The Coding Loci of Evolution and Domestication: Current Knowledge and Implications for Bio-Inspired Genome Editing

Virginie Courtier-Orgogozo, Arnaud Martin

The Coding Loci of Evolution and Domestication: Current Knowledge and Implications for Bio-Inspired Genome Editing

Virginie Courtier-Orgogozo[1,*] and Arnaud Martin[2]

1: Institut Jacques Monod, CNRS, UMR 7592, Université Paris Diderot, Paris, France.
2: Department of Biological Sciences, The George Washington University, Washington, DC, USA.
* To whom correspondence should be addressed. Tel: +33 1 57278043; Fax: +33 1 57278087; Email: virginie.courtier@normalesup.org

Summary statement
We review >1200 identified coding mutations underlying domestication and natural evolution. This study uncovers knowledge biases and the prevalence of null mutations, and brings insights for successful genome editing.

Summary
One promising application of CRISPR/Cas9 is to create targeted mutations to introduce traits of interest into domesticated organisms. However, a major current limitation for crop and livestock improvement is to identify the precise genes and genetic changes that must be engineered to obtain traits of interest. Here we discuss the advantages of bio-inspired genome editing, *i.e.* the engineered introduction of natural mutations that have already been associated with traits of interest in other lineages (breeds, populations, or species). To get a landscape view of potential targets for genome editing, we used Gephebase (www.gephebase.org), a manually-curated database compiling published data about the genes responsible for evolutionary and domesticated changes across Eukaryotes, and examined the >1,200 mutations that have been identified in the coding regions of more than 700 genes in animals, plants and yeasts. We observe that our genetic knowledge is relatively important for certain traits, such as xenobiotic resistance, and poor for others. We also note that protein-null alleles, often due to nonsense and frameshift mutations, represent a large fraction of the known loci of domestication (42% of identified coding mutations), compared to intraspecific (27%) and interspecific evolution (11%). While this trend may be subject to detection, publication, and curation biases, it is consistent with the idea that breeders have selected large-effect mutations underlying adaptive traits in specific settings, but that these mutations and associated phenotypes would not survive the vagaries of changing external and internal environments. Our compilation of the loci of evolution and domestication uncovers interesting options for bio-inspired and transgene-free genome editing.

Abbreviations
CRISPR: Clustered regularly interspaced short palindromic repeats

Introduction

For almost 100 years, functional genetics has relied on the study of a small number of laboratory model organisms amenable to manipulation, or of domesticated species where

selected phenotypes could bring an entry point into the identification of the causal genes. While classical laboratory genetics mostly focused on artificial targeted modifications of genes (via gene inactivation or gene overexpression) to understand the effect of various genes on biological processes, quantitative genetic studies examined the segregating genetic loci associated with existing, observable variation between populations or between lineages that have been artificially selected and their wild counterparts (Martin and Orgogozo, 2013; Rockman, 2012; Stern, 2000). Both approaches have improved our understanding of the connection between genes and phenotypes (Orgogozo et al., 2015).

In parallel, the recent development of CRISPR/Cas9 technology is dramatically expanding the landscape of possibilities in the field of genetics, enabling the targeted editing of DNA in a growing number of organisms (this Special Issue of JEB). It now becomes possible to introduce a mutation of interest in any species, provided that the organisms can be maintained for some time in the laboratory to allow delivery of the molecules required for CRISPR/Cas9 genome editing. The new CRISPR/Cas9 technology is transformative in at least three major ways. First, it allows fundamental research in a higher number of organisms. Expanding the number of species studied in the laboratory increases not only the taxonomic breadth of current research but also the scope of biological problems that can be addressed. For instance, some of the natural null alleles responsible for cavefish depigmentation have been reproduced by CRISPR/Cas9 knock-out, allowing a refined study of the pleiotropic effects of this variation on the organism's physiology and behavior (Klaassen et al., 2018; Ma et al., 2015). In butterflies, the wing patterning roles of *WntA* and *optix* were first discovered in natural populations based on linkage mapping studies, before being re-explored by laboratory CRISPR/Cas9 gene knockouts in additional lineages, which uncovered unexpected functions for both genes (Mazo-Vargas et al., 2017; Zhang et al., 2017). Second, CRISPR/Cas9 facilitates the production of genetically modified animals for various biomedical purposes, such as disease models or tissue donors (Tan et al., 2016). Third, with CRISPR/Cas9 editing we are entering a new era for crop and livestock improvement. At the onset of domestication, early domestic populations were obtained by selection of desirable traits among a pool of wild individuals, and later on, random mutagenesis was sometimes used to increase allelic diversity. Varieties obtained via such breeding techniques often suffered from deleterious alleles that were carried over together with the alleles associated with desired traits (Moyers et al., 2017). For example, many rice varieties carry the *sd1* allele associated with high grain yield, and because this allele is linked to an allele in *qDTY1.1* which decreases yield under drought, these varieties are also highly sensitive to drought (Vikram et al., 2015). In the past few decades, we have witnessed a second phase of domestication, where genetic manipulation techniques of specific genes using physical, chemical and biological means (*e.g.* T-DNA insertion/transposons) have contributed to improve crop species (Ma et al., 2016). However, with these first-generation genetic engineering tools, the integrations of transgenes into host genomes were random with respect to the site of insertion into the genome, were sometimes unstable, were limited to insertions of new pieces of DNA (mutation at a targeted site was not possible), and raised public concern (Voytas and Gao, 2014).

The new CRISPR/Cas9 technology and related techniques offer several advantages compared to first-generation genetic tools: genetic engineering (defined here as the manipulation of an organism's genome via molecular tools) is cheaper and easier, many distinct types of mutations are now feasible (specific nucleotide substitutions with no introduction of foreign DNA, altering the expression of existing genes, inserting small fragments of DNA, etc.), it is also possible to mutate several genes all at once (Niu et al.,

2017), targeted mutations can be obtained in species for which genetics was so far very difficult/impossible, and mutations can be introduced directly in the breed of interest without going through multiple rounds of crosses. The range of traits and species that can be potentially manipulated with CRISPR and other second-generation genetic tools thus far exceeds the ones available with first-generation tools.

Nowadays a major obstacle for crop and livestock improvement is to identify the precise genes and genetic changes that must be engineered to obtain particular traits of interest. In other words, the key for understanding biodiversity and improving varieties is to find the link between phenotype and genotype (Hufford et al., 2019). To summarize the breadth of knowledge about the genes and the mutations responsible for natural and domesticated variation, we created the database Gephebase, available at gephebase.org (Courtier-Orgogozo et al., 2019; Martin and Courtier-Orgogozo, 2017). This database compiles published genotype-to-phenotype relationships, defined here as a sequence variation causing an observable trait variation (Orgogozo et al., 2015), in a variety of organisms. It excludes human and animal clinical traits that are catalogued in other databases, and focuses on non-deleterious phenotypes (*i.e.* presumably adaptive or neutral traits). It synthesizes a rich body of literature in genetics aiming at identifying the "Loci of Evolution" or "QTN - Quantitative Trait Nucleotide" (Hoekstra and Coyne, 2007; Martin and Orgogozo, 2013; Rockman, 2012; Stern and Orgogozo, 2008) that underlie the evolution of "desirable" traits that are maintained in the gene pool by natural or anthropogenic selection. Each entry in Gephebase (contraction of *Ge*notype-*phe*notype data*base*) corresponds to a phenotypic variation associated with one or several linked mutations in an identified gene, manually curated from the literature, and supported by strong experimental evidence (*eg.* linkage mapping, association mapping, functional validation). This comparative dataset is inherently biased by detection methods towards large-effect loci and does not reflect the complete spectrum of mutational effects and polygenic architectures that occur in nature, notably by a lack of power in identifying small effects (Rockman, 2012). For instance, most of the economically-relevant cattle traits have complex genetic architectures that include multiple genes of small effects (Georges et al., 2018), and mutations that are predominantly regulatory and non-coding (Boitard et al., 2016). Nonetheless, cataloguing and analyzing the current state of knowledge from the characterized large-effect mutations represents an important avenue of research for both comparative genetics and bioinspiration.

In recent years it has been observed that mutations in the same genes often produce the same type of trait changes in various species, sometimes across very distant taxonomic ranges (Martin and Orgogozo, 2013; Stern, 2013). Adaptation to starch-rich food has occurred independently in various species via comparable genetic changes, the duplication of *amylase* genes in humans, dogs, rats, mice and pigs (Pajic et al., 2019). Other striking examples are the amino-acid replacement mutations M918T and L925I which, among other sites of the sodium channel gene *para* (*syn. kdr*), confer resistance to a wide range of pyrethroid insecticides (Dong et al., 2014). These mutations have evolved independently in more than 50 species, including various pests and parasites such as mosquitoes, kissing bugs, headlice, bedbugs, varroa mites (Durand et al., 2012; González-Cabrera et al., 2016; Kapantaidaki et al., 2018; Sierra et al., 2016). This repeatability of specific amino-acid replacements in response to recent pesticide pressure is so extreme that amphipod crustaceans that inhabit rivers exposed to agricultural run-off also evolved M918T and L925I resistance alleles (Major et al., 2018). It thus makes sense to think that a mutation conferring a trait of interest in a given species is likely to produce a similar phenotypic effect in other, distantly-related species (Lenser and Theißen, 2013). In this respect, the Gephebase

dataset constitutes a great resource to identify the most promising gene targets for applied gene editing.

Compared to cis-regulatory regions, coding regions are generally more conserved and easier to delimitate. They thus are preferred genome sites to be edited and copied from one species to another. To examine potential targets for bio-inspired genome editing, we focus here on the coding mutations compiled in Gephebase and especially the ones that abolish gene function, since they are easier to obtain via CRISPR/Cas9. How common are natural gene knockout mutations in Gephebase? Does the proportion of null mutations differ between domestication and natural evolution? Our overview of current genetics knowledge brings original insights for designing genome editing experiments in crop and livestock species.

