural or artificial), increased genetic drift due to both the reduction in effective population size and fast growth rate can have previously unforeseen consequences on genome evolution (Edmonds et al. 2004).

In natural environments, the term allele or mutation surfing was been coined to describe how mutations can spread at the front of an expanding population (Edmonds et al. 2004; Klopfstein et al. 2006; Excoffier et al. 2009; Peischl et al. 2013; Lotterhos and Whitlock 2014). Eventually, rapidly expanding population(s) can become genetically distinct from the core population purely because of nonadaptive demographic processes. Peischl et al. (2013) recently argued that the accumulation of deleterious mutations during range expansion has been largely unappreciated because most contemporary studies focus on the adaptive consequences of natural selection. As an example, they analyzed the distribution of deleterious mutations in human populations—a topic of considerable interest and debate (Klopfstein et al. 2006; Lohmueller et al. 2008; Fu et al. 2012; Do et al. 2015)—and reported an excess of deleterious alleles in non-African human populations, especially for private alleles (i.e., alleles exclusive to non-African humans). These results imply that many of the common contemporary deleterious mutations in Europeans arose during the out-of-Africa range expansion itself. Günther and Schmid (2010) observed a similar surfing effect of deleterious mutations in *Arabidopsis thaliana*, where individuals at the edge of the species distribution contained a significantly higher proportion of predicted deleterious amino acid polymorphisms than other more central accessions.

Analogous predictions can be made for populations that have experienced large demographic changes, but in the context of artificial selection. During the process of domestication and subsequent improvement, populations are expected to undergo multiple bottlenecks (and expansions), accompanied by strong artificial selection on numerous, genetically complex traits (Morrell et al. 2011). In turn, this implies that the process of artificial selection will have wide ranging repercussions on genome evolution. First, selection is predicted to be relaxed on characters that are important in the wild, but not under agricultural conditions (Lu et al. 2006). Second, linkage between desirable beneficial and unwanted deleterious mutations may hinder the ability of selection to efficiently fix beneficial mutations, while weeding out deleterious ones. Essentially, this occurs because selection acts on the net effect of both beneficial and deleterious mutations for a given genotype (Hill-Robertson effect; Felsenstein 1974; Cruz et al. 2008; Morrell et al. 2011). A corollary of this effect is that selection against deleterious mutations will be less effective in regions of the genome with reduced levels of recombination (Charlesworth et al. 1993), leading to a predicted enrichment in deleterious mutations, in these regions, following repeated bouts of selection, such as during domestication (Lu et al. 2006; Haddrill et al. 2007; Morrell et al. 2011; Mezmouk and Ross-Ibarra 2014). Yet, this has not been explicitly demonstrated in domesticated species, in part due to the extent of data required to show such a relationship, at least until the rise of next generation sequencing technology. The population genomic consequences of domestication, and in general artificial selection toward a new optimal fitness peak, are likely to be similar to the allele surfing effect described in the previous paragraph. Simply put, the combined effects of a reduction in effective population size and fast population growth during domestication can drag along nonadaptive mutations, especially those that arose during the process of domestication itself.

Few studies have explicitly examined the fate of deleterious mutations in domesticated species. Work in rice (Günther and Schmid 2010) has identified an excess of nonsynonymous sites in domesticated lines compared with wild ancestors, without explicit predictions about their fitness effects. Deleterious mutations in dogs (Cruz et al. 2008) and rice (Lu et al. 2006) have also been linked to the process of domestication. Yet, these early model studies only examined a small fraction of the genetic diversity present in either wild or domesticated individuals. In addition, new analytical methods (Adzhubei et al. 2010; Choi et al. 2012; Sim et al. 2012) now permit stronger inference regarding mutational effects. In this study, we set out to identify putative deleterious mutations and study their fate in a large panel of domesticated sunflowers and their wild relatives using transcriptome-wide data.

Common sunflower (*Helianthus annuus* L.), an annual, insect-pollinated plant, is considered one of the world's most important crops (Burke et al. 2002). Since the start of its domestication nearly 4,000 years ago in North America, it has been the focus of a vast amount of research both from an evolutionary and a crop improvement, domestication perspective (Harter et al. 2004; Smith 2006; Blackman et al. 2011). Although the domesticated sunflower and its wild progenitor are completely interfertile (Snow et al. 1998), several morphological and life history traits distinguish them (Burke et al. 2002). In short, domesticated sunflowers flower earlier, have a single large head that does not shatter, and have larger seeds compared with their wild, highly branched, counterpart (Burke et al. 2002). Following its initial domestication to primitive varieties (local domesticates called landraces), sunflowers were brought to Europe by naturalists in the sixteenth century (Putt 1997), and eventually to Russia where they underwent intensified crop improvement. Most modern cultivars (elite lines) trace their origin back to material from this period (Fernández-Martínez et al. 2009). Since then, the common modern selection practice of reintroducing wild alleles from intra or interspecific crosses into cultivars (Jarvis and Hodgkin 1999) has led to a significant fraction of the domesticated elite genome originating from other closely related species (Baute et al. 2015).

Given this long history of domestication followed by crop improvement in sunflowers, a certain number of plausible predictions regarding the fate of deleterious mutations in domesticated (landrace and elites) lines can be inferred. First, we expect genetic variability to be reduced due to the effect of domestication. Second, proportionally more deleterious mutations are expected in domesticated lines compared with wild individuals. Third, we expect variation in recombination rate along the sunflower genome to play a crucial role in modulating the efficacy of selection. As such, regions of the genome with an elevated deleterious mutation load should coincide with low recombining regions. Finally, as we predict that the accumulation of deleterious mutations during domestication is a general phenomenon, not due to idiosyncrasies of sunflower domestication, we expect patterns to hold true in other related domesticated species.

