Pervasive selection pressure in wild and domestic pigs

J. Leno-Colorado[1#], S. Guirao-Rico[2#], M. Pérez-Enciso[134], S. E. Ramos-Onsins[1*]

[1] Centre for Research in Agricultural Genomics (CRAG) Consortium CSIC-IRTA-UAB-UB, Bellaterra, Spain.

[2] Departament de Genètica, Microbiologia i Estadística & Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.

[3] Universitat Autònoma de Barcelona, Department of Animal Science, Bellaterra, Spain.

[4] Institut Català de Recerca I Estudis Avançats (ICREA), Barcelona, Spain.

[#]Equal contribution

*Corresponding author: sebastian.ramos@cragenomica.es

ABSTRACT

Animal domestication typically affected numerous polygenic quantitative traits, such as behaviour, development and reproduction. However, uncovering the genetic basis of quantitative trait variation is challenging, since it is probably caused by small allele-frequency changes. To date, only a few causative mutations related to domestication processes have been reported, strengthening the hypothesis that small effect variants have a prominent role. So far, the studies on domestication have been limited to the detection of the global effect of domestication on deleterious mutations and on strong beneficial variants, ignoring the importance of variants with small selective effects. In addition, very often, the study of the effects of selection are conducted on genome sequences that contain a non-negligible fraction of missing data, especially in non-model organisms. Hence, appropriate methods to account for these positions are needed. To overcome these difficulties, here we propose to estimate the proportion of beneficial variants using the asymptotic MacDonald-Kreitman (MK) method based on estimates of variability that summarizes the site frequency spectrum (SFS) while accounting for missing data and use them to perform an Approximate Bayesian Computation (ABC) analysis to infer the Distribution of Fitness Effects (DFE) of each population. We applied this approach to 46 genome sequences of pigs from three different populations, one wild and two domestics, with very different demographic histories and selective pressures. The obtained results showed that domestic and wild pig populations do not differ in nonsynonymous fixed mutations. Therefore, differences in α estimation among breeds are determined by their polymorphisms. The comparison of α for total and exclusive mutations suggests that the different domestic populations have suffered recent divergent changes in their functional versus neutral polymorphisms ratio, while the wild population is compatible with α=0. Besides, the DFE inferred with ABC indicates that both wild and domestic pigs display a large number of deleterious mutations at low frequency and a high number of neutral and/or nearly-neutral mutations that may have a significant effect on the evolution of domestic and wild populations. In addition, models not considering beneficial mutations have higher posterior probabilities, suggesting that beneficial mutations are difficult to detect or are scarce. Indeed, for all three populations, the median proportion of the strong favourable mutations are very low ($\leq$ 0.1%) in those models that includes positive selection, with the average values of weak beneficial mutations around 0.6% for wild boar and 0.8-1.0% for the domestic pigs. Lastly, the analysis based on exclusive mutations showed that recent demographic changes may have severely affected the fitness of populations, especially that of the local Iberian breed.

Keywords: Domestication, Distribution of fitness effects, Proportion of beneficial mutations, Approximate Bayesian calculation, Polygenic selection, Population genomics

INTRODUCTION

Domestic animal histories are evolutionary experiments that have often lasted for millennia resulting in dramatic phenotypic changes to suit human needs. In addition, domestic species can be structured into subpopulations (breeds) that are partly or completely genetically isolated and can display a wide catalogue of specific phenotypes. Therefore, they offer a very valuable material of utmost interest to study the interplay between demography and accelerated adaptation. However, as their demographic history can be quite complex, many events remain unknown or poorly documented nowadays.

The pig (*Sus scrofa*) is a particularly interesting species because of its domestication history and its relatively well-annotated genome. *S. scrofa* originated in Southeast Asia ~1-4 MYA and spread throughout Eurasia ~0.2-1.2 MYA, colonizing all climates except the driest (Frantz et al. 2013, Zhang et al. 2021). Subsequently, the pig was domesticated from local wild boars (WB) independently in both Asia and Europe ~9,000 years ago. To complicate the story, modern European domestic pig breeds were crossed with Asian domestic pigs during the late 17th century and onwards. In breeds such as Large White (LW), approximately 30% of the genome is estimated to be of Asian origin (Bosse, Megens, Madsen, et al. 2014). Nevertheless, some local European breeds, such as the Iberian breed (IB), were spared genetic contact with Asian pigs and no evidence of genetic introgression has been found in this breed (Alves et al. 2003, Esteve-Codina et al. 2013). Moreover, domestic breeds have different recent demographic histories. For instance, the IB breed suffered a dramatic reduction of its effective population size during the last century (Alves et al. 2006), whereas many commercial breeds such as Duroc or LW have been introgressed with Asian pigs (Bosse, Megens, Frantz, et al. 2014).

Differences in the effective population size, demographic histories and artificial selective pressures between pig breed or populations could result in differences among their evolutionary rates. In addition to possible differences in the evolutionary rates between populations, there may be differences in the evolutionary rate between genes within genomes. For instance, it is known that the strength of the selection is affected by the position of the genes in the networks in which they participate. Genes that are more central in a network and are more connected with other genes are more evolutionarily constrained, while peripheral genes are more prone to be under adaptive selection (Fraser et al. 2002; Hahn and Kern 2005; Montanucci et al. 2011; Alvarez-Ponce and

Fares 2012). Furthermore, it has been observed that the evolutionary rate, within a metabolic pathway, increases as we move downstream, possibly because upstream genes are more pleiotropic, since they are involved in more functions and hence, these genes are probably more conserved (Rausher, Miller, and Tiffin 1999; Riley, Jin, and Gibson 2003; Livingstone and Anderson 2009; Ramsay, Rieseberg, and Ritland 2009).

So far, the nature of the underlying genetic changes caused by domestication and ensuing artificial breeding is still under debate. While the most prevalent view is that regulatory changes have been targeted (Anderson 2013), several other studies underline the influence of protein coding changes (Rubin et al. 2012). Some authors have reported an increase in the rate of deleterious mutations in domestic pigs compared to their wild counterparts (Cruz, Vilà, and Webster 2008; Renaut and Rieseberg 2015; Pérez-Enciso et al. 2017; Leno-Colorado et al. 2017). Others, as in Makino et al. (2018) detected a general pattern of reduction of variability in domestic populations in relation to their wild counterpart, and a higher nonsynonymous/synonymous ratio across the frequency spectrum. These patterns were compatible with the effect of strong bottlenecks in domestic populations and the higher accumulation of deleterious mutations. Interestingly, the same authors observed the opposite trend in pigs (e.g., higher variability levels in domestic pigs compared to their wild counterpart). Moreover, most of these previous studies have focused on genes of major effect with clear signals of selective sweeps. In those studies, the hallmarks of positive selection were detected as valleys of reduced variation and/or population differentiation that spans relatively large regions (e.g., Amaral et al. 2011, Rubin et al. 2012, Frantz et al. 2013, Wilkinsonet al. 2013), but also by the presence of haplotype structure and homozygosity blocks (e.g., Fang et al. 2011, Bosse et al. 2012, Li et al. 2013). Some of these studies have detected recent breed specific signals of selection attributed to the domestication process (Li et al. 2014, Kim et al. 2015). Nevertheless, the signals were too scarce to explain the whole picture of the domestication process. Other studies have tried to elucidate the effect that domestication has at the genomic scale and on the fitness of individuals of domestic populations (e.g., Cruz et al. 2008, MacEachern et al. 2009, Kono et al. 2016, Perez-Enciso et al. 2016, Makino et al. 2018, Chen et al. 2018, Orlando and Librado 2019). For instance, an excess of deleterious variants has been observed in a number of domestic animal and plants (e.g., contrasting nonsynonymous versus synonymous polymorphism ratios, Chen et al. 2018, using the MacDonald framework, MacEachern et al. 2009, contrasting the ancestors with ancient DNA, Orlando and Librado 2019, combining the frequency of polymorphisms with

functional effects and divergence, Kono et al. 2016, Makino et al. 2018). Kono et al. (2016) and Perez-Enciso et al. (2016) found an excess of deleterious variants affecting phenotypes of interest, suggesting, as we previously mention above, that protein sequence may have a stronger influence than regulatory changes in the domestication process. Kono et al. (2016) also showed that null alleles are uncommon in domestic animal species (also reviewed by Anderson 2013), suggesting that phenotypic changes involved in domestication are produced by the accumulation of consecutive mutations that modify the gene functions under selection. Finally, the possible presence of beneficial mutations during the domestication process has also been reported (Perez-Enciso 2016).

Here, we are interested in determining the proportion and the selective effects of protein-coding variants in wild and domestic pig genomes to understand their role in the domestication process. Particularly, we aimed to test the role of both new and extant mutations in the domestication process and whether the phenotypes associated with domestic breeds are the product of a large number of variants with weak selective effects, as suggested by previous results. To achieve this, we have investigated the differential effects of selection on coding sequences at the different molecular scales (gene, metabolic pathway and whole-genome) in two domestic and one wild pig population using the McDonald-Kreitman framework (McDonald and Kreitman 1991, Eyre-Walker 2006, Fay 2011). We also have performed forward exploratory simulations and inferred the distribution of fitness effects (DFE) while taking into account the effect of different demographic scenarios. Interestingly, the analysis was performed using variability estimators that allow including positions with missing data (Ferretti, Raineri, and Ramos-Onsins 2012).

Our results support the hypothesis that changes in allele frequencies in coding variants with weak positive selective effect have been relevant for pig domestication, as evidenced by a relatively high number of nonsynonymous variants segregating at medium and high frequencies and by the obtained estimates of the DFE in domestic pig populations.

MATERIALS AND METHODS

Biological samples

We analyzed a sample of 46 pig (*Sus scrofa*) genomes (Table S1). These pigs correspond to European wild boars (WB, n = 20) and domestic pigs, which are represented by the Iberian Guadyerbas (IB, n = 6) and Large White (LW, n = 20) breeds. The two domestic breeds were selected because they have very different interesting features: IB is a local breed that has been under weak artificial selection intensity and with no documented evidence of Asian introgression. LW, in contrast, is a commercial breed undergoing strong artificial selection with a deliberate admixture with Asian pigs (Bosse, Megens, Madsen, et al. 2014; Groenen 2016). To analyze the divergence between the different breeds, we built the consensus ancestral reference sequence obtained from combining the genomic information from several *Sus* species (*S. barbatus*, *S. cebifrons*, *S. verrucosus*, *S. celebensi*, approximately 4.2 MYA of divergence) and the African warthog (*Phacochoerus africanus*, around ~10 MYA of divergence) and used as an outgroup, as in Bianco et al. (2015). All sequences (see reference numbers at Table S1) are available in public databases: those generated in several previous works (Rubin et al. 2012; Ramírez *et al.* 2014; Bianco *et al.* 2015; Frantz et al. 2015; Moon *et al.* 2015, Esteve-Codina *et al.* 2013, Leno *et al.* 2017) and those generated in this work (WBES0231, WBES0252, WBES0288, WBES0291 and WBES0297, see Table S1). They can be downloaded from the short read archive (SRA, http://www.ncbi.nlm.nih.gov/sra, see accession numbers in Table S1).

Mapping and genotyping analysis

For each pig genome, raw reads were mapped against the reference genome assembly (Sscrofa10.2, Groenen et al. 2012) using *BWA mem* option (H. Li and Durbin 2009). PCR duplicates were removed using *SAMtools rmdup* v. 0.1.19 (H. Li et al. 2009) and mapped reads were realigned around indels with the *GATK IndelRealigner* tool (McKenna et al. 2010). Genotype calling was performed with *SAMtools mpileup* and *bcftools call* v. 1.3.0 (H. Li et al. 2009) for each individual separately. We set a minimum (5x) and a maximum depth (twice the average sample's depth plus one) to call a SNP. Base quality was set to 20 (*P*-value=1e-2). Homozygous blocks (regions of contiguous positions with the same nucleotide as the reference genome) were also called, following the same criteria as with the SNPs (i.e., minimum and maximum coverage and

base quality) and using *samtools depth* utility, *BEDtools* (Quinlan 2014) and custom scripts (available at https://github.com/miguelperezenciso/NGSpipeline, Pérez-Enciso *et al*., 2017). This resulted in a *gVCF* file per individual with the information about variant calls and non-varying positions. Next, each *gVCF* file was converted into a fasta file and all fasta files were subsequently merged to obtain a multindividual *gVCF* file (Pérez-Enciso et al. 2017).

Analysis of the population structure of the samples

A principal component analysis (PCA) was performed using the total number of SNPs to explore the population structure. First, genotypes were converted to alternative allele frequency, being 0 for the homozygous reference genotype (0/0), 0.5 for the heterozygous genotype (0/1) and 1 for the homozygous alternative genotype (1/1). For cases of missing genotype (./.), these were replaced by the average SNP frequency across all individuals. We used the function *tcrossprod()* from R v. 3.3.1 (2016) to obtain the matrix of covariates from the matrix of frequencies. Finally, we obtained the principal components from the Eigen-value decomposition with the R function *eigen()*. The software ADMIXTURE (Alexander et al. 2009) was also applied to analyze the population structure. The more suitable K was estimated using cross-validation procedure (Alexander et al. 2009) and the Evanno's method (Evanno et al. 2005).

Estimation of codon bias

We have used the single reference sequence of *S. scrofa* to estimate the codon bias at gene and genome level, assuming that polymorphic variants are not going to strongly modify the proportion of codons at this species. We have estimated gene and genomic codon usage with the Major Codon Usage statistic (MCU, the frequency of major codons among all codons in a sequence) and the Effective number of codons (N_{cw}), using the python script following Fuglsang (2006). High values of MCU indicate a strong bias in codon usage, while low values suggest small bias. Instead, high values of N_{cw} indicate low codon bias because many codons are used. We have also calculated the correlation between MCU and α estimates (see below), considering all annotated coding regions and only coding regions having α values larger than zero.