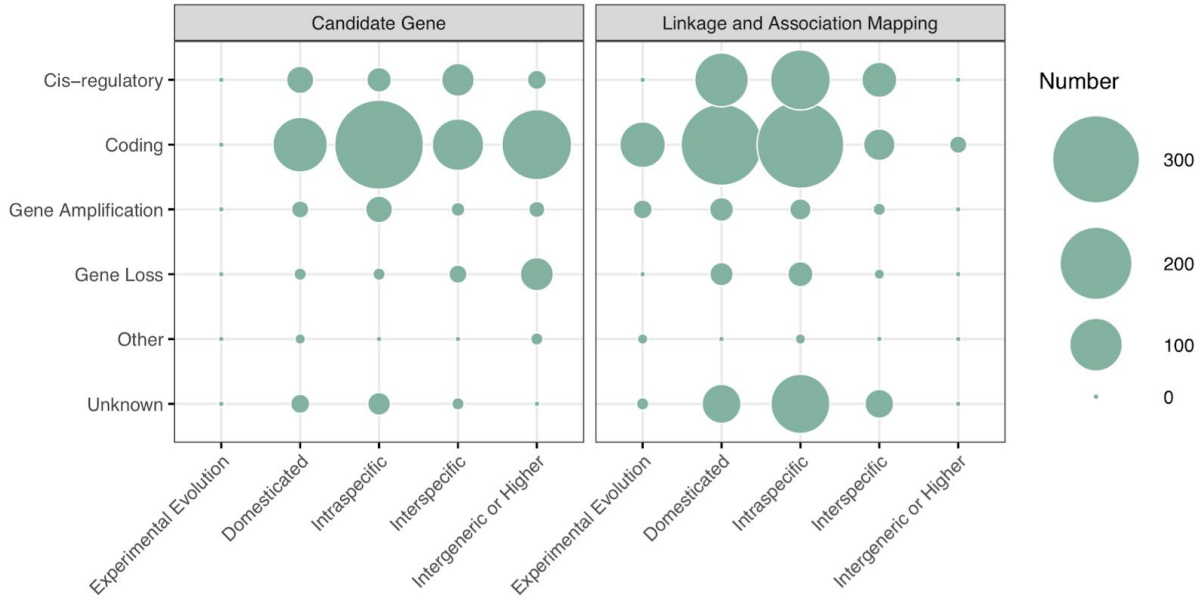

Fig. 1. Number of mutations according to Molecular Type and Taxonomic Distance. Mutations identified via a Candidate Gene approach are shown on the left and mutations identified by Linkage and Association Mapping are shown on the right. Circle areas are proportional to the number of mutations present in Gephebase in each category. “Other“ represent chimeric genes and supergenes. “Unknown” means that the locus has been narrowed down to a gene but that the type of mutation has not been determined.

Results

Distribution of the 1,281 coding changes present in Gephebase

As of May 2019, Gephebase contains 2,102 mutations, among which 1,281 are classified as coding (Fig. 1). These coding mutations affect the protein sequence of 1,147 genes in distinct lineages, corresponding to 390 gene families. Coding mutations are associated with 143 distinct phenotypic traits and the two most-represented traits, xenobiotic resistance and coloration, make up 543 coding mutations (Fig. 2A). About two thirds of the coding mutations compiled in Gephebase are involved in physiological variation, one third in morphological variation, and very few in behavioral variation (Fig. 2C). As expected, genetic model organisms gather a large proportion of accumulated data (Arnoult, 2014;

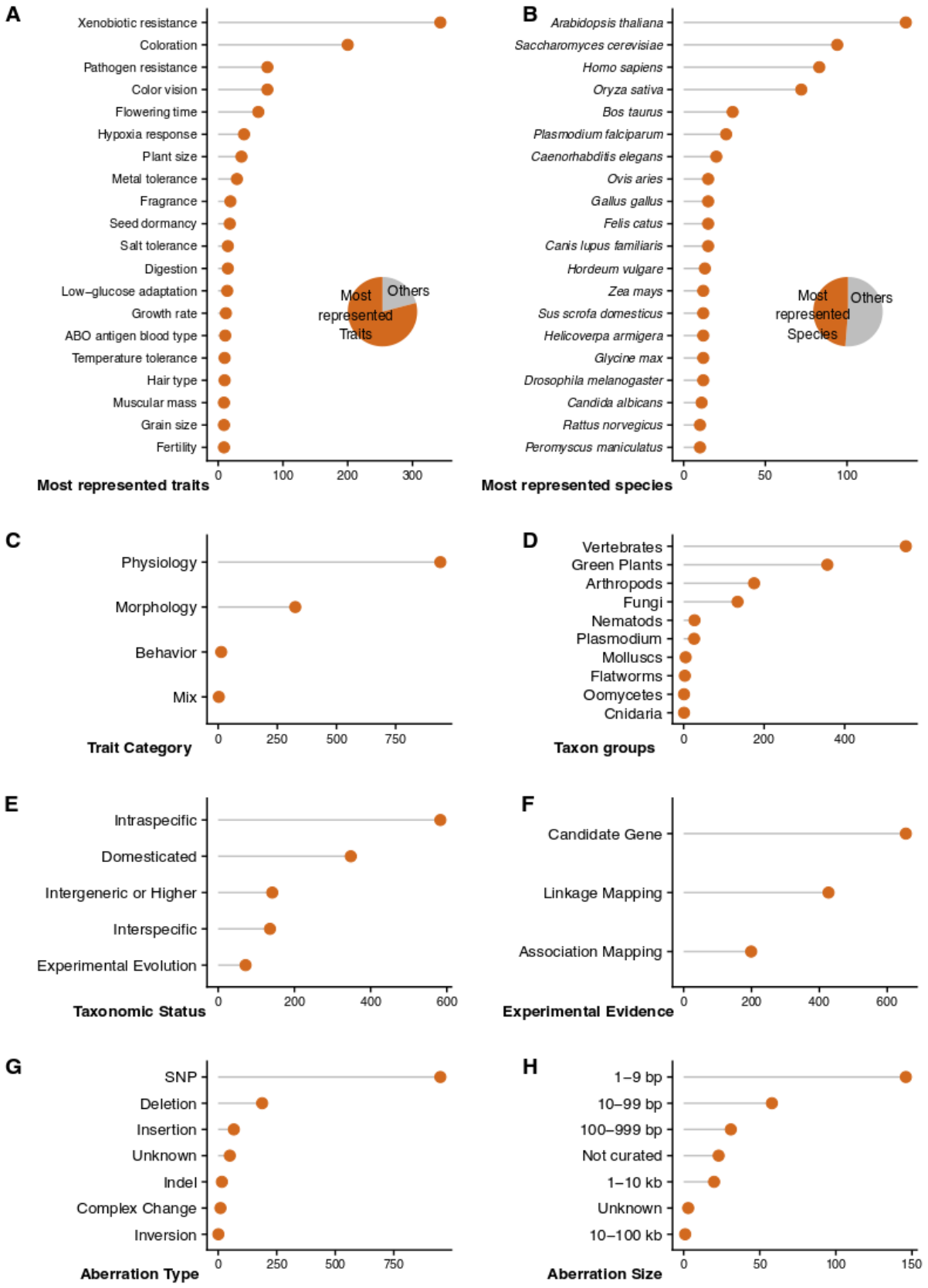

Fig. 2. Distribution of the 1281 coding changes present in Gephebase. The x-axis indicates the total number of mutations for each category. (A) Twenty most represented traits. (B) Twenty most represented species. (C) Trait categories. (D) Taxon groups. (E) Taxonomic Status. (F) Type of Empirical Evidence. (G) Aberration Type. Indels are cases involving either a deletion or an insertion

and where the direction of change is unclear. (H) Aberration size. Mutations classified as “SNP” or “Unknown” (panel G) are not represented in panel H. Note that the category “SNP” (single nucleotide polymorphism) includes single nucleotide substitutions between species for interspecific and intergeneric changes.

Courtier-Orgogozo et al., 2019), with the top species being *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Homo sapiens* and *Oryza sativa* (Fig. 2B). Fewer coding mutations underlying natural or domesticated phenotypic variation have been identified in the genetic model systems *Drosophila melanogaster and Caenorhabditis elegans* than in rice or cattle (Fig. 2B). This reflects the fact that those laboratory species are most commonly used for reverse genetics (artificial modification of a given gene) or for forward genetics in mutagenized stocks, two types of study that are not compiled in Gephebase, which only includes naturally-occurring variations. Conversely, forward genetics approaches aiming to identify causal loci are predominantly performed in agricultural species, where funding is more abundant. Overall, coding mutations have been curated in Gephebase in more than 340 species in various taxonomic groups (Vertebrates, Green Plants, Arthropods, Fungi, *Plasmodium*, Nematodes, Molluscs, Flatworms, Oomycetes, Cnidaria) (Fig. 2D). Most of the coding mutations in Gephebase are associated with intraspecific changes (45%) and domesticated evolution (27%) while interspecific changes represent 22% of the coding mutations (11% within the same genus and 11% in distinct genera) (Fig. 2E). Only 72 coding mutations from 15 published studies, all of them in yeast *Saccharomyces cerevisiae*, have been compiled in Gephebase in the Experimental Evolution category. Those mutations appeared in human-defined environments but humans did not pick the winners individually, unlike in the Domestication category. These cases represent 6% of the coding mutations of Gephebase (Fig. 2E, see Supplementary Text).

In Gephebase each mutation is attributed one type of Experimental Evidence among three possibilities, “Candidate Gene”, “Association Mapping” or “Linkage Mapping” (Martin and Orgogozo, 2013). The candidate gene approach is biased towards genes that have already been studied in other contexts and can be applied on phenotypic differences observed between distantly related species. Linkage mapping involves crosses between strains in the laboratory and then looks for associations between the genetic loci segregating in the resulting progeny population and the phenotypic traits. Association mapping, which relies on natural mixing of DNA sequences, is only applicable to polymorphic or intermixing populations. Phenotype-gene relationships that were found via association mapping are included in Gephebase only if they are supported by additional data (mutation in the same gene affecting the same trait in another organism, *in vitro* assay, etc.). In Gephebase, 51% of the coding mutations were identified via a Candidate Gene approach, 33% via Linkage Mapping and 16% via Association Mapping (Fig. 2F). Among the 1,281 coding mutations, 4% have not been narrowed down to the nucleotide level (“Unknown” cases in Fig. 2G). Note that Gephebase also contains 228 additional cases where the causal gene has been identified but the exact mutation has not been mapped and resides in the coding and/or cis-regulatory region (these mutations are classified as “Molecular Type: Unknown” and are thus not included in this meta-analysis as they are not classified as “coding”). The majority of the coding mutations (74%) are single nucleotide changes. Indels correspond to 21%, and the other coding mutations (1%) are inversions and complex changes (Fig. 2G). Among the coding mutations that are not classified as single nucleotide changes or as unknown, the

size of the aberration ranges from 1-9 bp (51% of the mutations) to 10-100 kb (<0.4%) (Fig. 2H).

Prevalence of gene loss-of-function alleles among identified domesticated variants

Compared to other mutations, null mutations abolish gene function and are thus likely to have larger and potentially more deleterious effects. The Domestication dataset in Gephebase relates to traits that have been artificially selected by breeders in plants, animals and yeasts (cultivated or bred by humans). We hypothesized that this Domestication dataset would comprise a large number of null alleles, based on the idea that breeders may have selected in animals and plants large effect mutations, and on the fact that null alleles that disrupt the protein coding region are easier to identify and validate with functional assays compared to subtle nucleotide changes. The wrinkled phenotype in Mendel's peas is such a case: it is due to a transposon insertion in a starch-branching enzyme causing a null phenotype that is maintained in gardens for its sweetness, but is not observed in the wild (Bhattacharyya et al., 1990). Similarly, the *myostatin* gene null allele causing the double-muscling phenotype of the Belgian Blue cattle has been selected despite its negative side-effects, such as increased difficulty in calving (Grobet et al., 1997; Kambadur et al., 1997; McPherron and Lee, 1997).

In Gephebase there are 348 coding mutations involved in Domestication: 42% of them are presumptive null alleles (*i.e.* they contain nonsense mutations or frameshift indels thought to abolish protein activity), 56% are not null. The 2% of mutations with unknown presumptive null status were excluded from our analysis. Null mutations represent 55% of the domesticated alleles identified in green plants (107/194) and 28% of the ones identified in Vertebrates (37/133). The fact that null alleles have been found at a higher frequency in domesticated plants than in domesticated animals (chi-square test, $p=10^{-6}$) may be related to higher levels of polyploidy (Wendel, 2000) and thus higher gene redundancy. In summary, Gephebase reveals that numerous instances of domestication have relied on the selection of coding mutations with gene inactivating effects.

