Universitat Autònoma de Barcelona

Effect of domestication in the pig genome

Jorge Leno Colorado

Effect of domestication in the pig genome

PhD. Thesis in Genetics

by

Jorge Leno Colorado

Directors

Dr. Miguel Pérez Enciso

Dr. Sebastián E. Ramos Onsins

Tutor

Dr. Mario Cáceres Aguilar

UNIVERSITAT AUTÒNOMA DE BARCELONA

Departament de Genètica, Facultat de Biociències

CENTRE DE RECERCA EN AGRIGÈNOMICA

Departament de Genètica Animal

Universitat Autònoma de Barcelona, June 2019

El **Dr. Miguel Pérez Enciso**, investigador ICREA del Departament de Ciència Animal i dels Aliments de la Universitat Autònoma de Barcelona (UAB), i el **Dr. Sebastián E. Ramos Onsins**, investigador del Centre de Recerca en Agrigenòmica (CRAG),

fan constar

que el treball de recerca i la redacció de la memòria de la tesi doctoral titulada "Effect of domestication in the pig genome" han estat realitzats sota la seva direcció per

JORGE LENO COLORADO

i certifiquen

Que aquest treball s'ha dut a terme al Departament de Genòmica Animal del Centre de Recerca en Agrigenòmica (CRAG), per a obtenir el grau de Doctor en Genètica per la Universitat Autònoma de Barcelona.

Bellaterra, a 30 de Maig de 2019

Dr. Miguel Pérez Enciso **Dr. Sebastián E. Ramos Onsins**

Contents

Summary

Animal domestication is an important process in the human history in which different traits of the animals were selected, such as faster growth or greater docility. To study domestication at the genetic level it is necessary to identify the markers related to this evolutionary process. Advances in sequencing technologies have improved the investigation of the genomics of domestication, which has allowed to determine the genetic changes that cause this transformation from wild to domestic species.

The main goal of this thesis is the evaluation of the domestication effect in the pig genome through the analysis of genetic diversity in domestic and wild populations.

In the first part, analyses of differentiation and linkage disequilibrium were performed to detect differences between domestic and wild pigs, using the pathway as the unit of analysis. Through the study of differentiation, using the Fst statistic, we obtained significant pathways related to behavior and development, which were some of the first selected traits in pigs. On the other hand, when performing the disequilibrium analysis, using the nSL statistic, we detected differences in pathways related to the reproduction of the animal, a recently selected trait. Besides, we made a co-association network using all pathways that are significantly different between domestic and wild pigs, obtaining three differentiated clusters, one related to growth and hormonal regulation, another with the sympathetic nervous system and the last with the reproduction.

In the second part, we performed an analysis of the strength of selection at the genome level in domestic and wild pigs, using two very different domestic populations, Iberian and Large White. Iberian breed is an autochthonous breed that has recently suffered a strong reduction in the effective population size, Large White is an international commercial breed that has been artificially improved and introgressed with Asian pigs. To analyze the strength of the selection we use the parameter α, which estimates the proportion of non-synonymous substitutions that are adaptive, using four different estimators of

variability, each focused on a part of the frequency spectrum: Fu&Li (only singletons), Watterson (whole spectrum giving more weight at low frequencies), Tajima (whole spectrum weighted uniformly) and Fay&Wu (increases the weight proportionally with the frequency). However, when analyzing the selection patterns, we did not find more common signals between the two domestic breeds than between domestic and wild ones. Instead, we found a larger effect of demography on the selection, Iberian has a very low variability due to its low population size, which is shown in the obtained selection patterns, which resemble a population reduction; while Large White has a larger variability, possibly due to the presence of Asian alleles in its genome, obtaining patterns that can be explained by the presence of both deleterious and beneficial mutations, together with a population expansion and/or migration.

Finally, we have developed a web-based application to analyze VCF files, which can help identify possible errors or biases, mainly related to the SNP coverage.

Resumen

La domesticación animal es un proceso realmente importante en la historia del hombre en el cual se seleccionaron diferentes rasgos de interés de los animales, como puede ser un crecimiento más rápido o una mayor docilidad. Para estudiar la domesticación a nivel genético es necesario identificar una serie de marcadores relacionados con este proceso evolutivo. Los avances en las tecnologías de secuenciación han mejorado considerablemente la investigación de la genómica de la domesticación, pudiendo determinar los cambios genéticos que causan esa transformación de especie salvaje a doméstica.

El objetivo principal de esta tesis es la evaluación del efecto de la domesticación en el genoma del cerdo mediante el análisis de la diversidad genética en poblaciones domésticas y salvajes.

En la primera parte se ha realizado un análisis de la diferenciación y del desequilibrio de ligamiento para detectar las diferencias entre cerdos domésticos y salvajes, utilizando la vía metabólica como unidad de análisis. Mediante el estudio de la diferenciación, utilizando el estadístico Fst, obtenemos una serie de rutas significativas relacionadas con el comportamiento y el desarrollo, que fueron algunos de los primeros rasgos seleccionados en cerdo. Sin embargo, al realizar el análisis del desequilibrio, mediante el estadístico nSL, detectamos diferencias en rutas relacionadas con la reproducción del animal, rasgo seleccionado recientemente. Por otro lado, realizamos una red de co-asociación entre todas las vías metabólicas significativamente diferentes entre cerdos domésticos y salvajes, obteniendo 3 clústeres diferenciados, uno relacionado con el crecimiento y la regulación hormonal, otro con el sistema nervioso simpático y el último con la reproducción.

En la segunda parte, realizamos un análisis de la fuerza de la selección a nivel genómico en cerdos domésticos y salvajes, utilizando dos poblaciones domésticas, Ibérico y Large White, las cuales son muy diferentes entre ellas. Mientras que Ibérico es una raza autóctona que ha sufrido recientemente una gran reducción del tamaño poblacional, Large White es una raza comercial

internacional que ha sido mejorada de manera artificial, además de introgresada con cerdos asiáticos. Para analizar la fuerza de la selección utilizamos el parámetro α, que estima la proporción de sustituciones no-sinónimas que son adaptativas, utilizando cuatro estimadores diferentes de la variabilidad, cada uno enfocado a una parte del espectro de frecuencias: Fu&Li (solo singletons), Watterson (todo el espectro dando más peso a las bajas frecuencias), Tajima (todo el espectro de manera uniforme) y Fay&Wu (incrementa el peso de manera proporcional a la frecuencia). Sin embargo, al analizar los patrones de selección no encontramos más señales comunes entre las razas domesticadas que al compararlas con la salvaje. En cambio, encontramos un mayor efecto de la demografía en la selección, Ibérico tiene una variabilidad muy baja debido a su bajo tamaño poblacional, lo cual se muestra en los patrones de selección obtenidos, que se asemejan a una reducción poblacional; mientras que Large White tiene una mayor variabilidad debido posiblemente a la presencia de alelos asiáticos en su genoma, obteniendo patrones explicados por la presencia tanto de mutaciones deletéreas como beneficiosas, además de una expansión poblacional y/o migración.

Por último, hemos desarrollado una aplicación web para poder analizar archivos VCF, la cual puede ayudarnos a identificar posibles errores o sesgos, principalmente relacionados con la cobertura del SNP.

