Detection of Domestication Signals through the Analysis of the Full Distribution of Fitness Effects using Forward Simulations and Polygenic Adaptation

Ioanna-Theoni Vourlaki[1†], David Castellano[2†], Ryan N. Gutenkunst[2*], Sebastian E. Ramos-Onsins[1*]

1. Centre for Research in Agricultural Genomics (CRAG), Autonomous University of Barcelona (UAB), Vall de la Moronta s/n, Cerdanyola del Valles, Barcelona, 08193, Spain.

2. Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ, USA.

[†] Equal contribution

*Co-corresponding authors

ABSTRACT

As a consequence of the process of domestication, wild and domestic individuals are adapted to very different environmental conditions. Although the phenotypic consequences of domestication are observable, the genetic causes are not evident in many cases. Artificial selection could be modifying the selection coefficients of new and standing variation in the population under domestication. Here, we aim to detect a genome-wide signal of domestication under a model of polygenic adaptation. We use forward simulations to investigate the 1D and 2D site frequency spectra (SFS) of mutations in two populations (Wild and Domestic) with divergent histories (demographic and selective) following a domestication split. We simulate ten different scenarios, varying the strength of selection upon beneficial mutations and the proportion of mutations whose selection coefficients change after domestication. First, we describe that in domesticated populations selection at linked sites needs to be invoked to explain the SFS of neutral mutations and that the mode of linked selection affecting the neutral SFS depends on the duration of the domestication bottleneck. Second, we find that some aspects of the full distribution of fitness effects (DFE), such as the shape and strength of the deleterious DFE, are accurately estimated in both populations when using only the 1D-SFS. However, the detection of significant differences in the beneficial DFE between populations remains challenging in most, but not all simulated scenarios when only the 1D-SFS is used. Third, when considering the 2D-SFS and a new joint DFE model, we are able to detect more subtle differences in the full DFE that are hidden in the 1D-SFS analysis. In conclusion, our work highlights the strengths and limitations of detecting a polygenic signal of domestication under a variety of domestication scenarios and genetic architectures.

INTRODUCTION

The increase in human population size and the emergence of modern society are linked to the domestication of plants and animals (Purugganan and Fuller 2009; Driscoll et al. 2009; Larson and Burger 2013; Amills et al. 2017; Stetter et al. 2018). Human civilization as we know it was possible thanks to the domestication of surrounding life forms, where people started to domesticate plants and animals such as wheat, dogs, pigs, or chickens (Avni et al. 2017; Dayan 1994; Redding 2015; Zeder 2012). Domestication is a process that allows humans and other species to obtain a mutualistic long-term relationship that implies an advantage for both species (Zeder *et al.* 2006). The domestication of fauna and flora by humans started approximately 10-15 thousand years ago and is still in progress (Larson et al. 2014; Zeder 2015). Although human civilization relies on domestication, we still lack a full genomic and evolutionary understanding of domestication. Domestication is a rapid process in terms of its evolutionary time scale, but it is not a discrete event, and it implies the gradual improvement of domesticated traits. The artificial selection generated by humans during domestication can be considered relatively stronger and hence faster than natural selection because the selective pressure imposed by humans tends to be very extreme. Moreover, domestication tends to be associated with bottlenecks; only a small number of individuals from the wild population become domesticated, which is expected to decrease the efficiency of natural selection (Wright *et al.* 2005). Another important difference between natural and artificial selection is that modern breeders usually apply truncation selection, that is, the selection of the top percentage individuals for the desired trait (Granleese *et al.* 2019). The prevalence of truncation selection in nature, or before the industrialization era, is unknown. Truncation selection is known to be an easy and efficient form of directional selection (Crow and Kimura 1979), and no severe accumulation of genetic load is expected in outcrossing species (Kondrashov 1988; Ohta 1989) if the population size remains large enough (Marsden *et al.* 2016). A recent comprehensive meta-analysis about the genetic cost of domestication (Moyers *et al.* 2018) found that harmful variants are more numerous (or segregate at higher frequencies) in domesticated populations compared to their wild relatives. This pattern is likely driven by a number of processes, which jointly decrease the efficacy of selection in domesticated populations as first suggested in rice genomes (Lu *et al.* 2006).

Selection, both natural and artificial, can occur either through a few loci with strong effects or via many loci with small effects (Jain and Stephan 2017a; b), depending on the genetic architecture of the trait and the strength of selection on it. Different patterns of genetic diversity around the selected loci are expected in response to these two models of selection (Stephan and John 2020). Classic hard selective sweeps have been reported in few candidate loci under important domesticated traits (Andersson 2012), such as the IGF2 gene region associated with lean domestic pigs (van Laere *et al.* 2003), the thyroid-stimulating hormone receptor (*TSHR*) in domestic chickens (Rubin *et al.* 2010) or the loci *sh4* and *qSW5* (Li et al. 2018; Shomura et al. 2008), which are involved in the traits of seed shattering and grain width in

domestic rice, respectively (Huang *et al.* 2012). These examples are consistent with a simple Mendelian genetic architecture where few loci determine most of the variance in the domesticated trait.

Polygenic adaptation, in contrast, describes a process in which a constellation of small changes in allele frequencies modify the differences in the trait under selection. A wide range of population genetics models and simulations describing polygenic adaptation have been investigated (e.g., Stephan 2016, de Vladar and Barton 2014). Some models analyze the polygenic response of a trait in the presence of mutation and stabilizing selection (Jain and Stephan 2015; de Vladar and Barton 2014), while others capture the response of a trait under mutation and stabilizing or directional selection after an environmental shift in a finite size population (Stephan and John 2020). The interplay between hard selective sweeps and polygenic adaptation has been theoretically studied by Höllinger et al. (2019) by measuring how selection and genetic drift favour sweeps over small allele frequency shifts and vice versa using a finite population size. Thornton (2019) simulated the dynamic effect of selected and neutral mutations at a single quantitative trait, showing that the new optimum trait in a population of finite size is reached before selective sweeps are completed in polygenic models with many involved loci. Stetter et al. (2018) showed the impact of different demographic conditions, such as population bottlenecks or exponential growth, on domestication in maize via polygenic adaptation and for traits with major effect loci. They performed an exhaustive simulation analysis using a limited number of QTLs under an additive model and observed the presence of selective sweeps, even for small effect size mutations, after sudden environmental changes in traits under stabilizing selection. They concluded that the effect size of new mutations, as well as demography, are the main parameters influencing the observed genetic architecture. In practice, polygenic adaptation is harder to detect than classic selective sweeps (Pritchard *et al.* 2010). However, polygenic adaptation has been detected in some particular studies in wild and domesticated populations. Genome-wide association studies (GWAS) combined with population genetic modeling could potentially lead to the detection of signals of polygenic adaptation (Berg and Coop 2014). Combining gene association and multivariate (redundancy analysis) methods have contributed to the identification of polygenic adaptation in the threatened fish species *Maccullochella peelii* (Harrisson *et al.* 2017). In *Drosophila*, Evolve and Resquencing (E&R) experiments that modified the regime of the temperature of populations have detected a polygenic adaptive response that could be explained by functional redundancy and quantitative traits (Barghi *et al.* 2019). Under climate change, the coral ecological divergence has been explained by the polygenic adaptation of many amino acid-changing variants (Rose *et al.* 2018). Small effects of polygenic variants seem to explain most of the existing variation in maize domesticated traits (Xue *et al.* 2016). It has been suggested that in animals, the initial domestication process may have acted on complex behavioral traits such as taming (Trut *et al.* 2009), which others have found to be affected by a large number of genes (Jasinska and Freimer 2009).

In this study, we wonder to what extent we are able to detect a genomic signal of domestication using a different approach: the comparison of the full distribution of fitness effects (DFE) on new and standing variation. To do that, we first compared the 1-dimensional (1D) and 2-dimensional (2D) unfolded site frequency spectra (SFS) of synonymous and nonsynonymous variants between domesticated and wild *in silico* populations. Second, we compared the inferred full DFE of new nonsynonymous mutations in those populations. The DFE of new deleterious mutations has been previously estimated contrasting the 1D-SFS of synonymous and nonsynonymous mutations of a multitude of species assuming that beneficial mutations only contribute to divergence but not to polymorphism because of their rapid fixation in the population (Barton and Zeng 2018; Boyko *et al.* 2009; Keightley and Eyre-Walker 2007; Kim et al. 2017. Tataru et al. (2017) proposed a model, polyDFE, for the inference of the full DFE and the proportion of adaptive substitutions (α) by using polymorphism data exclusively. Castellano et al. (2019) used polyDFE to compare the full DFE of new amino acid mutations across great apes, finding that the shape of the deleterious DFE is constant across this set of closely related species. Recently, using the 2D-SFS, a new method to jointly estimate the full DFE between two populations that have recently diverged and share many polymorphisms has been proposed (Huang *et al.* 2021). However, there are few studies comparing the DFE between domesticated and wild populations (Leno-Colorado *et al.* 2020). Here, we use forward simulations, considering the domestication process under different demographic and selective models. Several combinations of genetic architectures and selective effects have been simulated, from one considering a relatively small number of loci changing their selective effects to another of polygenic adaptation where many loci have divergent selective effects. In all cases, the selective effects of a proportion of existing variants can change (from deleterious to beneficial, and vice versa) in the domesticated population.

MATERIALS AND METHODS

Simulation of the Domestication Process with Selection and Demographic Changes

A simulation analysis of the domestication process is developed using the forward-in-time simulator SLiM2 (Haller and Messer 2016). This tool is very versatile and allows the introduction of a number of variable scenarios and parameters. The general model for the domestication process is developed in the SLiM script in the Zenodo database (10.5281/zenodo.7017885). Ten different scenarios of domestication are analyzed, and the parameters for each of the scenarios are shown in Table 1. All the options (flags) used for running the SLiM script are shown in a file in Zenodo (10.5281/zenodo.7017885). Briefly, the constructed model starts from a single panmictic population of *Ne* = 500 diploid individuals, with a genome containing 10 independent chromosomes, each with 1000 loci of 1500 base pairs of length, and each locus having one-third (4-fold) neutral synonymous positions and two-thirds of (0-fold) selected nonsynonymous positions interspersed along the locus. Each locus is separated from each other at different recombinational distances, following a convex curve in which the loci located at the telomeres are at a longer recombinational distance between loci (2.5e-2), while the loci near centromeres are closer in the recombinational distance (2.5e-4). The recombination within loci is fixed to a rate of 2.5e-7 between positions. Given the high computational burden of simulating entire chromosomes using variable recombination values and a considerable number of coding sites (1,500,000 coding sites per chromosome and 10 independent chromosomes) and the large number of mutations obtained, we perform a single run for each of the 10 scenarios. Note that the higher recombination among loci aims to mimic their real genetic distance separation but severely speeds down the simulation.

The demographic parameters for each scenario (Figure 2) are as follows: the initial blank population run for 10**Ne* generations to reach mutation-selection-drift equilibrium, then splits into two equal populations (outgroup population and the wild population), and after 10**Ne* generations, the wild population splits again into domestic and wild populations. Hereafter we refer to the Wild and Domestic populations. We aim to mimic a realistic animal domestication process, such as pig, where ancestral *Ne* estimates were around 10,000 (Groenen et al. 2012) and the domestication process occurred around 10,000 years ago (Zeder *et al.* 2006). The generation time was here assumed at 3 years per generation (Zhang et al. 2022). Two very different conditions for the bottleneck process are studied. The Domestic population suffers a bottleneck, reducing its population size temporarily to 50 diploid individuals, to recover again to *Ne* diploid individuals after the bottleneck. The bottleneck elapsed either 2*Ne**0.016 generations for scenarios with a short bottleneck or 2*Ne**0.161 generations for long bottleneck scenarios. Moreover, the simulation finished either in 2Ne*0.15 generations after the initiation of the short bottleneck process or 2Ne*0.005 generations of the long bottleneck. The selective effects produced by domestication are modelled by changing the fitness values of a proportion of the existing and new mutations in the domestic population (at the time of the split).

Most new nonsynonymous mutations occurring in ancestral and Wild populations are under negative selection (97.5%), and only 2.5% of the mutations have positive fitness effects. Domestic populations show different proportions of (new and standing) beneficial and deleterious variation depending on the scenario (Table 1). The negative effects in all scenarios and populations follow a gamma distribution with a shape value of 0.2 and a mean of S_d = -10 ($2N_e s_d$ when homozygote), while variants with positive effects follow an exponential distribution with a mean S_b = 1 or S_b = 10 ($2N_e s_b$ when homozygote) depending on the scenario (Table 1).

Types of Sites

The sites are initially divided into seven different types (named m_1 to m_7), being m_1 neutral (synonymous) and m_2 to m_7 functional (nonsynonymous) sites having a different selective effect when mutated (see Table 2). Mutations at m_5, m_6 and m_7 sites generate deleterious variants in the Wild population, and mutations at m_2, m_3 and m_4 sites generate beneficial mutations in the Wild population. The selection coefficient of mutations generated at m_2 (beneficial) or at m_5 (deleterious) sites are invariant for the Wild and Domestic populations. However, the mutations at m_3, m_4, m_6 and m_7 sites will change their selective effect in the Domestic populations relative to the Wild populations. That is, the new selective effect is drawn from the corresponding DFE section (positive or negative), independently of their value in the wild population. Those can be understood as sites with divergent selective effects. The selection coefficient of a given beneficial mutation at m_3 sites will remain beneficial in the Domestic population, but it will be different from the original beneficial effect at Wild. A mutation at m_4 sites will change its selection coefficient from beneficial in the Wild to deleterious in the Domestic population. Equivalently, the selection coefficient of a deleterious mutation at m_6 sites will remain negative in the Domestic population but it will be different from that found at Wild. A mutation at m_7 sites will change its selection coefficient from deleterious in the Wild to beneficial in the Domestic population. This hard-coding of selective effects on different sites allows us to gain insight into the relative importance of each mutation type for the domestication process.

Type of variants

The variants are classified into total (all observed variants), exclusive (variants that are present in a single population) and shared (variants that are present in both the Wild and Domestic populations). Note that exclusive and shared variants are not exactly coincident with new and standing variants, that is, new variants are those mutations that appear after the split between Wild and Domestic populations while standing variants are variants that appeared before the Domestication split. However, a new variant happening in the Wild population can be shared if a migration event transferred this variant to the Domestic population, and a standing variant can become exclusive if this variant disappears in one of the populations

(e.g., Lee and Coop 2017). The proportion of the different types of sites, polymorphisms and substitutions across scenarios can be found at Supplementary Figure S0 and Table S0.

Simulating the Domestication Process only with Demographic Changes

To disentangle selective from demographic effects in our comparative analysis of the SFS we used the *ms* (Hudson 2001) coalescent simulator to simulate samples according to demographic parameters under short and long bottleneck periods, with or without migration (Figure 2). The demographic parameters used here are the same as the ones used for simulating the different scenarios described above with SliM2. However, the parameters are re-scaled according to software requirements (that is, considering *Ne,* see the scripts at Zenodo). Here, the size of the simulated coding regions is 2000 bp. *ms* outputs are processed using an R script (10.5281/zenodo.7017885) to calculate the site frequency spectrum for total, shared, and exclusive variants, divergence, and the estimates of genetic diversity.