We found that 48% of Experimental Evolution coding mutations were null in our current compilation of studies (Fig. 3A), a proportion that exceeds the one found in Domestication but that is not significantly different due to small sample size (chi-square test, p=0.4). Further studies will help to characterize the spectrum of coding mutations at play during short-term evolution (Gresham et al., 2008).

Null mutations are more prevalent in Domestication than in Natural Evolution

To compare domestication versus natural evolution, we grouped here the Taxonomic Status categories, *Intraspecific*, *Interspecific*, and *Interspecific or Higher*, as cases of natural evolution. A previous meta-analysis based on a smaller dataset (331 mutations in total, both coding and cis-regulatory mutations) indicated that null mutations tend to contribute more to phenotypic variation in domesticated species than in nature (Stern and Orgogozo, 2008; Stern and Orgogozo, 2009). Using our updated dataset of 1281 mutations, we observed the same trend (Fig. 3A): 43% of 341 coding mutations associated with domestication are null, compared to 22% of the 837 mutations observed in natural contexts at all levels (Fig. 3A, chi-square test, $p=7.10^{-13}$).

This pattern could be explained by various forms of ascertainment bias, for instance if the domestication dataset is enriched with a few genes, each repeatedly investigated for loss-of-function alleles. For instance, following the initial mapping of a fragrant allele caused by the loss of *BADH2* activity (Bradbury et al., 2005), nine additional null alleles were

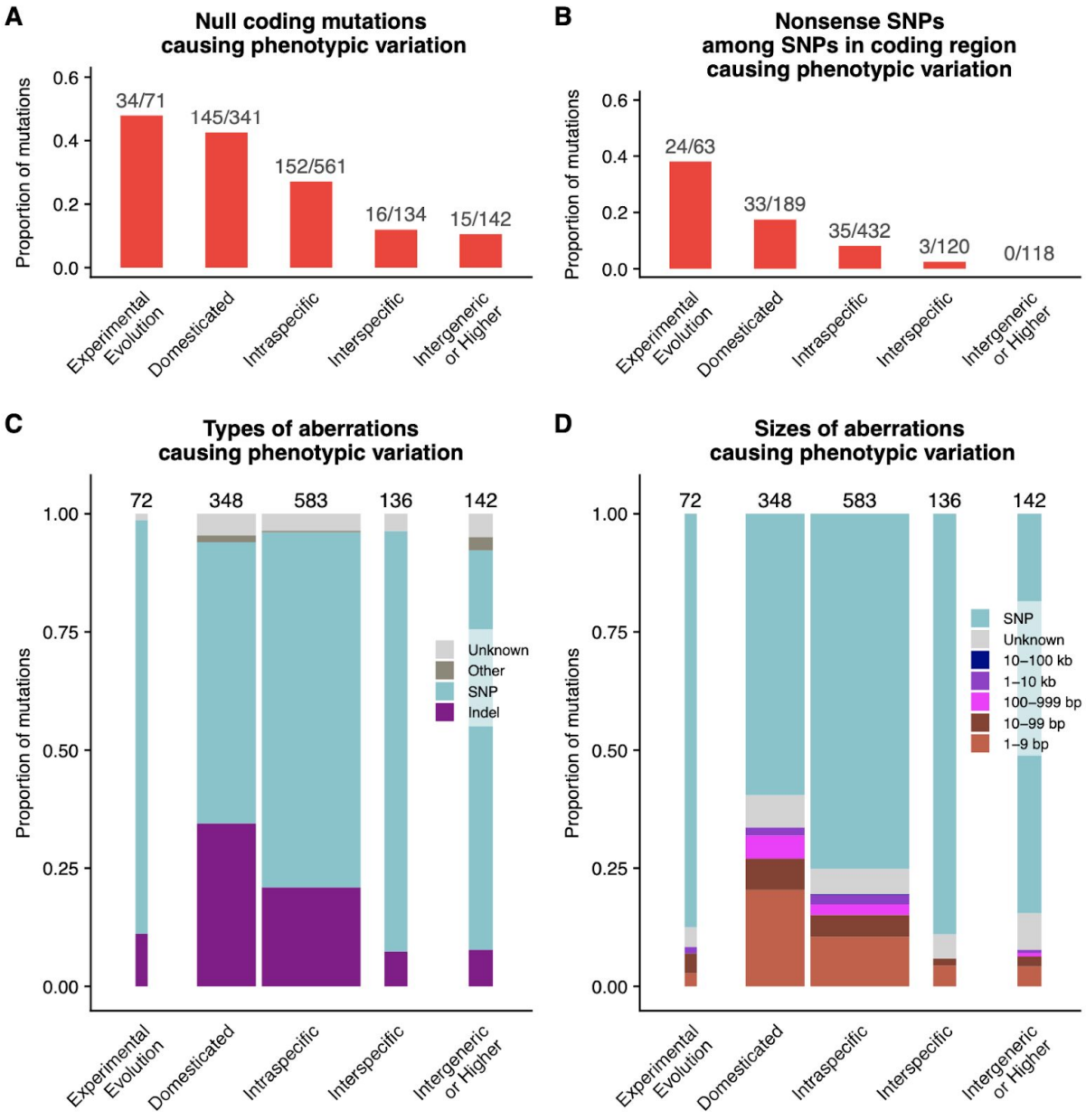

Figure 3. Coding disruptive mutations are common in artificial selection (Domestication), and decrease in proportion with evolutionary distance in natural contexts. (A) Proportion of null mutations among the coding mutations that are classified as null or not null (mutations for which the category null /not null has not been curated are not shown). (B) Proportion of nonsense mutations among single nucleotide coding changes. These cases form a subset of the null-coding mutations shown in A (which also include frameshift mutations). Single nucleotide changes located in introns that cause splicing errors are not shown in (B). (C) Types of aberrations. (D) Sizes of aberrations. The 72 cases of Experimental Evolution, representing <15 published studies, may not be representative. SNP: single nucleotide change. Numbers on top of the bars indicate the total number of mutations for each of the four categories, “Domesticated”, “Intraspecific”, “Interspecific”, and “Intergeneric or Higher”.

subsequently identified in various rice landraces by targeted sequencing (Kovach et al., 2009). We can alleviate this effect by excluding the cases marked as “Candidate Genes” (see Figure 2 in Courtier-Orgogozo et al. 2019). In this dataset corrected for ascertainment

bias, the trend is still observed, although it is barely significant, with 42% of 245 coding mutations associated with domestication being null, against 33% of the 309 mutations observed in natural contexts (chi-square test, 103/239 versus 101/295, p=0.036). To test whether this trend is genuine, it would be important to compile a larger number of coding mutations involved in natural evolution that have been identified by mapping. When excluding candidate genes, the trait "Coloration" is still represented by >30 mutations associated with natural evolution or with domestication and in this dataset, the proportion of null mutations for domestication (24/50) is also higher than for natural evolution (9/36, chi-square test, p=0.030).

Naturally selected null mutations rarefy with evolutionary distance

Certain cases of null mutations have been reported to contribute to natural evolution and have sometimes been observed to confer a selective advantage on a short term. For example, a population of *Arabidopsis thaliana* adapted to soils that are rich in salt via a premature stop codon in the gene *RAS1* (Ren et al., 2010). However, because null coding mutations are likely to abolish gene function in all tissues, they are likely to cause multiple defects, *i.e.* to cause pleiotropic effects (Stern, 2010; Stern and Orgogozo, 2008; Stern and Orgogozo, 2009). As a matter of fact, the *RAS1* mutation not only affects salt tolerance but also abscisic acid sensitivity during seed germination and early seedling growth. Over longer evolutionary time, when environmental and genetic contexts change, null mutations may become maladaptive and purged from the population. We therefore expect the proportion of null mutations to be lower between species than within species.

Our previous survey of 331 mutations, both coding and cis-regulatory, indicated that the proportion of null mutations is higher for intraspecific variation than for interspecific changes (Table 4 of (Stern and Orgogozo, 2008), and Fig. 3 of (Stern and Orgogozo, 2009). In our updated dataset of 1281 mutations, we observe the same trend: 27% of 561 coding mutations associated with intraspecific changes are null, and only 14% of 118 coding mutations classified as "Interspecific" are null (Fig. 3A, chi-square test, $p=2.10^{-4}$). The category "Intergeneric or Higher" was not compared in our meta-analysis because this approach, based almost exclusively on comparative genome sequence analysis, is biased towards the discovery of gene loss-of-function alleles (Fig. 1). In conclusion, the proportion of null mutations in the Coding Loci of Phenotypic Variation decreases progressively as we go from Domestication to Intraspecific Evolution, and from Intraspecific Variation to Interspecific Evolution.

Both nonsense mutations and disruptive structural variations decrease in representation over time

To further investigate the decrease with evolutionary distance in the proportion of disruptive mutations found in Gephebase, we examined three other aspects of the coding mutations besides null mutations. First, we observed that the proportion of nonsense mutations among all single nucleotide coding variants decrease with evolutionary time (Fig. 3B). Second, the proportion of indels was found to decrease with evolutionary time while the proportion of single nucleotide changes increases (Fig. 3C). The scarcity of indels in the Experimental Evolution dataset (Fig. 3C), compiling less than 15 studies in the yeast *S. cerevisiae*, may be linked to the scarcity of spontaneous indels in that species (Zhu et al., 2014). Third, we found that the sizes of the aberrations tend to decrease as we move from domestication to intraspecific cases and to interspecific changes (Fig. 3D).

A meta-analysis of >5,000 mouse gene mutants found that genes associated with physiological phenotypes were more likely to evolve via coding mutation, gene gain, or gene loss than genes associated with morphological traits, which involve more cis-regulatory mutations (Liao et al., 2010). To test whether the trends observed at four levels (null mutations, nonsense mutations, aberration type, and aberration size) are robust and valid for both morphological and physiological traits, we checked whether we could observe them for smaller datasets: for coding mutations involved in physiological evolution only (915 mutations, Fig. S1), for morphological evolution only (328 mutations, Fig. S2). To exclude potential biases associated with methods used to find the genetic loci, we also used a smaller dataset, where cases identified via Candidate Gene Approach were excluded (Fig. S3-4). In all datasets, the same global trend of decreasing proportion of disruptive mutations over evolutionary time was observed for physiological and morphological traits. In the last dataset (cases of physiological evolution identified by methods other than the Candidate Gene approach, 550 mutations), the trend can still be detected (Fig. S4).

Discussion

The importance of null mutations for artificial and natural selection

Among the 348 coding mutations compiled in Gephebase that are involved in Domestication, 42% of them are presumptive null alleles. They encompass nonsense mutations that generate truncated proteins, indels that cause a frameshift and mutations at splice sites that lead to aberrant transcripts, in addition to the mutations classified as gene losses in Gephebase. This 42% value is likely an overestimation of the fraction of null mutations in domesticated loci because null mutations are easier to identify and validate, whether genetic loci are found by linkage mapping, association mapping or a candidate gene approach. Typically, a genomic region identified by linkage mapping spans dozens of genes: if one of the genes is found to harbor a null mutation, it constitutes a privileged candidate gene to be tested further.