Estimation of levels and patterns of variability

Genetic diversity and divergence per pig population were estimated using *mstatspop* software (Nevado, Ramos-Onsins, and Perez-Enciso 2014; Bianco et al. 2015; Guirao-Rico et al. 2018, available from the authors, https://github.com/cragenomica/mstatspop). The *multi-VCF* file was converted into a *tfasta* (transposed *fasta*) file and *mstatspop* was run on i) the whole genome, ii) windows of 5-Mb size, and iii) gene coding regions. Note that, given the ubiquitous presence of missing data, the SFS with highest sample size that contained more coding SNPs, retained only around 20-25% of total coding variants. As the aim of this work is to detect the presence of (weak) beneficial selective effects and as not to lose power or bias the results of the analysis, we preferred to use an alternative method that consider the whole set of SNPs. We used four different estimators of nucleotide variability that takes into account missing data (Ferretti, Raineri, and Ramos-Onsins 2012): Watterson (Watterson 1975), Tajima (Tajima 1983), Fu&Li (Fu and Li 1993) and Fay&Wu's estimators (Fay and Wu 2000). The variability was estimated for total, shared and exclusive variants, being shared and exclusive nucleotide variability counted regarding to the total positions (i.e., total = shared + exclusive).

Filtering for artefactual effects

A preliminary analysis of the variability showed a moderate negative correlation (~ 0.3) between the levels of variability and divergence and the proportion of missing data for each gene. To eliminate this artifactual correlation, we plotted the estimators of variability and divergence versus the ratio of missing data and eliminated those genes that showed a ratio of missing data greater than 0.3. Since this filtering was not enough to completely remove the bias, we also removed genes with extreme values of variability and divergence (higher than 99% quantile of the total genes). The remaining ~13,500 genes (70% of the total annotated genes) showed a low or null correlation with missing data and were used in the present analysis (Table S2).

Estimation of the proportion of adaptive of variants

Under the neutral model, the majority of polymorphisms segregating in a population are neutral and only a small number of selected variants segregates for a short time on their way to loss or

fixation. Hence, most of the positive selected variants are only observed as fixed variants. In addition, functional positions (nonsynonymous positions) are constrained compared to non-functional positions (synonymous positions), and hence, their evolutionary ratios are smaller. In the neutral scenario, polymorphism and divergence (excluding the adaptive fixed variants) are proportional to the mutation rate and to the constriction factor in the case of nonsynonymous positions (McDonald and Kreitman 1991, Eyre-Walker 2006, Fay 2011). That is:

$$\frac{\theta_n}{\theta_s} = \frac{(1-\alpha)K_n}{K_s},$$

(Equation 1)

where θ_n is the nonsynonymous variability, θ_s is the synonymous variability, K_n is the nonsynonymous divergence, K_s is the synonymous divergence and α is the proportion of adaptive variants that have been fixed. To estimate the proportion of nonsynonymous substitutions that are adaptive (α) the previous expression is reordered (e. g., Eyre-Walker 2006):

$$\alpha = 1 - \frac{K_s}{K_n}\frac{\theta_n}{\theta_s}$$

(Equation 2)

A higher ratio of nonsynonymous to synonymous divergence versus polymorphisms suggests that positive selection has fixed adaptive variants ($\alpha > 0$) and the opposite case ($\alpha < 0$) suggests the presence of deleterious mutations segregating in the population.

If we consider that weak deleterious mutations are segregating in the population, we expect that their relative proportion will be higher at lower frequency variants and low or zero for fixed deleterious mutations. Following the same notation as in equation 2:

$$\frac{\theta_{in}(1-\beta_i)}{\theta_{is}} = \frac{(1-\alpha-\beta_d)K_n}{K_s},$$

(Equation 3)

where i refers to the frequency at which the calculation of variability is estimated, β_i is the proportion of weakly deleterious polymorphic mutations at frequency i, β_d is the proportion of weakly deleterious fixed mutations. $\beta_d < \beta_i$ was assumed at any frequency. Then, solving for the proportion of fixed adaptive variants (α):

$$\alpha = 1 - \beta_d - (1 - \beta_i)\frac{K_s}{K_n}\frac{\theta_{in}}{\theta_{is}}$$

(Equation 4)

We see that in case of calculating α without considering the effects of deleterious mutations, this would be underestimated depending on the frequency at which the estimates of variability are calculated. If we assume that the deleterious variants would hardly be fixed, a good estimator of α using equation 2 would be the one that estimates variability based on high frequencies, as it would hardly contain deleterious mutations. This is in agreement with the asymptotic arguments used in Messer and Petrov (2013) and implemented in Haller and Messer (2017).

Similarly, if we also consider that weak positively selected variants are segregating in the population, we expect that their relative proportion, compared to neutral ones, is higher at higher frequencies:

$$\frac{\theta_{in}(1 - \beta_i - \gamma_i)}{\theta_{is}} = \frac{(1 - \alpha - \beta_d - \gamma_d)K_n}{K_s},$$

(Equation 5)

where γ_i is the proportion of weakly advantageous polymorphic mutations at frequency i, and γ_d is the proportion of weakly advantageous fixed mutations. Again, solving for the proportion of fixed adaptive variants (α+γ_d):

$$\alpha + \gamma_d = 1 - \beta_d - (1 - \beta_i - \gamma_i)\frac{K_s}{K_n}\frac{\theta_{in}}{\theta_{is}}$$

(Equation 6)

In this case, the presence of adaptive variants segregating in the population would affect the estimates of variability based on high frequency variants when using equation 2, which would result in an underestimation of the proportion of fixed adaptive variants (α). Note that adaptive variants stabilized at intermediate frequencies, which can be an important source of adaptation considering the infinitesimal model, are not considered in this approach.

If we focus on the effects of polymorphic weakly selected mutations, equation 5 suggests that the ratio of nonsynonymous to synonymous polymorphisms would increase due to mutations having both positive and negative effects. It is expected that the number of mutations with negative selection coefficients would rapidly decrease as we move to intermediate and high frequencies, while the opposite trend is expected for mutations with positive selection coefficients. Hence, higher ratios of nonsynonymous to synonymous polymorphisms at higher frequencies may be explained by the presence of advantageous mutations segregating in the population.

Furthermore, in cases where two populations are from the same species and there are no fixed mutations between them (e.g., they have equal divergence ratios versus the outgroup), we can estimate the possible differential effect of the selection (positive and negative together) at any frequency between populations from the ratios of synonymous to nonsynonymous polymorphisms of the two populations ($R\_\beta\gamma_i$):

$$\frac{\theta_{in1}(1-\beta_{i1}-\gamma_{i1})}{\theta_{is1}} = \frac{\theta_{in2}(1-\beta_{i2}-\gamma_{i2})}{\theta_{is2}}$$

and

$$\frac{(1-\beta_{i1}-\gamma_{i1})}{(1-\beta_{i2}-\gamma_{i2})} = \frac{\theta_{is1}\theta_{in2}}{\theta_{in1}\theta_{is2}} = R\_\beta\gamma_i$$

(Equation 7)

In addition, a comparison of the $R\_\beta\gamma_i$ values calculated using different variability estimators (hereafter $R\_\beta\gamma_i$ pattern) can be used to inform about the effects of selection. For example, values over 1 indicate that the population 2 has a higher ratio of nonsynonymous to synomymous polymorphisms compared to population 1, either produced by an accumulation of deleterious or of beneficial polymorphisms. Importantly, different demographic effects (e.g., bottlenecks) together with the presence of mutations with small selective effects may also disturb the ratios of variability and hence must be considered when interpreting the results. We include a couple of possible scenarios that can account for possible patterns: (i) after split of two the two populations, both populations have the same population size, but population 1 is affected by the action of positive selection on a quantitative trait (polygenic effect), which causes an increase in the frequencies of some of its variants without getting fixed. Under this scenario, we expect a R_βγ >

1 when this is calculated based on high frequencies. (ii) after split of two the two populations, the population 2 remains equal population size as before the split and the population 1 suffers a reduction in its effective population size, which causes that the slightly deleterious mutations become effectively neutral. Then, R_βγ is expected to be > 1 when it is calculated based on low frequency variants.

The effect of linkage disequilibrium between selective (deleterious or adaptive) and neutral variants should not overly affect the expected estimate of the proportion of adaptive variants, as it would affect both synonymous and nonsynonymous positions in similar proportion. On the contrary, the interaction of variants with opposite selective effects would possibly reduce the effect of selection and would have a significant consequence on the estimation of adaptive fixed variants (Hill and Robertson 1966; Booker and Keightley 2018).

Bootstrap analysis

Nonparametric bootstrap analysis was performed to estimate the null distribution of the α statistic for each variability estimator and pig population. In each case, synonymous and nonsynonymous coding positions were randomly chosen with replacement and the α statistic was calculated as in equation 1. This process was repeated 100 times.

Simulations

We carried out forward simulations using the software *SLiM* (Haller and Messer 2017) in order to explore the interaction between the different selective effects and demographic factors affecting the evolution of pig populations during domestication. We explored the expected values of nucleotide diversity, divergence, α and $R\_\beta\gamma$ statistics under 63 different scenarios. For each scenario, we simulated three populations corresponding to wild, domestic and an outgroup species. We first simulated nine different scenarios that were classified into three main groups: i) standard neutral model (SNM); ii) a model with negative selection (NS) and iii) a model with positive selection (PS). For the models with selection, we let that selection operate from the ancestral species to the present time. Each group of scenarios (SNM, NS and PS) was simulated with a constant effective population size for the three populations or with a reduction or an expansion of

the effective population size in the branch leading to domestic pigs. A second group of simulations was performed under more complex scenarios. In those simulations, we incorporated the combined effect of negative and positive selective effects (using gamma and exponential distributions for the selective coefficients, respectively) plus demographic effects such as expansion and reduction of the effective population size in the domestic simulated populations and with or without migration from the wild into the domestic populations (in total 54 complex simulated scenarios). Figure S1 shows a general scheme for the simulated populations and Tables S3A-B show the parameter values used in these simulations. The obtained results were analyzed using *mstatspop* software (see above).

Approximate Bayesian computation (ABC) analysis

We used the ratio of the estimates of nucleotide variability ($\theta n/\theta s$) per nucleotide for nonsynonymous versus synonymous positions (Fu&Li, Watterson, Tajima and Fay&Wu) and of divergence (*Kn/Ks*) as statistics to infer the distribution of fitness effects (DFE) in coding regions. We compared four evolutionary models that differ in the shape of the DFE using the algorithm proposed by Tataru et al. (2017), which are the following: (i) model A: a model with a deleterious gamma DFE with the mean and the shape of the gamma distribution as model parameters, (ii) model C: a model with a gamma distribution of deleterious variants with two parameters (shape and mean) and an exponential distribution of beneficial variants with one parameter (mean), and the additional parameter of the proportion of beneficial versus deleterious variants, (iii) model DN: a model with a discrete distribution of a priori values of negative selective coefficients with the proportion of negatively selected mutations for each of the negative selective coefficients as parameters. and (iii) model D: a model with a discrete distribution of a priori values of possible selective coefficients (positive and negative) with the proportion of positively and negatively selected mutations for each selected coefficients as parameters. Some of the additional parameters, such as demographic or linkage effects, were considered as nuisances. Nuisance parameters mimic the demographic effects and other parameters such as linkage effect by using the difference between the observations for the neutral dataset (i.e., synonymous sites) and the expected under the neutral model. Others, such as errors in the polarity of unfolded mutations, were fixed.

Table S4 shows the parameters of each model and the prior distributions used in the analysis. We used *polyDFEv2* (Tataru et al 2019) to obtain the expected unfolded site-frequency spectrum (SFS). The code of *polyDFEv2* was slightly modified in order to print the SFS and the parameters for a large number of conditions, which are needed to perform the ABC analysis using summary statistics. For each model, one million iterations were run using different parameter conditions and the resulting SFS for each condition were kept to later calculate the ratios of variability, divergence and the α statistic. The ABC analysis was performed using the R library *abc* (Csillery et al. 2012). We performed a cross validation analysis to evaluate the ability of the approach to distinguish between models using the *cv4postpr()* function, as suggested in the *abc* library documentation. The confusion matrix indicated that these three models were quite distinguishable with a probability of true classification from model A versus C/DN/D of 0.69, from model C versus A/DN/D of 0.65, from model DN versus A/C/D of 0.83 and from model D versus A/C/DN of 0.80, using a tolerance value of 0.05 (Table S5). Posterior probabilities of each model given the observed data (i.e., the probability assigned to each model relative to the other models of the analysis), were obtained using the *postpr()* function and considering a multinomial logistic and a rejection approach. Additionally, a goodness of fit analysis, which evaluates whether the prior distribution for model parameters are realistic, was also performed. The best model was selected based on posterior probabilities. Once the best model was chosen, the ability to infer the parameters of the model was assessed using the *cv4abc()* function. Prediction errors for the parameter inference of each model are shown in Table S6 and Figure S2. The parameters of the best model were inferred with the *abc()* function using a local linear regression and a rejection approach. Posterior predictive simulations were performed using the α statistic to determine whether the simulated data generated from the estimated parameter of our best model resembled the observed data (1000 replicates). Finally, the α values can be simply estimated using equation 10 from Tataru et al. (2017), as the proportion of positive selective coefficients (s) values in the case of the discrete distribution.

Gene context and network topology analysis

We downloaded the complete list of pathways and genes of *S. scrofa* from KEGG v.20170213 (http://www.genome.jp/kegg/, Kanehisa et al. 2008). The list contained 471 pathways and 5,480 genes. The median and mean number of genes per pathway was 26 and 43, respectively, and ranged

from 1 to 949. We filtered the pathways according to their size, removing pathways with less than 10 and more than 150 genes in order to discard pathways that were not informative or too generic and complex. The final list contained 171 pathways and 3,449 genes.

To analyze the selection pressure of each gene according to its position in the pathway, we obtained different topological parameters. For that, we first downloaded the XML file of each pathway from KEGG v.20170213. These files were analyzed with the *iGraph* R package (Csardi G. and Nepusz T. 2006) to obtain the topological descriptors of each gene in each pathway. For each gene, three different measures were computed: *betweenness* (number of shortest paths going through a vertex), *in-degree* (number of in-going edges) and *out-degree* (number of out-going edges). These parameters are measures of the importance of a gene within a pathway: *betweenness* is a centrality feature, *in-degree* suggests the facility of a protein to be regulated and *out-degree* reflects the regulatory role of a protein. We tested whether negatively and positively selected genes differed in any of these statistics using a nonparametric Wilcoxon rank test, due to the extreme leptokurtic distributions involved.

Genomic context patterns

We have additionally tested whether there is a significant correlation between α and recombination, gene density, missing rate, %GC and CpG islands across genomes.

Testing the differences in the estimates of α using whole-genome data versus the mean of gene estimates.