Genetic Diversity Summary Statistics and the Fraction of Adaptive Substitutions ($\boldsymbol{\alpha}$)

The estimates of genetic diversity, population differentiation and divergence, and the parametric inference of the DFE and demographic patterns are computed in a sample of 20 diploid individuals (n=40 haploid chromosomes) per population. The 1D and 2D SFS are polarized using the simulated outgroup population. The number of mutations occurring in each population and the variants that become fixed and remained polymorphic in one or both populations is recorded. The fraction of adaptive substitutions ($\boldsymbol{\alpha}$) is estimated per population using the estimation methods proposed by Smith and Eyre-Walker (2002) and by the asymptotic McDonald & Kreitman test method (MKT, Messer and Petrov 2013), which corrects for the underestimation of α due to slightly deleterious mutations. The asymptotic MKT is computed using the web service available at http://benhaller.com/messerlab/asymptoticMK.html. Here, the default cutoff interval of x [0.1, 0.9] is used. PolyDFE (Tataru and Bataillon 2019) α estimates are computed using only polymorphisms counts to estimate the negative DFE (using model f1 from Table 5). Then, the expected number of non-adaptive nonsynonymous substitutions is contrasted with the observed number of nonsynonymous substitutions to estimate α. We also estimated α not using substitutions, but these estimates turn out to be very noisy (data not shown). Since we know the selection coefficients of mutations before and after the domestication split, all these summary statistics are investigated for different types of sites. The estimates of variability per site using shared (and exclusive) variants are calculated considering the total of positions under study (e.g., synonymous or m_1, and nonsynonymous or m_2…). That is, the sum of shared plus exclusive variability is equal to the total variability for such type of positions.

Distribution of fitness effects (DFE): Two complementary approaches

polyDFE: 1D-SFS

We use the polyDFEv2.0 framework (Tataru and Bataillon 2019) to estimate and compare the DFE across Wild-Domestic population pairs by means of likelihood ratio tests (LRTs). We use the R function *compareModels* (from https://github.com/paula-tataru/polyDFE/blob/master/postprocessing.R) to compare pairs of models. The inference is performed only on the unfolded SFS data (divergence counts to the outgroup are not fitted), and unfolded SFS data are fitted using a DFE model comprising both deleterious (γ-distributed) and beneficial (exponentially distributed) mutations. Note that in polyDFE s is defined to be the selection coefficient on the heterozygote (like in *dadi*), but the scaled selection coefficient is defined as $4N_e s$. The DFE of each Wild-Domestic population pair is inferred using the 1D-SFS of each population. polyDFE assumes that new mutations in a genomic region arise as a Poisson process with an intensity that is proportional to the length of the region and the mutation rate per nucleotide (μ). We assume that μ remains constant across simulations (as it is the case). Both an ancestral SNP misidentification error (ε) and distortion parameters (r_i) are estimated. However, we notice that the exclusion of ε does not affect the rest of estimated parameters because under the simulation conditions few sites are expected to be misidentified. The r_i parameters are fitted independently for each frequency bin (from n = 1 to n = 39), and they are able to correct any distortion that affects equally the SFS of synonymous and nonsynonymous variants (such as, in principle, demography or linked selection). To obtain the sampling variance of parameter estimates and approximate confidence intervals, we use a bootstrap approach. Bootstraps are generated by resampling the SNPs in chromosomal "chunks" or segments 100 times using an ad hoc R script (link here). These segments are non-overlapping and 100,000 bps long. Hence, we assume that SNPs are only independent across segments but not within segments. Model averaging provides a way to obtain honest estimates that account for model uncertainty. To produce the model average estimates of the full DFE we weight each competing model according to their AIC following the equation 6.1 shown in the polyDFEv2 tutorial. We use the R function getAICweights (from https://github.com/paula-tataru/polyDFE/blob/master/postprocessing.R to do the model averaging R) to obtain the AIC values.

dadi: 2D-SFS

dadi (Gutenkunst et al. 2009) is employed to infer the joint distribution of fitness effects (Jerison et al. 2014, Ragsdale et al. 2016, Huang et al. 2021) and the demographic history of all simulated population pairs. Our new model for the joint DFE between the two populations is a mixture of multiple components designed to mimic the selected mutation types in the simulations (Table 2; Figure 1). The major exception is that beneficial mutations are modelled to have a single fixed selection coefficient, rather than arising from a distribution. Let p_{+w} be the fraction of mutations that are positively selected in the Wild population, p_c be the fraction of mutations that change selection coefficient in the Domestic population, and p_{c+} be the fraction of those mutations that become beneficial in the Domestic population. To model mutation types m_2 and m_3, a proportion $p_{+w}(1-p_c) + p_{+w}\, p_c\, p_{c+}$ of mutations are assumed to have the same fixed positive

selection coefficient in both populations. To model m_4, a proportion $p_{+w}\ p_c\ (1-p_{c+})$ are assumed to have a fixed positive selection coefficient in the Wild population and a gamma-distributed negative selection coefficient in the Domestic population. To model m_5, a proportion $(1-p_{+w})(1-p_c)$ of mutations are assumed to have equal negative gamma-distributed selection coefficients in the two populations. To model m_6, a proportion $(1-p_{+w})\ p_c\ (1-p_{c+})$ are assumed to have independent gamma-distributed selection coefficients in the two populations. To model m_7, a proportion $(1-p_{+w})\ p_c\ p_{c+}$ mutations are assumed to have a gamma-distributed negative selection coefficient in the Wild population and a fixed positive selection coefficient in the Domestic population. All gamma distributions are assumed to have the same shape and scale . This model is implemented as in *dadi* as the function dadi.DFE.Vourlaki_mixture.

For inference, demographic models are first fit to neutral mutations from each simulation, then the new proposed joint DFE model is fit to the selected mutations. Demographic models (Figure 2) are estimated by running 100 optimizations per simulated data set. Then, the 2D-SFS for selected sites are precomputed conditional on the demography for 104 values of γ ($2N_A * s$, a population scaled selection coefficient for the heterozygote where $N_A$is the ancestral population size), 102 negative and 2 positives. For the negative part of the DFE, γ values were logarithmically equally spaced between -2000 and -10^{-4}. The expected DFE for selected sites can then be computed as a weighted sum over these cached spectra (Kim et al, 2017). The DFE parameters shape (α), scale (β), p_{+w}, p_c, and p_{c+} are then estimated by maximizing the Poisson likelihood of the simulated data, with the nonsynonymous rate of mutation influx θ fixed to twice that inferred for neutral sites in the demographic history fit. For the DFE inference, optimization is repeated until the best three results are within 0.5 log-likelihood units. Ancestral state misidentification is not modelled, because under the simulation conditions few sites are expected to be misidentified. Uncertainties of DFE parameter inferences are calculated by conventional bootstrapping, holding the demographic model fixed and dividing the simulated data into non-overlapping regions of 100,000 basepairs. Note that this procedure does not propagate uncertainty in demographic parameters through to the DFE parameters.

For the purpose of this work, *dadi* software is downloaded and installed according to the instructions provided at the following link: https://bitbucket.org/gutenkunstlab/dadi/src/master/. Since *dadi* operates as a module of Python, the Anaconda3 and Spyder (Python 3.7, Rossum and Drake 2006; Anaconda 2016; Raybaut 2009) versions are used in this study.

Data Availability Statement

The file Supplemental Material 1 includes Tables S0-S15. The file Supplemental Material 2 includes Figures S0-S3. The file Supplemental Material 2 includes Figures S4-S13 with log2 ratio of the simulated SFS for all scenarios versus the expected SFS under the same demographic scenario. The file Supplemental

Material 3 includes Figures S14-S23 with log2 ratio of the simulated SFS for shared variants in all scenarios versus the expected SFS under the same demographic scenario. The file Supplemental Material 3 includes Figures S24-S33 with log2 ratio of the simulated SFS for exclusive variants in all scenarios versus the expected SFS under the same demographic scenario. Additional information such as the entire outputs of the simulations, scripts and more detailed tables have been deposited in Zenodo database (10.5281/zenodo.7017885).

.

RESULTS

Studying the effect of domestication on the DFE of natural populations can be very challenging, especially if the available methods for inferring and comparing the DFE have not been benchmarked using exactly the same dataset. Within the scope of the present work, we conduct a simulation study using different combinations of parameters relevant to the domestication process. We aim to investigate the ability to detect the effect of domestication in the SFS and the full DFE of a domesticated population that is experiencing a large or small change in the number and selective effects of loci under domestication/selection after a short or a long bottleneck period, with or without migration. Hereafter, we refer to the Wild and Domestic populations.

The Wild populations have a constant DFE and constant population size. Beneficial mutations arise at Wild populations at a relatively low percentage (2.5%) following an exponential distribution, the rest of mutations are drawn from a gamma distribution with shape 0.2 and mean S_d=-10 (where $S_d=2N_es$, s is the selection coefficient in the homozygote, and N_e = 500 is the ancestral effective population size, see Material and Methods: Simulating the Domestication Process). All mutations, beneficial and deleterious, are co-dominant. The Domestic population originates from the Wild population through a bottleneck and a concomitant change of the selective effects in a proportion of nonsynonymous sites (Figure 2; Table 1).

The change in selective effects impacts new mutations in the Domestic population and pre-existing variants that originated before the domestication event. In other words, not only mutations that were deleterious (or beneficial) before the split can become beneficial (or deleterious) in the Domestic population, but even when the sign of the selection coefficient remains the same, the strength of selection can be modified. See Table 2 for all the combinations of changes in selective effects between Wild and Domestic populations. The simulated scenarios aim to cover a variety of possible changes in the genetic architecture (number of loci) and the strength of selection (selection coefficients) of the trait/s under domestication. Two selection coefficients for beneficial mutations are assumed: (i) strong, with a mean population-scaled selection coefficient of S_b=10 and (ii) weak, with S_b=1 when homozygote. Depending on the scenario, a selective change occurs only at a few (5%) or at a large proportion (25%) of positions in the Domestic population (Table 1, "%change" column). We leave three scenarios (scenario 1, 4 and 7) as negative controls; the DFE in the Domestic and Wild populations is the same. Among those positions showing an alteration of fitness effects, we also vary the proportion that is beneficial (Table 1, "%Positive" column). In all scenarios, the fraction of positions (and polymorphisms) with beneficial mutations in the Domestic population is equal or larger than in the Wild population and varies across scenarios (Table 1, see Figure S0 for the exact site composition of the Domestic populations in each scenario). Moreover, demographic changes affect only the Domestic population. Three demographic models are simulated: (i)

short bottleneck with migration, (ii) short bottleneck without migration and (iii) long bottleneck without migration. There is only migration from the Wild to the Domestic populations. Table 2 shows the different types of mutations according to their fitness in Wild and Domestic populations.

Descriptive Summary Statistics of the Simulated Populations

Approximately 120K mutations are observed under each of the scenarios in the Wild population, with variable numbers, between 53 to 60%, of polymorphic variants (Table S1A), while the rest are fixed (relative to the outgroup). The comparison among scenarios shows that those with a higher positive DFE mean (S_b=10, scenarios 1, 6, 7 and 10) exhibit a slightly higher proportion of fixations (~54K vs 47K) than those scenarios having an S_b=1 (first column in Table 3). This excess of fixed variants affects only nonsynonymous positions, while synonymous positions have the same number of fixations in all scenarios as expected (~22K, Table S1B) (Birky and Walsh 1988). Nucleotide diversity within the Wild and Domestic populations and the divergence (from the simulated outgroup) at nonsynonymous versus synonymous positions are shown in Table S2. The lower divergence ratio (D_n/D_s) versus polymorphism ratio (P_n/P_s) suggests an excess of deleterious polymorphisms at nonsynonymous sites, as expected given the simulated DFE parameters. The number of exclusive variants in the Domestic populations and shared variants between Wild and Domestic populations across all scenarios is shown in Table 3, as well as the mutations that were initially shared but that eventually got fixed in one of the populations.

In Domestic populations, we observe that the absolute number of exclusive polymorphic (SxD) and fixed (SfD) variants are more affected by the demographic bottleneck than by the strength of positive selection. As expected, the number of shared mutations (polymorphic in both populations, Ssh) is very high in scenarios with migration (scenarios 1-3) and very low in models with a long bottleneck and no migration (scenarios 7-10). The number of fixed variants in Wild which are still polymorphic in Domestic (SfWxD) is very low in scenarios with long bottlenecks without migration, while the number of fixed variants in Domestic which are still segregating in Wild (SfDxW) is the largest in scenarios with long bottlenecks with no migration.

The Fraction of Beneficial Substitutions: Comparing α across Scenarios and Types of Sites

Next, we ought to know how the strength of positive selection and the number of loci under positive selection affect the fraction of beneficial nonsynonymous substitutions (α) across scenarios (Table 4, see also Table S3 for the absolute number of fixations). At Wild, true α values correlate with the simulated conditions in Table 1; higher α values can be found in scenarios where strong selection is assumed (scenarios 1, 6, 7 and 10). In contrast, in Domestic (to observe the effects of domestication, α is calculated counting variants that are fixed in the Domestic and are polymorphic or absent in Wild), the highest α values are found in scenarios 1, 3, 5 and 9. These scenarios (except scenario 1) have a large fraction of

deleterious sites in the Wild population that become beneficial in the Domestic population (m_7 sites, Table 2). We also find that a substantial amount of the beneficial substitutions (25-60%) in those scenarios come specifically from m_7 sites.

In relation to understanding the contribution of exclusive variants (mostly new mutations that occurred after the domestication split) to current α estimates, we observe that exclusive beneficial substitutions are contributing very modestly to the total number of beneficial substitutions for short bottleneck scenarios (Table S3). This suggests that adaptive amino acid substitutions after short bottlenecks come mostly from standing variation that were present before the domestication split. In contrast, for long bottleneck scenarios, the number of exclusive fixed beneficial mutations is higher at the Domestic population than the Wild population and around 5 times larger than short bottleneck scenarios without migration (Table S3). This is because in small populations the time to fixation decreases for all mutations, neutral, beneficial and deleterious. For long bottleneck scenarios, exclusive beneficial substitutions, or new mutations, are playing an important role and they contribute substantially to the realized α at the end of the simulation.

True α vs. estimated α

In order to investigate if current α estimators can be used to extract meaningful conclusions from natural wild and domesticated populations, we estimate α for Wild and Domestic populations using different methods (Figure 3 and Tables S4-6 for confidence intervals). As expected, the standard MK test (Smith and Eyre-Walker 2002) underestimates α due to the segregation of deleterious mutations at low frequency. The estimates are improved when the asymptotic MK test is used (Haller and Messer 2017), which corrects for the excess of deleterious mutations at low frequencies. Nonetheless, the asymptotic MK test is not very accurate in scenarios with long bottlenecks. The asymptotic MK test detects positive selection in the Wild population only for scenarios with strongly beneficial mutations (scenarios 1, 6, 7 and 10). In the Domestic populations, significantly positive α values are estimated only for scenarios with short bottlenecks and strong positive selection (scenarios 1 and 6). Finally, we also estimate α using the polyDFE framework (Tataru et al. 2017). polyDFE α estimates tends to be less noisy and slightly more accurate than the asymptotic MK test. However, qualitatively speaking it suffers the same weaknesses than the asymptotic MK test under long bottlenecks. Our list of tested α estimators have very limited power to detect the subtle differences in α between these simulated population pairs.

Comparison of the Allele Frequency Distributions across Populations and Scenarios

Figure 4 shows the ratio (in $\log_2$ scale) between the observed derived allele frequency and the expected derived allele frequency, or unfolded site frequency spectrum (uSFS), under the same demographic model but without selection (obtained by coalescence simulation with *ms*, Hudson 2001). In

the absence of (direct or indirect) selection, the observed derived allele frequency and the expected derived allele frequency should be the same and the $\log_2$ ratio equal to zero. The small and uniform $\log_2$ ratio for synonymous mutations in Wild populations suggests that there is good agreement between the simulated data and our naive neutral expectation in the absence of selection. Thus, we do not find evidence of linked selection affecting the synonymous SFS in Wild populations. In contrast, the nonsynonymous derived allele frequency distribution in Wild populations shows an excess of low-frequency variants due to the direct effect of negative selection on those mutations and exhibits a considerable lack of intermediate- and high-frequency variants. However, in scenarios with strong positive selection (scenarios 1, 6, 7 and 10), beneficial nonsynonymous variants leave a signal in the SFS. Beneficial variants generate an increase of high-frequency variants that resembles (or matches) the neutral expectation at nonsynonymous positions (see Figures S4-S13 under these models).