Gene knockouts are not necessarily associated with trait loss (Appendix 2 of Stern Orgogozo, 2006). In certain cases, a gene loss-of-function can lead to the production of new metabolites, or novel chemical content. For example, the knock-out of fatty acid desaturase genes in soybean resulted in valuable seeds with olive-like fatty acid composition (Pham et al., 2012) and disruption of the *CD163* receptor gene in pigs made them resistant to porcine reproductive and respiratory syndrome virus (Whitworth et al., 2015). A common method to make plants resistant to pests and pathogens is to disrupt susceptibility genes, which are normally upregulated in response to pathogen effectors and which can create a favorable cellular environment for pathogens (Hilscher et al., 2017). In the case of polyploid organisms, the knockout of several paralogous genes is sometimes necessary to obtain the phenotype of interest. In sugarcane, a highly polyploid plant, TALEN-mediated targeting of a highly conserved region of the multiple caffeic acid *O*-methyltransferase (COMT) enzyme genes present in the genome was successfully used to reduce lignin content (Jung and Altpeter, 2016).

A higher proportion of harmful mutations in domestic lineages compared to natural evolution

Our investigation of various aspects of the mutations responsible for protein sequence changes (null mutations, nonsense mutations, aberration type, aberration size, splicing mutations) reveals that disruptive mutations tend to contribute less to phenotypic variation in natural changes within species than in domestication. This general trend is

observed not only with the total dataset, but also with restricted datasets (physiological changes only, morphological changes only, and when mutations identified via a candidate gene approach are excluded). Because Gephebase synthesizes very disparate studies, one can wonder whether this trend is artifactual. Most of the genetic loci associated with interspecific variation have been identified via a candidate gene approach, because no crosses are possible between distantly-related species. However, certain pairs of very closely related species can still produce viable hybrids in the laboratory, and they have allowed the identification of 333 coding mutations reported in Gephebase (Fig. S3C-D). When we compared only cases identified by linkage and association mapping, we still observed the same global tendency (Fig. S3), suggesting that the observed trend is genuine.

There are several explanations for the higher proportion of deleterious mutations in domestication than in natural evolution. First, domesticated lineages have usually experienced peculiar demographic histories such as smaller population sizes, bottlenecks and limited recombination, leading to a “cost of domestication”, with domesticate genomes harboring an elevated number of harmful mutations compared to their wild ancestors (Moyers et al., 2017). In yeasts, gene copy number variation and the rate of gene gain/loss was found to be higher in domesticated strains than in wild strains (Peter et al., 2018). Higher number of deleterious alleles have been observed in various species including maize, grape, sunflower, horse and dog but certain domesticates (carrot, potato, chicken, pig) show remarkable levels of genome-wide diversity, maybe because of their particular mating systems or the maintenance of important gene flow with their wild relatives (Bosse, 2019; Hufford et al., 2019).

Second, the environment of domesticates is very different from the wild and human-driven selection can be very strong, so that high selective pressures are presumably acting during the process of domestication. Many domesticated traits would dramatically impair survival and reproduction in the wild. Several domesticated yeast strains harbor null mutations in their two aquaporin gene copies, providing a growth advantage on high-sugar medium, a common environment for domesticated strains (Will et al., 2010). In contrast, functional aquaporin genes are found in wild strains, as the presence of these genes facilitate survival in freeze-thaw cycles, and is also in essentially all Eukaryotes (Will et al., 2010). High selective pressures can allow the selection of large-effect mutations with unintended pleiotropic effects. For instance, some of the bicolor coats found within the American Paint Horse breed are based on a heterozygous state of a *EDNRB* allele, in spite of having a lethal-recessive effect where homozygous foals do not survive the deleterious effects of the mutation on intestinal function (Metallinos et al., 1998). Similarly, the selection for the plain color of “barless” pigeon breeds used a missense allele of the *NDP* gene which makes this breed prone to vision defects (Vickrey et al., 2018). It is possible that subsequent mutations evolve in domesticates to compensate for the deleterious effects of disruptive mutations. The time frame for such secondary mutations to evolve is short but still possible. For example, the *DMRT3* nonsense mutation associated with “gaitedness” in horse has been followed by several subsequent mutations in various breeds, leading to multiple gaits such as diagonal and lateral gait (McCoy et al., 2019). In bacteria too, experimental evolution experiments show that null mutations are prevalent at the early stages of adaptation of populations to new environments, while their later evolution might involve more subtle mutations, finely tuning phenotypes (Hottes et al., 2013).

Third, the environment of domesticates is somewhat more constant than the natural environment. Noteworthy, a meta-analysis of genotype-by-environment studies in 11 plant species revealed that domesticates are less plastic than natural species (Des Marais et al.,

2013). Studies in yeasts illuminate how constant environments can favor a higher proportion of deleterious mutations than variable environments. Systematic gene deletion screens in *S. cerevisiae* found that about 20% of the genes have a measurable effect in one controlled laboratory setting (Giaever et al., 2002) and that 97% of the genes alter growth when tested in more than 1,000 chemical or environmental stress conditions (Hillenmeyer et al., 2008). Disruption of certain genes might thus be valuable in the special environment shaped by humans during the domestication process, although they would probably decrease fitness in some of the external conditions experienced by individuals in the wild. An illustrative case is the independent evolution of carotenoid levels in various vertebrate species. Intraspecific differences in carotenoid pigmentation in both sheep and cattle are due to coding knockouts of the *BCO2* gene, while interspecific differences in carotenoid levels between closely related species of fowls and warblers are due to cis-regulatory variants of that same gene (Eriksson et al., 2008; Toews et al., 2017). It is likely that the cis-regulatory changes observed in wild species allow the *BCO2* gene to be functional in certain tissues or external conditions, allowing wild species to display adaptive traits across the varying external conditions of their ecological niche.

In summary, the higher proportion of disruptive mutations in domestication compared to natural evolution can be explained by several, non exclusive factors: their peculiar demographic history, higher selective pressures and their relatively more constant environment.

Disruptive mutations are more prevalent in short-term than in long-term evolution

The proportion of disruptive mutations is found to decrease from intraspecific to interspecific evolution. A likely explanation for this trend is that over longer periods of time, disruptive mutations tend to be eliminated in favor of less pleiotropic, more tissue-specific mutations (Stern and Orgogozo, 2008; Stern and Orgogozo, 2009).

Two diploid genomes from the same species or population can present dozens to hundreds of differences in their total number of functional genes (Schrider and Hahn, 2010). These copy number variations are created by newly arising duplications in some genes and deletions in others. A deficit in deletions in gene coding regions has been observed in several populations in humans and *Drosophila* (Schrider and Hahn, 2010), implying that such deletions are more likely to be removed by purifying selection. Comparison of *de novo* mutations and segregating polymorphism in various species also shows that large-effect and potentially deleterious mutations arise frequently and are selected against in natural populations (Denver et al., 2005; Konrad et al., 2018; Robinson et al., 2018). Further examination of whole genome sequences across populations and species will undoubtedly help to better understand how deleterious mutations can be progressively eliminated over time during evolution.

Given that null mutations are presumably more deleterious and observed to be purged at the interspecific level, we recommend on a case by case basis that when several possibilities for gene editing are available, to introduce the mutations that are less disruptive at the gene level.

The advantages of Bio-Inspired Genome Editing

A priori knowledge about existing variation can guide the obtention of desirable traits for accelerated domestication. For example, dairy cattles typically grow horns and are subject to the expensive and distressing practice of disbudding or dehorning. Meanwhile, the cis-regulatory alleles responsible for hornlessness in meat cattle breeds are known

(Medugorac et al., 2012), and have been transferred to dairy cattle backgrounds by genome editing (Carlson et al., 2016; Young et al., 2019). The genome editing route is proposed as a faster, transgene-free path for improvement, compared to backcross introgressions of the existing allele into dairy cattle breeds (Mueller et al., 2019). This example in a farm animal illustrates how our knowledge of existing variants can inform or inspire genome editing for food production. The same principle extends to plants: for instance, a tomato resistance allele to powdery mildew, first identified as a coding frameshift in the *MLO1* gene from a cultivar from Ecuador (Bai et al., 2008), has now been reproduced using CRISPR (Nekrasov et al., 2017), leading its genome editors to state that "mutations can be readily introduced into elite or locally adapted tomato varieties in less than a year with relatively minimal effort and investment".

Compared to previous genetic engineering techniques, CRISPR/Cas9 facilitates the introduction into crop and livestock of natural mutations detected in other species or breeds (Hilscher et al., 2017; Van Eenennaam, 2017; Zhang et al., 2018). For instance, modern strains of barley cultivated in Europe were obtained by introgressing a powdery mildew resistance allele that originated in an ethiopian landraces and inspired similar breeding strategies in other crops (Kusch and Panstruga, 2017). Gene editing could be another, more efficient way to emulate such allele transfers, facilitating the assembly of traits derived from heirloom cultivars (Tieman et al., 2017). Furthermore, methods are being developed to avoid the introduction of foreign DNA during the CRISPR/Ca9 editing process (Liang et al., 2017; Woo et al., 2015). It would be interesting to see if introducing mutations that have already been found in nature in related organisms would raise fewer concerns in society than classical GMOs involving introductions of exogenous genes. Some experts also challenge the legitimacy of patents for natural mutations genetically engineered into crop and livestock (Barbier-Brygoo et al., 2015; Porter et al., 2016). The dwarf varieties of rice and wheat that contributed to the "Green Revolution" were due to single knockout mutations (Peng et al., 1999); CRISPR makes it trivial to reproduce dwarfism in other plants or cultivars, but should those lab-made mutants be under intellectual property? To this day, using CRISPR for generating any kind of edit is enough ground for filing a patent (Martin-Laffon et al., 2019).

Engineering natural mutations versus introducing human-designed exogenous genes

Several foreign genetic elements have been inserted in various plants and animals and produced the desired effect. For instance, fast-growing transgenic salmons were successfully engineered by integration of a continuously expressed Growth Hormone gene, under the control of an antifreeze-protein promoter from the ocean pout (Du et al., 1992). Nevertheless, mimicking natural mutations may limit the probability of deleterious, unwanted outcomes compared to the combination of various genetic elements from different species. For example, to increase anthocyanin levels in tomatoes, the transcription factor gene *ANT1* was constitutively expressed via the targeted insertion of an exogenous promoter upstream of the *ANT1* gene. But this exogenous piece of DNA was found to be deleterious for the cells, as they could not outcompete the non-transformed cells (Čermák et al., 2015). An antibiotic-resistance selection marker was integrated with the promoter and the genetically engineered cells could not be retrieved in the absence of kanamycin treatment, probably because anthocyanin accumulation represses tissue growth. A bio-inspired, genome editing and transgene-free alternative to this approach, we suggest, would be to emulate the existing genetic diversity of tomato cultivars. Specifically, a 4-bp frameshift indel mutation in the *SlMYBATV* transcription factor gene acts as a spectacular switch that boosts anthocyanin production in a wild tomato from the Galápagos Islands (Cao et al., 2017). This

knockout would be easy to reproduce using CRISPR, and could be combined to other anthocyanin variation loci that are currently under investigation. The literature starts to describe other successful cases where natural mutations first identified as a Loci of Evolution or Domestication have been redeployed by genome editing in livestock and crops (Carlson et al., 2016; Hilscher et al., 2017; Hufford et al., 2019; Van Eenennaam, 2017).