List of figures

List of tables

Chapter 1

General Introduction

General Introduction

Domesticated animals and the domestication process have been studied extensively during the last century in different fields of biology and archeology, since these animals have been selected for different traits of interest for human needs, such as less aggressiveness and greater reproductivity. The study of markers linked to this evolutionary process may shed light on its biological basis (Zeder 2006).

1.1. Molecular evolution and population genetics

Biological evolution can be defined as the process that converts one population in a new population or species through generations due to the occurrence of variations (Lewontin 1974). Therefore, for the evolution process to occur, there must be variation between individuals within the same population and at least part of this variation must be inheritable (Lewontin 1970; Endler 1986). For the appearance of this variation, it is necessary that there be error-prone replication of the DNA, which means that DNA copies will not be always identical to the original one. Any change that occurs in the DNA is known as mutation, which can be single-base substitutions, insertions, deletions or sequence rearrangements. Genetic variation is the ultimate responsible for phenotypic changes observed.

Mutations are random events that may be beneficial o harmful for the individual. Although most mutations are eliminated within a few generations, due to the high probability of losing it by chance when the copy number is low, some of them can increase their frequency through generations, and even become fixed. The change of these frequencies may be due to natural selection or genetic drift. Genetic drift is the change in the allele frequencies due to the random sampling of gametes. This process affects more profoundly small populations and rare alleles. On the other hand, the principle of natural selection increases the frequency of mutations with a high fitness (the ability of an individual to leave offspring) compared to those of low fitness. These two processes decrease the

genetic diversity. Thus, the genetic diversity in a population is determined by a balance between mutation, selection and genetic drift.

1.1.1. The (nearly) neutral theory of evolution

When the genetic diversity of populations began to be identified, Motoo Kimura realized that extant hypotheses so far did not suitably explain the variation patterns and protein substitutions rates, due to the large amount of genetic variation observed in nature and that genetic differences are accumulated linearly in time (Zuckerkandl and Pauling 1965). For this reason, he proposed an alternative theory, known as neutral theory of molecular evolution (Kimura 1968).

The main idea of the neutral theory is that most mutations are neutral, i.e. they do not have advantages or disadvantages on fitness, and most of the evolutionary changes are the result of genetic drift that acts on these neutral alleles. The implications of this theory are (Kimura 1968, 1983; Casillas and Barbadilla 2017):

- Deleterious mutations are eliminated rapidly, and the beneficial ones are fixed rapidly, therefore, most of the variability observed in a population is due to the neutral mutations.
- Polymorphism level (θ) of a population depends on the effective population size (N_e) and neutral mutation rate (μ_0) and it is defined as $\theta = 4N_e\mu_0$ for diploids.
- The rate at which a neutral mutation is fixed is equal to the neutral mutation rate ($K = \mu_0$), which is independent of the populations size.
- Polymorphisms are transients.

According to the neutral theory, the rate of fixation of the mutations is proportional to the mutation rate, which is the frequency in which new mutations are produced in each generation, therefore, the evolutionary rate is proportional to the generation time. Nevertheless, empirical observations demonstrated that is proportional to the absolute time. For this reason, Tomoko Ohta redefined the theory introducing the concept of nearly neutral mutations, which have a weak beneficial or harmful effect, i.e. a coefficient of selection (s) ~ 0 (Ohta 1973, 1992). This type of mutations would explain many mutations present in the genome. Therefore, the nearly neutral theory is distinguished by considering weak

selection and maintains that the interplay between natural selection and genetic drift is important for the evolution. The theory predicts that the evolution is faster in populations with a smaller population size than large populations because most of the nearly neutral mutations are eliminated in large populations, whereas behave as neutral in small populations and are randomly fixed. The implications of this model are (Ohta 1992; Ohta and Gillespie 1996; Casillas and Barbadilla 2017):

- Mutations with $s << 1/N_e$ are considered neutral and, therefore, they depend only on the genetic drift.
- Mutations with $s \sim 1/N_e$ will be nearly neutral, with a weak effect on the fitness and they depend on the balance between genetic drift and natural selection.
- Mutations with $s >> 1/10N_e$ will be strongly deleterious or beneficial (depending on s), and they depend mainly on natural selection.

1.1.2. Signatures of molecular selection

Selection leaves signatures in the genome that affect both within and between populations at different levels: variability, linkage disequilibrium and allele frequency. These signatures can be detected through different statistical tests, since positive and negative selection leave different footprints of selection in the genome (Nielsen 2005). Examples of different models of selection that affect the genome differently are the selective sweeps and negative selection, both decreases the variability but at different level. Negative selection acts on multiple loci, removing the deleterious mutations, which reduces the variability between populations. On the other hand, selective sweeps are processes that eliminate or reduce the variability of the neutral alleles linked to a new beneficial mutation as its frequency increases, which decreases the variability within a population. Moreover, selective sweeps cause the occurrence of new mutations at low frequency (Harris, Sackman, and Jensen 2018; Jensen 2014). Different models of selective sweeps are summarized in Figure 1.1. Table 1.1 shows the effect of positive and negative selection in the variability.

***Table 1.1**: Effect of different models of selection. Adapted from Nielsen 2005*

Model of Selection	Variability within populations	Variability between populations	Ratio of between to within variability	Frequency spectrum
Positive selection	May increase or decrease	Increased	Increased	Increases the proportion of high frequency variants
Negative selection	Reduced	Reduced	Reduced if selection if not too strong	Increases the proportion of low frequency variants
Selective sweep	Decreased	No effect on mean rate, but the variance increases	Increased	Mostly increases the proportion of low frequency variants

Population differentiation

Positive selection increases the degree of differentiation between populations in the positive selected loci. Comparison of allele frequencies can be used to infer the population demographic histories.

The most common measure of population differentiation is the Fst coefficient (Weir and Cockerham 1984), which is a statistic based on the allele frequency differentiation and that measures how different are two populations. Fst provides information about the demographic history of a population, Fst will be higher in the presence of selection than with random genetic drift (Helyar et al. 2011). The estimation of Fst throughout the whole-genome can give a pattern of differentiation for the population. Once the differentiated loci are available, it is possible to obtain the genes where these alleles are located. In principle, these genes will be enriched in selection targets.

This approach has been used in different studies in the pig, for example showing that signatures of differentiation between domestic and wild pigs are specific for each breed (Amaral et al. 2011) or obtaining signatures of diversifying selection for morphological traits in European pigs (Wilkinson et al. 2013).