The Domestic populations have suffered a shift in the selection coefficients at a sizeable number of mutations (both shared and exclusive) together with a demographic bottleneck, with or without migration from the Wild population. For synonymous variants, we show that the observed allele frequency distribution seems to be disturbed by indirect selection on nearby nonsynonymous mutations in most scenarios (Figure 4). In scenarios with long bottlenecks (scenarios 7-10), only the lowest and highest frequency bin are in excess, and all other frequency bins show a remarkable lack of synonymous variants. We find that the allele frequency distribution for nonsynonymous mutations in Domestic populations under long bottlenecks is very similar to those found for synonymous variants but very different to the expected in the absence of selection under a purely demographic model (Figure 4 and S1). Thus, under long bottlenecks allele frequency distributions seem deeply affected by (direct and indirect) selection. Similarly, synonymous variants also seem to be under indirect selection at scenarios with short bottlenecks having a large percentage (25%) of nonsynonymous mutations shifting their selection coefficients (scenarios 3, 5 and 6). Synonymous variants show an excess of singletons, a lack of low- to intermediate-frequency variants, and, again, a remarkable excess of mutations at high frequency. This result suggests that synonymous (and other non-beneficial nonsynonymous) mutations might be hitch-hiking with beneficial nonsynonymous mutations after the domestication split (Figures S4-S13) (Hartfield and Otto 2011). Interestingly, scenarios 3, 5, and 6 also show an increase of nonsynonymous variants at high frequency but not as strong as the one shown for synonymous variants. We show that for scenarios 3, 5 and 6, the allele frequency distribution for nonsynonymous mutations occurring at sites m_2, m_3 and m_7 (which are beneficial at the Domestic population) resemble that described for synonymous variants. However, this pattern is much weaker for deleterious nonsynonymous mutations occurring at deleterious sites (m_5 and m_6 sites, Figures S4-S13).

Finally, we analyze the uSFS for shared and exclusive variants. Here our aim is to determine the contribution of shared and exclusive variants to the observed uSFS and the process of domestication, particularly for selectively neutral synonymous variants that later will be important to correct (or infer) the demographic changes sufered by nonsynonymous variants. Again, we compare the uSFS of synonymous shared and exclusive variants in relation to the expected neutral shared and exclusive uSFS under the same demographic model in the absence of selection (Figures S14-S33 for all scenarios and Figures S2-S3 for the ratio of the demographic neutral pattern versus the Wright-Fisher Standard Neutral Model). In general, for short bottleneck scenarios (scenarios 1-6) we observe a modest impact of domestication on shared synonymous variants (Figures S14-S23). Except in scenario 3. There is an absence of intermediate frequency shared synonymous variants and an excess of low and high frequency shared synonymous variants in scenarios 3 relative to the expected in the absence of (linked) selection. Indeed, in scenario 3, the uSFS of shared synonymous variants is markedly disturbed, and surprisingly, all types of shared mutations, including deleterious mutations (m_5, m_6, m_4 in Domestic), exhibit this excess of high frequency variants that is not explained by demography alone (Figure S16). The high proportion of sites that change their selective effect together with migration from Wild to Domestic populations may explain this excess of shared polymorphisms at high frequency in scenario 3. Other scenarios which do not share this parameter combination do not show this pattern. In scenarios with long bottlenecks (scenarios 7-10) the SFS of shared synonymous variants in Domestic populations is massively disturbed. There is a large excess of shared singletons and a large absence of shared variants at the rest of frequency classes. In relation to exclusive mutations (Figures S24-S33), in scenarios with migration, short bottlenecks and a change in the DFE (scenario 2 and 3), there is an excess of exclusive synonymous variants at all frequencies (except singletons) suggesting that the contribution of exclusive variants to neutral diversity levels can be boosted in the presence of migration from the Wild population. For scenarios with short bottlenecks without migration (scenarios 4-6) there is a lack of exclusive synonymous variants in most frequency bins except for very low and high frequency bins where there is an excess. In scenarios with long bottlenecks (scenarios 7-10) the SFS of exclusive synonymous variants in Domestic populations is again massively disturbed. There is a large excess of almost fix mutations and a large absence of exclusive variants at the rest of frequency classes.

In conclusion, our hypothetical domestication processes (in the absence of free recombination) tend to affect more exclusive synonymous variants than shared synonymous variants. In the absence of migration, domestication tends to deplete the SFS of exclusive neutral mutations (and under long bottlenecks this depletion also affects shared synonymous variants). In the presence of migration, exclusive neutral variants become more abundant than expected in the absence of selection (and if the change in the DFE is big, also shared synonymous variants at high frequency become more abundant than expected).

Estimation of the DFE: detecting differences in DFE between Wild and Domestic populations

In the previous section we showed the pervasive impact of linked selection on synonymous polymorphisms. In the Domestic populations, the neutral SFS cannot be explained by the underlying true demography alone. The two methods we employ to estimate the DFE of nonsynonymous mutations either are agnostic to the underlying demography and just aim to correct for any distorter that affects equally the synonymous and nonsynonymous SFSs (polyDFE, Tataru and Bataillon 2019), or aim to first estimate the underlying demography using the synonymous SFS and then use that inferred demography to estimate the DFE parameters (*dadi*, Gutenkunst et al. 2009). Hence, the question is to what extent the nuisance *r* parameters from polyDFE or the inferred demographic model from *dadi* will be enough to recover the true DFE parameters in our simulations. In this work we are not particularly interested in recovering the true demography (given the pervasiveness of linked selection), but in recovering the true changes in the DFE.

polyDFE: 1D-SFS

We investigate whether polyDFE is able to capture differences in the DFE of Domestic and Wild populations across 10 possible domestication scenarios (Table 1). We run five nested models (Table 5) and compare them by means of likelihood ratio tests (LRTs) (Table 6).

Under an agnostic approach (not assuming any *a priori* information about our datasets), LRTs between a number of nested models allow us to address several relevant questions related to the DFE . First, we compare whether the estimated shape of the deleterious DFE is the same in both populations and whether the inclusion of beneficial mutations affects the estimation of the shape parameter. When comparing models that do not consider beneficial mutations (models f1 versus f10, first row at Table 6), the model with a different shape for Domestic and Wild populations is accepted for scenarios 1, 3, 6 and 10. This means that an artificial change in the shape of the deleterious DFE between Domestic and Wild populations is invoked. Scenarios 1, 6 and 10 include strong positive selection (S_b = 10), and a change in fitness effects occurs at 0%, 25%, and 5% of sites, in each scenario, respectively. Scenario 3 involves weak positive selection (S_b = 1), but many sites change their fitness effects (25%). In contrast, when comparing models that consider beneficial mutations (models f2 vs f20, third row at Table 6), a shared shape of the deleterious DFE is preferred in all scenarios (this is expected given the simulation parameters). This result suggests that not accounting for beneficial mutations can generate an artefactual change in the shape of the deleterious DFE between populations (as first noticed by Tataru *et al*. 2017). Thus, we find that when assuming no beneficial mutations the wrong shape is estimated (f1 vs f10) in the domesticated population when positive selection is strong (scenarios 1, 6, and 10) or when a large fraction of mutations become weakly beneficial (scenario 3).

Second, we find that including the positive DFE is only statistically significant for scenarios with strong positive selection (scenarios 1, 6 and 10) (Table 6 second and fourth row). Note, however, that all of our simulations include beneficial mutations. There is one scenario with strong positive selection that remains undetected, scenario 7. The distinct characteristic of this scenario is that although the domesticated population has undergone a long bottleneck, as in scenario 10, the fitness effects of mutations do not change between populations. Hence, our ability to detect beneficial mutations with polyDFE relies on the strength of positive selection but also on the proportion of sites that change their fitness effects and the duration of the domestication bottleneck. It is also worth mentioning that the power to detect beneficial mutations in the unfolded SFS increases when the shape of the deleterious DFE is jointly estimated (see the decrease in the p-value for scenarios 1, 6 and 10 for the comparisons of models f1 vs. f2 and models f10 vs. f20).

Finally, we wonder whether the differences in the beneficial DFE between Domestic and Wild populations are detectable or not. Interestingly, when comparing models f20 and f30 (last row at Table 6), our ability to detect differences in the beneficial DFE between Domestic and Wild populations is very limited in most scenarios. With the exception of scenario 6, where there is both a large proportion of mutations that show a change in fitness (25%) and strong positive selection (S_b = 10). The short domestication bottleneck in scenario 6 might also increase the power to detect this change in the positive DFE.

DFE parameter estimates with polyDFE

Since the true model generating the data in natural wild and domesticated populations is unknown, we first extract the AIC of each model and then compute the AIC-weighted parameters for all models (Table S7) (Tataru and Bataillon 2019; Castellano et al. 2019). Figure 5 shows the distribution of parameters estimated by polyDFE using bootstrap analysis. The mean deleterious coefficient (s_d) of the DFE is accurately estimated for both populations and all scenarios, except when there are long bottlenecks in Domestic populations, where the mean s_d is generally underestimated (except for scenario 9 which is slightly overestimated). Similarly, the shape of the negative DFE is generally well estimated except in scenario 6 and long-bottleneck scenarios, where the shape is overestimated. The estimation of the parameters of the positive DFE is more challenging, and for all scenarios, the estimates include high variance and, thus, a wide range of possible values. Moreover, in our particular case the true probability that a new mutation is beneficial (p_b) in Domestic populations is not easily defined because there are beneficial mutations in Domestic populations that were already segregating as effectively neutral or deleterious mutations just before the start of the domestication process. This violates the implicit model assumption of constant selection coefficients along generations. Plus, in some scenarios there is migration from the Wild to the Domestic population that will be re-introducing beneficial alleles. Supplementary Figure 0 shows the true composition of beneficial sites and segregating sites (m_2, m_3 and m_7) after the

domestication split which we believe is the best proxy of the current p_b in Domestic populations. However, for Wild populations p_b is well defined (due to the constant population size and DFE, and the lack of migration) but the estimates, as expected, tend to be noisy with very wide confidence intervals.

Figure 6 shows the inferred distribution of the AIC weighted DFE. Discrete DFE ranges tend to be less noisy than individual DFE parameters (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009). The discrete ranges for the deleterious DFE show an excellent agreement between the inferred and the simulated values especially for the fraction of effectively neutral mutations and for scenarios with short bottlenecks (scenarios 1-6), while for scenarios with long bottlenecks (scenarios 7-10) the inferred fraction of weakly and strongly deleterious mutations becomes noisy and sometimes unaccurate. The confidence intervals for the number of effectively beneficial mutations seem to capture the real values indicating a lack of power and not a bias. However, there are slight underestimations for scenarios 3 for both populations, scenario 5 for the Domestic population and scenario 7 and 9 for Wild populations.

dadi: 2D-SFS

Demography inference

A joint DFE between two diverged populations can be inferred with the *dadi* software suite using the 2D-SFS (Huang *et al.* 2021). This provides us with a potential advantage relative to the 1D-SFS approach because the joint DFE can quantify the fraction of mutations with divergent selective effects even when the underlying parameters describing the full DFE remain the same between populations. In contrast to polyDFE, *dadi* estimates first the demographic parameters to then estimate the DFE. The parameters of a demographic scenario describing two recent divergent populations are estimated by fitting a demographic model to the 2D-SFS of synonymous sites (Huang *et al.* 2021). Those parameters are estimated by maximizing a Poisson composite likelihood and contrasted with the corresponding expected SFS.

Supplementary Tables 12-13 show the estimated demographic parameters and their confidence intervals respectively. Here we describe the inferred parameters one by one:

- The population size of the wild (N_{ew}) before the domestication split is slightly but significantly underestimated in all scenarios, except for scenarios 1, 4 and 8 where it is well-recovered.
- The time of the domestication split (t_1) is in general well-recovered but in scenario 2 is slightly underestimated and in scenario 3 is slightly overestimated.
- The population size of the Domestic populations under the bottleneck (N_{e1d}) is overestimated only for scenario 2 and significantly underestimated for scenarios with long bottlenecks (scenarios 7-10). For the rest of scenarios N_{e1d} is well-recovered.
- The duration of the bottleneck (t_{bot}) is well-recovered in all scenarios except in scenarios with migration. In scenario 1 (no change in the DFE) the duration of the bottleneck is significantly

underestimated and very close to zero, this is expected given the high migration rate. In the scenarios with migration and a change in the DFE (scenarios 2 and 3) the duration of the bottleneck is substantially overestimated and in fact they are similar to the long bottleneck scenarios.

- The estimates of the population size after the bottleneck in Domestic populations (N_{e2d}) are more noisy and they are significantly overestimated in scenarios 2, 3 and 10, and significantly underestimated in scenarios 5 and 6.
- The time of the population size recovery (t_3), or the time since the end of the bottleneck, is well-recovered except in scenarios with migration and a change in the DFE. For scenarios 2 and 3 t_3 is significantly underestimated. The recovery time seems again affected by the migration parameter and the change in the DFE.
- Finally, for scenarios without migration from the Wild to the Domestic population the migration parameter (m) is consistently low and in all cases not significantly different from 0. In scenarios with migration (scenarios 1-3), there seems to be an interaction between the change in the DFE and the migration parameter. Scenario 1 which is a negative control (there is no change in the DFE between Domestic and Wild populations) the migration rate is significantly underestimated and half of its true value. However, in scenarios 2 and 3 which show a change of fitness effects in 5% and 25% of the sites, respectively, there is a large underestimation of the migration rate (13X lower than expected).

The previous comparison between the expected number of exclusive mutations in the absence of selection and the observed number of exclusive mutations in the presence of (indirect) selection showed a substantial excess of exclusive variants in scenarios 2 and 3. This emergent excess of exclusive variants might be "interpreted" by *dadi* as a low migration rate and a long but very mild bottleneck. Note that N_e estimates from levels of neutral diversity in these two scenarios are larger than N_e estimates from their paired Wild populations (Table S7). These results suggest that the linkage effects produced in the domestication process can bias some aspects of the inferred demographic model, especially in the presence of migration and when many sites change their fitness effect in the domesticated population.

Figure 7 reports the validation plots of the fitted model using the *dadi* framework. The narrow ranges of the distribution in the residual histogram plot (Figure 7, bottom right panels) reveals a good fit of the model under each distinct scenario (even in the absence of shared polymorphisms as in long bottleneck scenarios 7-10). Note that although scenarios 2 and 3 have the same demographic model than scenario 1, the 2D-SFS fills almost all the cell combinations (Figure 7, top left panels), while in scenario 1 (no change in the DFE) the 2D-SFS is concentrated in the diagonal. This spread of the synonymous 2D-SFS in scenarios 2 and 3 must be driven by the change in the DFE and selection at linked sites. This spread of the 2D-SFS is also observed in scenarios with short bottlenecks but without migration (scenarios 4-6).

Scenarios with long bottlenecks show no shared polymorphisms (scenarios 7-10) and all observed 2D-SFS cell combinations are in the margins. Particularly revealing for scenarios 2 and 3 are the model residuals of each 2D-SFS cell combination (Figure 7, bottom left panels). Here model residuals are computed as the expected counts derived from the model minus the observed counts in the (simulated) data. Cell combinations with more observed variants than expected are in blue, while cell combinations with less observed variants than expected are in red. In scenarios 2 and 3 there are less observed counts in the diagonal of the 2D-SFS than those expected by the model. In contrast, particularly evident for scenario 3, there are two blue colored vertical "stripes". The left stripe indicates an excess of observed mutations at low frequency in the Domestic population that can be at any frequency in the Wild population (this could correspond to synonymous variants in linkage with nonsynonymous mutations that were effectively neutral in Wild but that became more deleterious in Domestic). The second stripe indicate an excess of observed mutations at high frequency in the Domestic population that can be at any frequency in the Wild population (this could correspond to synonymous variants in linkage with nonsynonymous mutations that were effectively neutral in Wild but that became more beneficial in Domestic). In the rest of scenarios, the 2D-SFS model residuals show a random distribution of colours across cell combinations which is indicative of the absence of systematic biases.