Results

We analyzed transcriptome (RNAseq) data for 16 wild, 9 landraces, 10 elites, and 5 weedy *H. annuus* accessions comprising a total of 568 million Illumina paired-end (100 bp) reads. After aligning reads and parsing files according to specific quality thresholds (see methods), we detected 485,217 polymorphic sites (single nucleotide polymorphisms [SNPs]) in 25,112 genes of the *H. annuus* transcriptome. Result files were

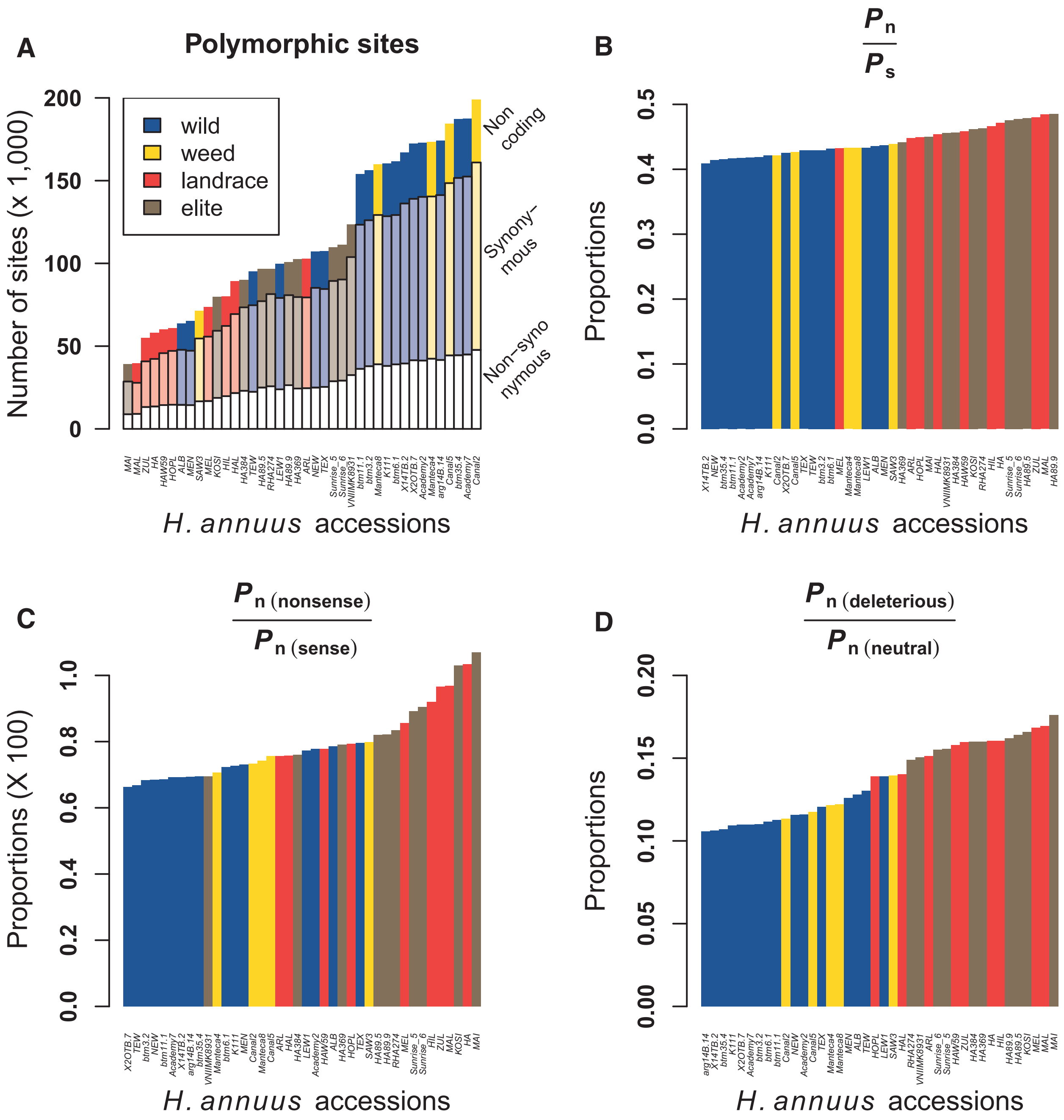

Fig. 1. (*A*) The number of coding synonymous, coding nonsynonymous, and noncoding mutations per individual. (*B*) The number of nonsynonymous mutations (P_n) divided by the number of synonymous mutations (P_s) per individual. (*C*) The number of nonsynonymous nonsense (i.e., alternative STOP codon, $P_{n(nonsense)}$) mutations divided by the number of sense nonsynonymous mutations ($P_{n(sense)}$) per individual. (D) The number of nonsynonymous deleterious mutations ($P_{n(deleterious)}$) divided by the number of neutral nonsynonymous mutations ($P_{n(neutral)}$) per individual. PROVEAN cutoff value <-2.5 for deleterious sites (see supplementary fig. S2, Supplementary Material online, for alternative cutoff values).

deposited in the Dryad Digital Repository (dx.doi.org/10.5061/dryad.8s459).

We recognize that our final data set likely contains a small fraction of false positives due to alignment and/or sequencing errors. Yet, given the large amount of data, high overall coverage, different quality threshold cut-offs tested, and visual inspection of a random subset of alignments, we are confident in the general patterns we observed here. In addition, it should be noted that the goal of this study is to uncover overarching principles about the genomic consequences of domestication rather than identify individual mutations responsible for specific traits.

As predicted, the total number of polymorphic sites varied greatly among classes (fig. 1*A*, one-way analysis of variance: $F_{3,36} = 11.2$, $P = 2 \times 10^{-5}$), with wild individuals containing the greatest number of sites, followed by elite and landrace individuals (post hoc Tukey tests, elite vs. landrace, $P = 0.28$; elite vs. wild, $P = 0.02$; landrace vs. wild, $P = 0.0002$). For their part, individual weed lines differed in their number of polymorphic sites, with individuals from California (*Canal2*, *Canal5*, *Manteca4*, and *Manteca8*) among the most polymorphic ones, while the sole weed representative from outside its North American native range (Australia, *SAW3*) among the least (fig. 1*A*). Among all polymorphic sites, 20% of sites contained noncoding mutations, 55% synonymous mutations, and 24% nonsynonymous mutations.