It is important to stress here that the transfer of foreign, beneficial mutations from one species to another is a common, natural route of evolution. Therefore, a given genomic region associated with a phenotype of interest that has transferred naturally between species represents a good target sequence to insert in yet another species to obtain a similar trait. An incredible number of transfers of DNA fragments from one species to another, via introgression between closely-related species or via horizontal gene transfer, through viruses or yet other unknown means between distantly related species, have been discovered in recent years (Daubin and Szöllosi, 2016; Gasmi et al., 2015; Matveeva and Otten, 2019; Parker and Brisson, 2019; Taylor and Larson, 2019; Wu et al., 2018). For example, adaptation to high altitude occurred independently in Tibetan human populations and in dogs via distinct alleles of the *EPAS1* gene, which have been transferred from archaic hominins to humans and from wolves to dogs, respectively (Witt and Huerta-Sánchez, 2019). Several metabolic pathways have been gained in diverse eukaryotic lineages by horizontal transfer of bacterial genes: cellulose biosynthesis in ascidians, amino acid and nitrogen metabolism, vitamin biosynthesis and iron acquisition, or carbohydrate metabolism in various arthropods and fungi (Husnik and McCutcheon, 2018). A remarkable case is the acquisition of the ability to synthesize carotenoids, which was acquired independently in aphids, gall midges, and spider mites, through the integration of a long piece of DNA encompassing enzyme genes from distinct fungi donors (Toews et al., 2017; Zhao and Nabity, 2017). Certain metabolic pathways and genes appear to transfer repeatedly between species, whereas others do not (Jain et al., 1999). Their rate of transfer might be related to their amenability to horizontal transfer or to what happens after they arrive in a new species. It is also possible that certain species are more prone (predisposed or “preadapted”, see Parker, 2016; Schilthuizen, 2019) than others to accept a given exogenous pieces of DNA and thus acquire the associated trait. General principles are starting to be uncovered among the characterized horizontal gene transfer events (Daubin and Szöllosi, 2016; Husnik and McCutcheon, 2018), and such trends may help to design effective ways to confer traits of interest via the introduction of exogenous genes.

As the number of identified Loci of Domestication and Evolution increases over the years, more and more possibilities for bio-inspired genome editing are becoming available.

Bias towards certain species is compensated by the existence of hotspot genes

Certain species are overrepresented in Gephebase, highlighting the fact that our understanding of the genotype-phenotype link comes from a handful of species. Nevertheless, such a bias towards certain taxonomic lineages may not be problematic for finding gene targets for genetic engineering and crop improvement. Indeed, during domestication the same traits have been selected in diverse plants (grain properties, color, flowering time, resistance to abiotic stress) and animals (meat, milk yield, coat color), and it has been found that similar traits have regularly evolved through independent mutations in the same genes (Martin and Orgogozo, 2013; Stern, 2013). Given the prevalence of such hotspot genes (Martin and Orgogozo, 2013), it is reasonable to expect genetic mutations identified in a model species to produce similar phenotypic outcomes in another species of interest, even distantly related. The development of CRISPR and the production of

comparable loss of function mutations in orthologous genes in various species in future years will undoubtedly help to grasp the extent to which similar mutations are expected to provoke similar phenotypic outcomes in other species.

Current genetics knowledge is biased towards certain traits

The phenotypic traits present in Gephebase, for which genes and mutations have been identified, are not representative of the observable diversity in Eukaryotes. Whereas Xenobiotic resistance and Coloration together represent more than 40% of the coding mutations compiled in Gephebase (Fig. 1A), other traits such as animal tameness or wood density are not found in Gephebase, as no gene has been firmly shown to be associated with such phenotypes yet. Given the prevalence of the xenobiotic resistance trait in Gephebase, it is not surprising that a large fraction of the crops and livestock that have been genetically engineered so far are organisms that are made resistant to pathogens or herbicides: various crops resistant to glyphosate through the introduction or modification of the *EPSPS* gene (Funke et al., 2006; Hummel et al., 2018; Zhang et al., 2018), chlorsulfuron-resistant maize via an amino acid substitution in the ALS gene (Svitashev et al., 2015), wheat lines carrying null mutations in *MILDEW RESISTANCE LOCUS* (*MLO*) (Wang et al., 2014), or cattle resistant to tuberculosis via inactivation of the *SP110* gene (Wu et al., 2015). In a recent review, Zhang et al. compiled 47 traits in various crops that have been improved by genome-editing techniques and 45% of them are resistances to pathogens or herbicides (Zhang et al., 2018). A similar percentage (6/13 traits) is found just for CRISPR/Cas9-based editing cases (Schindele et al., 2018; Zhang et al., 2018).

For increased yield, a key breeding trait, multiple genetic loci have been identified and constitute promising targets for genetic engineering (Li et al., 2016; Zhang et al., 2018). Genome editing can also be used to improve food quality, for example by inactivating the *BADH2* gene in rice and thus enriching its aromatic qualities (Shan et al., 2015). Other traits have not been manipulated yet by applied genetic engineering but are now at reach. Multiple genes have been identified whose alteration of the coding region leads to adaptation to high altitude in Vertebrates (humans, monkeys, mice, birds, lamas, snow leopards, etc.) (Witt and Huerta-Sánchez, 2019) (search for “altitude” in gephebase.org). They represent interesting targets for genetic engineering to create new varieties of livestock that are better adapted to high altitude.

While certain traits exhibit a large number of possible target genes for genome editing, others traits are poorly known regarding their genetic basis. One reason is that certain traits rely on a few large effect loci (such as resistance to particular herbicides or pathogens) while others rely on the accumulation of many small effect loci (for example adaptation to an arid environment). Mutations associated with the latter type of traits are, first, more difficult to identify and second, must be introduced all together to create the trait of interest, which can be laborious. It implies that certain desired traits are currently not reachable by genetic engineering given our current knowledge. By listing known cases of phenotypic changes explained by mostly large effect alleles, Gephebase provides a general overview of the traits that can be potentially attainable through genome editing.

Conclusion

Gephebase is a unique database that combines knowledge from fundamental genetics research and applied quantitative genetics in plants and animals. To improve crops and livestock via genome editing, we advise based on our meta-analysis of Gephebase to

(1) continue to investigate the genetic basis of natural evolution, (2) consider mimicking natural variants of relatively large-effects ("bio-inspiration"), and (3) create non-null mutations whenever possible.

Supplementary Material
Figures S1-4
Supplementary References

Supplementary files
File S1. Csv file exported from Gephebase in June 2019 containing all the mutations.
File S2. Csv file exported from Gephebase describing the 1281 coding mutations.
File S3. R script (R version 3.4 and R Studio version 1.2.1335) used to analyze the 1281 coding mutations and to create the figures.

References

Arnoult, L. A. (2014). La marche génétique de l'évolution. *Biol. Aujourdhui* **208**, 237–249.

Bai, Y., Pavan, S., Zheng, Z., Zappel, N. F., Reinstädler, A., Lotti, C., De Giovanni, C., Ricciardi, L., Lindhout, P. and Visser, R. (2008). Naturally occurring broad-spectrum powdery mildew resistance in a Central American tomato accession is caused by loss of Mlo function. *Mol. Plant. Microbe Interact.* **21**, 30–39.

Barbier-Brygoo, H., Chilliard, Y., Durand, Elmayan, T., Goldringer and Porter, J. (2015). Rapport de synthèse du groupe de travail sur la propriété intellectuelle sur les connaissances dans le secteur végétal.

Bhattacharyya, M. K., Smith, A. M., Ellis, T. N., Hedley, C. and Martin, C. (1990). The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell* **60**, 115–122.

Boitard, S., Boussaha, M., Capitan, A., Rocha, D. and Servin, B. (2016). Uncovering adaptation from sequence data: lessons from genome resequencing of four cattle breeds. *Genetics* **203**, 433–450.

Bosse, M. (2019). No "doom" in chicken domestication? *PLoS Genet.* **15**, e1008089.

Bradbury, L. M., Fitzgerald, T. L., Henry, R. J., Jin, Q. and Waters, D. L. (2005). The gene for fragrance in rice. *Plant Biotechnol. J.* **3**, 363–370.

Cao, X., Qiu, Z., Wang, X., Van Giang, T., Liu, X., Wang, J., Wang, X., Gao, J., Guo, Y. and Du, Y. (2017). A putative R3 MYB repressor is the candidate gene underlying atroviolacium, a locus for anthocyanin pigmentation in tomato fruit. *J. Exp. Bot.* **68**, 5745–5758.

Carlson, D. F., Lancto, C. A., Zang, B., Kim, E.-S., Walton, M., Oldeschulte, D., Seabury, C., Sonstegard, T. S. and Fahrenkrug, S. C. (2016). Production of hornless dairy cattle from genome-edited cell lines. *Nat. Biotechnol.* **34**, 479.

Čermák, T., Baltes, N. J., Čegan, R., Zhang, Y. and Voytas, D. F. (2015). High-frequency, precise modification of the tomato genome. *Genome Biol.* **16**, 232.

Courtier-Orgogozo, V., Arnoult, L., Prigent, S. R., Wiltgen, S. and Martin, A. (2019). Gephebase, a Database of Genotype-Phenotype Relationships for natural and domesticated variation in Eukaryotes. *BioRxiv* 618371.

Daubin, V. and Szöllosi, G. J. (2016). Horizontal gene transfer and the history of life. *Cold Spring Harb. Perspect. Biol.* **8**, a018036.

Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M. and Thomas, W. K. (2005). The transcriptional consequences of mutation and natural selection in

Caenorhabditis elegans. *Nat. Genet.* **37**, 544.

Dong, K., Du, Y., Rinkevich, F., Nomura, Y., Xu, P., Wang, L., Silver, K. and Zhorov, B. S. (2014). Molecular biology of insect sodium channels and pyrethroid resistance. *Insect Biochem. Mol. Biol.* **50**, 1–17.

Durand, R., Cannet, A., Berdjane, Z., Bruel, C., Haouchine, D., Delaunay, P. and Izri, A. (2012). Infestation by pyrethroids resistant bed bugs in the suburb of Paris, France. *Parasite* **19**, 381.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'Hom, B., Tixier-Boichard, M., Strömstedt, L., Wright, D., Jungerius, A., Vereijken, A. and Randi, E. (2008). Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS Genet.* **4**, e1000010.

Funke, T., Han, H., Healy-Fried, M. L., Fischer, M. and Schönbrunn, E. (2006). Molecular basis for the herbicide resistance of Roundup Ready crops. *Proc. Natl. Acad. Sci.* **103**, 13010–13015.

Gasmi, L., Boulain, H., Gauthier, J., Hua-Van, A., Musset, K., Jakubowska, A. K., Aury, J.-M., Volkoff, A.-N., Huguet, E. and Herrero, S. (2015). Recurrent domestication by Lepidoptera of genes from their parasites mediated by bracoviruses. *PLoS Genet.* **11**, e1005470.

Georges, M., Charlier, C. and Hayes, B. (2018). Harnessing genomic information for livestock improvement. *Nat. Rev. Genet.* 1.