We have studied the behaviour of the α statistic when it is estimated considering a single large dataset (*i.e.,* genome) or when it is estimated using the mean of many subsets (*i.e.,* genes). To do that, we made an R script (check_ratios_vs_meanratios.R) in which we simulated a hypothetical number of polymorphisms and substitutions per gene, following a Poisson distribution (we considered ~10x more substitutions than polymorphisms, 2x more nonsynonymous positions than synonymous and 10x more functional constraint at nonsynonymous versus synonymous). We estimated α per window and per total. The distribution of α per gene can be strongly skewed to

418 negative values when the windows become smaller, thus dragging the mean to negative values as
419 well.

420

421 All the scripts used are available at Zenodo database
422 (https://zenodo.org/record/6124306#.YlcVSy8RqLc).

423

424

RESULTS

Predominance of shared variants and global similar selective effects of mutations in genomic sequences of pig populations

We found a total of 6,684,142 SNPs in autosomes, with 149,440 SNPs located in coding regions. 12.5% of the SNPs in the coding regions are shared among the three populations, 32.2% are shared between at least two populations, 31.2% are exclusive to Large White (LW), 2.2% are exclusive to Iberian (IB) and 34.4% are exclusive to Wild boar (WB) (Table 1 and Table S7). The proportion of exclusive SNPs in each population is in accordance with its specific demographic history (Esteve-Codina et al. 2013, Bosse, Megens, Madsen, et al. 2014). Based on the PCA analysis and using the total number of SNPs, we found that the individuals of each breed cluster together and are well separated from other breeds (Figure S3A). The results from the ADMIXTURE analysis (Figure S3B) suggest that, K=2 is the most likely number of populations, where WB and IB are considered a single population. Under a K=2 scenario, only two LW and one WB individual show a significant percentage of admixture among groups. Nevertheless, for larger values of K, new groups emerge, being the IB one of these separated groups (Figure S3B). However, additional subgroups within WB and LW phenotypes appear and disappear when increasing the K value, making those subgroups apparently unreliable. Therefore, we decided to analyze separately the three breeds, LW, IB and WB, following the main patterns of population structure (Figure S3A) and the phenotypic features of the animals, which essentially separate domestic (commercial and local breed, separately) from wild animals.

446

447 **Table 1.** Number of synonymous and nonsynonymous SNPs according to its allelic status in each pig population. A:
448 Ancestral allele, F: Fixed allele, P: Polymorphic allele. IB: Iberian; LW: Large White; WB: Wild boar. SNPs that are
449 missing in any of the populations are not considered.

IB	LW	WB	Synonymous	Non-synonymous
F	F	F	20297	9342
P	P	P	11712	7597
A	A	F	0	0
A	F	A	0	0
F	A	A	3	5
A	A	P	30314	20988
A	P	A	26027	15035
P	A	A	1833	1588
A	F	F	0	0
F	A	F	1	0
F	F	A	1	0
A	P	P	10128	7930
P	A	P	1676	1254
P	P	A	700	363
A	F	P	11	1
A	P	F	0	2
F	A	P	30	30
P	A	F	0	0
F	P	A	8	4
P	F	A	1	1
F	P	P	4924	2378
P	F	P	242	139
P	P	F	81	52
F	F	P	1140	489
F	P	F	4911	2073
P	F	F	38	22
			114078	69293

450

451

452 For each breed, coding positions were classified as polymorphic, fixed (i.e., different allele from
453 the outgroup) or ancestral allele (i.e., same allele as in the outgroup), with the aim of identifying
454 those variants that appeared previously or posteriorly to the domestication process (Table 1).

Surprisingly, we found very few fixed mutations between populations, indicating that the phenotypic traits of each population are not associated with fixed coding variants. Similarly, we found very few fixed coding variants in domestic (IB or LW) versus wild (WB). There are few variants fixed in the domestic breeds that are polymorphic in the wild population, suggesting that these variants were previously present in wild breeds or, alternatively, were introgressed into WB from domestic breeds. Most of the variants that are exclusive of a single breed are polymorphic, which is in agreement with the recent origin of these variants. We found a large number of fixed variants in the IB that are polymorphic in LW and WB, likely due to a reduction of the effective population size of the IB breed. The ratio of nonsynonymous to synonymous polymorphism was always lower than one and showed similar values for the three populations regardless of the variability estimator used (Table S9). This result suggests that, on average, there are no differential effects of selection between domestic and wild populations, although this might not be the case when individual genes are considered.

Low codon bias at whole-genome scale

We estimated the level of codon bias at genome scale using MCU and N_{cw} statistics to control for the possible effect of selection on synonymous positions. Non-neutral synonymous mutations can have a large impact on the inference of the proportion of beneficial selection, and on the estimation of the Distribution of Fitness Effects (DFE). Indeed, the effect of bias in codon usage causes an overestimation of the beneficial proportion of variants that become fixed by increasing the ratio of synonymous polymorphisms versus synonymous fixations (Akashi, 1995, Matsumoto et al. 2016). For this species, we observed a low and large values of MCU and N_{cw}, respectively, indicating low levels of codon bias at genome scale (mean MCU=0.485, Figure S5). However, it should be mentioned that positive selection could be acting on synonymous positions of some specific genes. We therefore have assessed whether there was a correlation between MCU and a, considering all coding regions or only coding regions showing positive a values. We observed no correlation between MCU and a values when considering only genes with positive a values (Figure S5) and slightly negative correlation when considering all genes regarding their respective α values (data not shown).

485 **Limited influence of genomic context and the network topology on selective patterns**

486 The heterogeneity in the recombination rate, the gene density, the %GC and the distribution of
487 CpG islands across the genome can affect the local levels of variability. A previous study on the
488 IB breed detected a strong correlation between recombination and variability, although no
489 correlation was observed between variability and gene density or GC content (Esteve-Codina et
490 al. 2013). However, the effect of these factors on the estimation of the proportion of adaptive
491 nonsynonymous mutations (α) has not been previously studied. When we assess whether there is
492 a correlation between the estimated α and the above-mentioned factors, we found that there is no
493 correlation between the estimated α and recombination, gene density, %GC and CpG in any of the
494 three breeds (P-values > 0.01).

495

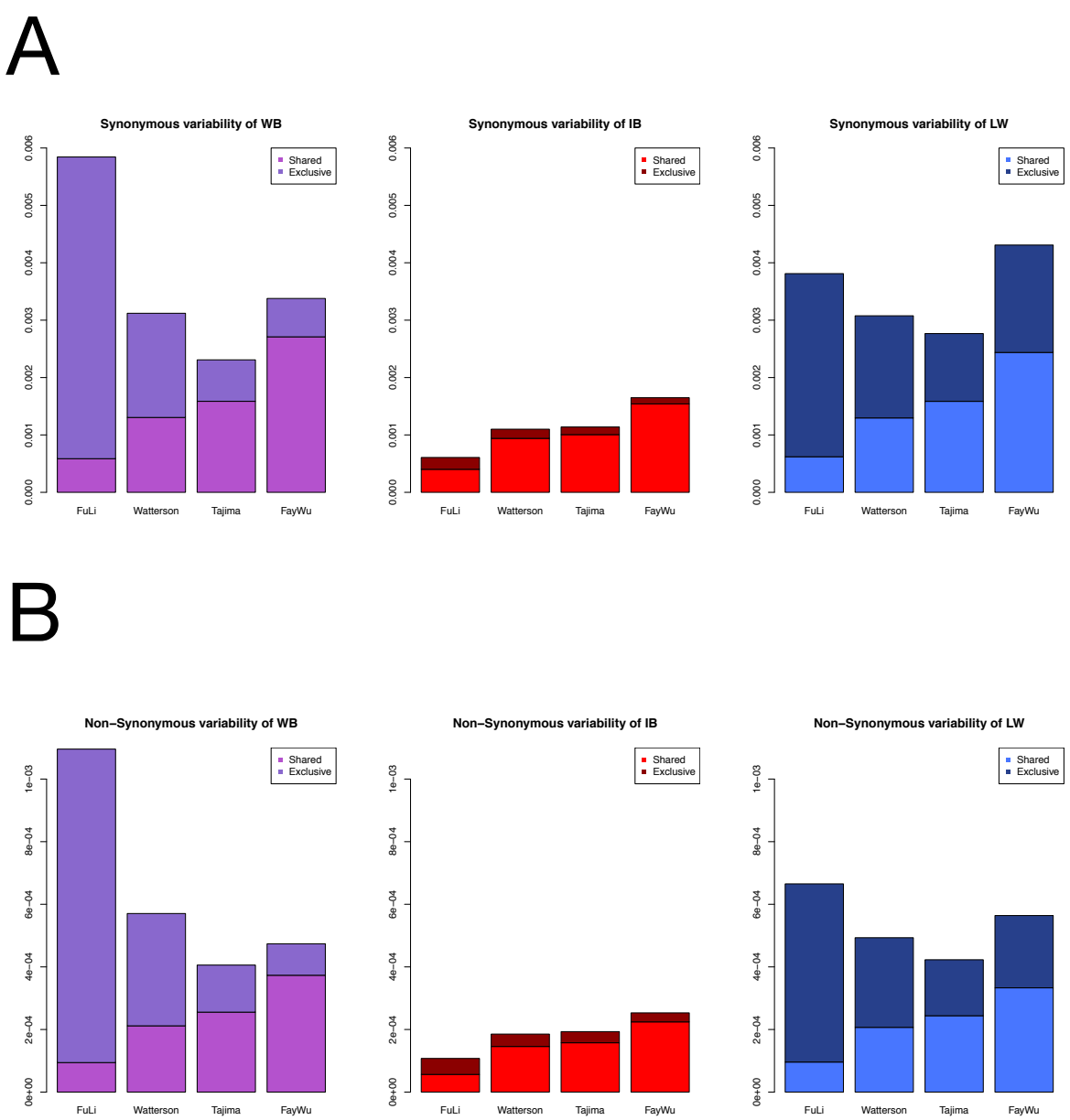
496 Next, we investigated the effect of gene network topology on the selective patterns. It has been
497 claimed that topology limits the 'evolvability' of genes and that highly connected genes are more
498 constrained and, consequently, less likely to be targets of positive selection. We compared the
499 network topology features (*betweenness*, *out-degree* and *in-degree*) of genes within pathways
500 regarding the estimates of α, grouping genes with positive versus negative α values. We found that
501 genes with negative α values show significant large values of the *betweenness* statistic in the three
502 pig breeds compared to genes with positive α values (P-value < 0.01; Figure S4). LW and WB
503 showed significant values (P-values < 0.01) of the *in-degree* statistic for genes with negative α
504 values compared to genes with positive a values. However, we did not observe significant
505 differences in the *out-degree* values between genes with negative and positive a values in any of
506 the three breeds (Figure S4). These results suggest that, in the three breeds, genes that are more
507 central in a pathway are more evolutionary constrained compared to peripheral genes. In addition,
508 in LW and WB, the genes that are more constrained tended to have a higher number of upstream
509 genes that regulated them, which is also in agreement with the central position of these genes in
510 the pathway. We did not observe significant differences in *in-degree* statistic in the IB breed
511 between genes with negative and positive a values, likely because of a relaxation of functional
512 constraints as a consequence of the reduction of its effective population size.

513

Levels of nucleotide variation at protein coding regions are compatible with the history of the surveyed pig populations and with the presence of positive selection

To assess the selective effect of domestication, we first studied the pattern of variation at synonymous and nonsynonymous positions using four estimators of variability that differentially weight the SNP frequencies (See Material and Methods). Ideally, it would be more informative to analyze the whole genome Site Frequency Spectrum (SFS). Unfortunately, the relatively high number of positions with missing data discourages their use. A possible alternative would be to obtain the SFS from a reduced number of samples, and therefore use only a partial number of SNPs for subsequent analysis. However, by using this alternative we can either lose power or introduce some sort of bias, hence, we preferred to analyze the whole set of SNPs using those estimates of variability based on different frequencies of the spectrum that account for missing data. Nevertheless, in order to clarify the patterns of the SFS for these populations, we estimated the SFS for a subset of SNPs (around 25-30% of the available coding variants, depending on the breed) for a projection of variants on 38 haploid samples in both LW, WB and on 10 haploid samples in IB (Figure S6). The SFS profile for both synonymous and nonsynonymous showed a rapid decrease in the number of variants from lower to higher frequencies. We observed a slight increase in the number of polymorphisms at the highest frequencies at both synonymous and nonsynonymous sites and no apparent signals of admixture (i.e., no sudden peaks at specific ranges of frequencies). Estimates of whole-genome variability levels per nucleotide using different estimators are shown in Figure 1 and detailed in Table S8. We have considered the synonymous positions as neutral reference since no strong bias in codon usage has been detected (Figure S5). We expect that, under the Standard Neutral Model (SNM), the values for the different estimates of variability should be similar whereas differences among them may indicate demographic and/or selective effects. We observed that i) the levels of variability are different for each estimator within breeds and ii) the levels of variability are different for the same estimator for different breeds. However, for each breed, we observed a similar ratio of nonsynonymous to synonymous polymorphisms regardless of the used estimator, suggesting that demographic effects are responsible for the differences in the levels of variability (Figure 1). The less variable population is the IB breed, which shows far fewer singletons compared to WB and LW, probably as a consequence of the known reduction of its population size. Note than in all the three populations, high-frequency variants are proportionally more abundant than those at intermediate frequencies,

545 which would be compatible with the accepted demographic history of the surveyed populations
546 (i.e., introgression in LW, bottleneck in IB and some population reduction and introgression in
547 WB) but also with the presence of pervasive positive selection in all three populations.



548

549 **Figure 1.** Estimates of the levels of variation at synonymous (A) and nonsynonymous (B) sites for each variability
550 estimators and pig population and where variants were classified as shared and exclusive variants. WB; Wild boar;
551 IB, Iberian; LW, Large White.