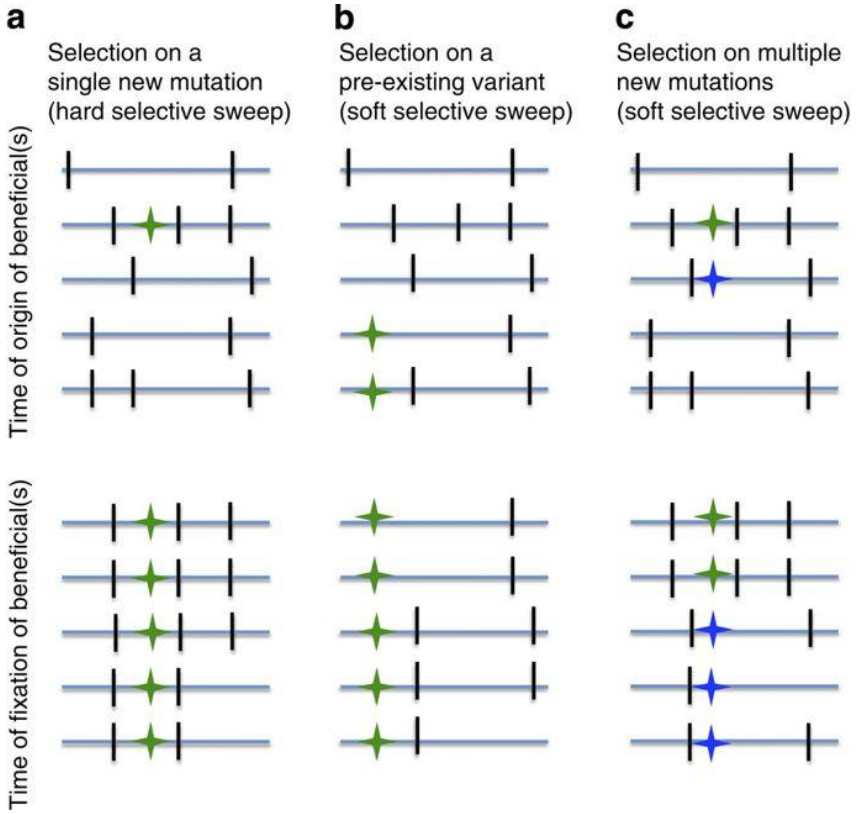

Figure 1.1: *Overview of selective sweep models, from the time of origin of beneficial mutation(s) to the time of the fixation. In the hard selective sweep (a), the beneficial mutation and the closely linked neutral mutations are fixed, whereas more distant neutral mutations can be brought only at intermediate frequency due to the recombination. However, there are different models of soft selective sweeps, i.e. when the beneficial mutation already exists (b), it may carry the haplotypes in which it is segregating at intermediate frequency, it can carry multiple haplotypes. Another model of soft sweep is when there are multiple new beneficial mutations (c), which can carry the haplotypes in which it emerged at an intermediate frequency. Adapted from Jensen (2014)*

Linkage disequilibrium

When a beneficial mutation increases its frequency in a population (there is positive selection), regions around this selected locus will be affected: variability will be reduced, linkage disequilibrium will increase, and the pattern of allele frequencies will be changed (Nielsen 2005).

Different methods that are used to measure the linkage disequilibrium are based in the haplotype homozygosity (HH), which is the probability that two chromosomes with the same polymorphic allele(s) are identical for all positions of a specific region.

The nSL (number of Segregating sites by Length) metrics detects positive selection based on an increase in the HH by measuring the length of the homozygous segment in terms of number of mutations (Ferrer-Admetlla et al. 2014). nSL is based on the iHS statistic but changing the way of measuring the length of the homozygous region. According to Ferrer-Admetlla et al. this simple change in the estimation results in a more robust to recombination/mutation rates and slightly more robust to changes in population size. This method is powerful to detect soft sweeps and even ongoing sweeps.

Frequency spectrum

As described above, selection affects the frequency of alleles and its distribution within a population. One of the most commonly used ways to study the impact of selection on the allele distribution is the frequency spectrum. The frequency spectrum is the count of the number of mutations at the different frequencies:

$$x_i = i/n$$

$$i = 1, 2, \ldots, n-1 \text{ (n = number of samples)}$$

In the case of a standard neutral model the frequency x_i is proportional to 1/i. However, with a negative selection, the allele frequency at low frequencies will be increased, but with a positive selection the allele frequency will increase at high frequencies. The selective sweeps affect the frequency spectrum skewing it towards rare alleles (Fig 1.2) (Nielsen 2005).

Consequently, many neutrality tests are based on the frequency spectrum to assess the goodness-of-fit of the standard neutral model. These neutrality tests compare two estimators of the variability θ, which is defined as $\theta = 4N_e\mu$ for diploids. With the standard neutral model, different unbiased estimators of θ will be equal. The typical estimators of the variability are θ_S (Watterson 1975), θ_π (Tajima 1983), θ_ξ (Fu and Li 1993), θ_H (Fay and Wu 2000). The difference between them is the weight that they give to each type of polymorphism according to their frequency, while θ_ξ is only based on singletons, θ_H give more weight to polymorphisms at high frequencies (ancestral polymorphisms). Although not as much as θ_ξ, θ_S and θ_π emphasize low-frequency alleles, with the difference between them that θ_S give less weight to high-frequency alleles than θ_π, which

weights uniformly (Figure 1.3) (Achaz 2009). In Table 1.2 we can observe the estimation of weights based on each estimator of θ. Mathematically, the estimator of θ, based on the weights of frequencies, can be calculated as follow:

$$\hat{\theta}_{\omega} = \frac{1}{\sum_i \omega_i} \sum_{i=1}^{n-1} \omega_i i \xi_i$$

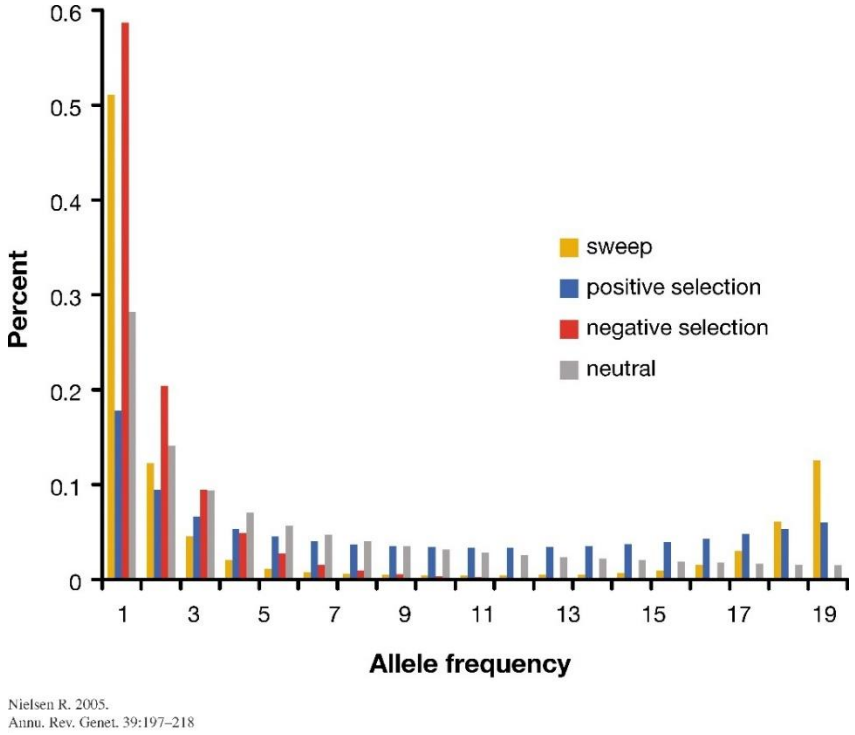

***Figure 1.2**: Frequency spectrum under selective sweep, neutrality, positive selection and negative selection. Adapted from Nielsen 2005.*