When looking at the total number of nonsynonymous versus synonymous mutations (P_n/P_s) per individual, the patterns in figure 1*A* were reversed (fig. 1*B*, beta regression: $Z_{(df=3)} = -9.2$, $P < 2 \times 10^{-16}$). Domesticated lines showed the greatest proportion of nonsynonymous mutations (Wilcoxon rank-sum test, elite vs. landrace: $W = 38$, $P = 0.6$; elite vs. wild: $W = 160$, $P = 4 \times 10^{-7}$; landrace vs. wild: $W = 141$, $P = 7 \times 10^{-6}$). Among nonsynonymous mutations, ~0.2% of those represented nonsense (alternative STOP codons) mutations, and there was an excess in their proportion compared with all nonsynonymous mutations in the domesticated lines (fig. 1*C*, beta regression: $Z_{(df=3)} = -5.2$, $P < 1.9 \times 10^{-7}$; Wilcoxon rank-sum test, elite vs. landrace: $W = 46$, $P = 0.97$; elite vs. wild: $W = 148$, $P = 0.0001$; landrace vs. wild: $W = 133$, $P = 0.0001$). Finally, we used PROVEAN (**Pro**tein **V**ariation **E**ffect **An**alyzer; Choi et al. 2012) to bioinformatically predict the deleterious effect of nonsynonymous mutations. According to the default deleterious threshold value (−2.5) for PROVEAN, we identified 14% of nonsynonymous mutations as deleterious (fig. 1*D*) and this varied among classes of individuals (beta regression: $Z_{(df=3)} = -11.5$, $P < 2 \times 10^{-16}$). Similar to the patterns in figure 1*B* and *C*, the proportion of deleterious mutations was the greatest in domesticated lines, but again there was no difference between landrace and elite individuals (fig. 1*D*, Wilcoxon rank-sum test, elite vs. landrace: $W = 40$, $P = 0.72$; elite vs. wild: $W = 160$, $P = 4 \times 10^{-7}$; landrace vs. wild: $W = 160$, $P = 4 \times 10^{-7}$). In addition, we calculated proportions as in figure 1*B*–*D* using nucleotide diversity (π) instead of actual count data and present these in supplementary figure S1, Supplementary Material online.

We also performed analyses of deleterious mutations using different parameters in order to verify the robustness of the results presented here. First, we used a range of threshold values (more tolerant or more stringent; supplementary fig. S2, Supplementary Material online) to identify deleterious mutations. Second, because landrace lines were sequenced at an earlier date than many of the wild and elite genotypes, read depth often was lower, leading to proportionally more missing data ($F_{(3,36)} = 5.8$, $P = 0.002$, mean fraction of missing data per SNP: Wild = 10%, weed = 19%, landrace = 33%, elite = 14%). As such, we performed analyses using a range of missing data thresholds (i.e., keeping only polymorphic sites with 1%, 5%, 10%, or 20% missing data per SNP; supplementary fig. S2, Supplementary Material online). Results were similar based on these different PROVEAN (supplementary fig. S2, Supplementary Material online) and missing data thresholds (supplementary fig. S3, Supplementary Material online). Third, we also identified deleterious mutations using another frequently used approach (SIFT; Ng and Henikoff 2003; Sim et al. 2012). Although the identity of the deleterious mutations identified by either SIFT or PROVEAN sometimes varied, given that the methods rely on differ assumptions, overall results were qualitatively similar, and the ranking of lines by $P_{n(\text{deleterious})}/P_{n(\text{neutral})}$ was nearly identical using either PROVEAN (fig. 1*D*) or SIFT (supplementary fig. S4, Supplementary Material online). In addition, mutations identified as deleterious by SIFT had a significantly lower PROVEAN score (more deleterious) than the ones identified as tolerated (i.e., neutral) by SIFT (supplementary fig. S5, Supplementary Material online).

Next, we quantified the change in the proportions of deleterious mutations for private mutations (i.e., mutations that were found exclusively in wild or domesticated lines). Note that weed lines were excluded from this analysis given that their evolutionary origin is variable (weed individuals are introgressed to varying extent with domesticated germplasm; Baack et al. 2008; Kane and Rieseberg 2008). In the wild lines, 40% of mutations were private, compared with 16% for the domesticated lines. Restricting our analyses to private mutations, we found that the patterns identified in figure 1*D* were more pronounced in all domesticated lines ($\chi^2_{(df=1)}$ tests, $P \ll 0.001$), but not in wild lines (fig. 2, where colored bars as in fig. 1*D* are superimposed with relative frequencies for private mutations exclusively).

Nonsynonymous deleterious mutations segregate at a lower frequency compared with nonsynonymous neutral mutations in both wild and weed lines (fig. 3*A* and *B*; Kolmogorov–Smirnov test, $P < 2.2 \times 10^{-16}$). Conversely, this pattern was reversed in the landrace (fig. 3*C*; Kolmogorov–Smirnov test, $P = 2 \times 10^{-8}$) and elite lines (fig. 3*D*; Kolmogorov–Smirnov test, $P = 6 \times 10^{-8}$). PROVEAN score (deleterious effect) was also positively correlated with allele frequency in the wild (i.e., more deleterious alleles are also rarer, Spearman rho, $P < 2 \times 10^{-16}$; supplementary fig. S6, Supplementary Material online).

We identified 131 putative regions of the *H. annuus* genome showing a significant excess of deleterious mutations (at $P < 0.05$ significance). These regions had a lower mean (1.15 centiMorgans [cM]/Megabase [MB]) and median (0.25 cM/MB) recombination rate than the balance of the genome (table 1; mean = 3.41 cM/MB, *t*-test, $P = 1.6 \times 10^{-6}$ and median = 0.42 cM/MB, Wilcoxon rank-sum test, $P = 2.5 \times 10^{-15}$). As an illustration of this genome-wide trend, there was a large excess of deleterious mutations in the low recombining regions of linkage group 10 (here, a putative centromeric region of the chromosome; fig. 4).

Finally, we also performed the analyses depicted in figure 1 in two other Compositae for which RNAseq data were available on National Center for Biotechnology Information (NCBI) Sequence Read Archive for at least one wild individual from the native range in addition to several domesticated individuals (table 2). Both cardoon and globe artichoke showed the expected pattern: A significant increase in the load of deleterious mutations in the domesticated lines compared with wild relatives (χ^2 test, $P \ll 0.01$; table 2, see also supplementary figs. S7 and S8, Supplementary Material online for details).