DFE parameter estimates with ∂a∂i

The estimated demographic parameters above are used for the inference of the DFE of nonsynonymous mutations using the 2D-SFS. Figure 8 shows a comparison between the 2D-SFS inferred by the most likely model vs. the observed 2D-SFS and the residuals of the model for each 2D-SFS cell combination. The residual histograms for nonsynonymous mutations (Figure 8, bottom right panels) show a nicely distributed residuals with a mean around zero for all scenarios with the exception of few very negative outliers in scenarios with long bottlenecks. We observe again the blue colored vertical "stripes" pattern found before for synonymous mutations at scenario 3. The left stripe could be generated by nonsynonymous mutations that have become more deleterious in the Domestic population (mutations at m_4 sites plus some mutations at m_6 sites). While the right stripe can be interpreted as nonsynonymous mutations that have become more beneficial in the Domestic populations (mutations at m_7 sites plus some mutations at m_3 sites). Both of these stripes are likely driven by direct selection on selectively divergent mutations falling at m_3, m_4, m_6 and m_7 sites, and also by other nonsynonymous mutations that might have a constant selection coefficient but turn out to be in linkage with those selectively divergent mutations.

Figure 9 and Supplementary Tables 14-15 show the inferred DFE parameters under the new joint DFE model. The inferred strength of negative selection (s_d) is overestimated between 1.5X to 3X times. For the shape parameter (b) there is also a general overestimation of this parameter, especially in scenario 3. Regarding the inference of the strength of positive selection note that with *dadi* the positive DFE is

precomputed with a unique point mass distribution of positive selection, or no beneficial mutations at all. Under the polyDFE framework beneficial mutations are drawn from an exponential distribution mimicking our forward simulations. However, with *dadi* we assume either no beneficial mutations or a given proportion of beneficial mutations with $S_b = 1$ or $S_b = 10$ when homozygote. Model comparison shows that *dadi* can distinguish between weak and strong positive selection in most cases (Table 7). However, for scenarios 2, 5 and 9 which show a high proportion of positive change on Domestic populations (25%), the likelihood is somewhat better considering a strongest selection coefficient ($S_b = 10$).

As we shown before with polyDFE, the estimation of the positive DFE and its changes between populations is very challenging mainly because beneficial mutations tend to be rare, and the true p_b in Domestic populations is not well defined due to the action of positive selection on standing variation and the re-introduction of beneficial mutations due to migration. Two violations that we believe migh be commonplace in nature. This is the reason why we propose a new joint DFE model which might be able to tackle some of this complexity (Figure 1 and Materials and Methods for a full description). Under this new joint DFE model, we can measure the fraction of mutations that change their selection coefficient (p_c). This fraction p_c is accurately estimated in scenarios with shared polymorphisms (scenarios 2, 3, 5 & 6), and it is not statistically different from zero for scenarios 1, 4, and all the long bottleneck scenarios (scenarios from 7 to 10). Scenarios 1, 4 and 7 are negative controls without a change in the DFE between Wild and Domestic populations. However, scenarios 8, 9 and 10 do suffer a change in the DFE but remain undetected very likely due to the lack of shared polymorphisms. Even in long bottleneck scenarios where p_c seems to be poorly estimated, confidence intervals seem to capture the real values indicating a lack of power and not a bias. The estimated proportion of positive mutations in Wild populations (p_{+w}) is very close to the true value in scenarios where strong positive selection is driven these mutations (scenarios 1, 6, 7 and 10). In scenarios with weak positive selection, p_{+w} tends to be overestimated, particularly in scenarios with migration and a change in the DFE (scenarios 2 and 3). Note that in Figure 9 is illustrated the proportion of nonsynonymous mutations that change their selection coefficient toward positive values in the Domestic populations as the product of $p_c * p_{c+}$. Unfortunately, this product, which aims to estimate the fraction of m_7 sites in our simulations, is always not significantly different from zero. Overall, the new joint DFE model seems able to capture the fraction of mutations that change their selection coefficient in the presence of shared polymorphisms, but it is not able to discern how many delererious mutations have become beneficial. We believe p_c could be used in future studies to quantify the number of mutations with divergent fitness effects between natural wild and domesticated populations.

DISCUSSION

The literature about the genomic basis of domestication (e.g., Ross-Ibarra et al. 2007, Flood and Hancock 2017; Moyers et al. 2018; Flori et al. 2019; Frantz et al. 2020; Leno-Colorado et al. 2020) revolves around three main questions: (1) Is domestication driven by few loci of large effect or by many loci of weak effect (polygenic adaptation)? (2) Is domestication dominated by standing variation (variation that was present before the start of domestication) or by new exclusive mutations in the domesticated populations? (3) Does domestication have a genetic cost due to the bottleneck and inbreeding (reviewed by Moyers et al. 2018)? In this work, we ask an intimately related question: Can we investigate domestication as a change in the full distribution of fitness effects comparing the SFS of wild and domesticated populations? This question is motivated by empirical results. For example, a small number of fixed functional variants have been detected in domestic pig populations in relation to wild populations (Leno-Colorado et al. 2020, but see Groenen et al. 2012; Rubin et al. 2010). Although frequent introgression has been described between pig populations (Frantz et al. 2015; Ramírez et al. 2015), there are clear phenotypic features that distinguish wild from domestic populations. The absence of major genetic signals to explain most traits under domestication makes us think that polygenic adaptation might be a plausible hypothesis. We have simulated potential domestication scenarios involving both a population bottleneck and a change in the selection coefficients in new and standing variation in the domesticated population. Thus, in our simulations, we consider that a hypothetical environmental change (*i.e.,* domestication), makes a number of deleterious mutations (and effectively neutral mutations) in the wild populations to become beneficial in the domesticated populations. Also, beneficial mutations in the wild populations can become deleterious or effectively neutral in the domesticated populations. However, since the rate of new beneficial mutations in the wild populations is relatively small (2.5% of all new mutations), then beneficial variants that become deleterious are a minority of the minority and we expect a second-order effect of those mutations in our simulations. We also investigate the impact of migration from the wild to the domesticated population on our ability to measure a change in the full distribution of fitness effects.

First, we start describing the fraction of adaptive substitutions (α) across scenarios and populations. We wonder which is the contribution of new beneficial mutations and mutations that were already segregating at the beginning of domestication to α. We are also interested in quantifying the contribution of deleterious polymorphisms in the wild populations that become beneficial in the domesticated populations. Most contemporary adaptive fixations came from variation that predates the initiation of domestication. In other words, we find that in all scenarios, the largest proportion of adaptive substitutions came from polymorphisms that were already beneficial before the domestication split. However, in some scenarios, deleterious polymorphisms that became beneficial in domesticated populations can also explain a large fraction of all adaptive substitutions (Table 4, Table S3). Interestingly, we find that the contribution

of new beneficial variants (exclusive of the domesticated populations) on α depends on the duration of the bottleneck. Most beneficial mutations that have occurred after the domestication split, and only in the domesticated population, have not reached fixation in scenarios with short bottlenecks, and hence their exclusion from the computation of α affects very little the realized α values. In contrast, new exclusive beneficial mutations in scenarios with long bottlenecks contribute substantially to the current fraction of adaptive substitutions in domesticated populations. We hypothesize this is due to the reduction in the time to fixation expected in small populations. It is also important to note that α in domesticated populations is not much larger than in wild populations even when around 5%-10% of the mutations are beneficial in the domesticated population (as in scenarios 3, 5 and 9, see Figure S0) compared to the 2.5% in the wild populations. In fact, in many scenarios, α in domesticated populations can be smaller than in wild populations. These modest α values in domesticated populations may be due to the bottleneck associated with the domestication process, but also to the short time since the initiation of domestication. The small effective population size, on one hand, reduces the efficacy of positive and negative selection (and α) in domesticated populations but on the other hand, accelerates the fixation of mutations (neutral, deleterious, and beneficial). Unfortunately, we find that our list of α estimators (MKT, MKTa, polyDFE), when applied to our simulations, have little power to detect statistically different α values between populations. We do not recommend extracting strong conclusions from applying α estimators alone to natural wild and domesticated populations.

Second, we study the impact of direct and indirect selection and demography on the 1D-SFS. We compare the 1D-SFS of synonymous (neutral) and nonsynonymous (selected) variants in our simulated scenarios against the expected 1D-SFS for (only) neutral mutations under the same demographic model. We show that in domesticated populations selection at linked sites needs to be invoked to explain the 1D-SFS of synonymous mutations. We hypothesize that the mode of linked selection affecting the 1D-SFS depends on the duration of the bottleneck. For those scenarios with long bottlenecks, we think that background selection is the main distorter of the synonymous 1D-SFS. Under long bottlenecks, haplotypes carrying deleterious variants (and neutral variants in linkage) can raise in frequency (Hartfield and Otto 2011). After the bottleneck these haplotypes will decrease in frequency dragging with them neutral variation. These anti-sweeps (or sweeps toward extinction) might produce the observed synonymous 1D-SFS in our simulations which show an excess at low frequency variants and a lack of variation in the rest of frequency classes (Figure 4). The 1D-SFS of synonymous and nonsynonymous mutations is very similar in our simulations with long bottlenecks but very different from that expected in the absence of selection (Figures S1-3). This makes us think that a particularly strong form of background selection might be taking place. Surprisingly, the synonymous 1D-SFS also seems affected by indirect selection, or selection at linked sites, in short bottleneck scenarios when a substantial fraction of standing deleterious (and effectively neutral) variation become beneficial (Figure 4, scenarios 2, 3, 5, 6). Short bottlenecks are expected to have

a modest impact on the population frequencies of standing variation, particularly in the presence of migration from the wild population to the domesticated population. In these scenarios, we think that the shift in the selection coefficients in the domesticated populations is setting the ideal conditions for the emergence of soft sweeps (Hermisson and Pennings 2005; Stetter et al. 2018). This multitude of soft sweeps could explain the excess of synonymous variants at a high frequency that we observe in our simulations (Figure 4). Note that in scenarios 3, 5 or 6 the excess of (neutral) synonymous variants at high frequency is even larger than the excess found for (selected) nonsynonymous variants. We hypothesize that soft sweeps might be behind this distortion of the synonymous 1D-SFS in our short bottleneck simulations. This detailed analysis of the 1D-SFS is a prerequisite to interpret the comparison of the full DFE using polyDFE and *dadi*, because both methods contrast the (1D or 2D) SFS of synonymous and nonsynonymous mutations to infer the DFE parameters.

Third, the comparison of the full DFE for domesticated and wild *in silico* populations finds that polyDFE seems to provide more accurate estimates of the deleterious DFE than the new joint DFE model build on *dadi*. The discretized DFE analysis with polyDFE shows that the confidence intervals for the number of effectively beneficial mutations seem to capture the real values in most scenarios indicating a lack of power and not a bias. However, due to the lack of power with polyDFE is not possible to detect statistically significant differences between populations for the beneficial DFE, except when selection is strong and many sites change their selection coefficients (as in scenario 6). The new joint DFE model seems to accurately estimate the strength of positive selection across most scenarios, but when a very large fraction of mutations become beneficial it tends to invoke stronger positive selection. We find that, even in the demographically and selectivelly "stable" wild populations, the inference of the probability that a new mutation is beneficial (called p_b in polyDFE and p_{+w} in *dadi*) can be extremely challenging with both methods. More importantly, for domesticated populations, the meaning of this probability is unclear because the selection coefficients of new and standing variation are not static, migration from the wild to the domesticated population can re-introduce beneficial mutations and our domesticated populations are far from demographic equilibrium. We find that joint DFE models offer one key advantage in the study of domestication. This is the computation of the fraction of mutations that *change* their selection coefficient in the domesticated population relative to the wild population (p_c) and from those how many became beneficial during domestication (p_{c+}). We find that in the presence of shared polymorphisms p_c is well estimated, but p_{c+} turned out to be very noisy. Thus, our new joint DFE model contrasting the 2D-SFSs is able to quantify a signal of domestication, understood as the fraction of mutations with divergent selective effects, when two populations have diverged recently enough to share many polymorphisms. Given the strong distortion of the synonymous SFS that we described before it is not surprising that polyDFE only detects a significant difference in the positive DFE in one scenario or that *dadi*'s deleterious DFE and p_{c+} estimates are unaccurate and noisy. polyDFE does not estimate the underlying demography, but it corrects

for the potential impact of demographic changes on the estimation of the DFE. The way polyDFE corrects for demographic changes, and/or linked selection, is through the nuisance parameters r_i which contrast the relative observed synonymous SFS to the relative expected neutral SFS under the Wright-Fisher model and then uses these inferred r_i parameters to correct the nonsynonymous SFS and estimate the DFE (Eyre-Walker *et al*. 2006; Tataru et al. 2017). It is likely that the strong impact of selection at linked sites in our simulations is making the r_i parameters overcorrect the nonsynonymous SFS. This might be hampering our ability to detect differences in the positive DFE between populations. In contrast, *dadi* infers first the demographic parameters with synonymous mutations. The inference of the demographic model with *dadi* shows that under some domestication scenarios, even in the presence of realistic recombination rates, important aspects of the inferred demographic model, such as the migration rate or the duration of the bottleneck, can be substantially biased.

Finally, we want to discuss the limitations and future directions of this work. This work is not an exhaustive benchmarking across all one hundred and twenty-eight possible domestication scenarios that we could have been simulated given the value of the different parameters that we play with at Table 1. We do not investigate the prevalence of hard and soft sweeps in our simulated scenarios but this could certainly enrich the study of domestication under polygenic adaptation. Our work also posits the question of how can we infer (or at least correct) the demographic changes in the presence of pervasive selection at linked sites and changes in the selection coefficients of standing variation. Machine learning approaches might be a natural choice for such complex scenarios involving rich non-equilibrium dynamics. But perhaps methods that first estimate demography using patterns of linkage disequilibrium, which seem more robust to linked selection than SFS statistics (Ragsdale and Gutenkunst 2017; Novo *et al.* 2022), in combination with SFS methods can provide more general, accurate enough, and faster estimates than machine learning tools. We have not even incorporated the complexity introduced by the distribution of dominance (but see Arunkumar *et al.* 2015) or the impact of population size changes in the wild populations. Nonetheless, the limits to comprehend the domestication process by comparing the DFE between wild and domesticated populations has become more evident.

Acknowledgments

S.E.R.O. is supported by grants AGL2016-78709-R (MEC, Spain) and PID2020-119255GB-I00 (MICINN, Spain), by the CERCA Programme/Generalitat de Catalunya and acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D 2016-2019 and 2020-2023 (SEV-2015-0533, CEX2019-000917) and the European Regional Development Fund (ERDF). This research was supported in part by NSF Grant No. PHY-1748958 and the Gordon and Betty Moore Foundation Grant No. 2919.02. ITV is supported by a predoctoral fellowship funded by MCIN/AEI/ 10.13039/501100011033 through the Grant BES-2017-081139 and by "ESF Investing in your future". RNG and DC were supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM127348 to RNG).

LITERATURE CITED

Amills M., Megens H. J., Manunza A., Ramos-Onsins S. E., and Groenen M. A. M., 2017 A genomic perspective on wild boar demography and evolution, pp. 376–387 in Ecology, Conservation and Management of Wild Pigs and Peccaries, Cambridge University Press.

Andersson L., 2012 Genetics of animal domestication, pp. 260–274 in Biodiversity in Agriculture: Domestication, Evolution, and Sustainability, Cambridge University Press.