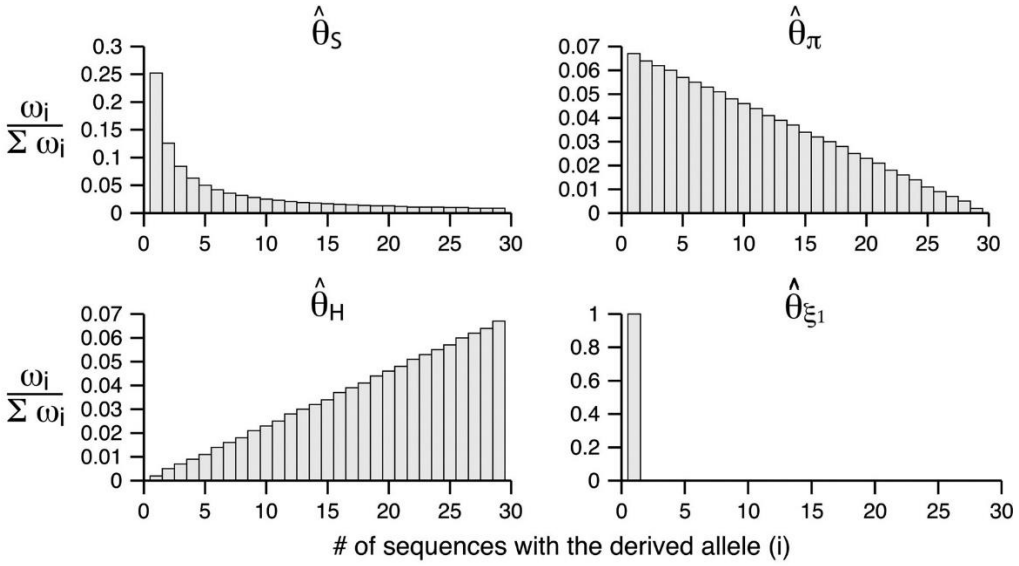

***Figure 1.3**: Distribution of weights by the frequency of the derived allele, based on each estimator of the variability. Adapted from Achaz 2009*

The most common neutrality test based on the frequency spectrum is the Tajima's D test (Tajima 1989). In this test, the number of polymorphic sites (θ_S) and the average pairwise differences between all sequences of the samples (θ_π) are compared. Fu and Li extended this test with the Fu and Li's D and Fu and Li's F (Fu and Li 1993), which compare the number of singletons (θ_ξ) with the number of all derived alleles (θ_S) and with the mean pairwise differences between sequences (θ_π), respectively. Fay&Wu's suggested a test (Fay and Wu 2000) to give more weight to the high-frequency alleles, comparing the average pairwise differences between all sequences of the samples (θ_π) with the number of high-frequency alleles (θ_H) (Achaz 2009; Nielsen 2005; Casillas and Barbadilla 2017).

***Table 1.2**: Estimation of ω for each estimator of the variability. Adapted from Achaz 2009.*

Estimators	ω
$\hat{\theta}_s$	$\omega_i = i^{-1}$
$\theta\pi$	$\omega i{=}n{-}i$
$\hat{\theta}_\xi$	$\omega_1 = 1, \omega_{i>1} = 0$
$\hat{\theta}_H$	$\omega_i = i$

McDonald-Kreitman test

The combined analysis of polymorphic mutations (polymorphisms) and fixed mutations (divergence) is a powerful method to study the effect of the selection in the genome. One of the most common metrics that use these principles is the McDonald-Kreitman test.

The McDonald-Kreitman test (McDonald and Kreitman 1991) is used to detect the selection footprint at the molecular level, exploiting the fact that the ratio between synonymous and non-synonymous mutations is modified when the loci is under selection. To perform this test, a 2x2 contingency table with the number of synonymous and non-synonymous polymorphisms (P_s and P_n) and divergence sites (D_s and D_n) is constructed. If selection only affects non-synonymous mutations, negative selection will decrease the number of these mutations relative to synonymous ones, however, positive selection will increase it. The effect in fixed mutations is stronger than in the polymorphic ones. When all mutations are neutral or strongly deleterious, the ratio D_n/D_s is equal to the ratio P_n/P_s, however, in the presence of positive selection the divergence ratio is greater than the polymorphic one due to the fact that adaptive mutations rapidly reach fixation (Casillas and Barbadilla 2017).

An extension of this test is α, which is an estimation of the proportion of non-synonymous substitutions that are adaptive (N. G. C. Smith and Eyre-Walker 2002):

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s}$$

Under neutrality, α is expected to be 0, α > 0 indicates positive selection and a negative α is due to the segregation of slightly deleterious mutations. The estimation of α can be underestimated if the population size is stable, because slightly deleterious mutations contributes more to non-synonymous polymorphism than non-synonymous divergence compared with synonymous mutations. On the other hand, if there is an expansion of the population size, the slightly deleterious mutations could be fixed and contribute more with the non-synonymous divergence than with non-synonymous polymorphisms, overestimating the α (Eyre-Walker 2006). Table 1.3 summarizes the expected

effect of different models of selection on P_s, P_n, D_s, D_n, estimators of variability (θ) and α, compared with the neutral model.

***Table 1.3**: Expected effect of different selection models in different parameter with respect to the neutral model*

Parameters		Positive selection (1)	Selective sweep (2)	Strong negative selection (3)	Background selection (4)
P_s		No effect	Decreases	No effect	Decreases
P_n		Decreases	Decreases	Decreases	Decreases
D_s		No effect	Increases	No effect	Decreases
D_n		Increases	Increases	Decreases	Decreases
θ_ξ	Syn	No effect	Decreases	No effect	Decreases – No effect
	nonSyn	Decreases	Decreases	Decreases- No effect	Decreases – No effect
θ_S	Syn	No effect	Decreases	No effect	Decreases
	nonSyn	Decreases	Decreases	Decreases	Decreases
θ_π	Syn	No effect	No effect	No effect	Decreases
	nonSyn	Decreases	No effect	Decreases	Decreases
θ_H	Syn	No effect	Increases	No effect	Decreases
	nonSyn	Increases	Increases	Decreases	Decreases
α		Positive	0 - slightly positive	Negative	slightly negative - 0

(1) Positive selection at genome level without linked positions to the beneficial mutation: We only could observe effect in the non-synonymous (causal) positions. The increase in D_n is greater than the decrease in P_n, for this reason α will be positive. All the non-synonymous variability will decrease.

(2) Selective sweep (positive selection with linked position to the beneficial mutation): It will affect both neutral and causal mutations. Although the 2 types of polymorphism decrease, the P_n does so more strongly, in the same way D_n increases more than the D_s. The decrease in P_n is the same as the increase in D_n, being α near to zero, unless multiple selective sweeps occur in the same region, in which case D_n will increase more and α will be more positive.

(3) Strong negative selection at genome level without linked positions to the deleterious mutation: It only affects the causal mutations, decreasing P_n and D_n, but D_n will do more, which causes a negative α. Although all the variability will be diminished, the one based on singletons (θ_ξ) will do it very weakly without almost

changes. The more weight is given to high-frequency alleles the greater the decrease, so the decrease will be $\theta_H > \theta_\pi > \theta_S > \theta_\xi$.

(4) Background selection (negative selection with linked alleles to the deleterious mutation): It will affect both neutral and causal mutations. Although both polymorphisms and divergence will decrease, the D_n *will do it slightly more. In the same way as with the unlinked negative selection, the more weight is given to the alleles in high frequency the more the variability will decrease. In this case, the synonymous and non-synonymous variability will decrease, but in a greater proportion the non-synonymous due to the causal mutations.*

1.1.3. Analysis of variability to study the selection footprint

Advances in sequencing techniques during the last century allowed to detect and investigate the mutations and selection signatures presents in the genome.