552

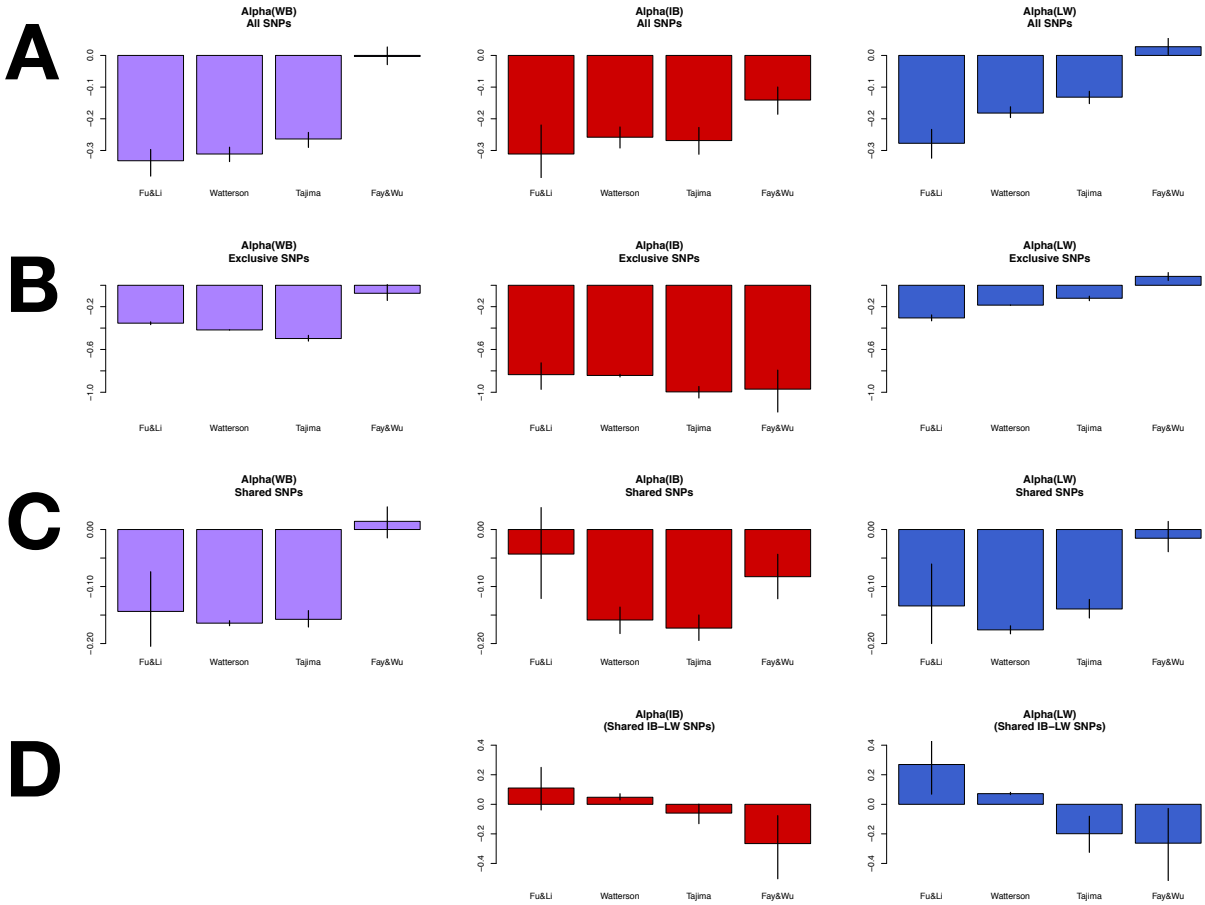
α's values and $R\_\beta\gamma$ ratios based on all SNPs might reflect a differential effect of selection due to domestication

The differential effect of selection in the domestic and wild populations can be studied by comparing their respective α values. Figure 2A and Table S9 show the genome-wide α values calculated using the four variability estimators for each population. As expected, the α values are negative when α is calculated using the estimate of variability based on low-frequency variants ($\alpha_{Fu\&Li}$), probably reflecting the relatively high proportion of deleterious versus neutral mutations that are segregating at low frequencies. We observed a similar value of $\alpha_{Fu\&Li}$ in all populations, suggesting a similar proportion of segregating deleterious mutations, irrespective of the domestication process or other demographic events (Figure 2A). Moreover, we observed milder negative values of α, or even positive for LW when α is calculated based on variants at high frequencies (Figure 2A), according to expectations, which point to a progressive elimination of deleterious mutations as we move towards higher frequencies. Nevertheless, the pattern of α (i.e., the comparative a value calculated using the four different variability estimators within each population) is very different in each population. WB and LW show positive or null α values when it is calculated based on high frequencies (Table S9). Instead, IB show very low negative α values for all estimators of variability. We found a compatible pattern when using the reduced subset of SNPs for the SFS estimation (Figure S6-A), where it can be observed that the estimates of a in all three populations are very similar among them ($a \sim -0.05$), although their confidence intervals are quite wide.

The differences in the ratio of synonymous to nonsynonymous variability between the two different breeds is summarized by the $R\_\beta\gamma$ ratio (Figure 3). We observed that the largest deviations from $R\_\beta\gamma = 1$ are observed when the ratio was calculated based on high-frequency variants ($\alpha_{Fay\&Wu}$). Although the ratio of the two populations is difficult to interpret because of their different underlying demographic histories, some trends can be observed. WB shows an excess of nonsynonymous variants segregating at intermediate frequencies (WB-IB, WB-LW), which might be explained by a past bottleneck that increased deleterious mutations at intermediate frequencies. In addition, the $R\_\beta\gamma$ ratio in IB-LW shows an incremental pattern of this ratio from low to high

582 frequencies, which is compatible with an increase of nonsynonymous beneficial variants on their
583 way to fixation in LW.

584



585

586 **Figure 2.** Estimates of α for each pig population based on different variability estimators. Total variants (A), exclusive
587 variants (B), shared variants (C) and shared variants between IB and LW (D). Bootstrap intervals at 95% are indicated
588 by a line at each bar. WB; Wild boar; IB, Iberian; LW, Large White.

589

590 **α's and $R\_\beta\gamma$ ratios based on exclusive and shared polymorphisms might reflect changes in**
591 **selective patterns before and after domestication**

592 We observed a high ratio of nonsynonymous to synonymous singletons ($\alpha_{Fu\&Li}$, Figure 2B) when
593 the analysis was performed based on exclusive polymorphisms, suggesting that they have
594 deleterious effects in all populations. Nevertheless, the values of α calculated based on

595 intermediate frequency variants (α_{Tajima}) in the WB and IB populations are lower than to those
596 based on low-frequency variants, which point to a change in the selective pressure, maintaining
597 nonsynonymous variants at relatively high frequencies. Nevertheless, the $\alpha_{Fay\&Wu}$ values (-0.075,
598 -0.971 and 0.083 for WB, IB and LW, respectively, Table S9) show a similar trend in relation to
599 that based of Total SNPs, that is, close to 0 or positive for WB and LW, but strongly negative for
600 IB. Concordant estimates are observed in the analysis of the SFS based on a reduced number of
601 SNPs (0.155, -0.913 and 0.277 for WB, IB and LW, respectively, Figure 6B), with the difference
602 that a clear positive and not 0 α values is observed in WB. The R_βγ statistic shows the same
603 pattern as that calculated using all SNPs but with all over one (Figure 3). That indicates that WB
604 has a higher proportion of nonsynonymous polymorphisms compared to IB, in contrast to what is
605 observed when the analysis is performed based on all SNPs. This would suggest a recent change
606 in the constraint of nonsynonymous positions likely at IB breed, as this ratio in IB-LW is also
607 affected. This is also in agreement with the low α value in IB breed at exclusive variants regarding
608 to Total SNPs.

609

610 On the other hand, the α values based on shared variants are in general more moderate (closer to
611 zero) than those based on exclusive variants (Figure 2C), likely because shared nonsynonymous
612 polymorphisms are older and hence, expected to be more functionally constrained than the
613 exclusive ones. Additionally, the values of α based on singletons ($\alpha_{Fu\&Li}$) are less negative than
614 those based on intermediate-frequency variants. The α estimates based on shared variants in the
615 analysis of the reduced subset of SNPs are very similar to Total SNPs and very close to zero (Figure
616 6C). The $R\_\beta\gamma$ statistic for shared variants shows similar patterns than those observed for all
617 variants but with values much closer to 1, indicating a small or moderate selective effect on the
618 shared variants compared to all variants (Figure 3).

619

620 When we calculated the α values from shared variants only between the two domestic breeds, we
621 found an inverse pattern regarding to that calculated from all SNPs in each population, with high
622 positive values of α based on low frequencies and very negative values when α is calculated based
623 on high-frequency variants (Figure 2D). This could be due to i) the active elimination of new
624 nonsynonymous variants to preserve differences among domestic breeds ($\alpha_{Fu\&Li}$) and ii) the

presence of nonsynonymous variants targeted by the process of domestication that shifts them toward high frequencies ($\alpha_{Fay\&wu}$). Nevertheless, we cannot discard that this excess of nonsynonymous variants at high frequencies and the lack of nonsynonymous singletons at low frequency could be due to a more complex and not previously explored demographic scenario.

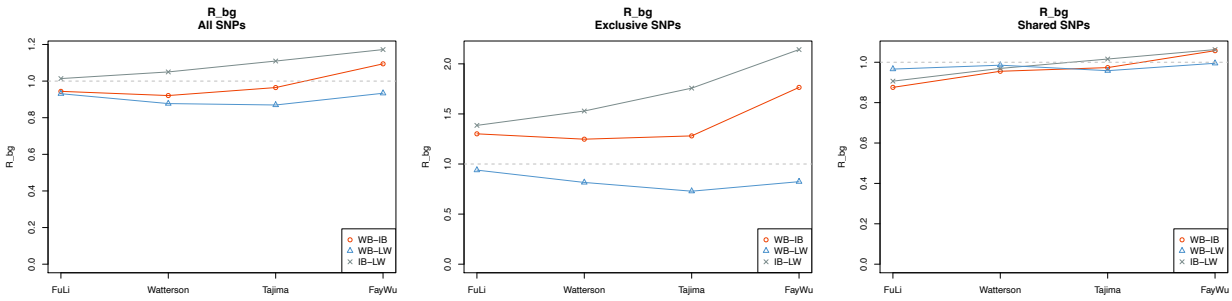


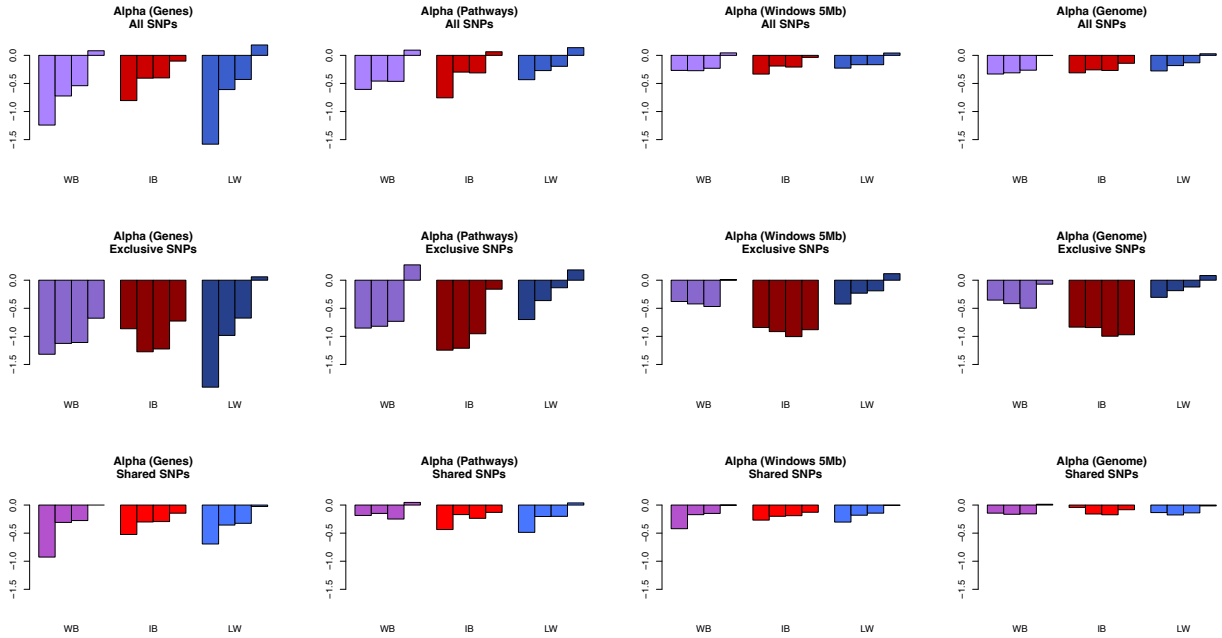
Figure 3. Estimates of $R\_\beta\gamma$ for all (left), exclusive (centre) and shared (right) variants. WB; Wild boar; IB, Iberian; LW, Large White.

Values of α are dependent of the molecular scale but the patterns of the estimated α's are similar across the different molecular scales

In addition to the genome-wide analysis, α was calculated using three additional molecular scale levels: i) gene level, ii) genes within windows of 5 Mb, and iii) genes within the same pathway. Figure 4 shows the median of the distributions of the α values for each scale level. When the analysis was performed based on all SNPs, the pattern of α values estimated at the genome-wide level are concordant with those estimated at the gene level, genes within windows and genes within pathways for each breed. However, differences in the value of α within each breed are notorious depending on the scale level examined. The median estimates of α are generally lower at the gene scale level and most of them are very negative, while at the genome-wide scale, the α values are closer to zero. However, the distribution of α values can have a large variance at the gene scale since few variants are used for its estimation. We identified the regions and pathways that showed extreme α values (Table S10 and S11). We found a large number of genes showing $\alpha = 1$ (highest value) because the number of polymorphic nonsynonymous variants per gene was zero. We also found a moderately high correlation of α values between breeds (rho ~ 0.7, Pearson correlation

649 when considering pathways) suggesting that in general, these breeds are under similar selective
650 effects. When considering shared and exclusive variants, we generally observed the same pattern,
651 from genes to whole-genome, that is, larger α values at the gene level and closer to zero α values
652 at the larger scale. The differences in α values could be explained because of the distribution of
653 this ratio statistic (*i.e.*, skewed distribution to negative values) and the uneven distribution of the
654 functional variants, in which the mean can be displaced to more negative values (see Materials and
655 Methods).

656



657

658 **Figure 4.** Estimates of the median values of α based on different variability estimators and for each pig population at
659 different molecular scales and for all, exclusive and shared variants. Within each population, the order of different α's
660 is: Fu&Li, Watterson, Tajima and Fay&Wu. WB; wild boar; IB, Iberian; LW, Large White.