Arunkumar R, Ness RW, Wright SI, Barrett SC. 2015 The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. Genetics. 199(3):817-29. doi: 10.1534/genetics.114.172809.

Avni R., M. Nave, O. Barad, K. Baruch, S. O. Twardziok, et al., 2017 Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 357: 93–97. https://doi.org/10.1126/science.aan0032

Barghi N., R. Tobler, V. Nolte, A. M. Jakšić, F. Mallard, et al., 2019 Genetic redundancy fuels polygenic adaptation in Drosophila. PLoS Biology 17. https://doi.org/10.1371/journal.pbio.3000128

Barton H. J., and Zeng K., 2018 New methods for inferring the distribution of fitness effects for INDELs and SNPs. Molecular Biology and Evolution 35: 1536–1546. https://doi.org/10.1093/molbev/msy054

Berg J. J., and Coop G., 2014 A Population Genetic Signal of Polygenic Adaptation. PLoS Genetics 10. https://doi.org/10.1371/journal.pgen.1004412

Boyko A. R., Boyko R. H., Boyko C. M., Parker H. G., Castelhano M., et al., 2009 Complex population structure in African village dogs and its implications for inferring dog domestication history. Proceedings of the National Academy of Sciences of the United States of America 106: 13903–13908. https://doi.org/10.1073/pnas.0902129106

Castellano D., MacIà M. C., Tataru P., Bataillon T., and Munch K., 2019 Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. Genetics 213: 953–966. https://doi.org/10.1534/genetics.119.302494

Coffman A. J., Hsieh P. H., Gravel S., and Gutenkunst R. N., 2016 Computationally efficient composite likelihood statistics for demographic inference. Molecular Biology and Evolution 33: 591–593. https://doi.org/10.1093/molbev/msv255

Crow J. F., and Kimura M., 1979 Efficiency of truncation selection. Proceedings of the National Academy of Sciences of the United States of America 76. https://doi.org/10.1073/pnas.76.1.396

Dayan T., 1994 Early domesticated dogs of the near east. Journal of Archaeological Science 21. https://doi.org/10.1006/jasc.1994.1062

de Vladar H. P. de, and Barton N., 2014 Stability and response of polygenic traits to stabilizing selection and mutation. Genetics 197: 749–767. https://doi.org/10.1534/genetics.113.159111

Driscoll C. A., Macdonald D. W., and O'Brien S. J., 2009 From wild animals to domestic pets, an evolutionary view of domestication. Proceedings of the National Academy of Sciences of the United States of America 106 Suppl 1: 9971–9978. https://doi.org/10.1073/pnas.0901586106

Eyre-Walker A, Keightley PD. 2007 The distribution of fitness effects of new mutations. Nat Rev Genet. 8(8):610-8. doi: 10.1038/nrg2146.

Eyre-Walker, A, and Keightley P D, 2009 Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol. Biol. Evol. 26: 2097–2108.

Flood P. J., and Hancock A. M., 2017 The genomic basis of adaptation in plants. Current Opinion in Plant Biology 36.

Flori L., Moazami-Goudarzi K., Alary V., Araba A., Boujenane I., et al., 2019 A genomic map of climate adaptation in Mediterranean cattle breeds. Molecular Ecology 28. https://doi.org/10.1111/mec.15004

Frantz L. A. F., Bradley D. G., Larson G., and Orlando L., 2020 Animal domestication in the era of ancient genomics. Nature Reviews Genetics 21.

Frantz L. A. F., Schraiber J. G., Madsen O., Megens H. J., Cagan A., et al., 2015 Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nature Genetics 47. https://doi.org/10.1038/ng.3394

Godambe V. P., 1985 The Foundations of Finite Sample Estimation in Stochastic Processes. Biometrika 72: 419. https://doi.org/10.2307/2336094

Granleese T., Clark S. A., and van der Werf J. H. J., 2019 Genotyping strategies of selection candidates in livestock breeding programmes. Journal of Animal Breeding and Genetics 136: 91–101. https://doi.org/10.1111/jbg.12381
Groenen M. A. M., Archibald A. L., Uenishi H., Tuggle C. K., Takeuchi Y., et al., 2012 Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491. https://doi.org/10.1038/nature11622
Gutenkunst R. N., R. D. Hernandez, Williamson S. H., and Bustamante C. D., 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5. https://doi.org/10.1371/journal.pgen.1000695
Haller B. C., and Messer P. W., 2016 SLiM: An Evolutionary Simulation Framework. http://benhaller.com/slim/SLiM_Manual.pdf 3.4: 1–600.
Haller B. C., and Messer P. W., 2017 AsymptoticMK: A web-based tool for the asymptotic McDonald-Kreitman test. G3: Genes, Genomes, Genetics 7. https://doi.org/10.1534/g3.117.039693
Harrisson K. A., Amish S. J., Pavlova A., Narum S. R., Telonis-Scott M., et al., 2017 Signatures of polygenic adaptation associated with climate across the range of a threatened fish species with high genetic connectivity. Molecular Ecology 26. https://doi.org/10.1111/mec.14368
Hartfield M, Otto SP. 2011 Recombination and hitchhiking of deleterious alleles. Evolution. 65(9):2421-34. doi: 10.1111/j.1558-5646.2011.01311.x.
Hartl D. L., Moriyama E. N., and Sawyer S. A., 1994 Selection intensity for codon bias. Genetics 138. https://doi.org/10.1093/genetics/138.1.227
Hermisson J., and Pennings P. S., 2005 Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169. https://doi.org/10.1534/genetics.104.036947
Höllinger I., Pennings P. S., and Hermisson J., 2019 Polygenic adaptation: From sweeps to subtle frequency shifts. PLoS Genetics 15. https://doi.org/10.1371/journal.pgen.1008035
Huang X., Fortier A. L., Coffman A. J., Struck T. J., Irby M. N., et al., 2021 Inferring Genome-Wide Correlations of Mutation Fitness Effects between Populations. Molecular Biology and Evolution 38. https://doi.org/10.1093/molbev/msab162
Huang X., Kurata N., Wei X., Wang Z. X., Wang A., et al., 2012 A map of rice genome variation reveals the origin of cultivated rice. Nature 490. https://doi.org/10.1038/nature11532
Hudson R. R., 2001 Two-locus sampling distributions and their application. Genetics 159. https://doi.org/10.1093/genetics/159.4.1805
Jain K., and Stephan W., 2015 Response of polygenic traits under stabilizing selection and mutation when loci have unequal effects. G3: Genes, Genomes, Genetics 5. https://doi.org/10.1534/g3.115.017970
Jain K., and Stephan W., 2017a Modes of rapid polygenic adaptation. Molecular Biology and Evolution 34.
Jain K., and Stephan W., 2017b Rapid adaptation of a polygenic trait after a sudden environmental shift. Genetics 206. https://doi.org/10.1534/genetics.116.196972
Jasinska A. J., and Freimer N. B., 2009 The complex genetic basis of simple behavior. Journal of Biology 8.
Jerison E. R., Kryazhimskiy S., Desai M. S., 2014 Pleiotropic consequences of adaptation across gradations of environmental stress in budding yeast. arXiv:104.09.7839. https://doi.org/10.48550/arXiv.1409.7839
Keightley P. D., and Eyre-Walker A., 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177. https://doi.org/10.1534/genetics.107.080663
Kim B. Y., Huber C. D., and Lohmueller K. E., 2017 Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics 206. https://doi.org/10.1534/genetics.116.197145
Kondrashov A. S., 1988 Deleterious mutations and the evolution of sexual reproduction. Nature 336.
Laere A. S. van, Nguyen M., Braunschweig M., Nezer C., Collette C., et al., 2003 A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425. https://doi.org/10.1038/nature02064
Larson G., and Burger J., 2013 A population genetics view of animal domestication. Trends in Genetics 29.
Larson G., Piperno D. R., Allaby R. G., Purugganan M. D., Andersson L., et al., 2014 Current perspectives and the future of domestication studies. Proceedings of the National Academy of Sciences of the United States of America 111.

Lee KM, Coop G. 2017 Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data. Genetics. 207(4):1591-1619. doi: 10.1534/genetics.117.300417.
Leno-Colorado J., Guirao-Rico S., Pérez-Enciso M., and Ramos-Onsins S. E., 2020 Pervasive selection pressure in wild and domestic pigs 1. bioRxiv.
Li N., Xu R., Duan P., and Li Y., 2018 Control of grain size in rice. Plant Reproduction 31.
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(3):126-31. doi: 10.1016/j.tig.2006.01.004.
Marsden C. D., del Vecchyo D. O., O'Brien D. P., Taylor J. F., Ramirez O., et al., 2016 Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proceedings of the National Academy of Sciences of the United States of America 113. https://doi.org/10.1073/pnas.1512501113
Messer P. W., and Petrov D. A., 2013 Frequent adaptation and the McDonald-Kreitman test. Proceedings of the National Academy of Sciences of the United States of America 110. https://doi.org/10.1073/pnas.1220835110
Moyers BT, Morrell PL, McKay JK. 2018 Genetic Costs of Domestication and Improvement. J Hered. 109(2):103-116. doi: 10.1093/jhered/esx069.
Novo I, Santiago E, Caballero A. 2022 The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. PLoS Genet. 18(1):e1009764. doi: 10.1371/journal.pgen.1009764.
Ohta T., 1989 The mutational load of a multigene family with uniform members. Genetical Research 53. https://doi.org/10.1017/S0016672300028020
Pritchard J. K., Pickrell J. K., and Coop G., 2010 The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. Current Biology 20.
Purugganan M. D., and Fuller D. Q., 2009 The nature of selection during plant domestication. Nature 457.
Ragsdale A. P., and Gutenkunst R. N., 2017 Inferring demographic history using two-locus statistics. Genetics 206. https://doi.org/10.1534/genetics.117.201251
Ragsdale A. P., Coffman A. J., Hsieh P., Struck T. J., Gutenkunst R. N., 2016 Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. Genetics 203. http://doi.org/10.1534/genetics.115.184812
Ragsdale AP, Gutenkunst RN. 2017 Inferring Demographic History Using Two-Locus Statistics. Genetics. 206(2):1037-1048. doi: 10.1534/genetics.117.201251.
Ramírez O., W. Burgos-Paz, E. Casas, M. Ballester, E. Bianco, et al., 2015 Genome data from a sixteenth century pig illuminate modern breed relationships. Heredity 114. https://doi.org/10.1038/hdy.2014.81
Raybaut P., 2009 Spyder Documentation. Spyder Project.
Redding R. W., 2015 The Pig and the Chicken in the Middle East: Modeling Human Subsistence Behavior in the Archaeological Record Using Historical and Animal Husbandry Data. Journal of Archaeological Research 23. https://doi.org/10.1007/s10814-015-9083-2
Rose N. H., Bay R. A., Morikawa M. K., and Palumbi S. R., 2018 Polygenic evolution drives species divergence and climate adaptation in corals. Evolution 72. https://doi.org/10.1111/evo.13385
Ross-Ibarra J., Morrel P. L. l, and Gaut B. S., 2007 Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceedings of the National Academy of Sciences of the United States of America 104. https://doi.org/10.1073/pnas.0700643104
Rossum G. van, and Drak F. L. e, 2006 Python Reference Manual. October 22.
Rubin C. J., Zody M. C., J. Eriksson, J. R. S. Meadows, E. Sherwood, et al., 2010 Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464. https://doi.org/10.1038/nature08832
Shomura A., Izawa T., Ebana K., Ebitani T., Kanegae H., et al., 2008 Deletion in a gene associated with grain size increased yields during rice domestication. Nature Genetics 40. https://doi.org/10.1038/ng.169
Smith N. G. C., and Eyre-Walker A., 2002 Adaptive protein evolution in Drosophila. Nature 415. https://doi.org/10.1038/4151022a
Stephan W., 2016 Signatures of positive selection: From selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. Molecular Ecology 25. https://doi.org/10.1111/mec.13288

Stephan W., and John S., 2020 Polygenic adaptation in a population of finite size. Entropy 22.
Stetter M. G., Thornton K., and Ross-Ibarra J., 2018 Genetic architecture and selective sweeps after polygenic adaptation to distant trait optima. PLoS Genetics 14. https://doi.org/10.1371/journal.pgen.1007794
Tataru P., and Bataillon T., 2019 PolyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. Bioinformatics 35. https://doi.org/10.1093/bioinformatics/bty1060
Tataru P., Mollion M., Glémin S., and Bataillon T., 2017 Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. Genetics 207. https://doi.org/10.1534/genetics.117.300323
Thornton K. R., 2019 Polygenic adaptation to an environmental shift: Temporal dynamics of variation under Gaussian stabilizing selection and additive effects on a single trait. Genetics 213. https://doi.org/10.1534/genetics.119.302662
Trut L., Oskina I., and Kharlamova A., 2009 Animal evolution during domestication: The domesticated fox as a model. BioEssays 31.
Wright S. I., Bi I. V., Schroeder S. C., Yamasaki M., Doebley J. F., et al., 2005 Evolution: The effects of artificial selection on the maize genome. Science 308. https://doi.org/10.1126/science.1107891
Xue S., P. J. Bradbury, Casstevens T., and Holland J. B., 2016 Genetic architecture of domestication-related traits in maize. Genetics 204. https://doi.org/10.1534/genetics.116.191106
Zeder M. A., 2012 The domestication of animals. Journal of Anthropological Research 68. https://doi.org/10.3998/jar.0521004.0068.201
Zeder M. A., 2015 Core questions in domestication research. Proceedings of the National Academy of Sciences of the United States of America 112. https://doi.org/10.1073/pnas.1501711112.
Zeder M. A., Emshwiller E., Smith B. D., and Bradley D. G., 2006 Documenting domestication: The intersection of genetics and archaeology. Trends in Genetics 22.
Zhang M, Yang Q, Ai H, Huang L. 2022 Revisiting the evolutionary history of pigs via De Novo mutation rate estimation in a three-generation pedigree. Genomics Proteomics Bioinformatics. S1672-0229(22)00014-6. doi: 10.1016/j.gpb.2022.02.001.
Birky CW Jr, Walsh JB. Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci U S A. 1988 Sep;85(17):6414-8. doi: 10.1073/pnas.85.17.6414.

Table 1

Parameters for each analyzed scenario

Scenario	Bottleneck	Positive DFE1 (S_b)	Domestication		Positive DFE2 (S_b)	m W->D
	t_bot (2Ne)	**exp mean (2Nes)**	**% Change**	**% Positive**	**exp mean (2Nes)**	**(2Nem)**
1	0.016	10	0	2.5	10	200
2	0.016	1	5	25.0	1	200
3	0.016	1	25	10.0	1	200
4	0.016	1	0	2.5	1	0
5	0.016	1	25	25.0	1	0
6	0.016	10	25	2.5	10	0
7	0.161	10	0	2.5	10	0
8	0.161	1	5	2.5	1	0
9	0.161	1	25	25.0	1	0
10	0.161	10	5	10.0	10	0

Ne: Ancestral effective population size, S_b: Population selection coefficient of new beneficial mutations $2N_es$ in homozygotes, %change: percentage of sites that change their selection coefficient in the Domestic population, %Positive: from the percentage of sites that change their selection coefficient (% change) this is the percentage that change toward positive values in the Domestic population. Exp mean: Mean value of the exponential distribution for new positive derived variants in the Domestic population. m W->D: migration rate of Wild to Domestic population.