Microsatellites and mitochondrial DNA

Initial studies used microsatellites (Megens et al. 2008) and mitochondrial DNA (mtDNA) to reconstruct the demographic history of species.

The mtDNA sequence can be obtained quickly and with a low cost. The study of mitochondrial markers has been useful for different phylogenetic studies, for example proofing the entry of the pig in Europe (Larson et al. 2007) or to reveal that the pig was domesticated independently in different regions (Giuffra et al. 2000; Larson et al. 2005). The problem of this technique is that only the maternal locus is inherited and that there is no recombination, which may prevent the identification of some complex demographic events such as the hybridization of different regions.

Microsatellites are genetic markers that were used widely in population genetics, due to its high polymorphism, reproducibility and automation for detection. Microsatellites has been used in different studies to evaluate genetic relationships (Usha, Simpson, and Williams 1995), diversity (Ljungqvist, Åkesson, and Hansson 2010) and population structure (Haasl and Payseur 2011). In the pig species, for example, it has been possible to detect the strong genetic divergence between domestic and wild pigs (Giuffra et al. 2000; Larson et al. 2005). However, it has certain disadvantages, such as the high cost, the

time of obtaining and the difficulty to detect the limits of the lengths of repetition. These limitations highlighted the need for new techniques of high-density genotyping (Zhang and Hewitt 2003). Nowadays, microsatellites are used to answer very specific questions, such as parentage ascertainment or forensics.

Single Nucleotide Polymorphisms (SNPs)

The SNP is the position of a single nucleotide in which a mutation is present when comparing different sequences of the same population (Vignal et al. 2002) and are the best studied mutations. SNPs can appear within a gene or in non-coding regions of the DNA. There are two types of SNPs in the coding regions: (i) synonymous substitution if it does not affect the protein sequence and (ii) non-synonymous substitution if it changes the amino acid of the protein sequence by another (missense mutation) or if it changes the amino acid by a stop codon (nonsense mutation).

Despite the low polymorphism of a single SNP compared to microsatellites (Helyar et al. 2011), the number of SNPs that can be found in the genome gives the same or more information. For this reason, high-density SNP chips were developed and used in a great number of researches, improving the knowledge of the differences in the genome between populations and species. These chips have been used in several studies with different purposes such as the detection of linkage disequilibrium, QTL (quantitative trait loci) mapping and studies of association of the genome (Ponsuksili et al. 2011; Wang et al. 2012; Bosse et al. 2012; Yang et al. 2017). The problem of the commercial high-density SNP chips is that they have been designed using a small group of individuals of selected populations, which can give a bias if we analyze different populations, the so-called SNP ascertainment bias (Albrechtsen, Nielsen, and Nielsen 2010).

The first high-density chip of *Sus scrofa* included ~60,000 SNPs and was generated using different commercial populations and wild boars (Ramos et al. 2009). The most recent high-density SNP chip contained more than 650,000 SNPs. However, an increase in the number of SNP does not lead to a much more accurate prediction for breeding values, for example in cattle the increment of

accuracy when using 500,000 markers instead 50,000 is only the 1.6% (VanRaden et al. 2011).

Next Generation Sequencing (NGS)

The arrival of the Next Generation Sequencing (NGS) have revolutionized the biology, becoming possible to sequence the whole-genome in different species. Meaningful analysis of NGS data depends mainly on the accurate SNP and genotype calling, i.e. in identifying the variants and determining the genotype of each individual in this site with reliability (Nielsen et al. 2011).

The use of sequence data is a topic of high interest due to the drastic decrease in time and cost of genomic studies and the theoretical possibility of finding all causal mutations in the genome. Being able to analyze the complete genome of complete populations is highly interesting for the study of population genetics since we can look directly at the genome for the causal mutations instead of having to rely on indirect signals based on the linkage disequilibrium.

The pig is one of the sequenced species (Groenen et al. 2012), which has helped to conduct new studies of populations genetics, such as the estimation of genetic diversity, the study of homozygous regions (ROHs) and the detection of footprints of selection (Larson and Burger 2013; Bosse et al. 2012; Esteve-Codina et al. 2011; Rubin et al. 2012).

Nevertheless, the way in which NGS data is generated and analyzed is relevant since this technology is highly error-prone, especially in the detection of rare mutations. High error rates are due to different factors, like the base-calling and alignment errors. For instance, there are problems that have been detected in NGS data related to GC content and non-random reading errors (Minoche, Dohm, and Himmelbauer 2011; Lou et al. 2013). In addition to sequencing errors, the reliability of the SNPs detected with NGS depends on the coverage of the sample. In the case of low coverage in the samples, accurate SNP and genotype calling is difficult, and there is often considerable uncertainty associated with the results. It is crucial to quantify this uncertainty since it will affect the subsequent analyses, such as the identification of rare mutations and the estimation of allele frequencies (Nielsen et al. 2011).

1.2. Domestication

Domestication of animals and plants is one of the main achievements in the human history. There are different ways to define the domestication but can be understood as a relation of mutualism between the domesticator (human) and the domesticated (animal or plant) in order to obtain an evolutionary advantage in both parts. Domestication had a significant effect not only on domesticated animals and plants but also on human evolution and the environment in general. Domesticated animals have a variety of shared traits among them, behavioral (e.g., decreased aggressiveness), morphological (e.g., brain size) and physiological changes (e.g., increased growth and prolificacy), which have been selected for human benefit, which differentiates them from their wild ancestors (Zeder 2015).

Understanding how animals and plants respond to human manipulation, or domestication, is very relevant for different purposes such as improving existing crops and livestock and to be able to domesticate new species. Therefore, exploring the basic concepts of domestication provides an great opportunity to examine the interaction between humans and the environment, and how the evolution of the human culture interacts with the biological evolution (Zeder 2015).

1.2.1. Different ways for the animal domestication

In general it is thought that domestication is gradual over time and is characterized by a relationship that gradually intensifies between animals and humans (Vigne 2011). Thanks to the study of genetic sequences it has been possible to determine that domestication occurred by multiple independent events, as demonstrated in pigs (Larson et al. 2005), goats (Luikart et al. 2001), sheep (Pedrosa et al. 2005), horses (Vilà et al. 2001), and cows (Hanotte et al. 2002).

In addition to having multiple events of domestication, it has been observed that domestication occurred in three different ways (Zeder 2012; Larson and Fuller 2014):

- Commensal way. This type of domestication is not sought by humans consciously. This process occurs due to the change of the environment of the animals by humans, which causes some animals to be attracted by different elements of the human environment, such as food waste. Following this, the reciprocal relationship between animals and humans laid the foundations of domestication, in particular, captivity and controlled breeding was developed. In this case, therefore, animal domestication is a co-evolutionary event in which one species adapts to the environment of another species that is in a process of evolution. In most cases, once the animals are part of the human society, phenotypic differences with their wild ancestral can be so large as to give them a separate taxonomic name (Gentry, Clutton-Brock, and Groves 2004).
- Prey way. Although this type of domestication was initiated by humans, the intention was not to domesticate the animals, but to obtain resources more efficiently. This approach was followed mainly with medium and large herbivores, which were hunted until this moment. The human being probably changed his way of hunting to have more availability of the prey and thus increase the supply of the resource. In this way, they managed herds and controlled the diet and reproduction of the animals (Zeder 2012).
- Directed way. This type of domestication was intended by humans. Once there were already domesticated animals and plants, this made humans more likely to domesticate other species, such as the horse, the donkey and the camel, which went from being dams to transportation sources. Although there are some species that have never been domesticated, such as gazelles (Zeder 2006) and zebras (Diamond 2002), most of the domesticated animals of the last centuries were domesticated deliberately, examples of these animals are many small pets like hamsters (Fritzsche et al. 2006) and some marine species (Duarte, Marba, and Holmer 2007).