González-Cabrera, J., Rodríguez-Vargas, S., Davies, T. E., Field, L. M., Schmehl, D., Ellis, J. D., Krieger, K. and Williamson, M. S. (2016). Novel mutations in the voltage-gated sodium channel of pyrethroid-resistant Varroa destructor populations from the southeastern USA. *PloS One* **11**, e0155332.

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., DeSevo, C. G., Botstein, D. and Dunham, M. J. (2008). The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet.* **4**, e1000303.

Grobet, L., Martin, L. J. R., Poncelet, D., Pirottin, D., Brouwers, B., Riquet, J., Schoeberlein, A., Dunner, S., Ménissier, F. and Massabanda, J. (1997). A deletion in the bovine myostatin gene causes the double–muscled phenotype in cattle. *Nat. Genet.* **17**, 71.

Hilscher, J., Bürstmayr, H. and Stoger, E. (2017). Targeted modification of plant genomes for precision crop breeding. *Biotechnol. J.* **12**, 1600173.

Hoekstra, H. E. and Coyne, J. A. (2007). The locus of evolution: evo devo and the genetics of adaptation. *Evol. Int. J. Org. Evol.* **61**, 995–1016.

Hottes, A. K., Freddolino, P. L., Khare, A., Donnell, Z. N., Liu, J. C. and Tavazoie, S. (2013). Bacterial adaptation through loss of function. *PLoS Genet.* **9**, e1003617.

Hufford, M. B., Berny Mier y Teran, J. C. and Gepts, P. (2019). Crop Biodiversity: An Unfinished Magnum Opus of Nature. *Annu. Rev. Plant Biol.* **70**, 727–751.

Hummel, A. W., Chauhan, R. D., Cermak, T., Mutka, A. M., Vijayaraghavan, A., Boyher, A., Starker, C. G., Bart, R., Voytas, D. F. and Taylor, N. J. (2018). Allele exchange at the EPSPS locus confers glyphosate tolerance in cassava. *Plant Biotechnol. J.* **16**, 1275–1282.

Husnik, F. and McCutcheon, J. P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67.

Jain, R., Rivera, M. C. and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci.* **96**, 3801–3806.

Jung, J. H. and Altpeter, F. (2016). TALEN mediated targeted mutagenesis of the caffeic acid O-methyltransferase in highly polyploid sugarcane improves cell wall composition for production of bioethanol. *Plant Mol. Biol.* **92**, 131–142.

Kambadur, R., Sharma, M., Smith, T. P. and Bass, J. J. (1997). Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. *Genome Res.* **7**, 910–915.

Kapantaidaki, D. E., Sadikoglou, E., Tsakireli, D., Kampanis, V., Stavrakaki, M., Schorn, C., Ilias, A., Riga, M., Tsiamis, G. and Nauen, R. (2018). Insecticide resistance in trialeurodes vaporariorum populations and novel diagnostics for kdr mutations. *Pest Manag. Sci.* **74**, 59–69.

Klaassen, H., Wang, Y., Adamski, K., Rohner, N. and Kowalko, J. E. (2018). CRISPR mutagenesis confirms the role of oca2 in melanin pigmentation in Astyanax mexicanus. *Dev. Biol.* **441**, 313–318.

Konrad, A., Flibotte, S., Taylor, J., Waterston, R. H., Moerman, D. G., Bergthorsson, U. and Katju, V. (2018). Mutational and transcriptional landscape of spontaneous gene duplications and deletions in Caenorhabditis elegans. *Proc. Natl. Acad. Sci.* **115**, 7386–7391.

Kovach, M. J., Calingacion, M. N., Fitzgerald, M. A. and McCouch, S. R. (2009). The origin and evolution of fragrance in rice (Oryza sativa L.). *Proc. Natl. Acad. Sci.* **106**, 14444–14449.

Kusch, S. and Panstruga, R. (2017). mlo-based resistance: an apparently universal "weapon" to defeat powdery mildew disease. *Mol. Plant. Microbe Interact.* **30**, 179–189.

Lenser, T. and Theißen, G. (2013). Molecular mechanisms involved in convergent crop domestication. *Trends Plant Sci.* **18**, 704–714.

Li, M., Li, X., Zhou, Z., Wu, P., Fang, M., Pan, X., Lin, Q., Luo, W., Wu, G. and Li, H. (2016). Reassessment of the four yield-related genes Gn1a, DEP1, GS3, and IPA1 in rice using a CRISPR/Cas9 system. *Front. Plant Sci.* **7**, 377.

Liang, Z., Chen, K., Li, T., Zhang, Y., Wang, Y., Zhao, Q., Liu, J., Zhang, H., Liu, C. and Ran, Y. (2017). Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. *Nat. Commun.* **8**, 14261.

Liao, B., Weng, M. and Zhang, J. (2010). Contrasting genetic paths to morphological and physiological evolution. *Proc. Natl. Acad. Sci. U. S. OF* **107**, 7353–7358.

Ma, L., Jeffery, W. R., Essner, J. J. and Kowalko, J. E. (2015). Genome editing using TALENs in blind Mexican cavefish, Astyanax mexicanus. *PLoS One* **10**, e0119370.

Ma, X., Zhu, Q., Chen, Y. and Liu, Y.-G. (2016). CRISPR/Cas9 platforms for genome editing in plants: developments and applications. *Mol. Plant* **9**, 961–974.

Major, K. M., Weston, D. P., Lydy, M. J., Wellborn, G. A. and Poynton, H. C. (2018). Unintentional exposure to terrestrial pesticides drives widespread and predictable evolution of resistance in freshwater crustaceans. *Evol. Appl.* **11**, 748–761.

Martin, A. and Courtier-Orgogozo, V. (2017). Morphological evolution repeatedly caused by mutations in signaling ligand genes. In *Diversity and Evolution of Butterfly Wing Patterns*, pp. 59–87. Springer.

Martin, A. and Orgogozo, V. (2013). The Loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evol. Int. J. Org. Evol.* **67**, 1235–1250.

Matveeva, T. V. and Otten, L. (2019). Widespread occurrence of natural genetic transformation of plants by Agrobacterium. *Plant Mol. Biol.* 1–23.

Mazo-Vargas, A., Concha, C., Livraghi, L., Massardo, D., Wallbank, R. W., Zhang, L., Papador, J. D., Martinez-Najera, D., Jiggins, C. D. and Kronforst, M. R. (2017). Macroevolutionary shifts of WntA function potentiate butterfly wing-pattern diversity. *Proc. Natl. Acad. Sci.* **114**, 10701–10706.

McCoy, A. M., Beeson, S. K., Rubin, C.-J., Andersson, L., Caputo, P., Lykkjen, S., Moore, A., Piercy, R. J., Mickelson, J. R. and McCue, M. E. (2019). Identification and validation of genetic variants predictive of gait in standardbred horses. *PLoS*

Genet. **15**, e1008146.
McPherron, A. C. and Lee, S.-J. (1997). Double muscling in cattle due to mutations in the myostatin gene. *Proc. Natl. Acad. Sci.* **94**, 12457–12461.
Medugorac, I., Seichter, D., Graf, A., Russ, I., Blum, H., Göpel, K. H., Rothammer, S., Förster, M. and Krebs, S. (2012). Bovine polledness–an autosomal dominant trait with allelic heterogeneity. *PloS One* **7**, e39477.
Metallinos, D. L., Bowling, A. T. and Rine, J. (1998). A missense mutation in the endothelin-B receptor gene is associated with Lethal White Foal Syndrome: an equine version of Hirschsprung disease. *Mamm. Genome* **9**, 426–431.
Moyers, B. T., Morrell, P. L. and McKay, J. K. (2017). Genetic costs of domestication and improvement. *J. Hered.* **109**, 103–116.
Mueller, M. L., Cole, J. B., Sonstegard, T. S. and Van Eenennaam, A. L. (2019). Comparison of gene editing versus conventional breeding to introgress the POLLED allele into the US dairy cattle population. *J. Dairy Sci.*
Nekrasov, V., Wang, C., Win, J., Lanz, C., Weigel, D. and Kamoun, S. (2017). Rapid generation of a transgene-free powdery mildew resistant tomato by genome deletion. *Sci. Rep.* **7**, 482.
Niu, D., Wei, H.-J., Lin, L., George, H., Wang, T., Lee, I.-H., Zhao, H.-Y., Wang, Y., Kan, Y. and Shrock, E. (2017). Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* **357**, 1303–1307.
Orgogozo, V., Morizot, B. and Martin, A. (2015). The differential view of genotype–phenotype relationships. *Evol. Popul. Genet.* 179.
Pajic, P., Pavlidis, P., Dean, K., Neznanova, L., Romano, R.-A., Garneau, D., Daugherity, E., Globig, A., Ruhl, S. and Gokcumen, O. (2019). Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628.
Parker, J. (2016). Myrmecophily in beetles (Coleoptera): evolutionary patterns and biological mechanisms. *Myrmecol. News* **22**, 65–108.
Parker, B. J. and Brisson, J. A. (2019). A Laterally Transferred Viral Gene Modifies Aphid Wing Plasticity. *Curr. Biol.*
Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K. and Llored, A. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339.
Pham, A.-T., Shannon, J. G. and Bilyeu, K. D. (2012). Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* **125**, 503–515.
Porter, J. R., Durand, J.-L. and Elmayan, T. (2016). Edited plants should not be patented. *Nature* **530**, 33–33.
Ren, Z., Zheng, Z., Chinnusamy, V., Zhu, J., Cui, X., Iida, K. and Zhu, J.-K. (2010). RAS1, a quantitative trait locus for salt tolerance and ABA sensitivity in Arabidopsis. *Proc. Natl. Acad. Sci.* **107**, 5669–5674.
Robinson, J. A., Brown, C., Kim, B. Y., Lohmueller, K. E. and Wayne, R. K. (2018). Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. *Curr. Biol.* **28**, 3487–3494.
Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evol. Int. J. Org. Evol.* **66**, 1–17.
Schilthuizen, M. (2019). *Darwin comes to town: How the urban jungle drives evolution*. Picador.
Schindele, P., Wolter, F. and Puchta, H. (2018). Transforming plant biology and breeding with CRISPR/Cas9, Cas12 and Cas13. *FEBS Lett.* **592**, 1954–1967.
Schrider, D. R. and Hahn, M. W. (2010). Gene copy-number polymorphism in nature. *Proc.*