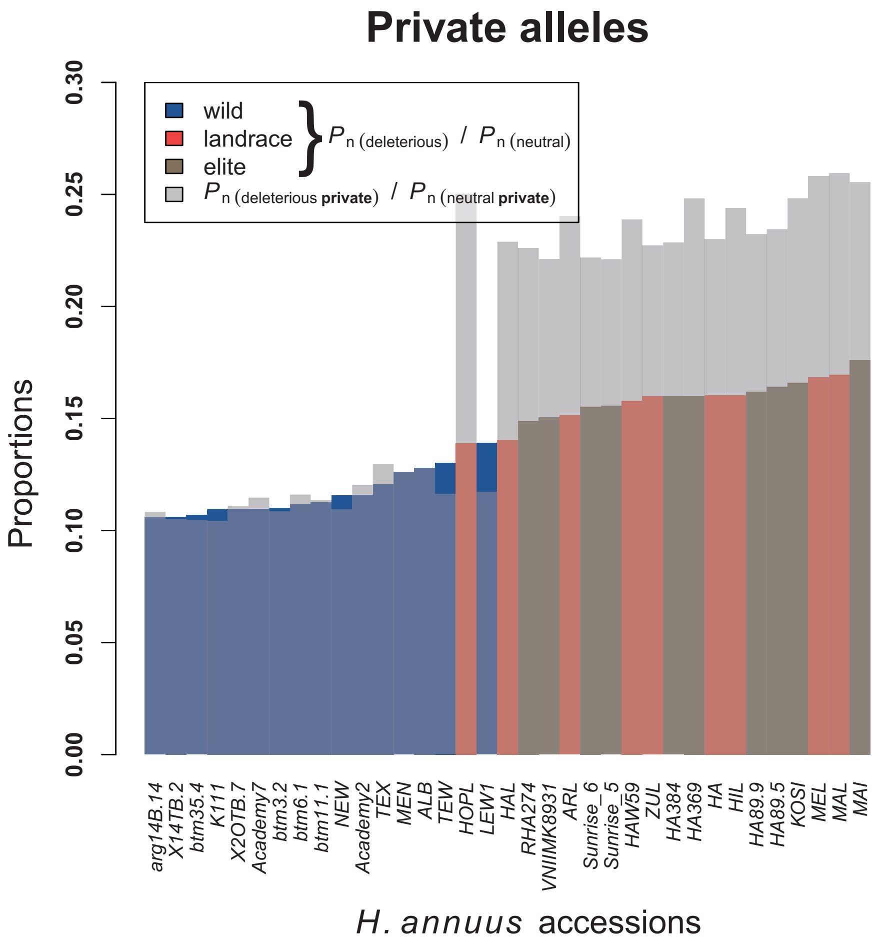

Fig. 2. Changes in the frequency of deleterious mutations when restricting analyses to mutually exclusive mutations (i.e., private mutations found in either the wild or domesticated lines). All changes in frequency between gray (private) and colored (private and shared) bars are significant in the domesticated lines (χ^2 tests, $P \ll 0.001$).

Discussion

Accumulation of Deleterious Mutations during Domestication

Under mutation–selection balance, deleterious mutations are expected to be held at low frequency in large sexual populations (Hartl and Clark 1997). However, during domestication, this balance may be disrupted and deleterious mutations may rise in frequency, despite their negative effects on fitness. Thus, plant and animal domestication may have previously unforeseen impacts on genome evolution that may in turn have phenotypic repercussions (Lu et al. 2006; Cruz et al. 2008; Günther and Schmid 2010). Here, we show that the distribution of different kinds of mutations in wild and domesticated sunflowers is largely consistent with an excess of deleterious mutations arising as a consequence of domestication over the last 4,000 years. In wild ancestral populations, there exists a large amount of genetic variability and predictably artificial selection has greatly reduced this variability, by as much as 80% in domesticated lines (fig. 1A), confirming previous results (Blackman et al. 2011). The novelty of the current work is to show that although only a fraction of the variability remains, nonsynonymous (fig. 1B), nonsense (fig. 1C), and more specifically deleterious mutations (fig. 1D) are enriched in these domesticated lines. Finally, we also present evidence that low recombining regions of the genome harbor a greater load of deleterious mutations than expected, as predicted by theory.

In addition, the other Compositae crops we analyzed appear to show the same general deleterious genomic effect of domestication. Admittedly, the paucity of publicly available data for these crops and the use of a single wild variety make any strong conclusion speculative. Clearly, there is a need to explore these questions using a more thorough sampling design.

A caveat of our current work resides in the fact that the deleterious effect of a mutation is based on bioinformatic predictions. Although this approach is widely used and has been shown to be highly effective in identifying different types of mutations (Ng and Henikoff 2003; Adzhubei et al. 2010; Choi et al. 2012; Sim et al. 2012), not all mutations identified through this approach will necessarily have a deleterious effect. Mutations annotated as deleterious could also be locally adaptive under certain conditions and need to be better characterized. In this study, we set out to uncover overarching principles about the genomic consequences of