Table 2

Types of mutations in simulated scenarios

	Wild	Domestic
m_1	Neutral	No change, remain Neutral
m_2	Beneficial	No change, remain Beneficial
m_3	Beneficial	Change to a different Beneficial effect
m_4	Beneficial	Change to Deleterious
m_5	Deleterious	No change, remain Deleterious
m_6	Deleterious	Change to a different Deleterious effect
m_7	Deleterious	Change to Beneficial

Table 3

Number of fixed, exclusive and shared variants observed in the Domestic populations for each scenario

Scenarios	SfWD	SfW	SfD	SxW	SxD	Ssh	SfWxD	SfDxW
1	55446	0	0	14440	16539	51081	281	211
2	47645	0	0	25526	35707	46442	813	344
3	47193	0	0	28436	36185	44599	765	499
4	47377	5	24	50933	42248	20691	900	2385
5	47785	47	132	51511	40213	16641	783	3274
6	54503	62	139	48257	38018	13427	884	2966
7	54909	312	4267	56671	4351	4	9	9260
8	47221	318	4552	61805	4228	3	9	10928
9	47582	270	4362	61077	4362	4	17	11058
10	55340	449	4301	56019	4293	1	16	9351

SfWD: Fixed variant in the species in relation to the outgroup. SfW: Exclusive fixed variant in Wild. SfD: Exclusive fixed variant in Domestic. SxW: Exclusive polymorphism in Wild, SxD: Exclusive polymorphisms in Domestic. Ssh: Shared polymorphic variants. SfWxD: Fixed variants in Wild and polymorphic in Domestic. SfDxW: Fixed in Domestic and polymorphic in Wild.

Table 4

True α for Total, Shared and Exclusive mutations in relation to all nonsynonymous fixations.

	All variants				Shared Variants			Exclusive variants		
	Wild	Domestic			Wild	Domestic		Wild	Domestic	
	α_{total}*	α_{total}†	α_{m2+m3}§	α_{m7}¶	α_{total}*	α_{total}†	α_{m7}¶	α_{total}*	α_{total}†	α_{m7}¶
1	0.211	0.156	0.156	0.000	0.211	0.156	0.000	.	.	.
2	0.056	0.063	0.045	0.017	0.056	0.063	0.017	.	.	.
3	0.061	0.106	0.081	0.025	0.061	0.106	0.025	.	.	.
4	0.059	0.060	0.060	0.000	0.059	0.061	0.000	.	.	.
5	0.061	0.105	0.041	0.064	0.061	0.105	0.063	0.097	0.101	0.089
6	0.213	0.089	0.078	0.011	0.213	0.089	0.011	0.175	0.085	0.000
7	0.208	0.055	0.055	0.000	0.208	0.070	0.000	0.219	0.026	0.000
8	0.060	0.040	0.039	0.001	0.060	0.047	0.001	0.112	0.026	0.001
9	0.058	0.094	0.034	0.060	0.058	0.101	0.062	0.072	0.080	0.058
10	0.212	0.055	0.050	0.005	0.212	0.072	0.005	0.254	0.023	0.004

* fix (m_2, m_3, m_4)/fix(Nsyn); † fix (m_2, m_3, m_7)/fix(Nsyn); § fix(m_2,m_3)/fix(Nsyn); ¶ fix(m_7)/fix(Nsyn), where numerator and denominator belong to Total, shared or exclusive.. s: scenario.. § The variants that are fixed at Domestic and at Wild are not counted in the Domestic α calculation, as they are considered occurred before the split of the populations.

Table 5

List of nested polyDFE models and (co)estimated parameters. Independently estimated parameters for the domesticated and wild population (Var). Jointly estimated parameters for the Domestic and Wild populations (Fix)

Model name	Deleterious DFE		Beneficial DFE		Population mutation rate	Nuisance parameters	Polarization parameter
	β	S_d	p_b	S_b	θ	r_i	ε
f1	Var	Var	None	None	Var	Var	Var
f10	Fix	Var	None	None	Var	Var	Var
f2	Var	Var	Var	Var	Var	Var	Var
f20	Fix	Var	Var	Var	Var	Var	Var
f30	Fix	Var	Fix	Fix	Var	Var	Var

Table 6

Likelihood ratio test p-value for each scenario. Significant and marginally significant comparisons (< 0.1) are highlighted in bold.

Models	Domestication Scenarios									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
f1 vs. f10										
(constant vs. variable shape in deleterious DFE models)	**0.007**	0.533	**0.018**	0.167	0.239	**0.000**	0.592	0.308	0.252	**0.008**
f1 vs. f2										
(variable shape while comparing deleterious DFE vs. full DFE)	**0.095**	0.990	0.996	0.649	0.969	**0.007**	0.816	0.902	0.846	0.147
f2 vs. f20										
(constant vs. variable shape under full DFE model)	0.442	0.539	0.380	0.423	0.351	0.317	0.321	0.550	0.398	0.137
f10 vs. f20										
(constant shape while comparing deleterious DFE vs. full DFE)	**0.002**	0.987	0.246	0.334	0.900	**0.000**	0.844	0.730	0.775	**0.030**
f30 vs. f20										
(constant shape while comparing constant positive DFE vs. variable positive DFE)	0.597	0.919	0.064	0.342	0.826	**0.011**	0.823	0.846	0.466	0.658

Table 7

Comparing models with and without positive selection from *dadi* likelihoods

Scenario	True S_b	AIC_0	AIC_1	AIC_{10}	Lower AIC Model
1	10	5514.9	5360.1	5325.4	$S_b = 10$
2	1	8277.9	8221.5	8208.1	$S_b = 10$
3	1	13578	13456	13517	$S_b = 1$
4	1	8866.7	8810.3	8833.6	$S_b = 1$
5	1	8540.8	8520.1	8475.3	$S_b = 10$
6	10	8513.0	8368.4	8264.4	$S_b = 10$
7	10	1125.8	1076.4	1073.7	$S_b = 10$
8	1	1081.9	1053.1	1065.5	$S_b = 1$
9	1	1192.5	1183.1	1182.9	$S_b = 10$
10	10	1236.4	1125.2	1121.5	$S_b = 10$

AIC_0: Akaike Information Criteria ($2k-2\ln L_{S0}$) for a model 0, considering $S_b = 0$. k=5 after fixing demographic parameters.
AIC_1: Akaike Information Criteria ($2k-2\ln L_{S1}$) for a model 1, considering $S_b = 1$. k=5 after fixing demographic parameters.
AIC_{10}: Akaike Information Criteria ($2k-2\ln L_{S10}$) for a model 10, considering $S_b = 10$. k=5 after fixing demographic parameters.

FIGURE LEGENDS

Figure 1. Joint DFE models simulated and fit. **A:** Illustration of the joint DFE model used in the SLiM simulations, with mutation types illustrated. **B**: Illustration of the joint DFE model used in the dadi inferences, in which a fixed positive selection coefficient is a assumed.

Figure 2. Diagram of the demographic model. New: Effective population size of the Wild population. Ne_{1d}: Effective population size of the Domestic population at bottleneck. Ne_{2d}: Effective population size of the Domestic population after a bottleneck. t_1: Number of generations from the Wild-Outgroup split to the Domestic-Wild split. t_{bot}: Number of generations in the bottleneck period. t_2: Number of generations from the bottleneck to the present. t_3: Number of generations from the beginning of the bottleneck to the present. m: Wild to Domestic migration rate from the bottleneck to the present (migration occurs along t_2).

Figure 3. Comparison of true and inferred α values across different methods. Asymptotic: MKTa (Messer and Petrov 2013), MKT: α calculation using Smith and Eyre-Walker (2002). polyDFE: α inference using the algorithm from Tataru et al. (2017). True α is depicted with a solid black line. Confidence intervals at 95% are obtained by bootstrap.

Figure 4: Log_2 ratio of the simulated SFS (with selection) versus the expected SFS (in the absence of selection) but under the same demographic model.

Figure 5: Sampling distributions of estimated DFE parameters from polyDFE obtained by 100 bootstrap replicates. Dashed vertical lines show the true simulated values of the parameters and the coloured vertical lines show the median inferred value across replicates. For s_b we show two dashed lines; one in grey when $S_b = 10$, and the other in black when $S_b = 1$. We show the true p_b found in Wild populations, which is the lower bound of what can be found in Domestic populations. To obtain s_d and s_b from S_d and S_b values we use π at synonymous sites and the true simulated mutation rate (1e-6) to get the realized N_e after the bottleneck and/or the action of linked selection. Note that in polyDFE s is defined to be the selection coefficient on the heterozygote (like in *dadi*), but the scaled selection coefficient is defined as $4N_es$.

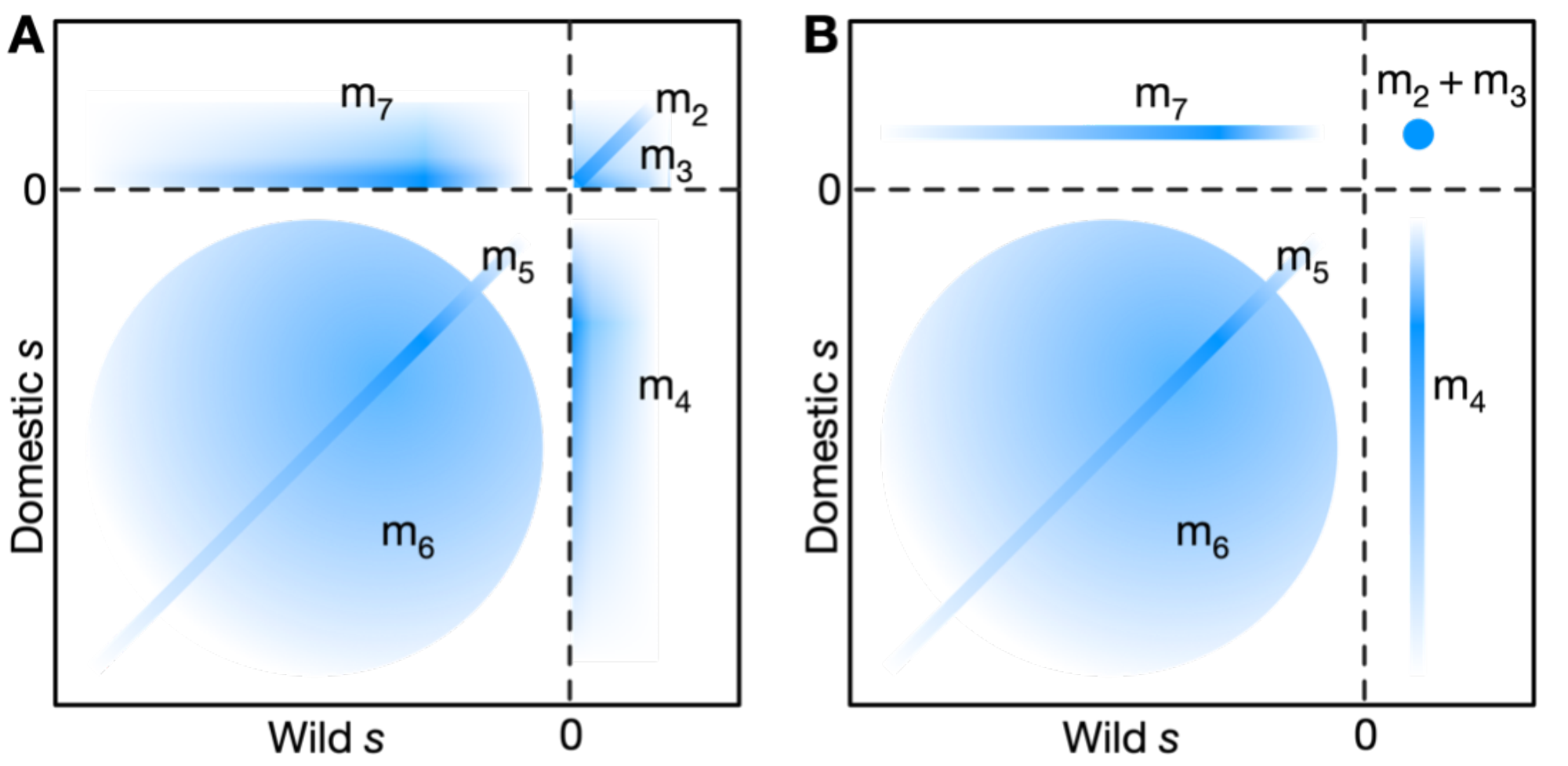
Figure 6: Discretized full DFE inferred by polyDFE, showing the Wild and Domestic populations under each scenario.

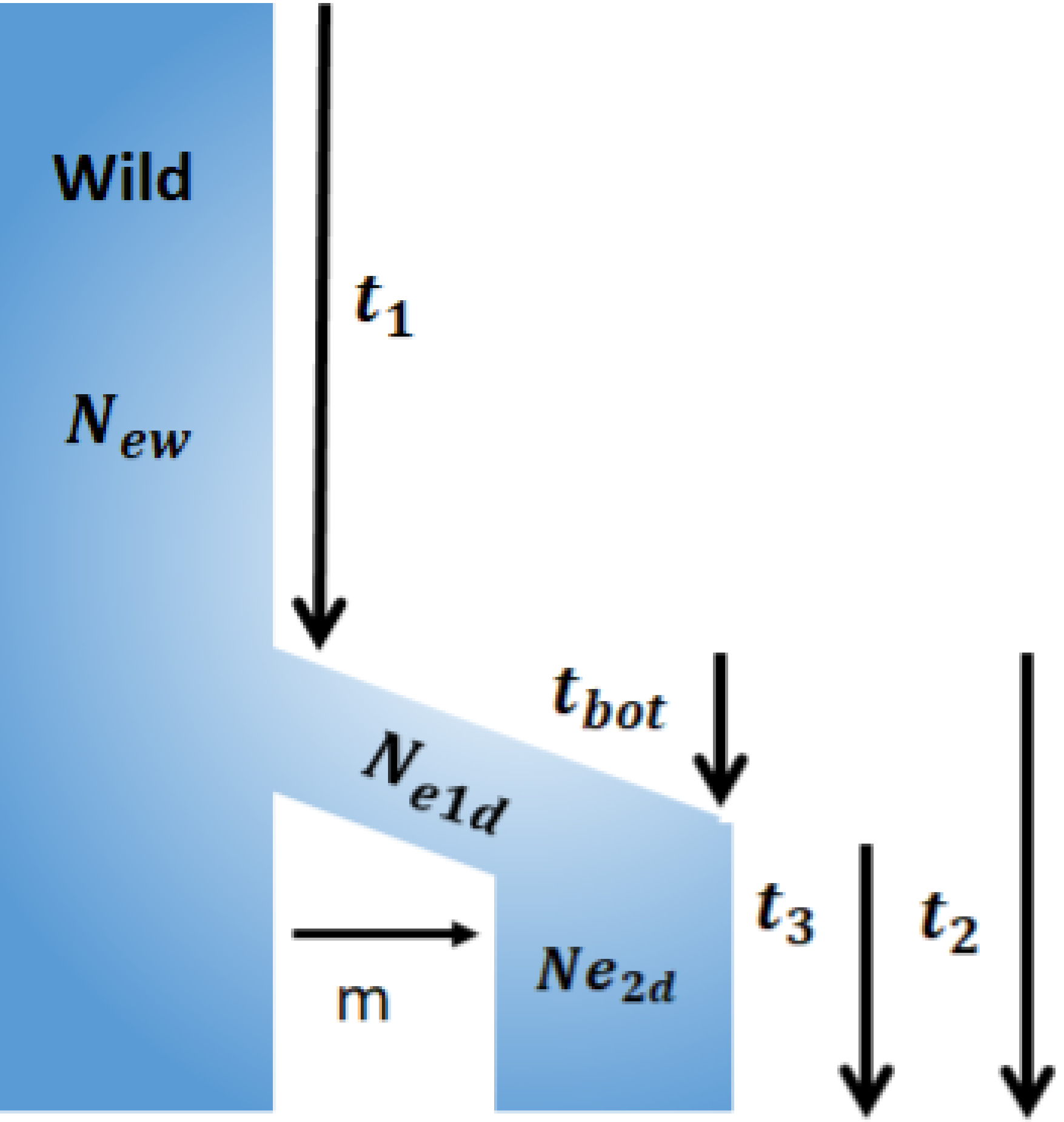
Figure 7: Estimation of demographic parameters from synonymous mutations under each scenario. Each plot includes four subfigures or panels. The one entitled "data" (top left panels) represents the two-