661

662 **Simulated data under different scenarios that include the joint effect of demography and**
663 **selective events were more concordant to the observed data**

664 We used computer simulations to study how the different demographic and selective events
665 occurred during domestication process shaped nucleotide variation present in these populations.
666 We simulated populations mimicking the process of domestication using *SLiM* software (Haller
667 and Messer 2017), coupled with several demographic events, including changes of the population

size and/or migration. We analyzed the genome-wide patterns of α and the $R\_\beta\gamma$ statistic produced by 63 simulated scenarios that included different demographic events and selective forces acting separately (simple scenarios) or jointly (complex scenarios). The results of the simulation study are summarized in Figures S7-S48. The observed patterns of α based on all variants in the surveyed populations are not compatible with simple scenarios that only consider demographic or positive selection forces (Figure S7). Rather, α patterns from simulated data (irrespective of the magnitude of α) fit a scenario with a predominant effect of negative selection (Figure S7). However, the $R\_\beta\gamma$ statistic do not fit any of the simulated simple scenarios (Figure S8). When more complex scenarios were considered (i.e., including a bottleneck, positive/negative selection and/or migration, Figures S9-S14), the general α patterns generated by those scenarios that include both negative and positive selection resembled those observed in WB and LW (with negative α's at low frequency values to slightly positive α values at high frequency). The scenarios that also include some migration events are the ones that showed more concordance for these two breeds (Figures S12, S14). On the other hand, the IB population is more compatible with a scenario without positive selection and with a recent population size reduction (Figure S13). The trends in the $R\_\beta\gamma$ statistic are, in broad strokes, concordant with the conclusions extracted from the comparison between the observed and simulated patterns of α (Figures S15-S20).

The observed patterns of α values based on exclusive variants are similar to those based on total variants but only for the LW population (Figure S21). These patterns cannot be fully explained by any of the complex simulated scenarios that are concordant when considering all variants, although surprisingly, they would be more compatible with those including a population size reduction (Figures S24-S28). The observed $R\_\beta\gamma$'s are compatible with the scenarios that combine both types of selection and a population size reduction (WB-IB) or with scenarios that include migration (WB-LW; Figures S29-S34). Finally, the observed patterns of α and $R\_\beta\gamma$ statistics calculated from shared variants are also compatible with scenarios which includes both types of selection (Figures S35-S48), and being quite compatible with those including expansion demographic events. Overall, the simulations data showed that complex scenarios, including demography, migration, and positive and negative selection, may be necessary to explain the observed data.

Table 2. Posterior Probabilities for each ABC model (multinomial logistic method with tolerance 0.01) and for each pig population for Total variants, Exclusive variants. and Shared variants.

Posterior probabilities for model comparison (multinomial logistic method)

TOTAL	modelA	modelC	modelDN	modelD
WB	0.000	0.039	0.623	0.338
IB	0.000	0.031	0.968	0.001
LW	0.000	0.195	0.526	0.279
EXCLUSIVE	**modelA**	**modelC**	**modelDN**	**modelD**
WB	0.007	0.011	0.512	0.470
IB	0.004	0.001	0.995	0.000
LW	0.001	0.111	0.113	0.775
SHARED	**modelA**	**modelC**	**modelDN**	**modelD**
WB	0.001	0.334	0.405	0.260
IB	0.000	0.413	0.576	0.011
LW	0.001	0.447	0.462	0.090

Models that assume a discrete distribution of beneficial and deleterious mutations would fit better the observed data

We used an approximate Bayesian computation (ABC) analysis to infer the DFE separately for each population using the ratios of nonsynonymous to synonymous variants (i.e., polymorphism and divergence) obtained from the whole-genome analysis (see Materials and Methods). Four different models implemented in *polyDFE2* software (Tataru et al. 2019) were tested. These models overcome the inference of the demographic parameters (and others such as linkage effects) by the inclusion of nuisance parameters (see Material and Methods). The four models were: model A, which assumes a gamma distribution of deleterious mutations; model C, which assumes a gamma distribution for deleterious mutations and an exponential distribution for beneficial mutations; model DN, that assumes a discrete distribution of only deleterious and neutral mutations, and model D, that assumes a discrete distribution of deleterious, neutral and beneficial, mutations. Goodness of fit (GoF) analysis revealed that the simulated data under the different models fits differentially to the observed data, although the used range of parameters for priors are compatible with the observed data for all the four models (Table S12). Posterior probabilities showed that the DN is the most likely model for all three populations when using total number of

718 variants (Table 2). The posterior probability for this model is especially high for the IB breed
719 (0.97). Nevertheless, note that the model D is just below the DN model by less than half of the
720 probability in the case of WB and LW. Finally, the posterior predictive analysis indicated that the
721 observed α values for the three populations are within the range determined by the minimum and
722 maximum simulated α values (i.e., Q1-1.5*IQR, Q3+1.5*IQR, respectively, being IQR the
723 Interquantile Range Q3-Q1) under both models DN and D, although not always inside the Q1 and
724 Q3 quantiles (Figure 5). The mean parameters of the DFE inferred for each population are shown
725 in Table 3 (see also Table S13). The obtained results indicated that the DFE is quite similar among
726 all three populations, which is not entirely surprising because they share a long-term history.
727 According to model DN, and despite there is a lot of uncertainty in the inferred estimates (Table
728 S13), the obtained results show that the DFE contains a large fraction of very deleterious variants,
729 with approximately 75% of the variants being strongly deleterious (S =−2000), and with
730 approximately 12% of the variants being neutral or slightly deleterious (approx. 4%). The model
731 D infers a higher proportion of weak deleterious mutations compared to the neutral ones, although
732 the sum of both is similar to model DN. Finally, the inferred contribution of positive selection is
733 relatively low (around 0.7-1.1%).

734

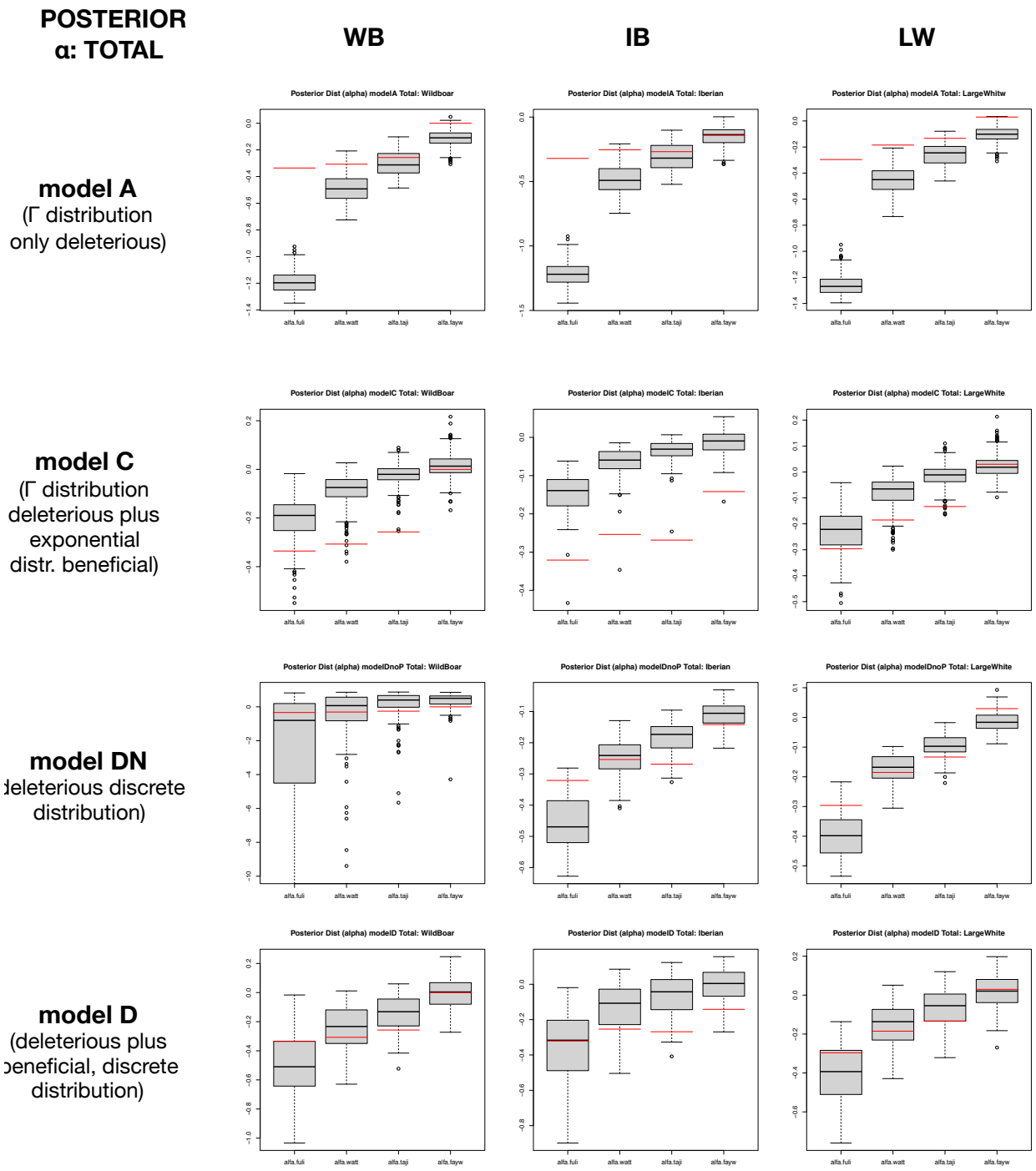


Figure 5. Posterior distributions of the α values for total variants based on different variability estimators (Fu&Li, Watterson, Tajima and Fay&Wu). Box plots indicate simulated distributions of α values. Red lines indicate observed α values.

The differential patterns of DFE based on exclusive versus shared variants may indicate selective differences after the split of the populations

We are aware that the inference of the SFS is going to be highly distorted by choosing only a subsection of the variants (e.g., exclusive variants are mostly very recent and have no time to reach high frequencies and shared variants show no or few singletons per populations). However, the nuisance parameters incorporated in polyDFE should account for this effect (Tataru et al. 2017). The classification of polymorphisms in exclusive or shared are dependent of the relationship between two populations, and are a priori not related to the selective effect of these polymorphisms across their frequencies, although shared (mostly older) and exclusive (mostly recent) variants are chronologically related to the selection of variants. Then, we considered that the inference of DFE from exclusive and shared polymorphisms can give some clues about recent and past events related to the domestication processes. The models with higher posterior probabilities in the case of WB and IB breeds based on exclusive variants are the same than those for total variants (Table 2). However, for LW and based on exclusive variants, the model D has higher posterior probabilities, in contrast to what was obtained based on total variants (Table 2). The posterior predictive simulations showed that the models DN and D yielded similar estimated *a* values to those from the observed data, with the IB breed exhibiting the posterior distributions of *a* values more distant to the observed data (Figure 6). The estimates of the parameters of the models indicate that, for all populations, the exclusive segregating variants exhibit less strongly deleterious effects compared to those based on total and shared variants (Table 3). Indeed, the IB breed shows significantly lower proportions of strong deleterious mutations, according to its assumed recent population decline. As in the analysis based on total variants, posterior predictive analysis based on exclusive variants showed that models DN and D are those generating *a* values more similar to the observed ones, but in this case, the observed *a* values for the IB breed were slightly closer to the simulated *a*'s, compared to those based on total variants (Figure 6).The results obtained based on shared variants also show that the DN is the most likely model for all populations, although the model C shows closer probabilities (Table 2), especially in the case of LW, which might suggest that shared variants may have played a significant role as a substrate for adaptive process. However, posterior predictive distribution of *a* values for this breed under this model resembled less the observed data compared to those for the most likely models (Figure 7).

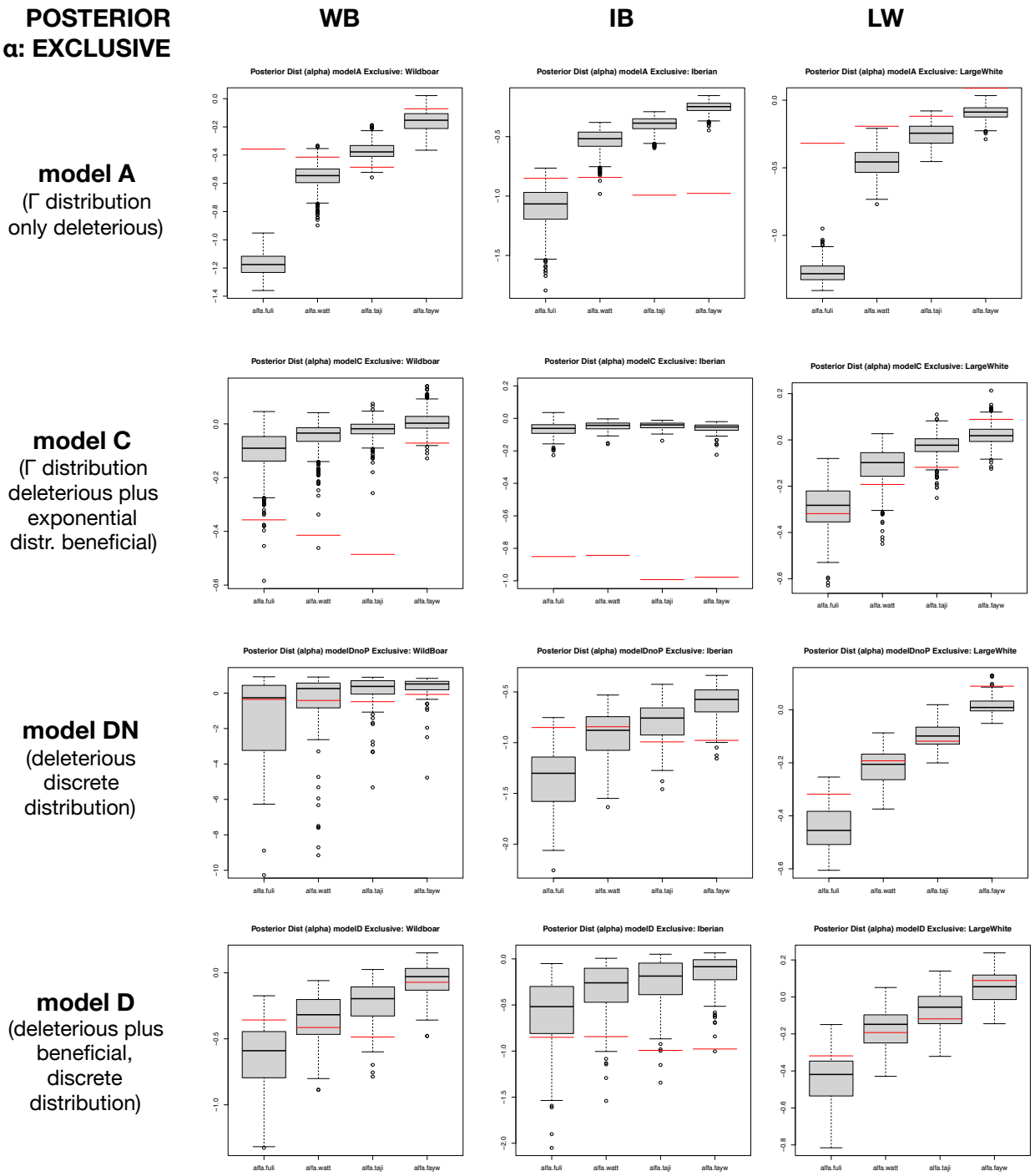


Figure 6. Posterior distribution of the α values for exclusive variants based on different variability estimators (Fu&Li, Watterson, Tajima and Fay&Wu). Box plots indicate the simulated distributions of α values. Red lines indicate observed α values.