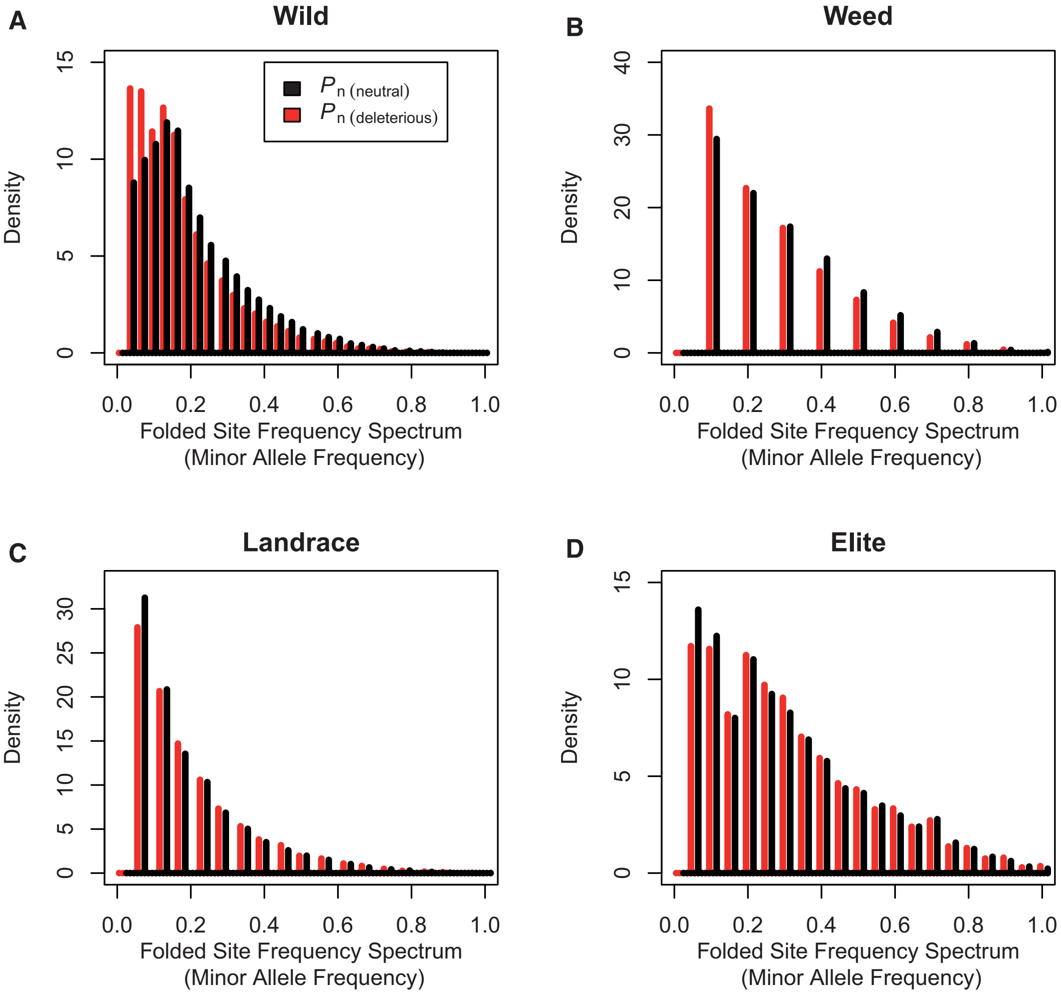

Fig. 3. Folded site frequency spectra for nonsynonymous neutral ($P_{n(neutral)}$, black) and nonsynonymous deleterious ($P_{n(deleterious)}$, red) mutations for each of the four classes (wild, weed, landrace, and elite lines). Note here that minor allele frequencies were calculated compared with the major allele in the reannotated reference transcriptome, as described in the Methods section.

domestication rather than identify individual mutations responsible for specific traits and their effect on fitness.

In addition, our current approach does not explicitly account for the ancestral state of a particular mutation. Recent work in human population genetics (Simons et al. 2014; Do et al. 2015) indicates that the algorithms predicting the functional effects of deleterious variants are dependent on the ancestral/derived state of the reference allele (i.e., if the reference carries the derived allele, it is more likely to be classified as benign than if it carries the ancestral allele). In the future, this will have to be explicitly accounted for by carefully choosing an outgroup and by employing approaches that predict functional effects in a way that is independent of the ancestral/derived status of the reference, such as in Simons et al. (2014) and Do et al. (2015). Nevertheless, this bias should not affect the well-known pattern of reduction in genetic diversity in domesticated lines (fig. 1*A*, supplementary fig. S1*A*, Supplementary Material online, and Blackman et al. 2011), nor should it affect P_n/P_s ratios in figure 1*B*. Finally, in figure 1*C*, the fact that the reference is closely related to the domesticated lines should reduce the number of nonsense mutations identified in the domesticated lines (i.e., the open reading frames [ORFs] in the domesticated lines should be more similar to the reference than to the more distantly related wild lines), thus making our estimates conservative.

No Difference between Landrace and Elite Lines

Intriguingly, we did not observe a difference between landrace and elite lines, even though the latter have gone through a second stage of domestication (i.e., improvement). If anything, it appears that landraces have a greater load of deleterious mutations than elite lines (fig. 1*D*). Although a greater load of deleterious mutations may have been predicted in elite compared with landrace lines due to the increased artificial selection pressure, there are several factors that may explain this pattern. Modern selection practices (e.g., the practice of reintroducing wild alleles from intra or interspecific crosses into cultivars; Jarvis and Hodgkin 1999) may effectively lead to the removal of deleterious alleles, thus counteracting the deleterious effects of domestication (Baute et al. 2015). This would also explain why elite lines harbor more variable mutations than landrace lines (fig. 1*A*). In addition, the maintenance in seed banks of landrace lines collected as long as 60 years ago implies that these must have

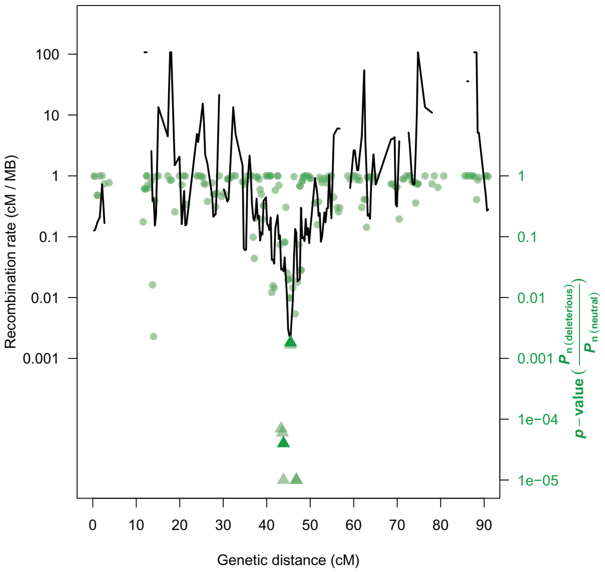

Fig. 4. Chromosome (linkage group) 10. Example of relationship between recombination rate (left *y*-axis, cM/MB) and deleterious load (significance *P* value for $P_{n(\text{deleterious})}/P_{n(\text{neutral})}$ evaluated per 1 cM window compared with the balance of the genome). Triangles correspond to *P* values significant after *q*-value correction (*q*-value < 0.05).

been grown and propagated more than a dozen times in order to keep viable seeds (sunflower seed viability starts dropping rapidly after 4–5 years). Such a process might allow deleterious mutations to accumulate. Yet given the small number of generations since seed banks have been established, it is unlikely to play a large role here. Clearly, domestication is a multifaceted process that can have unforeseen consequences on genome evolution.

Patterns in Weed Lines

Although domesticated plants are usually selected for a particularly well-defined suite of characters (Burke et al. 2002), weed populations may have a different evolutionary history depending on their location and level of weediness. In the United States, weed and wild sunflower populations are known to continuously exchange genes (Kane and Rieseberg 2008). As a consequence, the California weed sunflowers resemble wild populations present in the area, with a large number of total mutations (fig. 1*A*), but a low ratio of deleterious ones (fig. 1*D*). In contrast, the weed representative from Australia (SAW3) is unlikely to have hybridized with wild populations (sunflowers are native to North America), which would explain its closer resemblance to domesticated lines in terms of lower number of mutations, but higher proportion of deleterious ones (fig. 1*A* and *D*).

Origin of Deleterious Mutations

Given the relatively recent history of sunflower domestication compared with its divergence from its sister taxon *Helianthus argophyllus* (~1 Ma; Kane et al. 2009), it is probable that a substantial fraction of mutations in domesticated lines has arisen from the genome-wide accumulation of wild standing variation (Lu et al. 2006). Indeed, more than 80% of the mutations identified in the domesticated lines are also present in the wild populations. Yet, in addition to standing genetic variation, it appears that mutations that arose more recently, during the process of domestication itself, may be preferentially enriched for deleterious alleles. As such, when we restricted analyses to mutations that were mutually exclusive (i.e., private to either the wild or the domesticated lines), the proportion of deleterious mutations sharply increased in domesticated lines (fig. 2). These results are in agreement with the expansion load model, which predicts that private de novo mutations arising during the phase of population expansion itself are more often deleterious than rare genetic variants already present in the source population (Peischl et al. 2013). Alternatively, frequent bottlenecks during domestication could increase the frequency of deleterious alleles already present in the wild, but too rare to be detected with the current sampling design.