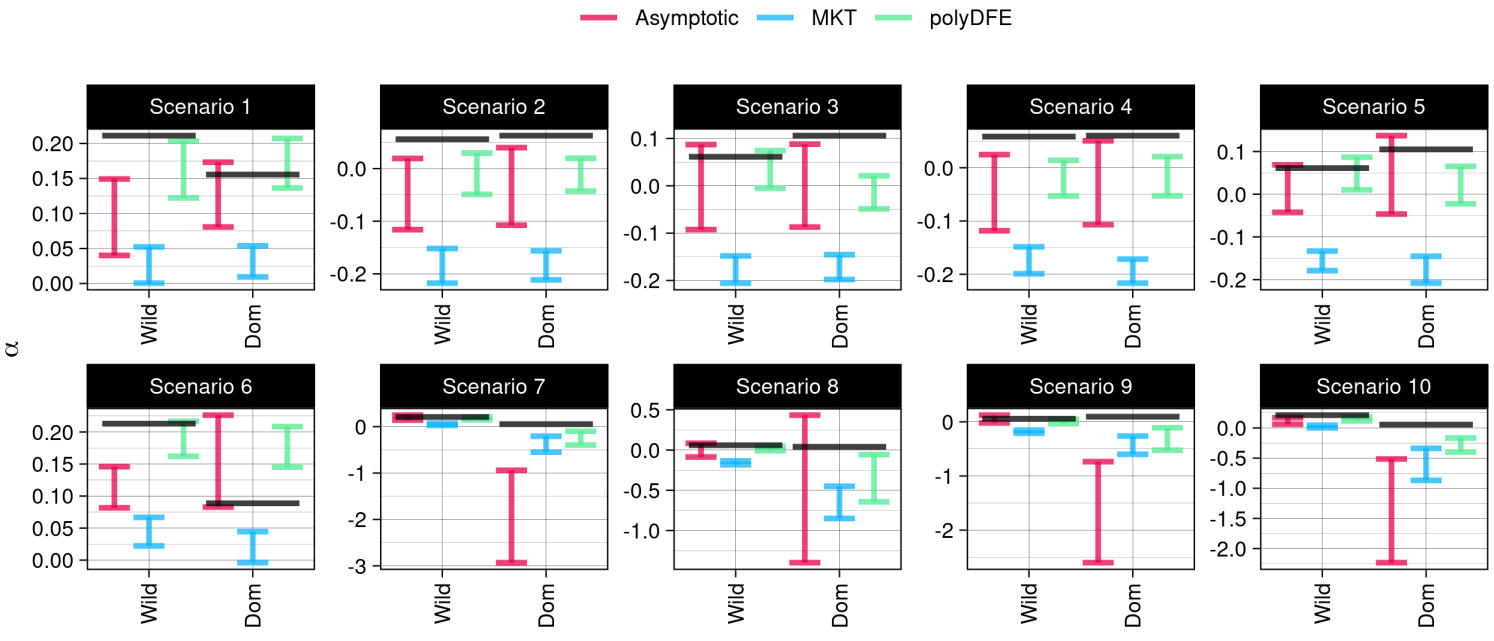
dimensional simulated SFS for both populations, while the right one labeled "model" (top right panels) depicts the expected 2D-SFS for the inferred dadi model. The two plots at the bottom illustrate the residual performance (bottom left panels) and the residual histogram (bottom right panels) of the model.

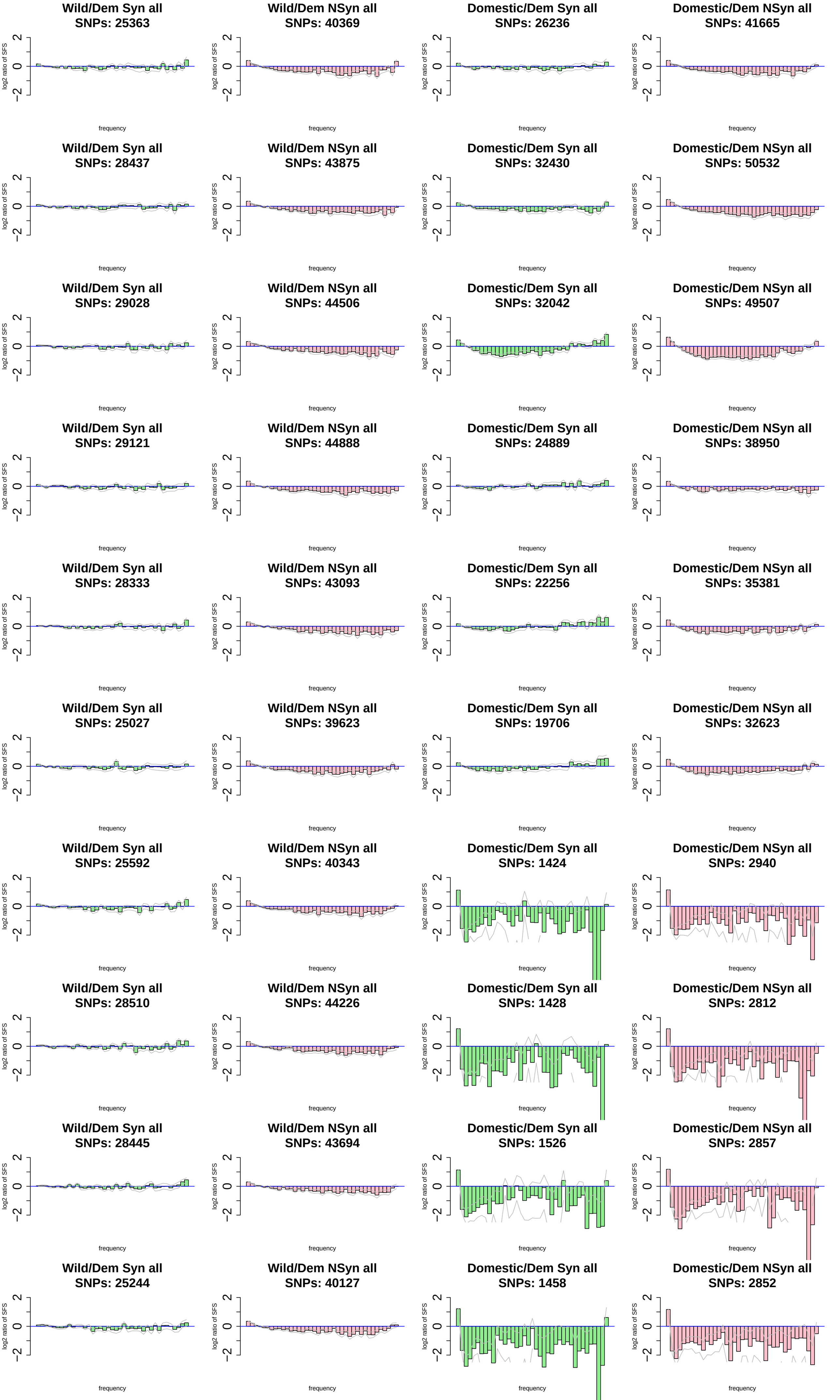
Figure 8: Estimation of the DFE from nonsynonymous mutations under each scenario. Each plot includes four subfigures or panels. The one entitled "data" (top left panels) represents the two-dimensional simulated SFS for both populations, while the right one labeled "model" (top right panels) depicts the expected 2D-SFS. The two plots at the bottom illustrate the residual performance (bottom left panels) and the residual histogram (bottom right panels) of the model.

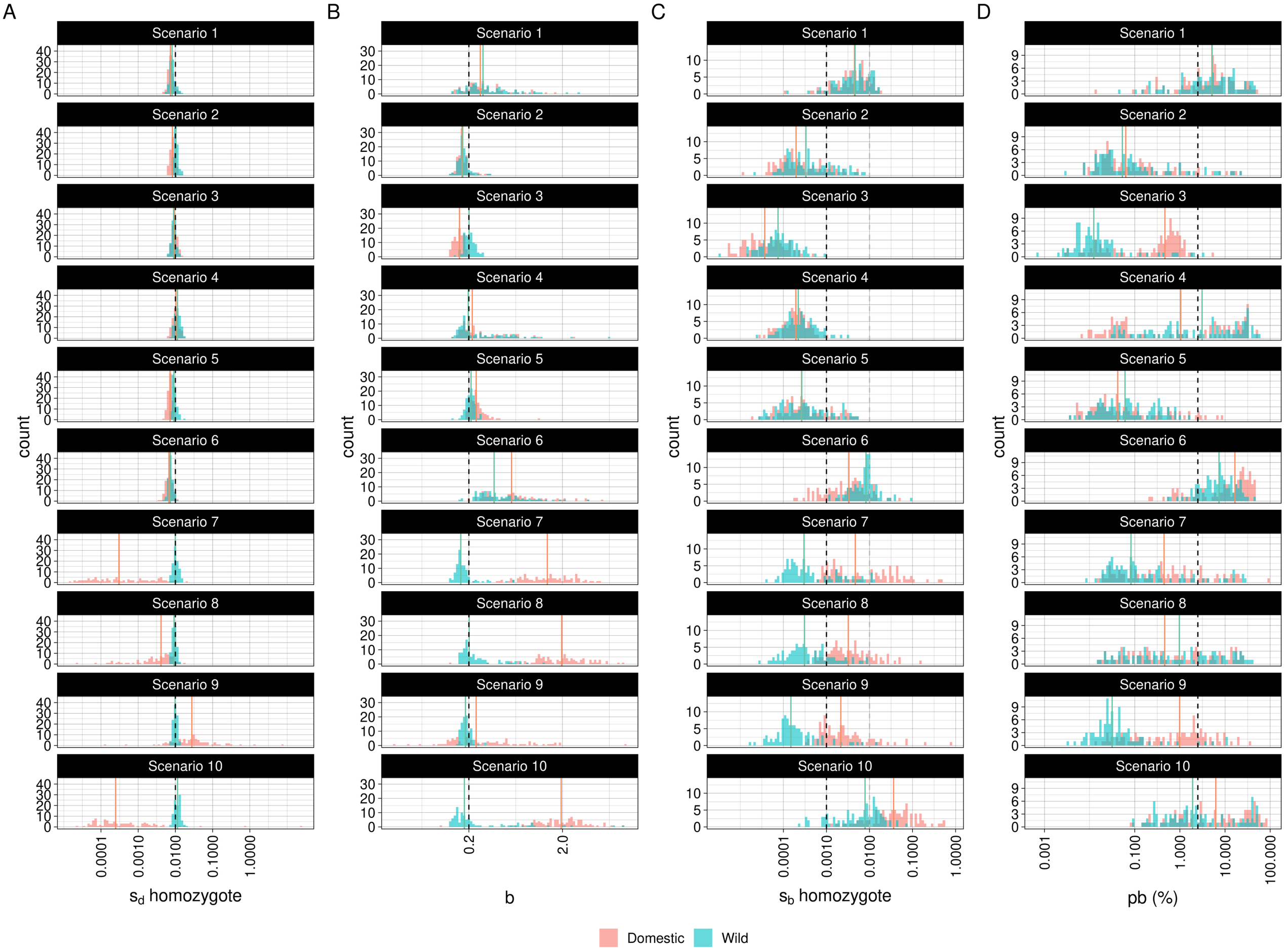
Figure 9: Confidence intervals and values for the inferred parameters versus the real ones (vertical dashed lines and crosses) for each population and scenario for s_d (A), the shape parameter b (B) and the DFE parameters describing divergent selective effects (C). To obtain s_d from S_d values we use π at synonymous sites in the inferred ancestral population size and the true simulated mutation rate (1e-6) to get the realized N_e after the bottleneck and/or the action of linked selection. Note that in *dadi s* is defined to be the selection coefficient on the heterozygote (like in polyDFE), but the scaled selection coefficient is defined as $2N_es$.

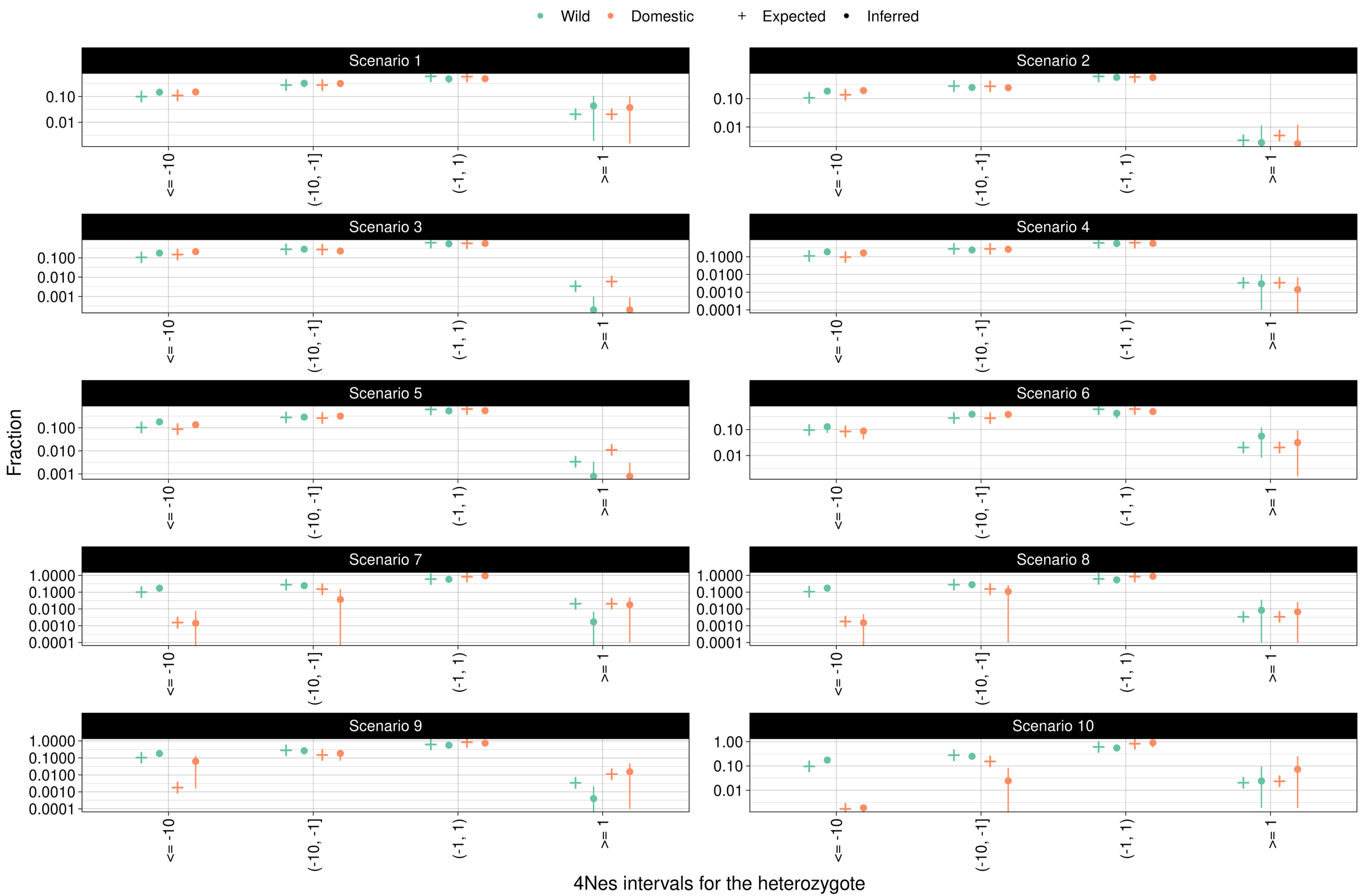

A
m7
m2
m3
0
m5
m4
m6
Domestic s
Wild s
0
B
m7
m2 + m3
0
m5
m4
m6
Domestic s
Wild s
0