1.2.2. Origins of domestication traits

Traits modified by the domestication process can be broadly classified as those related directly to the domestication process (say changes in temperament, lack

of fear to humans) and those that were deliberately targeted by humans (say a continuous oestrus in small ruminants) (Olsen and Wendel 2013).

Captive animals can accumulate previously deleterious mutations because of the relaxation of natural selection, an example could be new phenotypes that can please people such as fancy coat colors that would be eliminated in the wild environment (Fang et al. 2009; Cruz, Vilà, and Webster 2008; Zeder 2012). Due to these changes, genomes of domestic animals differentiated further from their wild ancestors. The most early and universal traits of domestication were likely those related to behavior, specifically, those related to the docility of the animal with the human. Many of the selected traits have an economic motivation, such as the production of milk and wool or egg laying. On the other hand, other traits are aesthetic changes without any extra function, like the coat color. However, there are some phenotypic changes that are not deliberated selected but are consequence of the domestication, such as dental irregularities caused by nutritional stress and changes in the morphology due to being used in the transport of heavy loads (Dobney and Ervynck 2000; Rossel et al. 2008).

Due to the advances in sequencing techniques, it has been possible to demonstrate that gene flow is common both between different domestic populations of the same species and between domestic and wild species (Marshall et al. 2014; L. A. F. Frantz, Schraiber, et al. 2015). This hybridization often significantly affected the genomes and phenotypes of domestic animal populations, as has been observed in several cases such as in pigs (Ottoni et al. 2013; Marshall et al. 2014; L. A. F. Frantz, Schraiber, et al. 2015), commercial chickens (Eriksson et al. 2008), bovid species (Hanotte et al. 2002; Verkaar et al. 2004), cats (Pierpaoli et al. 2003) and horses (Jordana, Pares, and Sanchez 1995). The hybridization between populations is also generalized in plants like grapes (Myles et al. 2011), apples (Cornille et al. 2012), maize (van Heerwaarden et al. 2011) and rice (Nuijten et al. 2009).

Domestication, in addition to affecting animals and plants at the genotypic and phenotypic level, also has an impact on the environment (Zeder 2015). In particular, the activities to improve the performance and provision of resources of economic interest can have strong impacts on natural environment. Examples of these activities are the burning of vegetation to increase the abundance of plants

and herbivorous animals of economic importance and the modification of landscapes to improve water supply or expand the habitat zones of domesticated plants and animals (B. D. Smith 2011).

1.3. Pig as model of domestication

1.3.1. Pig demographic history

The pig (*Sus scrofa*) is an *Eutherian* mammal, member of the *Suidae* family, which belongs to the *Cetartiodactyla* order, originated 20-30 Mya (L. A. F. Frantz et al. 2016). *Sus scrofa* is the only member of the *Suidae* family that has been domesticated. The pig emerged in the Southeast Asia ca. 4 Mya, during the Pliocene, and migrated towards the west colonizing almost the Eurasia mainland and North Africa ca. 1.2 Mya. The European and Asian wild boars diverged soon after the colonization, which resulted in differences between the two groups at the genomic and phenotypic levels, in addition to those related to the effective population size and the demographic history. Within Asia it is observed a clear differentiation between northern (North China, Japan and Tibet) and southern (South China) populations of wild boars, with an estimated divergence of 0.6 Mya (Groenen et al. 2012; L. A. F. Frantz et al. 2013).

During the Pleistocene there was a replacement of most *Sus* species by *Sus scrofa*. Due to this replacement and the colonization of all Eurasia and North Africa, the pig population expanded before it suffered a population bottleneck 20,000 years ago, coinciding with the last glacial maximum (Groenen et al. 2012). Low temperatures caused isolation of wild boars in different regions of Europe (Iberia, Italy and Balkans) (Scandura et al. 2008), however, in Asia it affected mainly the northern than the southern populations of wild boars (L. A. F. Frantz et al. 2013; L. A. F. Frantz, Madsen, et al. 2015). This severe bottleneck followed by recent inbreeding in European pigs resulted in a much lower genetic diversity in the European wild boars than in the Asian ones (Groenen et al. 2012). While this difference in the diversity is also observed between the domestic breeds of both continents, the recent gene flow of Asian domestic pigs to European domestic breeds led to greater nucleotide diversity in European domestic pigs

than in European wild boars (Giuffra et al. 2000; L. A. F. Frantz, Schraiber, et al. 2015).

1.3.2.Pig domestication

Sus scrofa is one of the first species used in agriculture that was domesticated and, of the domesticated species, is one of the most economically important around the world. There is evidence that the pig was domesticated in both Asia and Europe throughout multiple independent domestication events on each continent ca. 9,000 years ago (Larson et al. 2007; Megens et al. 2008; Groenen et al. 2012), several domestication centers across Eurasia have been proposed, from where pig migrated after domestication, as reviewed in Ramos-Onsins et al. (2014, Figure 1.4). As described above, domestication is not an instantaneous process, but a long-term period in which hybridization between wild and domestic animals exists. In the pig in particular, it has been reported admixture between wild boars and domestic pigs (D. J. Goedbloed et al. 2013; D. J. Goedbloed et al. 2013b; A. C. Frantz, Massei, and Burke 2012). Nowadays, domestic pig consists of many breeds that have been separated and isolated for a long time, which has caused genetic differences between them.

During the 18th and 19th centuries, there were important changes in the pig production such as the intensive production of confinement that increases the size of flocks and herds (White 2011; Wealleans 2013). These changes, together with the arrival of the Asian pigs in Europe and their crossing with some European breeds to improve the growth and the litter size, resulted in a high performance of these breeds (Giuffra et al. 2000; Megens et al. 2008; Ai et al. 2015). The crossing with Asian pigs has caused that most European breeds have ~20-30 % of their genome introgressed with Asian genes. Selective strategies were used in these breeds, resulting in a wide variety of improved breeds that are currently used internationally in animal production. One example of these international commercial breeds is Large White, which is estimated that approximately 30% of its genome is of Asian origin (Bosse et al. 2014). These breeds are phenotypically different between them and they have been selected for different traits, such as

meat quality, growth, reproductive performance or immunity (White 2011; Amaral et al. 2011; Wilkinson et al. 2013).