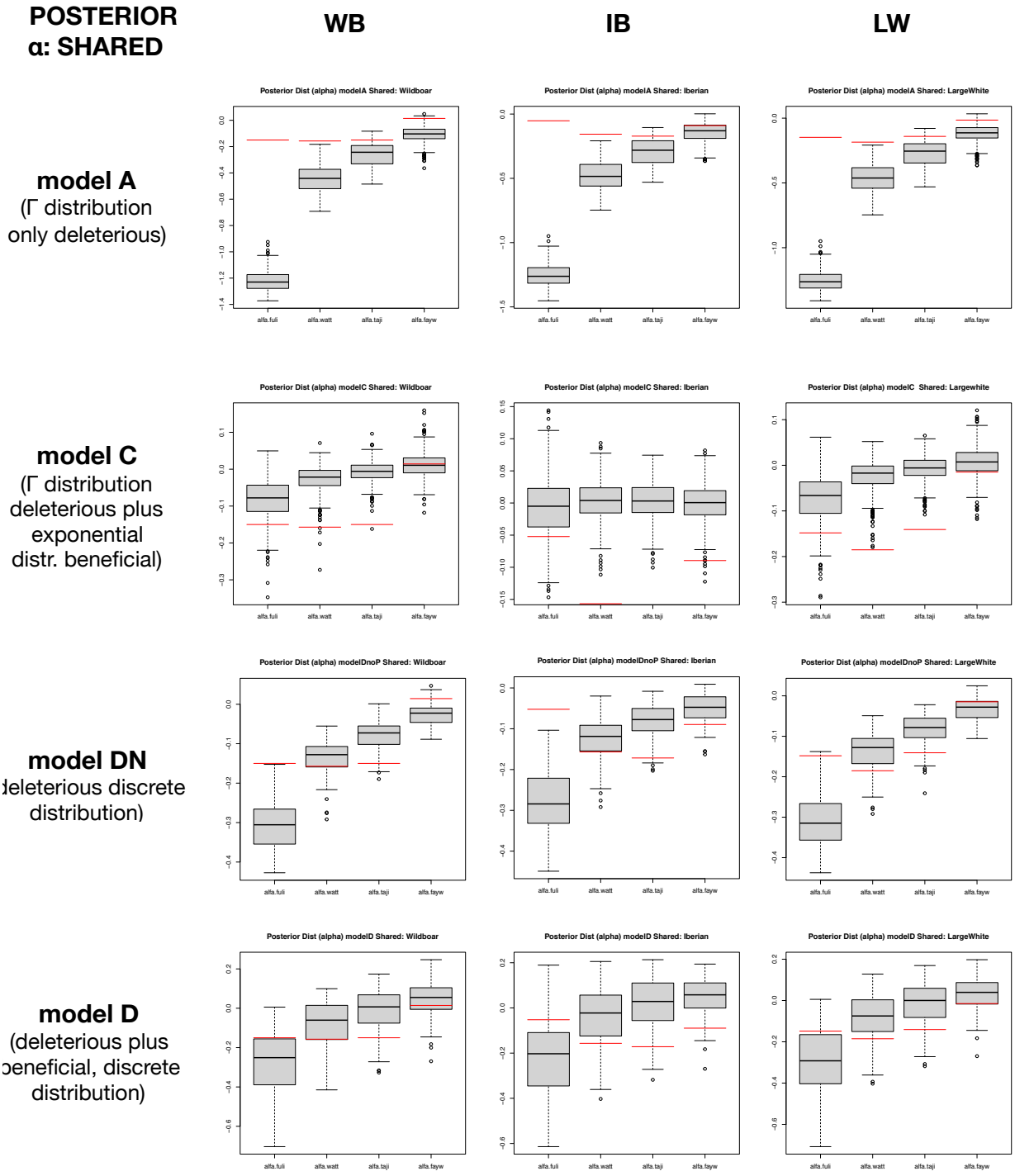


Figure 7. Posterior distribution of the α values for shared variants based on different variability estimators (Fu&Li, Watterson, Tajima and Fay&Wu). Box plots indicate the simulated distributions of α values. Red lines indicate observed α values. .

Table 3. Inferred selective parameters (weighted mean) for each ABC model and pig population. (A) Total variants. (B) Exclusive polymorphisms. (C) Shared polymorphisms. Note that for model DN and D, the proportion of each discrete S value is relative to the total S values, considering negative inferred values as zero.

Table 3. Mean Inference of parameters for each of the four analyzed models

Mean	model A		model C				model DN					model D						
TOTAL	Sd	b	Sd	b	pb	Sb	p1 (S=-2000)	p2 (S=-200)	p3 (S=-20)	p4 (S=-2)	p5 (S=0)	p1 (S=-2000)	p2 (S=-200)	p3 (S=-20)	p4 (S=-2)	p5 (S=0)	p6 (S=2)	p7 (S=20)
WB	24634.19	0.179	4724.54	2.514	0.145	0.009	0.736	0.070	0.024	0.045	0.125	0.741	0.057	0.023	0.098	0.074	0.006	0.001
IB	27016.47	0.180	6625.76	3.064	0.144	0.011	0.767	0.049	0.016	0.045	0.123	0.757	0.057	0.018	0.061	0.098	0.008	0.001
LW	20351.15	0.189	6563.28	1.940	0.137	0.017	0.772	0.058	0.019	0.027	0.125	0.771	0.050	0.019	0.076	0.073	0.010	0.001

Mean	model A		model C				model DN					model D						
EXCLUSIVE	Sd	b	Sd	b	pb	Sb	p1 (S=-2000)	p2 (S=-200)	p3 (S=-20)	p4 (S=-2)	p5 (S=0)	p1 (S=-2000)	p2 (S=-200)	p3 (S=-20)	p4 (S=-2)	p5 (S=0)	p6 (S=2)	p7 (S=20)
WB	33447.89	0.170	3694.78	0.821	0.153	0.002	0.723	0.055	0.018	0.091	0.113	0.709	0.064	0.020	0.135	0.065	0.006	0.001
IB	1834.41	0.216	253.88	44.58	0.151	0.038	0.584	0.091	0.027	0.223	0.074	0.599	0.101	0.028	0.142	0.119	0.009	0.001
LW	22064.47	0.187	8311.62	1.046	0.133	0.027	0.751	0.077	0.025	0.025	0.123	0.762	0.058	0.020	0.091	0.058	0.010	0.001

Mean	model A		model C				model DN					model D						
SHARED	Sd	b	Sd	b	pb	Sb	p1 (S=-2000)	p2 (S=-200)	p3 (S=-20)	p4 (S=-2)	p5 (S=0)	p1 (S=-2000)	p2 (S=-200)	p3 (S=-20)	p4 (S=-2)	p5 (S=0)	p6 (S=2)	p7 (S=20)
WB	25128.11	0.182	5223.06	5.852	0.146	0.016	0.790	0.042	0.013	0.021	0.133	0.773	0.052	0.017	0.061	0.085	0.012	0.001
IB	29540.82	0.182	9042.17	11.758	0.147	0.005	0.804	0.035	0.010	0.020	0.131	0.783	0.051	0.016	0.045	0.089	0.014	0.001
LW	24929.43	0.185	5169.65	5.573	0.137	0.020	0.798	0.038	0.014	0.022	0.129	0.777	0.052	0.017	0.059	0.084	0.011	0.001

Sd: 4Ns mean value for mutations with negative effects. b: shape of the gamma distribution for mutations with negative effect. pb: proportion of beneficial mutations. Sb: 4Ns mean value for mutations with positive effects. p1 (S=-2000): proportion of functional variants having 4Ns=-2000, p2 (S=-200): proportion of functional variants having 4Ns=-200, p3 (S=-20): proportion of functional variants having 4Ns=-20, p4 (S=-2): proportion of functional variants having 4Ns=-2, p5 (S=0): proportion of functional variants having 4Ns=0 (neutral), p6 (S=+2): proportion of functional variants having 4Ns=+2, p7 (S=+20): proportion of functional variants having 4Ns=+20. Nuisance parameters are not shown.

DISCUSSION

The study of the genetic effects produced by domestication can be challenged for many reasons. First, the current domesticated species have been severely manipulated by humans, which means that most of individuals were not crossed randomly, and consequently, different complex domestication scenarios can be found such as a high degree of structuration, a fast creation of new lineages from highly inbreeding crosses, a forced introgression between far related populations or from close species and a very divergent selective events across time and space, among others (e.g., Mignon-Grasteau et al. 2005, Ross-Ibarra et al. 2007, Ramos-Onsins et al. 2014, Gaut et al. 2018). Moreover, in animals, the polygenic nature of the domestication traits often precludes identifying their underlying genes since the domestic phenotypes might be probably caused by subtle allele-frequency changes of variants distributed throughout the genome, and hence, very difficult to be detect. In addition, the study of the effects of selection using genome sequences that contain a nonnegligible fraction of missing data, such as those from non-model organisms, is challenging and needs the use of appropriate methods to account for these positions. Statistics that exploit the frequency of the variants while accounting for missing data are particularly appropriate for such analyses (Ferretti, Raineri, and Ramos-Onsins 2012). Despite all this inconvenience, domestic populations are an excellent model to study the effect of strong and recent selection (e.g., Doebley, Gaut and Smith 2006, Groenen 2016).

One of the main goals of this work is to provide a novel approach that combines the use of different estimators of variability that account for missing data with the asymptotic approach proposed by Messer and Petrov (2013) and Uricchio et al. (2019) in order to take into account some of these issues. This approach is designed to be used as an alternative in case of the estimation of the full SFS is compromised by large amounts of missing data. Although it can be less precise (we used only four statistics to capture the entire trend of α's across the SFS), it allows analyzing a larger number of positions, helps to reduce the variance (by summarizing the SFS into few statistics) and facilitates the visual interpretation. To illustrate the utility of the proposed approach, we used this methodology to perform an exhaustive comparative study of the observed patterns of functional versus neutral diversity and divergence in domestic pigs and wild boars. In addition, and to delve deeper into the domestication process of this species, we also performed a forward simulation

819 study, including several diverse evolutionary scenarios and the inference of DFE parameters given
820 different selective models. Note that the DFE was inferred using Bayesian calculations (ABC)
821 instead of exact Bayesian or Likelihood methods since despite ABC requires additional steps and
822 validation analysis and is in general less precise, it allows contrasting models and inferring
823 parameters from complex datasets or data containing missing information (Beaumont et al. 2002).
824 Finally, we have also analysed exclusive and shared polymorphisms separately to extract
825 information about the new and the past domestication events of the demographic history of these
826 populations, but also about the role of population admixture.

827

828 **General selection pressures on pigs and the process of domestication**

829 Like us, others have been already performed several analyses to shed light on the process of
830 domestication using the MacDonald and Kreitman extension methods or using other estimates
831 such as variability or divergence at functional or synonymous positions (MacEachern et al. 2009,
832 Kono et al. 2016, Makino et al. 2018). As in Makino et al. (2018), we do not observe an increase
833 in functional diversity in domestic versus wild populations. This may be explained by several
834 recent events occurring in these populations: (i) differences in the recent history our local and
835 commercial domesticated populations (i.e., high inbreeding degree in Iberian local pigs and recent
836 gene flow from Asian pigs into the commercial pigs); (ii) demographic effects in the wild boar
837 population that may have reduced their diversity (Groenen et al 2016) or have increased the
838 confidence intervals of the patterns of α's (Figures S5-S46); (iii) differential adaptive forces in
839 local (IB) versus commercial pigs (LW), with a recent high selective pressure in this last
840 population.

841 Accordingly, with their recent history, the IB breed shows the lowest levels of synonymous and
842 nonsynonymous variation among the breeds studied. Note that the two domestic breeds analyzed
843 here have very different recent histories: the IB is a local Spanish breed (Guadyerbas) that suffered
844 a strong bottleneck during the 1970s (Esteve-Codina et al. 2013) and with no evidence of
845 introgression whereas the LW breed was admixed with pigs of Asian origin (Bosse, Megens,
846 Madsen, et al. 2014). Therefore, our observations are perfectly compatible with the small effective
847 population size and the close relatedness of the individuals expected for this population. However,
848 for the other two breeds, the obtained results do not seem to conform to what was expected. since

849 we detected very similar levels of variability between LW and WB, even though we expected to
850 find higher levels of variation in the first due to the documented introgression of Asian germplasm
851 into LW (approximately 20–35% of the genome has been estimated to be of Asian origin; Groenen
852 et al. 2012, Bosse et al. 2012, Bosse, Megens, Madsen, et al. 2014, Frantz et al. 2015, Bianco et
853 al. 2015, Ai et al. 2015). Interestingly, the high levels of variability were observed for variants that
854 belonged to different frequency ranges in these two populations: singletons in WB and in high-
855 frequency derived alleles in LW. Although these differences may be mainly due to the effects of
856 gene flow in LW, we cannot discard an important effect of the selective programs applied to this
857 commercial breed.