Table 1. Recombination Rate in Regions with Significantly Elevated Load of Deleterious Mutations Compared with the Balance of the Genome.

	Recombination Rate (cM/MB)		
	Regions with elevated load (131 regions)	Balance of the genome	*P* values
Mean	1.15	3.41	*t*-test, 2.53×10^{-15}
Median	0.25	0.42	Wilcoxon test, 1.6×10^{-6}

Variation in Recombination Rate

The exact phenotypic and fitness consequences of an increase in load of deleterious mutations remain to be explicitly tested. In the future, crop breeding programs may wish to focus on removing these mutations (or complementing them via hybridization) to further improve yields. The fact that these mutations tend to accumulate in low recombining regions of the genome (fig. 4 and table 2, but see Mezmouk and Ross-Ibarra 2014) will render this task more difficult. Indeed, it is precisely because the efficacy of selection (either natural or artificial) is reduced in low recombining regions of the genome that these mutations have not been purged from the genome despite thousands of years of domestication (McMullen et al. 2009). Nevertheless, there are several avenues of research which could be explored to specifically target these recalcitrant regions, including genome editing (Perez-Pinera et al. 2012), recombination rate modifiers (Li et al. 2007), or mining the diversity of wild relatives (McCouch et al. 2013).

Conclusion

When a desired genome is selected for propagation, all mutations, beneficial, neutral, or deleterious, shift in frequency, and this sometimes can have unforeseen consequences. Artificial selection and the population genetic environment in which it is performed can thus interfere with natural selection. As such, there appears to be a genetic cost to domestication

Table 2. Load of Deleterious Mutations for Two Other Closely Related Compositae.

Scientific Name	Common Name	Wild/Domesticated	$P_{n\ (\mathrm{total})}$	$P_{n\ (\mathrm{deleterious})}$	$P_{n\ (\mathrm{deleterious})}/P_{n\ (\mathrm{neutral})}$	χ^2 test (χ^2 statistic, df, *P*-value)[a]
Cynara carduncului var. sylvestris	Wild cardoon/globe artichoke	Wild (Sicily, core population)	12,069	1,900	0.187	$\chi^2 = 34$, df = 2, 3×10^{-8}
C. carduncului var. altilis	Cardoon	Domesticated	7,786	1,458	0.23	
C.carduncului var. altilis	Cardoon	Domesticated	6,844	1,239	0.221	
C. carduncului var. sylvestris	Wild cardoon/globe artichoke	Wild (Sicily, core population)	11,344	1,465	0.146	$\chi^2 = 113$, df = 5, 2×10^{-16}
C. carduncului var. scolymus	Globe artichoke	Domesticated	12,868	1,832	0.166	
C. carduncului var. scolymus	Globe artichoke	Domesticated	11,847	1,889	0.19	
C. carduncului var. scolymus	Globe artichoke	Domesticated	10,521	1,781	0.204	
C. carduncului var. scolymus	Globe artichoke	Domesticated	10,320	1,648	0.19	
C. carduncului var. scolymus	Globe artichoke	Domesticated	9,143	1,537	0.202	

NOTE.—The wild cardoon/globe artichoke samples (*C. carduncului* var. sylvestris) are the same individual as domesticated cardoon and globe artichoke are thought to have arisen from the same wild cardoon species (Scaglione et al. 2012). Numbers differ slightly for *C. carduncului* var. sylvestris because quality thresholds are based on a different number of individuals for the cardoon (3) and globe artichoke (6) comparisons.

[a] χ^2 test comparing the ratios $P_{n\ (\mathrm{deleterious})}/P_{n\ (\mathrm{neutral})}$ in wild (expected) versus domesticated (observed) group.

and our analyses suggest that a significantly biased proportion of amino acid substitutions occurring during domestication are deleterious, especially in regions of the genome with low levels of recombination. Although challenging, the removal or complementation of these deleterious mutations represents a well-defined target of future crop improvement efforts.

Methods

We analyzed transcriptome (RNAseq) data for 16 wild, 9 landraces, 10 elites, and 5 weed accessions (supplementary table S1, Supplementary Material online). All reads were sequenced on an Illumina (San Diego, CA) GAII or HiSeq next-generation sequencing platform (paired-end reads, 2 × 100 bp, non-normalized libraries). Note that wild and weed individuals were sequenced as part of an interspecific study on genomic islands of divergence in wild sunflowers and are reported in detail in Renaut, Grassa, Yeaman, Moyers, et al. (2013). Transcriptomes for landrace and elite lines are described in Baute et al. (2015). Sequences are publicly available on NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/ [last accessed May 7, 2015], see supplementary table S1, Supplementary Material online, for details). Once sequencing files were acquired, all further analyses were performed using custom R (version 3.0.2, R Core Team 2013) scripts, publicly available on GitHub (https://github.com/seb951/helianthus_deleterious_domestication, last accessed May 7, 2015).