R. Soc. B Biol. Sci. **277**, 3213–3221.
Shan, Q., Zhang, Y., Chen, K., Zhang, K. and Gao, C. (2015). Creation of fragrant rice by targeted knockout of the Os BADH 2 gene using TALEN technology. *Plant Biotechnol. J.* **13**, 791–800.
Sierra, I., Capriotti, N., Fronza, G., Mougabure-Cueto, G. and Ons, S. (2016). Kdr mutations in Triatoma infestans from the Gran Chaco are distributed in two differentiated foci: Implications for pyrethroid resistance management. *Acta Trop.* **158**, 208–213.
Stern, D. L. (2000). Evolutionary developmental biology and the problem of variation. *Evol. Int. J. Org. Evol.* **54**, 1079–1091.
Stern, D. L. (2010). *Evolution, Development, and the Predictable Genome*. 1st ed. Roberts & Company Publishers.
Stern, D. L. (2013). The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764.
Stern, D. L. and Orgogozo, V. (2008). The loci of evolution: How predictable is genetic evolution ? *Evol. Int. J. Org. Evol.* **62**, 2155–2177.
Stern, D. and Orgogozo, V. (2009). Is Genetic Evolution Predictable? *SCIENCE* **323**, 746–751.
Svitashev, S., Young, J. K., Schwartz, C., Gao, H., Falco, S. C. and Cigan, A. M. (2015). Targeted mutagenesis, precise gene editing, and site-specific gene insertion in maize using Cas9 and guide RNA. *Plant Physiol.* **169**, 931–945.
Tan, W., Proudfoot, C., Lillico, S. G. and Whitelaw, C. B. A. (2016). Gene targeting, genome editing: from Dolly to editors. *Transgenic Res.* **25**, 273–287.
Taylor, S. A. and Larson, E. L. (2019). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat. Ecol. Evol.* **3**, 170.
Tieman, D., Zhu, G., Resende, M. F., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S. O., Taylor, M. and Zhang, B. (2017). A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391–394.
Toews, D. P., Hofmeister, N. R. and Taylor, S. A. (2017). The evolution and genetics of carotenoid processing in animals. *Trends Genet.* **33**, 171–182.
Van Eenennaam, A. L. (2017). Genetic modification of food animals. *Curr. Opin. Biotechnol.* **44**, 27–34.
Vickrey, A. I., Bruders, R., Kronenberg, Z., Mackey, E., Bohlender, R. J., Maclary, E. T., Maynez, R., Osborne, E. J., Johnson, K. P. and Huff, C. D. (2018). Introgression of regulatory alleles and a missense coding mutation drive plumage pattern diversity in the rock pigeon. *Elife* **7**, e34803.
Vikram, P., Swamy, B. M., Dixit, S., Singh, R., Singh, B. P., Miro, B., Kohli, A., Henry, A., Singh, N. K. and Kumar, A. (2015). Drought susceptibility of modern rice varieties: an effect of linkage of drought tolerance with undesirable traits. *Sci. Rep.* **5**, 14799.
Voytas, D. F. and Gao, C. (2014). Precision genome engineering and agriculture: opportunities and regulatory challenges. *PLoS Biol.* **12**, e1001877.
Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C. and Qiu, J.-L. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat. Biotechnol.* **32**, 947.
Wendel, J. F. (2000). Genome evolution in polyploids. In *Plant molecular evolution*, pp. 225–249. Springer.
Whitworth, K. M., Rowland, R. R., Ewen, C. L., Trible, B. R., Kerrigan, M. A., Cino-Ozuna, A. G., Samuel, M. S., Lightner, J. E., McLaren, D. G. and Mileham, A. J. (2015). Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. *Nat. Biotechnol.* **34**, 20.
Will, J. L., Kim, H. S., Clarke, J., Painter, J. C., Fay, J. C. and Gasch, A. P. (2010).

Incipient balancing selection through adaptive loss of aquaporins in natural Saccharomyces cerevisiae populations. *PLoS Genet.* **6**, e1000893.
Witt, K. E. and Huerta-Sánchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philos. Trans. R. Soc. B* **374**, 20180235.
Woo, J. W., Kim, J., Kwon, S. I., Corvalán, C., Cho, S. W., Kim, H., Kim, S.-G., Kim, S.-T., Choe, S. and Kim, J.-S. (2015). DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nat. Biotechnol.* **33**, 1162.
Wu, H., Wang, Y., Zhang, Y., Yang, M., Lv, J., Liu, J. and Zhang, Y. (2015). TALE nickase-mediated SP110 knockin endows cattle with increased resistance to tuberculosis. *Proc. Natl. Acad. Sci.* **112**, E1530–E1539.
Wu, D.-D., Ding, X.-D., Wang, S., Wójcik, J. M., Zhang, Y., Tokarska, M., Li, Y., Wang, M.-S., Faruque, O. and Nielsen, R. (2018). Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nat Ecol Evol* **2**, 1139–1145.
Young, A. E., Mansour, T. A., McNabb, B. R., Owen, J. R., Trott, J. F., Brown, C. T. and Van Eenennaam, A. L. (2019). Genomic and phenotypic analyses of six offspring of a genome-edited hornless bull. *Nat. Biotechnol.* 1–8.
Zhang, L., Mazo-Vargas, A. and Reed, R. D. (2017). Single master regulatory gene coordinates the evolution and development of butterfly color and iridescence. *Proc. Natl. Acad. Sci.* **114**, 10707–10712.
Zhang, Y., Massel, K., Godwin, I. D. and Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biol.* **19**, 210.
Zhao, C. and Nabity, P. D. (2017). Phylloxerids share ancestral carotenoid biosynthesis genes of fungal origin with aphids and adelgids. *PloS One* **12**, e0185484.
Zhu, Y. O., Siegal, M. L., Hall, D. W. and Petrov, D. A. (2014). Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci.* **111**, E2310–E2318.

Acknowledgements
We thank JEB for organizing the “Genome editing for comparative physiology” symposium and for inviting us to write this review. We thank Matt Hahn for suggesting relevant papers and Patrice Dumas, Fabien Duveau and Olivier Tenaillon for comments on a previous draft.

FOOTNOTES
Competing interests
The authors declare no competing or financial interests.

Funding
The research leading to this paper has received funding from the European Research Council under the European Community’s Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 337579) to VCO. AM is supported by the National Science Foundation [IOS-1656553 and IOS-1755329].

Supplementary Material

The Coding Loci of Evolution and Domestication: Current Knowledge and Implications for Bio-Inspired Genome Editing (Courtier-Orgogozo and Martin)

Supplementary Figures 1-4
Supplementary References

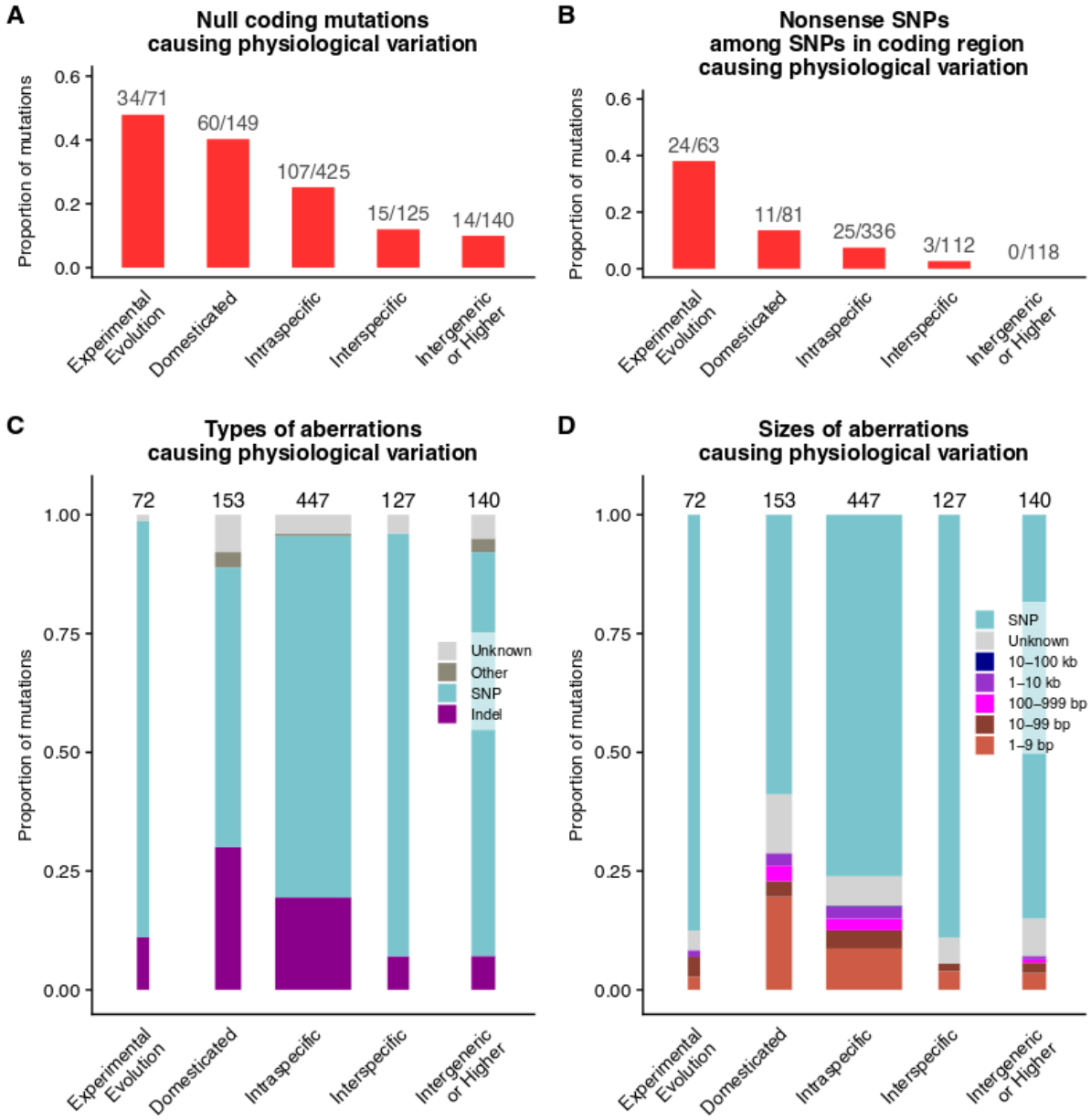

Fig. S1. The proportion of null and disruptive mutations among the coding mutations causing physiological variation decreases with evolutionary time. Same legend as Fig. 3.