858

859 **Domestication hallmarks at pig coding regions**

860 Another main goal of this work was finding the hallmarks of positive selection produced as a
861 consequence of the domestication since this process implies a process of positive (human
862 mediated) selection for traits that benefit both humans and the species of interest. The paucity of
863 fixed variants found at coding positions in the three breeds indicates that the observed heritable
864 phenotypic differences among the breeds are either due to: i) very few selective sweeps, ii) positive
865 selection at noncoding functional regions that were not analyzed in here, iii) changes in the
866 frequencies of nonsynonymous variants without being fixed. If the first hypothesis is true, we
867 expect that domestication process should fix the adaptive variants for those genes underlaying the
868 phenotypes of interest. However, we found no fixed variants between domestic breeds and wild
869 pigs. Although this might be a consequence of some genetic exchange among populations, we
870 found that the individuals were classified in groups by their location and according to their
871 respective phenotype (Figure S3A), which suggest that the domestication features of the different
872 breeds, even if admixed, are maintained. When we checked the α values for those genes that were
873 previously reported to show signals of positive selection using other approaches (Groenen 2016),
874 we found that these genes show little or no nonsynonymous polymorphisms or fixed variants
875 (Table S14). This absence of variability is typical from regions under selective sweeps, although
876 not necessarily implicating that these genes are the targets of domestication since there are no
877 variants fixed or close to fixation at their coding regions. We only found significant values of α
878 over zero at the gene KIT in the IB breed, the genes IGF2R and JMJD1C in the LW breed and the

879 gene LRRTM3 in the WB population. This low number of genes with positive α values would
880 make the first hypothesis unlikely. The second hypothesis implies that the functional regions
881 implicated in domestication would be out of coding regions (promotors, enhancers, and others.
882 (e.g., Li et al 2018, Rubin et al 2012, Anderson 2012). However, although being a promising
883 hypothesis, we did not analyze those regions because it requires a very accurate analysis of
884 homology and their associated functionality, which is very complicated at the genome level,
885 especially for non-model species with a high proportion of missing data. The third hypothesis
886 suggests that the domesticated phenotype is caused by a moderate change in the frequency of a
887 relatively large number of variants with small selective effects. In this case, depending on the size
888 of the selective effect there would only be changes in the frequencies of the variants without
889 reaching fixation. In the last case, the functional variants involved in domestication should be
890 segregating in the analyzed populations. These positively selected variants segregating at high
891 frequencies, together with the presence of deleterious mutations also segregating at low
892 frequencies, would be reflected as an excess of non-neutral polymorphism compared to divergence
893 (*i.e.*, negative α statistic at high frequencies). Hence, in cases where there is a significant
894 proportion of positive selection variants that have not yet being fixed, we expect to observe a trend
895 in the α slope showing more negative α values at intermediate-high frequencies. However, we did
896 not observe this pattern in any of the three populations examined when the analysis was performed
897 based on all coding positions, although it was observed for α values estimated based on exclusive
898 and shared mutations, which suggest that different types of variants (total, shared and exclusive)
899 could be capturing different aspects of the domestication process (demography versus selection).

900 The estimation of the DFE from the ABC analysis showed that the most likely evolutionary model
901 for all three populations based on total variants was that consisting in a discrete DFE without
902 significant positive selection effect (model DN; Table 2), showing a clear genome-wide effect of
903 the action of purifying selection. We also observed a reduced effect of purifying selection in IB
904 and in less extend in WB when the analysis was performed based on exclusive variants, which
905 suggest a reduction of the population size of these two populations. However, for LW and when
906 the analysis was performed based on exclusive positions, the most likely model was that with a
907 discrete distribution that includes the effect of positive selection (model D; Table 2), which may
908 reflect the increase of new Asian variants which increased in frequency by artificial selection..
909 Nevertheless, the differences between models including or excluding beneficial variants were

910 relatively small, suggesting that, in general, a few proportion of beneficial mutations contributed
911 to the domestication process. In fact, under model D, the estimated global proportion of beneficial
912 mutations (weak and strong) was relatively small and slightly higher when the analysis is based
913 on shared variants (0.1% based on total and exclusive SNPs and 1.4% based on shared SNPs;
914 Table 3) and similar in wild and domestic populations. Nevertheless, this proportion of mutations
915 may be substantial in absolute numbers (*i.e.,* several thousand mutations).

916 Although an excess of nonsynonymous shared variants compared to the synonymous ones can be
917 explained by some demographic scenarios such as bottlenecks, they may also reflect biological
918 constraints at the species level. For instance, the phenotypic variation in a polygenic selective
919 scenario could be caused by subtle changes in the frequencies of many genes (in an infinitesimal
920 scenario) which would result in the observed phenotypic differences among the breeds. On the
921 other hand, exclusive variants may reflect recent and breed-specific selective hallmarks and hence,
922 would be responsible for the observed differences between domesticated and wild breeds. In both
923 cases, shared and exclusive polymorphisms are contributing to the differences in the SFS between
924 functional and non-functional positions. Nevertheless, the differences of the DFE when the
925 analysis was performed based on total or shared variants is very small, suggesting that exclusive
926 variants would be more informative to detect the effects of the change of selective effects.

927 In addition, our simulated domestication scenarios indicate that the effect of positive selection
928 irrespective of being either strong and affecting a small percentage of variants or weak and
929 affecting a large percentage of variants is not reflected as marked changes in the estimated patterns
930 of α. This could be due to the short time since the change in the fitness effects of variants occurred
931 but also by the interaction of positive and negative selection and demographic processes in the
932 case of the complex scenarios, which are the most compatible with the observed data. In fact, the
933 observed α patterns are compatible with the simulated demographic effects (population size
934 reduction in WB and IB and gene flow in LW) but also, as in the ABC analysis results, with the
935 effect of positive selection in LW when the analysis was based on exclusive variants.

936 We are aware that the evolutionary models used here are very simple and contain few parameters
937 and the real observations contain high heterogeneity that could not be fitted to these models. The
938 reasons for this heterogeneity may be technical (e.g., not adequate filtering of raw sequences),
939 conceptual (undetected correlations that distort model assumptions) or biological (too simplistic

models to explain the real data). In any case, the model that assumes a discrete distribution of deleterious mutations (model DN) seems to generally explain better the observed data, together with the model D (model DN but including the effect of beneficial selection) in a minor degree and also that exclusive variants seemed to be more informative to detect the changes of the selective effects.

Final remarks

The observed patterns of variability are compatible with the presence of deleterious mutations segregating in all three breeds but also with weak signals of positive selection. In addition, when the variants are split into shared and exclusive, we observed patterns that are in line with the simulated data under different demographic scenarios with the joint action of positive and negative selection. We found a clear effect of deleterious mutations at low-frequency variants and a possible mild effect of positive selection at higher frequencies. However, additional analyses contrasting evolutionary models that consider the effects of standing variation, whose effect change under the domestication process, may shed more light and will help to understand the patterns of variation shaped by the domestication process.

ACKNOWLEDGMENTS

We acknowledge L. Silió and M. C. Rodríguez for helpful comments on the manuscript. 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). S.G-R. was supported by a Beatriu de Pinós postdoctoral fellowship (AGAUR; 2014 BP-B 00027).

CONFLICT OF INTEREST DISCLOSURE

967 The authors of this article declare that they have no financial conflict of interest with the content
968 of this article. Sebastian E. Ramos-Onsins is one of the PCIEvolBiol recommenders.

969

970 **LITERATURE CITED**

971 Achaz G. 2009. “Frequency Spectrum Neutrality Tests: One for All and All for One.” *Genetics*
972 183: 249–58. https://doi.org/10.1534/genetics.109.104042.

973 Ai H., Fang X., Yang B., Huang Z, Chen H, Mao L, et al. 215 "Adaptation and possible ancient
974 interspecies introgression in pigs identified by whole- genome sequencing." *Nat Genet.*
975 2015;47:217–25.

976 Akashi H. 1995. "Inferring weak selection from patterns of polymorphism and divergence at
977 “silent” sites in Drosophila DNA". Genetics 139(2):1067–1076.

978 Alexander D.H., Novembre J. and Lange K. (2009) Fast model-based estimation of ancestry in
979 unrelated individuals. Genome Research, 19:1655–1664.

980 Alvarez-Ponce D., and Fares M. A. 2012. “Evolutionary Rate and Duplicability in the
981 Arabidopsis Thaliana Protein-Protein Interaction Network.” *Genome Biology and Evolution*
982 4 (12): 1263–74. https://doi.org/10.1093/gbe/evs101.

983 Alves E,, Ovilo C., Rodríguez C., and Silió L. 2003. “Mitochondrial DNA sequence variation
984 and phylogenetic relationships among Iberian pigs and other domestic and wild pig
985 populations.” *Animal Genetics* 34:319-324.

986 Alves E., Fernández A. I., Barragán C., Ovilo C., Rodríguez C., and Silió L.. 2006. “Inference of
987 Hidden Population Substructure of the Iberian Pig Breed Using Multilocus Microsatellite
988 Data.” *Spanish Journal of Agricultural Research* 4 (1): 37–46.
989 http://www.inia.es/gcontrec/pub/alves-fernandez-barragan-..._1141287881015.pdf.

990 Amaral A.J., Ferretti L., Megens H.J., Crooijmans R.P.M.A., Ni H, Ramos-Onsins S.E., Perez-
991 Enciso M., Schook L.B., Groenen M.A.M.. 2011 "Genome-wide footprints of pig
992 domestication and selection revealed through massive parallel sequencing of pooled DNA."
993 *PLoS One.* 2011;6(4):e14782. doi:10.1371/journal.pone.0014782

994 Andersson L. 2012 "How selective sweeps in domestic animals provide new insight into

995 biological mechanisms." J Intern Med. 2012;271:1–14.

996 Andersson L. 2013 "Molecular consequences of animal breeding." *Curr Opin Genet Dev*.
997 23(3):295-301. doi: 10.1016/j.gde.2013.02.014.

998 Beaumont M.A., Zhang W., Balding D.J. Approximate Bayesian computation in population
999 genetics. Genetics. 2002;162(4):2025-2035.

1000 Bianco E., Nevado B., Ramos-Onsins S. E., and Pérez-Enciso M. 2015. "A Deep Catalog of
1001 Autosomal Single Nucleotide Variation in the Pig." *PLoS ONE* 10 (3): 1–21.
1002 https://doi.org/10.1371/journal.pone.0118867.

1003 Booker T.R, and Keightley P.D.. 2018. "Understanding the Factors That Shape Patterns of
1004 Nucleotide Diversity in the House Mouse Genome." *Molecular Biology and Evolution* 35
1005 (12): 2971–88. https://doi.org/10.1093/molbev/msy188.

1006 Bosse M., Megens H.-J., Frantz L.A.F., Madsen O., Larson G., Paudel Y., Duijvesteijn N. et al.
1007 2014. "Genomic Analysis Reveals Selection for Asian Genes in European Pigs Following
1008 Human-Mediated Introgression." *Nature Communications* 5: 4392.
1009 https://doi.org/10.1038/ncomms5392.

1010 Bosse M., Megens H.-J., Madsen O., Frantz L.A.F., Paudel Y., Crooijmans R.P.M.A. and
1011 Groenen M.A.M. 2014. "Untangling the Hybrid Nature of Modern Pig Genomes: A Mosaic
1012 Derived from Biogeographically Distinct and Highly Divergent Sus Scrofa Populations."
1013 *Molecular Ecology* 23 (16): 4089–4102. https://doi.org/10.1111/mec.12807.

1014 Bosse M., Megens H.-J., Madsen O., Paudel Y., Frantz L.A.F., Schook L.B, Crooijmans
1015 R.P.M.A and Groenen M.A.M. 2012. "Regions of Homozygosity in the Porcine Genome:
1016 Consequence of Demography and the Recombination Landscape." *PLoS Genetics* 8 (11).
1017 https://doi.org/10.1371/journal.pgen.1003100.

1018 Chen J., Ni P., Li X., Han J., Jakovlić I., Zhang C., Zhao S. 2018 "Population size may shape the
1019 accumulation of functional mutations following domestication." *BMC Evolutionary Biology*
1020 18:4 DOI 10.1186/s12862-018-1120-6

1021 Cruz F., Vilà C., and Webster M.T. 2008. "The Legacy of Domestication: Accumulation of
1022 Deleterious Mutations in the Dog Genome." *Molecular Biology and Evolution* 25 (11):
1023 2331–36. https://doi.org/10.1093/molbev/msn177.

1024 Csardi G., and Nepusz T. 2006. “The Igraph Software Package for Complex Network Research.”
1025 *InterJournal, Complex Systems* 1695. http://igraph.org.

1026 Csilléry K., François O. and Blum M.G.B. 2012 " abc: an R package for approximate Bayesian
1027 computation (ABC) " Methods in Ecology and Evolution 2012, 3, 475–479.

1028 Doebley J.F., Gaut B.S. and Smith B.D. 2006. "The molecular genetics of crop domestication".
1029 Cell. 29;127(7):1309-21. doi: 10.1016/j.cell.2006.12.006.

1030 Duret L, Semon M, Piganeau G, Mouchiroud D and Galtier N. 2002. "Vanishing GC-rich
1031 isochores in mammalian genomes". Genetics 162(4):1837–1847.

1032 Esteve-Codina A., Paudel Y., Ferretti L., Raineri E., Megens H.-J., Silió L., Rodríguez M.C.,
1033 Groenen M.A.M., Ramos-Onsins S.E., and Pérez-Enciso M. 2013. “Dissecting Structural
1034 and Nucleotide Genome-Wide Variation in Inbred Iberian Pigs.” *BMC Genomics* 14 (148):
1035 1. https://doi.org/10.1186/1471-2164-14-148.

1036 Evanno G., Regnaut S., Goudet J. 2005. "Detecting the number of clusters of individuals using
1037 the software STRUCTURE: a simulation study". Mol Ecol. 14(8):2611-20.

1038 Eyre-Walker A. 2002. “Changing Effective Population Size and the McDonald-Kreitman Test.”
1039 *Genetics* 162: 2017–24. http://www.genetics.org/content/genetics/162/4/2017.full.pdf.

1040 Eyre-Walker A. 2006 "The genomic rate of adaptive evolution." *Trends Ecol Evol* 21(10):

1041 569–575.

1042 Fang M., Larson G., Ribeiro H.S., Li N., Andersson L. 2011 "Contrasting mode of evolution at a
1043 coat color locus in wild and domestic pigs." *PLoS Genet.* 5:e1000341.

1044 Fay J.C. and Wu C-I. 2000. “Hitchhiking Under Positive Darwinian Selection.” *Genetics* 155:
1045 1405–13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461156/pdf/10880498.pdf.

1046 Fay J.C., Wyckoff G.J., Wu C.I. 2001 "Positive and negative selection on the human genome."
1047 Genetics 158(3):1227–1234.

1048 Ferretti L., Raineri E. and Ramos-Onsins S.E. 2012. “Neutrality Tests for Sequences with
1049 Missing Data.” *Genetics* 191 (4): 1397–1401. https://doi.org/10.1534/genetics.112.139949.