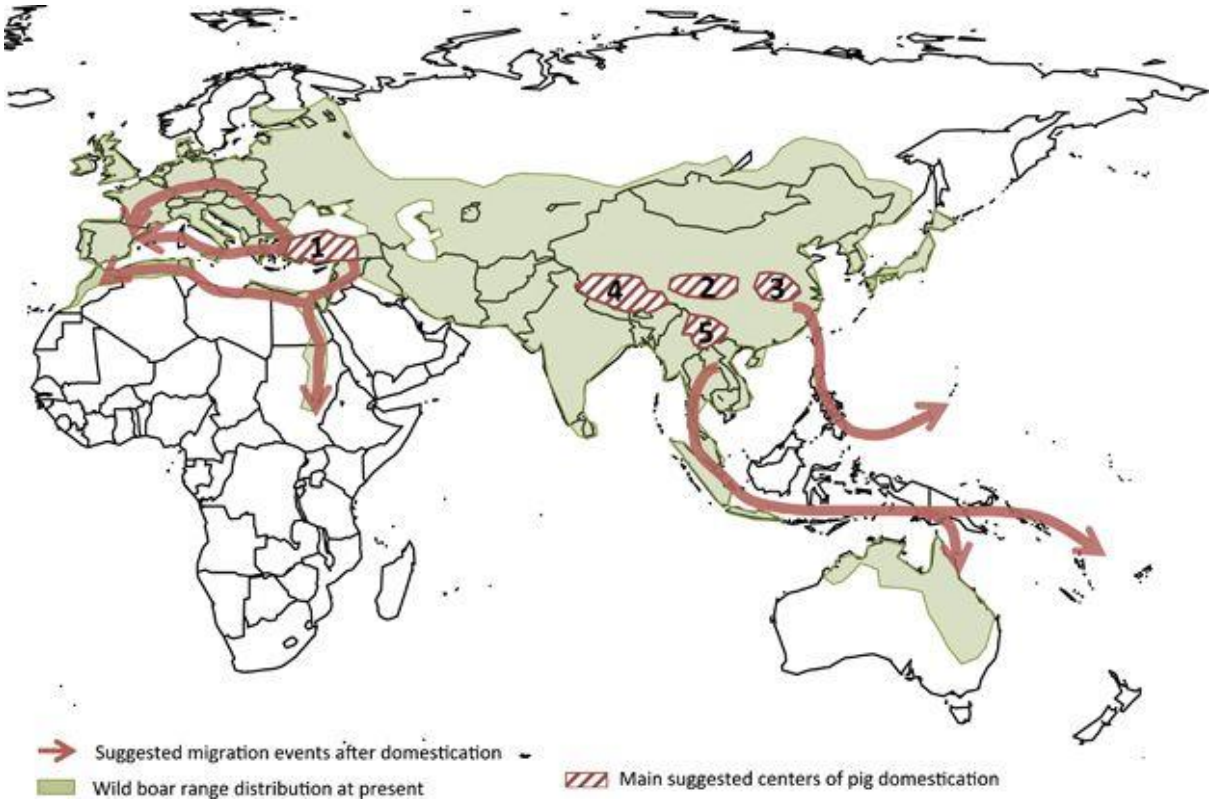

***Figure 1.4**: Map with the suggested pig domestication centers in Eurasia, indicated with lined areas, and the posterior migration, shown with arrows. Ramos-Onsins et al. 2014*

As described above, the pig species suffered multiple natural demographic events (reductions, expansions, migrations, etc.) and artificial (human-mediated) ones interacting with new environments (isolated domestication by regions, introgression, admixture, etc.). All these events caused a wide diversity of phenotypes adapted to different conditions around the world. It is expected that these phenotypic changes will be related to genotypic changes. Therefore, the study of the diversity in pigs of different regions and breeds, in both domestic and wild populations, can give important information for the study of evolution and adaptation in the species, as demonstrated in different works (Larson et al. 2005; Stoneking and Krause 2011; Bianco et al. 2015; Groenen 2016). In particular, the analysis of variability and homozygous regions allow to detect possible regions under differentiated selection between domestic and wild pigs, thus giving

information of possible signs of domestication (Amaral et al. 2011; Rubin et al. 2012; Groenen 2016).

1.4. Pathway analysis

Most of studies aiming at finding genes involved in selection are based on the detection of outlier loci, obtaining a set of genes putatively under selection. However, genes do not act in isolation but interact with other genes to carry out a biological process within a cell that leads to a product or a change in the cell, this process is generally named as pathway. Some of the most common pathways are involved in metabolism, signal transmission and regulation of gene expression.

Pathway analysis is a good option to increase the biological interpretability, reduce the complexity and increase the explanatory power of the analysis (Daub et al. 2013). In cows, this type of analysis has been used to identify pathways related with milk production (Buitenhuis et al. 2014) and pathways associated to metabolic traits of dairy cows (Ha et al. 2015).

Pathway analysis requires information about the different pathways and the interaction networks. For this purpose, there are several pathways databases with the functionality, structure, interactions and information about the known pathways. Examples of pathway databases are KEGG (http://www.genome.jp/kegg/, Kanehisa et al. 2008) and Reactome (http://www.reactome.org/, Matthews et al. 2009). In addition to these databases, the NCBI Biosytems database (http://www.ncbi.nlm.nih.gov/biosystems/, Geer et al. 2010) centralizes existing pathway databases, containing records from different source databases as those previously described, among others.

In addition of the possibility of analyzing the pathway as a unit, it must be considered that the position of genes within a pathway affects to the strength of selection. Those genes that are more central and linked in a pathway tend to be more evolutionarily constrained than peripheral genes, which as they interact with the environment tend to have a greater adaptive selection (Fraser et al. 2002; Hahn and Kern 2005; Montanucci et al. 2011; Alvarez-Ponce and Fares 2012). Furthermore, several studies have demonstrated that downstream genes have

higher evolution rate than the upstream, due probably to the high pleiotropy of the upstream genes, since they participate in multiple functions and processes (Rausher, Miller, and Tiffin 1999; Riley, Jin, and Gibson 2003; Livingstone and Anderson 2009; Ramsay, Rieseberg, and Ritland 2009).

Chapter 2
Objectives

Objectives

The main objective of this thesis is to evaluate the effect of domestication on the genome of the pig by the analysis of the genetic diversity in wild and domestic populations, and to understand the biological processes in which these signals of domestication are involved.

To achieve this broad objective, we propose these specific objectives using genome-wide sequence in all cases:

- To analyze possible events of selection in swine, taking the pathway as the unit of analysis, and to detect the differences between domesticated and wild pigs.
- To study the selection strength in domestic pigs vs. wild boars.
- To relate the selection pressure of genes with their topological properties in the pathway in which they are involved.
- To develop a tool for the diagnosis of VCF files, which can help to identify systematic biases in the SNP calling process.

Chapter 3

A pathway-centered analysis of pig domestication and breeding in Eurasia

A Pathway-Centered Analysis of Pig Domestication and Breeding in Eurasia

J. Leno-Colorado[1], N.J. Hudson[2], A. Reverter[3], M. Pérez-Enciso[1,4*]

[1] Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Spain

[2] School of Agriculture and Food Science, University of Queensland, Queensland 4343, Australia

[3] CSIRO Agriculture and Food, Queensland Bioscience Precinct, 306 Carmody Rd., St. Lucia, Brisbane, Queensland 4067, Australia

[4] ICREA, Carrer de Lluís Companys 23, Barcelona 08010, Spain

* Author for correspondence: Miguel Pérez-Enciso, Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain; miguel.perez@uab.es

G3: GENES, GENOMES, GENETICS. 2017; 7 (7) : 2171–2184.