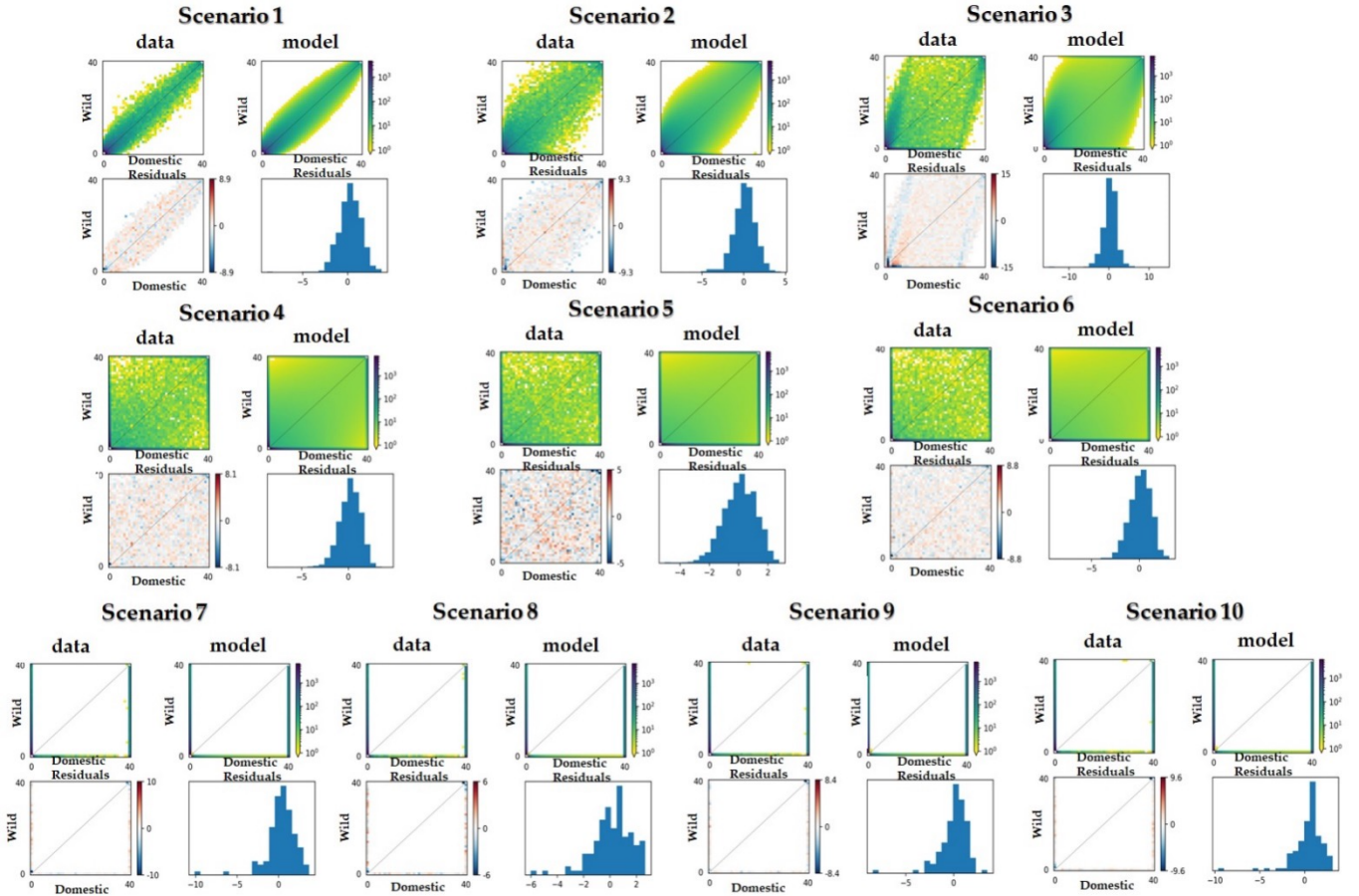

Wild
t_1
N_{ew}
t_{bot}
N_{e1d}
t_3
t_2
m
Ne_{2d}

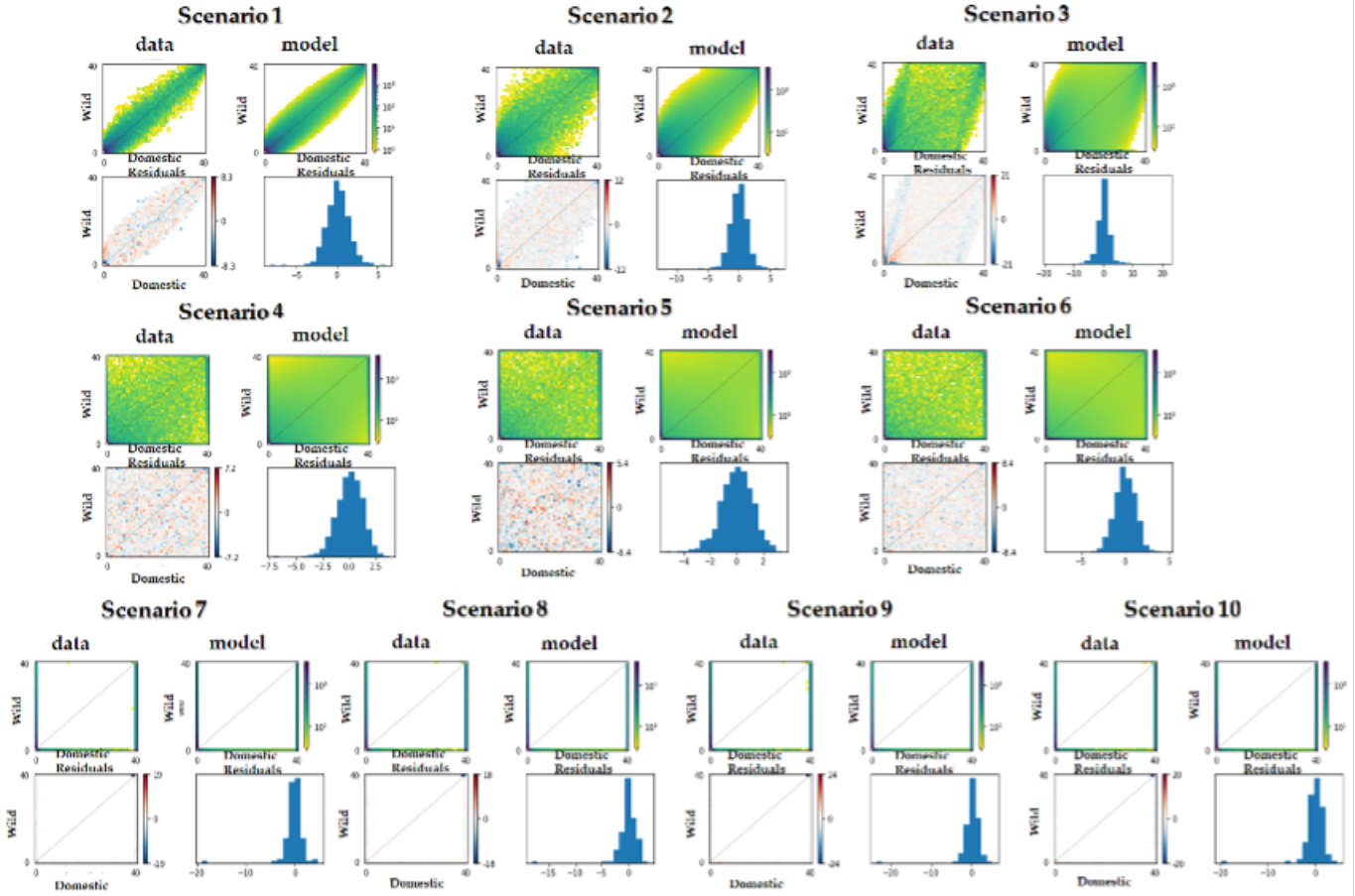

Asymptotic
MKT
polyDFE
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
Scenario 6
Scenario 7
Scenario 8
Scenario 9
Scenario 10
α
Wild
Dom

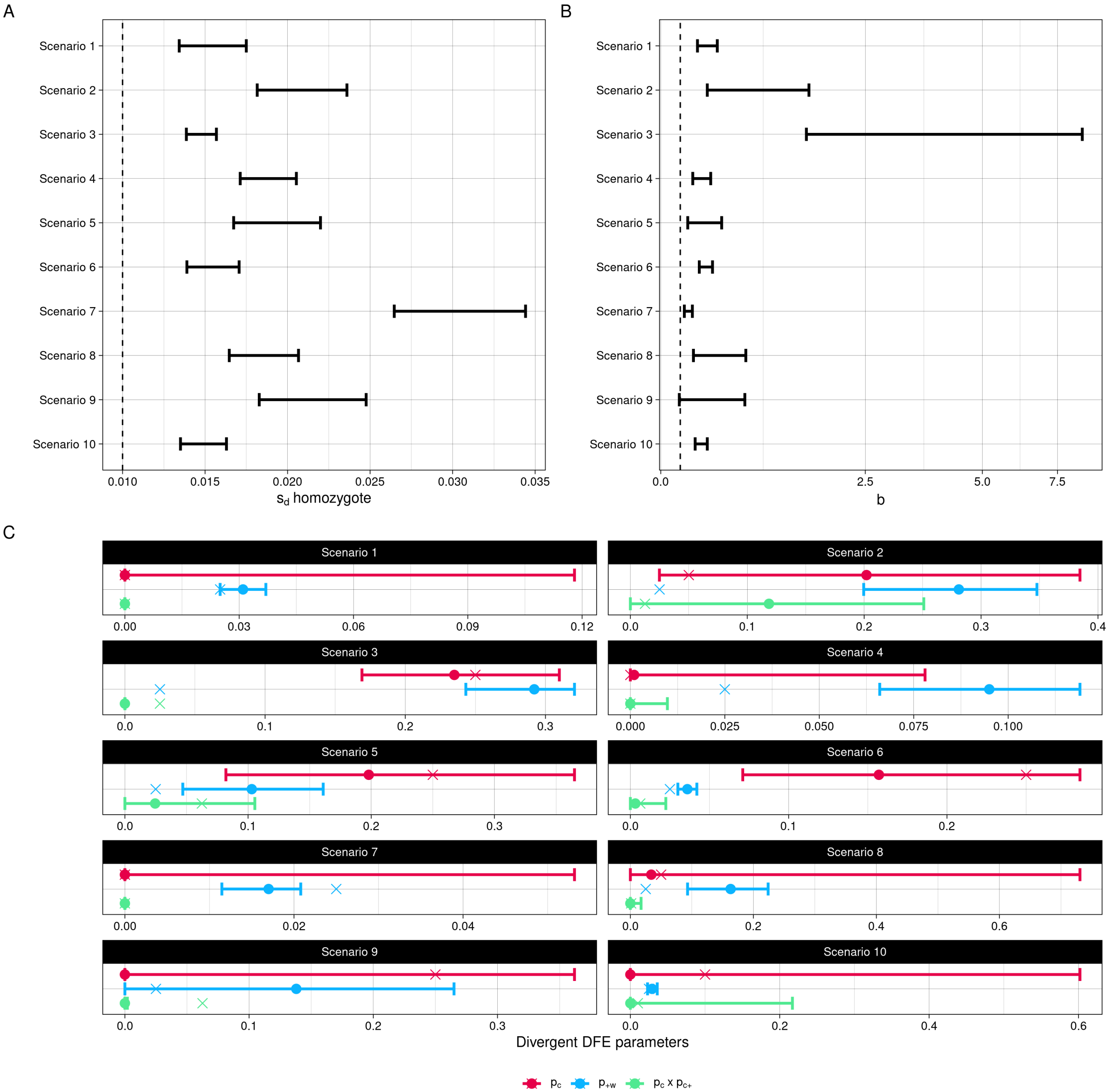

Wild/Dem Syn all
SNPs: 25363
Wild/Dem NSyn all
SNPs: 40369
Domestic/Dem Syn all
SNPs: 26236
Domestic/Dem NSyn all
SNPs: 41665
Wild/Dem Syn all
SNPs: 28437
Wild/Dem NSyn all
SNPs: 43875
Domestic/Dem Syn all
SNPs: 32430
Domestic/Dem NSyn all
SNPs: 50532
Wild/Dem Syn all
SNPs: 29028
Wild/Dem NSyn all
SNPs: 44506
Domestic/Dem Syn all
SNPs: 32042
Domestic/Dem NSyn all
SNPs: 49507
Wild/Dem Syn all
SNPs: 29121
Wild/Dem NSyn all
SNPs: 44888
Domestic/Dem Syn all
SNPs: 24889
Domestic/Dem NSyn all
SNPs: 38950
Wild/Dem Syn all
SNPs: 28333
Wild/Dem NSyn all
SNPs: 43093
Domestic/Dem Syn all
SNPs: 22256
Domestic/Dem NSyn all
SNPs: 35381
Wild/Dem Syn all
SNPs: 25027
Wild/Dem NSyn all
SNPs: 39623
Domestic/Dem Syn all
SNPs: 19706
Domestic/Dem NSyn all
SNPs: 32623
Wild/Dem Syn all
SNPs: 25592
Wild/Dem NSyn all
SNPs: 40343
Domestic/Dem Syn all
SNPs: 1424
Domestic/Dem NSyn all
SNPs: 2940
Wild/Dem Syn all
SNPs: 28510
Wild/Dem NSyn all
SNPs: 44226
Domestic/Dem Syn all
SNPs: 1428
Domestic/Dem NSyn all
SNPs: 2812
Wild/Dem Syn all
SNPs: 28445
Wild/Dem NSyn all
SNPs: 43694
Domestic/Dem Syn all
SNPs: 1526
Domestic/Dem NSyn all
SNPs: 2857
Wild/Dem Syn all
SNPs: 25244
Wild/Dem NSyn all
SNPs: 40127
Domestic/Dem Syn all
SNPs: 1458
Domestic/Dem NSyn all
SNPs: 2852
log2 ratio of SFS
frequency
2
0
−2


A
B
C
D
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
Scenario 6
Scenario 7
Scenario 8
Scenario 9
Scenario 10
count
0.0001
0.0010
0.0100
0.1000
1.0000
s_d homozygote
0.2
2.0
b
s_b homozygote
0.001
0.100
1.000
10.000
100.000
pb (%)
Domestic
Wild


Wild
Domestic
Expected
Inferred
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
Scenario 6
Scenario 7
Scenario 8
Scenario 9
Scenario 10
Fraction
<= -10
(-10, -1]
(-1, 1)
>= 1
4Nes intervals for the heterozygote


Scenario 1
data
model
Scenario 2
data
model
Scenario 3
data
model
Scenario 4
data
model
Scenario 5
data
model
Scenario 6
data
model
Scenario 7
data
model
Scenario 8
data
model
Scenario 9
data
model
Scenario 10
data
model
Wild
Domestic
Residuals


Scenario 1
data
model
Scenario 2
data
model
Scenario 3
data
model
Scenario 4
data
model
Scenario 5
data
model
Scenario 6
data
model
Scenario 7
data
model
Scenario 8
data
model
Scenario 9
data
model
Scenario 10
data
model
Wild
Domestic
Residuals


A
Scenario 1
Scenario 2
Scenario 3
Scenario 4
Scenario 5
Scenario 6
Scenario 7
Scenario 8
Scenario 9
Scenario 10
0.010
0.015
0.020
0.025
0.030
0.035
s_d homozygote
B
0.0
2.5
5.0
7.5
b
C
Divergent DFE parameters
p_c
p_{+w}
p_c x p_{c+}