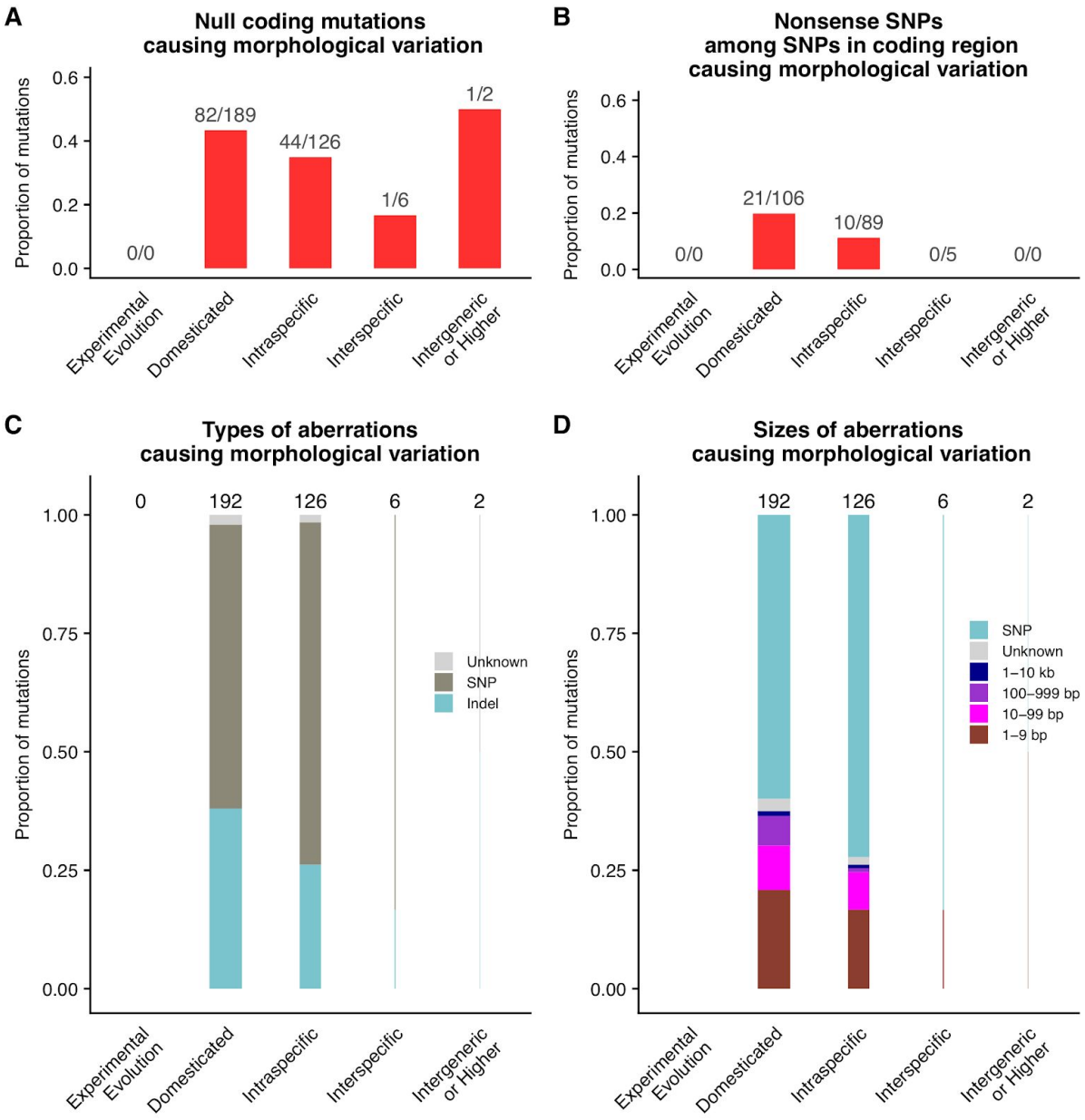

Fig. S2. The proportion of null and disruptive mutations among the coding mutations causing morphological variation decreases with evolutionary time. Same legend as Fig. 3. Cases curated as "Experimental Evolution", "Interspecific" and "Intergeneric or Higher" are too few to derive relevant estimates.

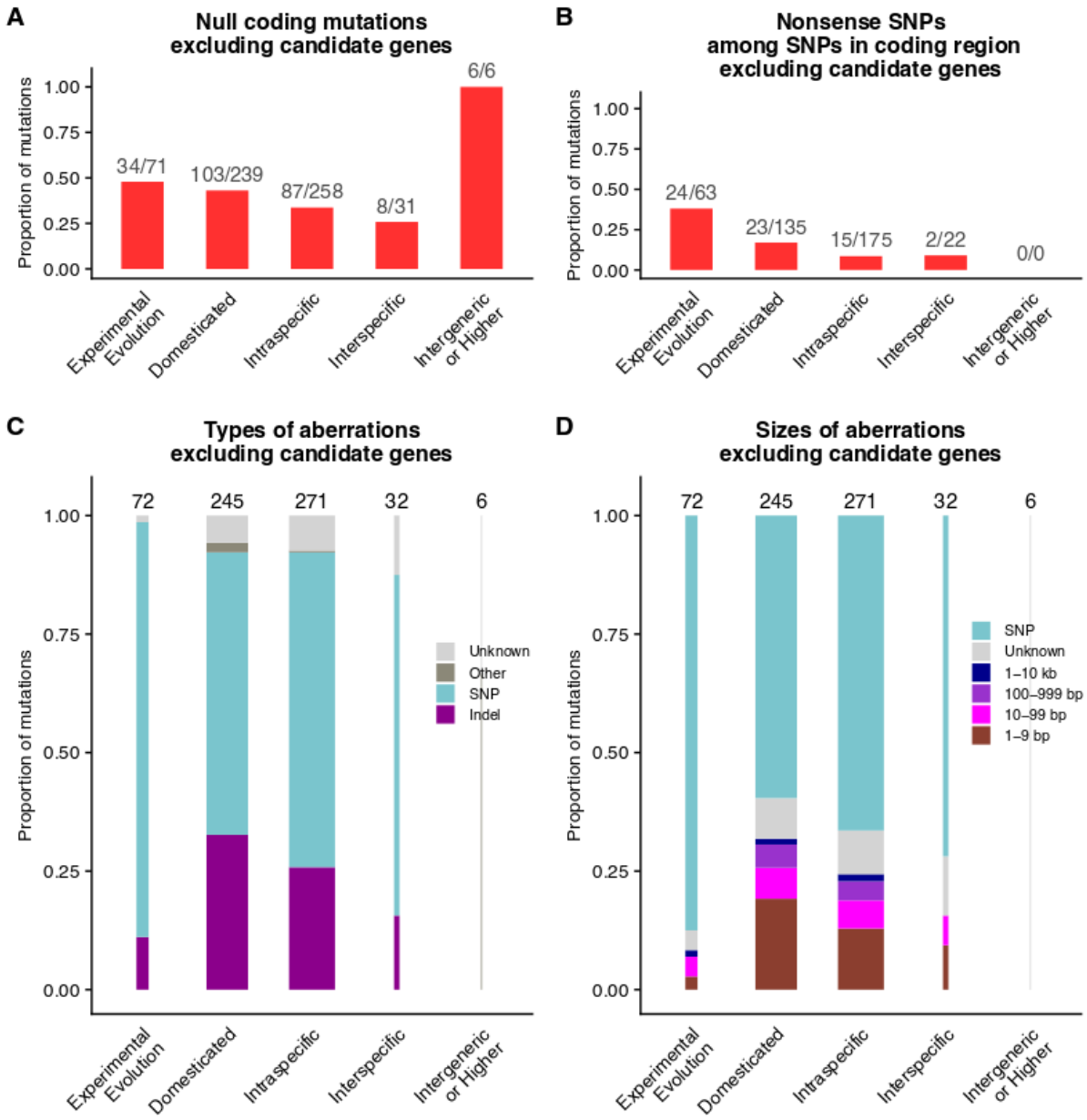

Fig. S3. The proportion of null and disruptive mutations among the coding mutations identified via methods distinct from the Candidate Gene Approach decreases with evolutionary time. Same legend as Fig. 3. The six cases curated as “Intergeneric or Higher” correspond to one study where authors performed a phylogenetic- and genome-wide screen for genes that have been inactivated repeatedly during evolution, in significant association with two metabolic phenotypes, the loss of the ability to synthesize vitamin C, and low levels of biliary phospholipids (Hiller et al., 2012).

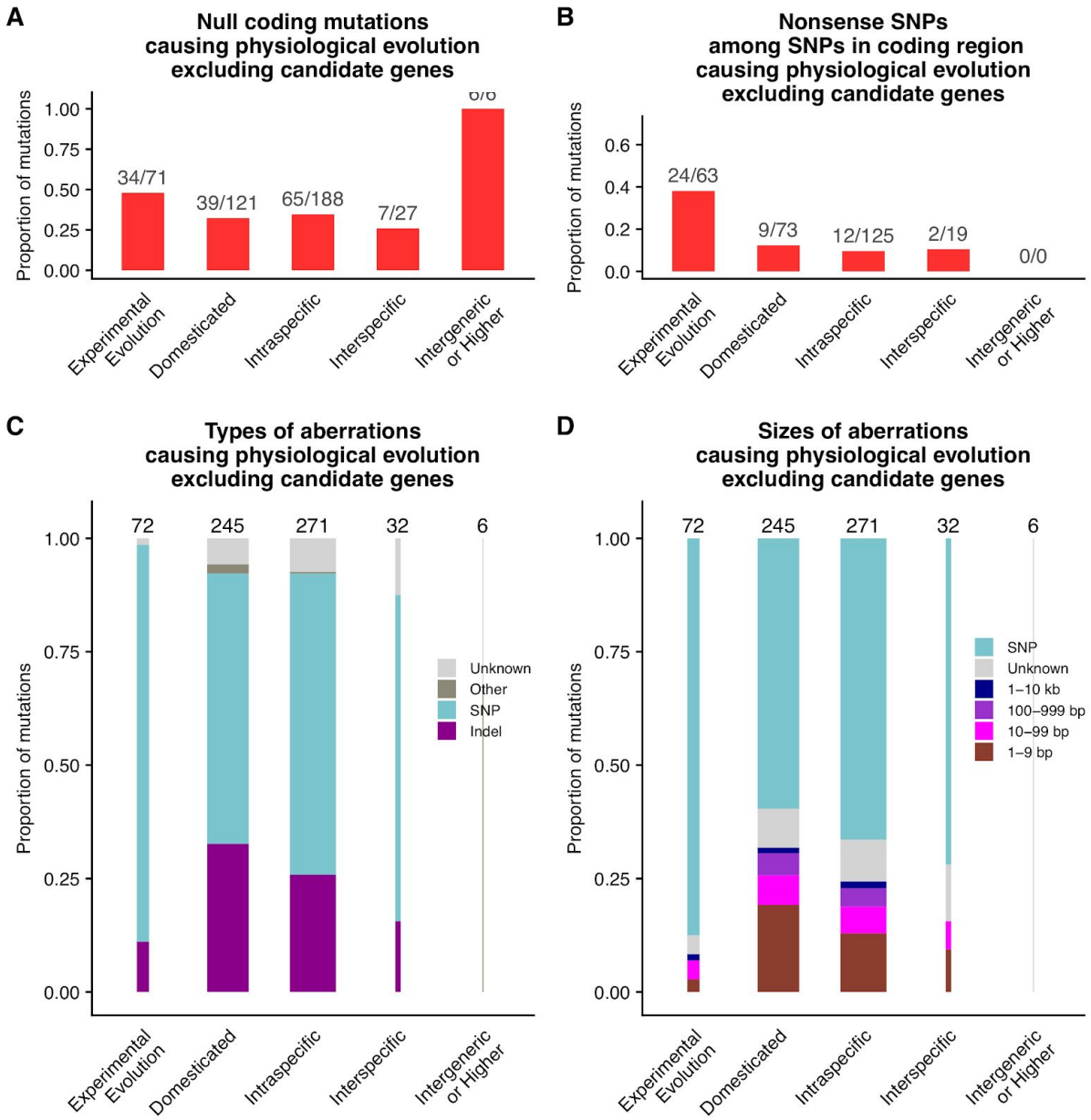

Fig. S4. The proportion of null and disruptive mutations among the coding mutations associated with physiological evolution identified via methods distinct from the Candidate Gene Approach tends to decrease with evolutionary time. Same legend as Fig. 3. The six cases curated as "Intergeneric or Higher" correspond to one study where authors performed a phylogenetic- and genome-wide screen for genes that have been inactivated repeatedly during evolution, in significant association with two metabolic phenotypes, the loss of the ability to synthesize vitamin C, and low levels of biliary phospholipids (Hiller et al., 2012).

Supplementary Reference

Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R. and Bejerano, G. (2012). A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep.* **2**, 817–823.