Alignments and Variant Calling

Reads were aligned against a reference transcriptome using the Burrows–Wheeler Aligner (BWA V.0.7.5A-R405, ALN and SAMPE commands; Li and Durbin 2009). Following alignments, we used the Indel Realigner from the genome analysis toolkit (GATK; McKenna et al. 2010) to correct alignment errors near indels. The reference data set consisting of 51,468 contigs (51.3 Mbp) is available on DRYAD (Renaut, Grassa, Yeaman S, Lai, et al. 2013) and described in a previous publication (Renaut, Grassa, Yeaman, Moyers, et al. 2013). Briefly, it was generated by sequencing four *H. annuus* libraries prepared from plants grown under different environmental conditions and assembled de novo using TRINITY (Grabherr et al. 2011). These plants came from a single elite individual (HA412), which is the focus of the sunflower genome project (Kane et al. 2011). It was excluded from the current analysis, given that libraries were composed of a mix of tissues different from all other sunflower individuals analyzed here (young leaf tissue). SAMTOOLS v.0.1.19-44428cd (MPILEUP, BCFTOOLS; Li et al. 2009) was used to call SNPs. Variant Calling Format files (.vcf files) were then parsed in order to remove sites according to specific quality thresholds. These thresholds were determined based on previous experience with similar RNAseq data sets in sunflowers (Renaut et al. 2012; Renaut, Grassa, Yeaman, Moyers, et al. 2013). First, phred-scaled genotype likelihoods below 15 (which corresponds to a genotyping accuracy of at least 95%) were considered as missing. Then, sites missing genotypes in more than 50% of all individuals sequenced were removed. Note that different missing data threshold (1%, 5%, 10%, 20%) was also tested and results are reported in supplementary figure S3, Supplementary Material online. SNPs with low expected heterozygosity ($H_e < 0.095$, or minor allele frequency <5%) were removed, given that they likely represent sequencing errors. SNPs with high observed heterozygosity ($H_o > 0.6$, or more than 60% of individuals heterozygous) were removed because they likely represent paralogous sequence variants. Once a list of high-quality SNPs had been determined, the reference transcriptome was reannotated at polymorphic loci using a "majority rule" consensus. Accordingly, for polymorphic sites, the transcriptome was modified with this major allele, and this reannotated transcriptome was used as the reference in the analyses of deleterious variants as described further down. This step is necessary in order to avoid the inherent bias of the transcriptome toward domesticated lines (the reference was built from an elite individual). In addition, we searched for deleterious variants after reannotating the transcriptome using exclusively domesticated lines or exclusively wild lines to assess reference bias. In both cases, overall results were quantitatively similar to the ones reported here (data not shown). Finally, we calculated minor allele frequency per line (wild, weed, landrace, and elite) separately in order to identify which mutations were private to wild and domesticated (landrace and elite) lines.

Protein Coding Evolution

For analyses of protein evolution, ORFs were identified from our reference transcriptome using the program GETORF in European Molecular Biology Open Software Suite (Rice et al. 2000). We showed previously that this approach works well at identifying coding regions in sunflowers (Renaut, Grassa, Yeaman, Moyers, et al. 2013). The longest open-ended ORF (minimum length of 300 nt) was kept as the most probable translated region of the gene. On the basis of these ORFs, every SNP was then be classified as "noncoding," "coding synonymous," "coding nonsynonymous," and "nonsense" (alternate STOP codon) for each individual. Nonsynonymous SNPs were then further categorized as "neutral" or "deleterious" based on bioinformatics predictions (see below).

Predicting Effect of Deleterious Mutations

We used a new bioinformatics tool, PROVEAN v.1.1.4 (Choi et al. 2012), in order to predict the effect of deleterious mutations for the subset of nonsynonymous SNPs. This alignment-based method first identifies closely related protein sequence homologs through position-specific iterated BLAST (PSI-BLAST) searches. It then compares alignment scores between the reference and its homologs before and after the introduction of an amino acid variation in the query sequence. Here, protein sequence homolog searches were done against NCBI nr database, but PROVEAN source code was modified to restrict searches to green plants (Viridiplantae). This bioinformatics approach has been shown to perform well in separating disease-associated variants from common polymorphisms in human protein variations (Choi et al. 2012), compared with other common bioinformatics predictors used mainly in human genetics such as POLYPHEN-2 (Adzhubei et al. 2010). Finally, it can be implemented easily for any species, as long as a set of reference proteins (here, the longest ORF) and a list of variable sites are available. To confirm our results, we used another program (SIFT; Ng and Henikoff 2003; Sim et al. 2012), which is frequently used to study the role of deleterious mutations in human diseases and tends to be more sensitive, but less specific than PROVEAN (Choi et al. 2012). Note however that our current approach does not explicitly account for the ancestral state of an allele. Recent work (Simons et al. 2014; Do et al. 2015) has shown that without such knowledge, the functional effect of nonsynonymous changes may be biased in a way that is dependent on the ancestral/derived state of the reference allele.

To determine the relative diversity of the four classes of sunflower accessions (wild, weed, landrace, and elites), we also calculated nucleotide diversity (π, calculated in SITES, Hey Lab Distributed Software, http://genfaculty.rutgers.edu/hey/%E2%80%A8software#SITES/, last accessed May 7, 2015) per class and for synonymous, nonsynonymous, nonsense, and nonsynonymous deleterious mutations separately. We then calculated the same ratios as presented in figure 1B–D, but using nucleotide diversity (π) instead of actual count data, and present these results in supplementary figure S1, Supplementary Material online.

Using previously calculated recombination rates based on physical and genetic map integration (Renaut, Grassa, Yeaman, Moyers, et al. 2013), we also tested whether regions of the sunflower genome which harbor an excess of deleterious mutations tended to recombine less often, as would be predicted by theory (Morrell et al. 2011). Regions of the genome showing an excess of deleterious mutations were first identified through a sliding window analysis and significance tested through a resampling approach similar to the one described in Hohenlohe et al. (2009) and Renaut, Grassa, Yeaman, Moyers, et al. (2013). Briefly, in sliding windows of 1 cM, we calculated the proportion of deleterious over neutral nonsynonymous mutations ($P_{\mathrm{n\ (deleterious)}}$ /$P_{\mathrm{n\ (neutral)}}$). To assess significance in each window, we randomly sampled with replacement from across the genome the same number of markers present in that window and recalculated the $P_{\mathrm{n\ (deleterious)}}$ /$P_{\mathrm{n\ (neutral)}}$ proportion. This was done 100,000 times per window and thus provides a null distribution of expected values for each genomic region, accounting for the number of markers. Significance (P) values therefore represent the fraction of the distribution exceeding the expected value. Adjacent windows with significant P values represent a single region with an excess of deleterious mutations.

Accumulation of Deleterious Mutations in Other Compositae

We expanded our analyses to other closely related domesticated species for which full transcriptome data were available publicly (Scaglione et al. 2012) in order to confirm that the rise in the proportion of deleterious mutations during domestication was not due to idiosyncrasies unique to the common sunflower. On the basis of literature searches, we identified two species of Compositae where RNAseq data was available for at least one wild individual from the native range and several domesticated individuals: Cardoon—*Cynara cardunculus* var. sylvestris (wild) compared with *C. cardunculus* var. altilis (domesticated), and globe artichoke—*C. cardunculus* var. sylvestris (wild) compared with *C. cardunculus* var. scolymus (domesticated). For other Compositae crops, transcriptome sequences either were not available from the native range or the sequencing error rate was too high and/or sequence depth too low for the planned analyses (Barker et al. 2008; Hodgins et al. 2014). Once we contacted the authors to verify the origin of the *Cynara* samples, we downloaded data files from NCBI Sequence Read Archive. Both these species have also been domesticated (Scaglione et al. 2012), albeit not to the same extent as sunflowers (Dempewolf et al. 2008). We ran our bioinformatics pipeline to align sequences, call polymorphic sites, and identify nonsynonymous and deleterious mutations for each of the two *Cynara* species. Note that here, the reference transcriptome was constructed from two domesticated and a wild individual as described in Scaglione et al. (2012).