1050 Fisher, R. A. 1919. “XV.—The Correlation between Relatives on the Supposition of Mendelian

1051 Inheritance." *Transactions of the Royal Society of Edinburgh* 52 (02): 399–433.
1052 https://doi.org/10.1017/S0080456800012163.

1053 Frantz L.A.F., Schraiber J.G., Madsen O., Megens H.-J, Bosse M., Paudel Y., Semiadi G. et al.
1054 2013. "Genome Sequencing Reveals Fine Scale Diversification and Reticulation History
1055 during Speciation in Sus." *Genome Biology* 14 (9): R107. https://doi.org/10.1186/gb-2013-
1056 14-9-r107.

1057 Frantz L.A.F., Schraiber J.G., Madsen O., Megens H.-J., Cagan A., Bosse M., Paudel Y.,
1058 Crooijmans R.P.M.A., Larson G., and Groenen M.A.M. 2015. "Evidence of Long-Term
1059 Gene Flow and Selection during Domestication from Analyses of Eurasian Wild and
1060 Domestic Pig Genomes." *Nature Genetics* 47 (10): 1141–48.
1061 https://doi.org/10.1038/ng.3394.

1062 Fraser H. B, Hirsh A.E., Steinmetz L.M., Scharfe C., and Feldman M.W. 2002. "Evolutionary
1063 Rate in the Protein Interaction Network." *Science* 296 (5568): 750–52.
1064 https://doi.org/10.1126/science.1068696.

1065 Fu Y.-X. and Li W.-H. 1993. "Statistical tests of neutrality of mutations". Genetics. 133(3):693-
1066 709. doi: 10.1093/genetics/133.3.693.

1067 Gaut B.S., Seymour D.K., Liu Q., Zhou Y. 2018. "Demography and its effects on genomic
1068 variation in crop domestication". Nat Plants. 4(8):512-520. doi: 10.1038/s41477-018-0210-
1069 1.

1070 Groenen M.A.M. 2016. "A Decade of Pig Genome Sequencing: A Window on Pig
1071 Domestication and Evolution." *Genetics, Selection, Evolution : GSE Sel Evol* 48 (23): 1–9.
1072 https://doi.org/10.1186/s12711-016-0204-2.

1073 Groenen, M.A.M., Alan L. Archibald, Hirohide Uenishi, Cristopher K. Tuggle, Yasuhiro
1074 Takeuchi, Max F. Rothschild, Claire Rogel-Gaillard, et al. 2012. "Analyses of Pig Genomes
1075 Provide Insight into Porcine Demography and Evolution." *Nature* 491 (7424): 393–98.
1076 https://doi.org/10.1038/nature11622.

1077 Guirao-Rico S., Ramirez O., Ojeda A., Amills M., and Ramos-Onsins S.E.. 2018. "Porcine Y-
1078 Chromosome Variation Is Consistent with the Occurrence of Paternal Gene Flow from Non-
1079 Asian to Asian Populations." *Heredity* 120 (1): 63–76. https://doi.org/10.1038/s41437-017-

0002-9.

Hahn M.W. and Kern A.D.. 2005. “Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks.” *Molecular Biology and Evolution* 22 (4): 803–6. https://doi.org/10.1093/molbev/msi072.

Haller B. C., and Messer P.W. 2017. "AsymptoticMK: A web-based tool for the asymptotic McDonald–Kreitman test". *G3: Genes, Genomes, Genetics* 7(5), 1569–1575. https://doi:10.1534/g3.117.039693.

Haller B.C. and Messer P.W. 2017. “SLiM 2: Flexible, Interactive Forward Genetic Simulations.” *Molecular Biology and Evolution* 34 (1): 230–40. https://doi.org/10.1093/molbev/msw211.

Hill, W.G. and Robertson A. 1966. “The Effect of Linkage on Limits to Artificial Selection.” *Genetical Research* 8 (3): 269–94. http://www.ncbi.nlm.nih.gov/pubmed/5980116.

Kanehisa M., Araki M., Goto S., Hattori M., Hirakawa M., Itoh M., Katayama T. et al. 2008. “KEGG for Linking Genomes to Life and the Environment.” *Nucleic Acids Research* 36 (SUPPL. 1): 480–84. https://doi.org/10.1093/nar/gkm882.

Kim H., Song K.D., Kim H..J., Park W., Kim J., Lee T. et al. 2015 "Exploring the genetic signature of body size in Yucatan miniature pig." *PLoS One*. 10:e0121732.

Kono T.J.Y., Fu F., Mohammadi M., Hoffman P.J., Liu C., Stupar R.M., Smith K.P., Tiffin P., Fay J.C., Morrell P.L. 2016 "The Role of Deleterious Substitutions in Crop Genomes." *Mol. Biol. Evol.* 33(9):2307–2317. https://doi:10.1093/molbev/msw102

Leno-Colorado J., Hudson N.J., Reverter A., and Pérez-Enciso M.. 2017. “A Pathway-Centered Analysis of Pig Domestication and Breeding in Eurasia.” *G3* 7 (7): 2171–84. https://doi.org/10.1534/g3.117.042671.

Li C., Zou C., Cui Y., Fu Y., Fang C., Li Y., Li J., Wang W., Xiang H., Li C. 2018 "Genome-wide epigenetic landscape of pig lincRNAs and their evolution during porcine domestication." Epigenomics. 2018 Dec;10(12):1603-1618. doi: 10.2217/epi-2017-0117.

Li H., and Durbin R. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” Journal Article. *Bioinformatics* 25 (14): 1754–60.

https://doi.org/10.1093/bioinformatics/btp324.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G.R., Durbin R. and Subgroup Genome Project Data Processing. 2009. “The Sequence Alignment/Map Format and SAMtools.” Journal Article. *Bioinformatics* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Li M., Tian S., Jin L., Zhou G., Li Y., Zhang Y., Wang T. et al. 2013. “Genomic Analyses Identify Distinct Patterns of Selection in Domesticated Pigs and Tibetan Wild Boars.” *Nature Genetics* 45 (12): 1431–38. https://doi.org/10.1038/ng.2811.

Li M., Tian S., Yeung C.K., Meng X., Tang Q., Niu L., Wang X., Jin L., Ma J., Long K., Zhou C., Cao Y., Zhu L., Bai L., Tang G., Gu Y., Jiang A., Li X., Li R. 2014 "Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication." *Sci Rep.* 4:4678. doi: 10.1038/srep04678.

Livingstone K. and Anderson S.. 2009. “Patterns of Variation in the Evolution of Carotenoid Biosynthetic Pathway Enzymes of Higher Plants.” *Journal of Heredity* 100 (6): 754–61. https://doi.org/10.1093/jhered/esp026.

MacEachern S., McEwan J., McCulloch A., Mather A., Savin K., Goddard M. 2009 "Molecular evolution of the Bovini tribe (Bovidae, Bovinae): is there evidenceof rapid evolution or reduced selective constraint in Domestic cattle?" *BMC Genomics* (10) *179;* https:// doi: 10.1186/1471-2164-10-179.

Matsumoto T., John A., Baeza-Centurion P., Li B., Akashi H. 2016. "Codon Usage Selection Can Bias Estimation of the Fraction of Adaptive Amino Acid Fixations". Mol Biol Evol. 2016 Jun;33(6):1580-9. doi: 10.1093/molbev/msw027.

Makino T., Rubin C.J., Carneiro M., Axelsson E., Andersson L., Webster M.T. 2018 "Elevated Proportions of Deleterious Genetic Variation in Domestic Animals and Plants." *Genome Biol Evol.* 10(1):276-290. https://doi: 10.1093/gbe/evy004.

McDonald J.H. and Kreitman M.. 1991. “Accelerated Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351: 652–54.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K. et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-

1137 Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.
1138 https://doi.org/10.1101/gr.107524.110.

1139 Messer P.W. and Petrov D.A. 2013. "Frequent Adaptation and the McDonald-Kreitman Test."
1140 *Proceedings of the National Academy of Sciences of the United States of America* 110 (21):
1141 8615–20. https://doi.org/10.1073/pnas.1220835110.

1142 Mignon-Grasteau S., Boissy A., Bouix J., Faure J.-M., Fisher A.D., Hinch G. N., Jensen P., Le
1143 Neindre P., Morméde P., Prunet P., Vandeputte M. and Beaumont C. 2005. "Genetics of
1144 adaptation and domestication in livestock." Livestock Production Science 93:3–14.

1145 Montanucci L., Laayouni H., Dall'Olio G.M., and Bertranpetit J. 2011. "Molecular Evolution
1146 and Network-Level Analysis of the N-Glycosylation Metabolic Pathway across Primates."
1147 *Molecular Biology and Evolution* 28 (1): 813–23. https://doi.org/10.1093/molbev/msq259.

1148 Moon S., Kim T.-H., Lee K.-T., Kwak W., Lee T., Lee S.-W., Kim M.-J. et al. 2015. "A
1149 Genome-Wide Scan for Signatures of Directional Selection in Domesticated Pigs." *BMC*
1150 *Genomics* 16 (1): 1–12. https://doi.org/10.1186/s12864-015-1330-x.

1151 Nevado B., Ramos-Onsins S.E. and Perez-Enciso M. 2014. "Resequencing Studies of Nonmodel
1152 Organisms Using Closely Related Reference Genomes: Optimal Experimental Designs and
1153 Bioinformatics Approaches for Population Genomics." *Molecular Ecology* 23 (7): 1764–79.
1154 https://doi.org/10.1111/mec.12693.

1155 Orlando L. and Librado P. 2019 "Origin and Evolution of Deleterious Mutations in Horses."
1156 *Genes* 10, 649; doi:10.3390/genes10090649

1157 Pérez-Enciso M., de los Campos G., Hudson N., Kijas J. and Reverter A. 2017. "The
1158 'Heritability' of Domestication and Its Functional Partitioning in the Pig." *Heredity* 118:
1159 160–68. https://doi.org/10.1038/hdy.2016.78.

1160 Quinlan A.R. 2014. "BEDTools: The Swiss-Army Tool for Genome Feature Analysis." *Current*
1161 *Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47
1162 (January): 11.12.1-11.12.34. https://doi.org/10.1002/0471250953.bi1112s47.

1163 Ramírez O., Burgos-Paz W., Casas E., Ballester M., Bianco E., Olalde I., Santpere G. et al. 2014.
1164 "Genome Data from a Sixteenth Century Pig Illuminate Modern Breed Relationships."
1165 *Heredity* 114 (2): 175–84. https://doi.org/10.1038/hdy.2014.81.

Ramos-Onsins S.E., Burgos-Paz W., Manunza A. and Amills M. 2014. "Mining the pig genome to investigate the domestication process". Heredity (Edinb) 113(6):471-84. doi: 10.1038/hdy.2014.68.

Ramsay H., Rieseberg L.H. and Ritland K. 2009. “The Correlation of Evolutionary Rate with Pathway Position in Plant Terpenoid Biosynthesis.” *Molecular Biology and Evolution* 26 (5): 1045–53. https://doi.org/10.1093/molbev/msp021.

Rausher M.D, Miller R.E., and Tiffin P. 1999. “Patterns of Evolutionary Rate Variation Among Genes of the Anthocyanin Biosynthetic Pathway.” *Molecular Biology and Evolution* 16 (2): 266–74. https://watermark.silverchair.com/mbev_16_02_0266.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAdwwggHYBgkqhkiG9w0BBwagggHJMIIBxQIBADCCAb4GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMB_IVBQyX1n-F9I4pAgEQgIIBj8r0BkU_bdeIWEuY_7bUxMFToA8Yo-26snF_yPY.

Renaut S., and Rieseberg L.H. 2015. “The Accumulation of Deleterious Mutations as a Consequence of Domestication and Improvement in Sunflowers and Other Compositae Crops.” *Molecular Biology and Evolution* 32 (9): 2273–83. https://doi.org/10.1093/molbev/msv106.

Riley R.M., Jin W. and Gibson G.. 2003. “Contrasting Selection Pressures on Components of the Ras-Mediated Signal Transduction Pathway in Drosophila.” *Molecular Ecology* 12 (5): 1315–23. https://doi.org/10.1046/j.1365-294X.2003.01741.x.

Ross-Ibarra J., Morrell P.L. and Gaut B.S. 2007. "Plant domestication, a unique opportunity to identify the genetic basis of adaptation". Proc Natl. Acad. Sci. USA. 2007 doi: 10.1073/pnas.0700643104.

Rubin C.-J., Megens H.-J., Martinez Barrio A., Maqbool K., Sayyab S., Schwochow D., Wang C. et al. 2012. “Strong Signatures of Selection in the Domestic Pig Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (48): 19529–36. https://doi.org/10.1073/pnas.1217149109.

Tajima F. 1983. “Evolutionary Relationship of DNA Sequences in Finite Populations.” *Genetics* 105: 437–60. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1202167/pdf/437.pdf.

1195 Uricchio L.H., Petrov D.A., Enard D. 2019 "Exploiting selection at linked sites to infer the rate
1196 and strength of adaptation." *Nat Ecol Evol*. 3(6):977-984. doi: 10.1038/s41559-019-0890-6.

1197 Watterson, G.A. 1975. "On the Number of Segregating Sites in Genetical Models without
1198 Recombination." *Theoretical Population Biology* 7 (2): 256–76.
1199 https://doi.org/10.1016/0040-5809(75)90020-9.

1200 Wilkinson S., Lu Z.H., Megens H.-J., Archibald A.L., Haley C., Jackson I.J., Groenen M.A.,
1201 Crooijmans R.P., Ogden R., Wiener P. 2013 "Signatures of diversifying selection in
1202 European pig breeds." PLoS Genet. 2013 Apr;9(4):e1003453. doi:
1203 10.1371/journal.pgen.1003453.

1204 Zeder M.A. 2012. "The Domestication of Animals." *Journal of Anthropological Research* 68
1205 (2): 161–90. https://doi.org/10.3998/jar.0521004.0068.201.

1206 Zhang M., Yang Q., Ai H., Huang L. 2021. Revisiting the evolutionary history of pigs via de
1207 novo mutation rate estimation by deep genome sequencing on a three-generation pedigree.
1208 bioRxiv2021. 2003.2029.437103.

1209

1210 **SUPPLEMENTARY MATERIAL**

1211 See Supplementary Material file added to see the Tables (S1-S14) and Figures (S1-S48).