Supplementary Material

Supplementary table S1 and figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank G.J. Baute for helpful discussion about sunflower domestication and two anonymous reviewers for insightful suggestions. We thank Y. Choi for help with running PROVEAN analyses. This work was supported by funding from the National Science Foundation Plant Genome Research Program (DBI-0820451), Genome Canada, and Genome BC.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* 7:248–249.

Baack EJ, Sapir Y, Chapman MA, Burke JM, Rieseberg LH. 2008. Selection on domestication traits and quantitative trait loci in crop-wild sunflower hybrids. *Mol Ecol.* 17:666–677.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol.* 25:2445–2455.

Baute GJ, Kane NC, Grassa CJ, Lai Z, Rieseberg LH. 2015. Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytol.* 206:830–838.

Blackman BK, Scascitelli M, Kane NC, Luton HH, Rasmussen DA, Bye RA, Lentz DL, Rieseberg LH. 2011. Sunflower domestication alleles support single domestication center in eastern North America. *Proc Natl Acad Sci U S A.* 108:14360–14365.

Burke JM, Tang S, Knapp SJ, Rieseberg LH. 2002. Genetic analysis of sunflower domestication. *Genetics* 161:1257–1267.

Charlesworth D, Morgan MT, Charlesworth B. 1993. Mutation accumulation in finite populations. *J Hered.* 84:321–325.

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.

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

Dempewolf H, Rieseberg LH, Cronk QC. 2008. Crop domestication in the Compositae: a family-wide trait assessment. *Genet Resour Crop Evol.* 55:1141–1157.

Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat Genet.* 47:126–131.

Edmonds CA, Lillie AS, Cavalli-Sforza LL. 2004. Mutations arising in the wave front of an expanding population. *Proc Natl Acad Sci U S A.* 101:975–979.

Excoffier L, Foll M, Petit RJ. 2009. Genetic consequences of range expansions. *Annu Rev Ecol Evol Syst.* 40:481–501.

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

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

Fernández-Martínez JM, Pérez-Vich B, Velasco L. 2009. Sunflower. In: Vollmann J, Rajcan I, editors. Oil crops. New York: Springer. p. 155–232.

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, et al. 2012. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.

Günther T, Schmid KJ. 2010. Deleterious amino acid polymorphisms in *Arabidopsis thaliana* and rice. *Theor Appl Genet.* 121:157–168.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* 8:R18.

Harter AV, Gardner KA, Falush D, Lentz DL, Bye RA, Rieseberg LH. 2004. Origin of extant domesticated sunflowers in eastern North America. *Nature* 430:201–205.

Hartl DL, Clark AG. 1997. Principles of population genetics. Sunderland (MA): Sinauer Associates.

Hodgins KA, Lai Z, Oliveira LO, Still DW, Scascitelli M, Barker M, Kane NC, Dempewolf H, Kozik A, Kesseli RV, et al. 2014. Genomics of Compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives. *Mol Ecol Res.* 14:166–177.

Hohenlohe P, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2009. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD Tags. *PLoS Genet.* 6:e1000862.

Jarvis DI, Hodgkin T. 1999. Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems. *Mol Ecol.* 8:S159–S173.

Kane NC, Gill N, King MG, Bowers JE, Berges H. 2011. Progress towards a reference genome for sunflower. *Botany* 89:429–437.

Kane NC, King MG, Barker MS, Raduski A, Karrenberg S, Yatabe Y, Knapp SJ, Rieseberg LH. 2009. Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* 63:2061–2075.

Kane NC, Rieseberg LH. 2008. Genetics and evolution of weedy *Helianthus annuus* populations: adaptation of an agricultural weed. *Mol Ecol.* 17:384–394.

Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. *Mol Biol Evol.* 23:482–490.

Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* 336:435–440.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Li J, Hsia A, Schnable P. 2007. Recent advances in plant recombination. *Curr Opin Plant Biol.* 10:131–135.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451:994–997.

Lotterhos KE, Whitlock MC. 2014. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol Ecol.* 23:2178–2192.

Lu J, Tang T, Tang H, Huang J, Shi S, Wu CI. 2006. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* 22:126–131.

McCouch S, Baute GJ, Bradeen J, Bramel P, Bretting PK, Buckler E, Burke JM, Charest D, Cloutier S, Cole G, et al. 2013. Agriculture: feeding the future. *Nature* 499:23–24.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, et al. 2009. Genetic properties of the maize nested association mapping population. *Science* 325:737–740.

Mezmouk S, Ross-Ibarra J. 2014. The pattern and distribution of deleterious mutations in maize. *G3* 4:163–171.

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

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

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst.* 23:263–286.

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

Perez-Pinera P, Ousterout DG, Gersbach CA. 2012. Advances in targeted genome editing. *Curr Opin Chem Biol.* 16:268–277.

Putt ED. 1997. Early history of sunflower. In: Schneiter AA, editor. Sunflower technology and production. Madison (WI): American Society of Agronomy/Crop Science Society of America/Soil Science Society of America. p. 1–19.

R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/.

Renaut S, Grassa C, Moyers B, Kane N, Rieseberg L. 2012. The population genomics of sunflowers and genomic determinants of protein evolution revealed by RNAseq. *Biology* 1:575–596.

Renaut S, Grassa CJ, Yeaman S, Lai Z, Kane NK, Moyers BT, Bowers JE, Burke JM, Rieseberg LH. 2013. Data from: genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun. Available from: http://dx.doi.org/10.5061/dryad.9q1n4.

Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat Commun.* 4:1827.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16:276–277.

Scaglione D, Lanteri S, Acquadro A, Lai Z, Knapp SJ, Rieseberg L, Portis E. 2012. Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotechnol J.* 10:956–969.

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

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40:W452–W457.

Smith BD. 2006. Eastern North America as an independent center of plant domestication. *Proc Natl Acad Sci U S A.* 103:12223–12228.

Snow A, Moran-Palma P, Rieseberg L, Wszelaki A, Seiler G. 1998. Fecundity, phenology, and seed dormancy of F1 wild-crop hybrids in Sunflower (*Helianthus annuus*, Asteraceae). *Am J Bot.* 85:794.