Published 2017 July 1. doi: 10.1534/g3.117.042671

Abstract

Ascertaining the molecular and physiological basis of domestication and breeding is an active area of research. Due to the current wide distribution of its wild ancestor, the wild boar, the pig (*Sus scrofa*) is an excellent model to study these processes, which occurred independently in East Asia and Europe ca. 9,000 yr ago. Analyzing genome variability patterns in terms of metabolic pathways is attractive since it considers the impact of interrelated functions of genes, in contrast to genome-wide scans that treat genes or genome-windows in isolation. To that end, we studied 40 wild boars and 123 domestic pig genomes from Asia and Europe when metabolic pathway was the unit of analysis. We computed statistical significance for differentiation (Fst) and linkage disequilibrium (nSL) statistics at the pathway level. In terms of Fst, we found 21 and 12 pathways significantly differentiated at a q-value < 0.05 in Asia and Europe, respectively; five were shared across continents. In Asia, we found six significant pathways related to behavior, which involved essential neurotransmitters like dopamine and serotonin. Several significant pathways were interrelated and shared a variable percentage of genes. There were 12 genes present in >10 significant pathways (in terms of Fst), comprising genes involved in the transduction of a large number of signals, like phospholipase PCLB1, which is expressed in the brain, or ITPR3, which has an important role in taste transduction. In terms of nSL, significant pathways were mainly related to reproductive performance (ovarian steroidogenesis), an important target trait as well during domestication and modern animal breeding. Different levels of recombination cannot explain these results, since we found no correlation between Fst and recombination rate. However, we did find an increased ratio of deleterious mutations in domestic vs. wild populations, suggesting a relaxed functional constraint associated with the domestication and breeding processes. Purifying selection was, nevertheless, stronger in significantly differentiated pathways than in random pathways, mainly in Europe. We conclude that pathway analysis facilitates the biological interpretation of genome-wide studies. Notably in the case of pig, behavior played an important role, among other physiological and developmental processes.

Keywords: behavior, domestication, pathway analysis, pig, genome sequence data

Introduction

Plant and animal domestication were cornerstone events in mankind's recent history (Diamond 2002). By ensuring a continuous and reliable supply of food, domestication allowed a steady increase in human population size that eventually resulted in the first urban societies, thereby facilitating the technology development that characterizes the human species. Although domestication has received considerable interest for many years from multiple disciplines, modern large-scale genomic technologies are shedding new light in a process where many unknowns still remain. This endeavor is largely facilitated in species, such as the pig, where a modern equivalent of the wild ancestor is available for comparison.

There is some ambiguity in defining what is domestication (Zeder 2015), given that many concurrent processes have occurred in the transition between the wild specimens and the individuals bred by humans, and that domestication likely involved gradual discontinuities in gene flow between domestic and wild populations instead of a sudden stop (Frantz et al. 2015). Nevertheless, in animals, there are some shared characteristics among the major domestic species since they have been selected to meet human preferences: Domestic animals have modified behavior, and growth and reproductive distinctive features compared to their wild ancestors. The genetic bases of these traits are clearly polygenic, as is evident for the numerous QTLs that have been identified (www.animalgenome.org/cgi-bin/QTLdb/SS/index). This complicates the discovery of genes underlying their phenotypic variability because small effect sizes are difficult to detect.

Traditionally, studies looking for selective signals have analyzed individual SNPs or carried out a genomic scan in windows of contiguous SNPs of arbitrary size (e.g., Amaral et al. 2011; Burgos-Paz et al. 2013; Rubin et al. 2012). Since genes do not act in isolation but in concerted action with other genes, we argue, as other studies have done (Daub et al. 2013), that analyzing genomic variability patterns from a metabolic pathway point of view should facilitate the biological interpretation of the results. Compared to a genome window analysis, this approach could improve power when individual gene signals are weak. By adding up these individually weak signals in a pathway framework, a global significant statistic can eventually be obtained. Note that a pathway analysis differs from analyzing a posteriori a list of statistically significant

genes using, by instance, gene ontology tools, since here we predefine a list of genes and we then study the collective behavior of genes of the whole list. The criticism by Pavlidis et al. 2012 is then less applicable when a prior hypothesis exists. For gene-set approaches, see review in e.g. Mooney et al. 2014.

Previous studies do show that taking into account how genes interact along metabolic pathways is enlightening. For instance, using a pathway approach, Daub et al. (2013) discovered that adaptation signals in humans, measured by increased differentiation, are enriched for pathogen resistance pathways. However, none of the genes were statistically outliers so probably this observation would have been overlooked had genes been analyzed individually. In cattle, Ha et al. (2015) identified several pathways associated to a number of key metabolites in dairy cows and Buitenhuis et al. (2014) revealed pathways associated to milk production. The main difference between those studies is the criterion that they used to merge individual signals, as numerous variants have been proposed [e.g., Wang et al. (2010)]. Here we used Fisher' statistics to combine several independent Fst values for each SNP into a gene P-value and, subsequently, those gene P-values were combined into a pathway P-value. We argue that combining signals from multiple SNPs should be less prone to false discoveries than taking the single most outlier signal for each gene.

Despite numerous studies in pig domestication [e.g., for a review see Ramos-Onsins et al. (2014)] so far, to our knowledge, pathway analysis has not been applied to improve our understanding of the domestication or breeding processes in this species. The fact that most phenotypic characteristics have a polygenic basis makes pathway analysis an attractive approach, provided that the pathway as a whole better explains the genetic basis of the trait than do individual genes. Here, we have used sequence data from 163 domestic and wild pigs to study how potentially selective processes associated with domestication and ensuing breeding have modeled the pig genome, when viewed from a pathway point of view. By using sequence instead of chips, we further avoid the issue of ascertainment bias, and provide a comprehensive, unbiased portray of nucleotide diversity. Note that a comparison between domestic and wild specimens necessarily confounds domestication and modern breeding signals, and truly disentangling genetic changes due to domestication from those caused by ensuing breeding requires ancient DNA studies at population scale, which is currently unrealistic despite some recent advances (Ramírez et al. 2014). Since we were

predominantly interested in the shared signal left by domestication and breeding across breeds, we tried to minimize the specific breed effects. To this end, we combined genomes from several domestic breeds, sampling evenly the number of specimens per breed.

Material and methods

Pig samples

We analyzed a sample of 163 wild and domestic pig (*Sus scrofa*) genomes (Supplementary Table S3.1, Supplementary Figure S3.1). The 163 pigs were classified into Asian domestic pigs (ASDM, n = 60), Asian wild boars (ASWB, n = 20), European domestics (EUDM, n = 63) and European wild boar (EUWB, n = 20). Asian domestics represented 10 Chinese breeds (Meishan, Bamaxiang, Hetao, Laiwu, Luchuan, Minzhu, Sichuan, Tibetan, Wuzhishan and Yunnan), which were chosen to represent the different geographic locations in China, six samples from each breed. ASWB comprised 10 boars from South China and 10 from the North (North China, Korea and East Russia). European domestic pigs were from all major breeds (Duroc, Landrace, Large White, Pietrain), plus the local breeds Iberian, Mangalica and the American miniature pig Yucatan, of Iberian descent (Burgos-Paz et al. 2013), 10 genomes per breed were chosen except for Mangalica and Iberian, where only 5 and 8 were available, respectively. The 20 European wild boars were from Spain, France, Netherlands, Switzerland, Italy, Greece, Tunisia and Near East. The domestic breeds used in this study are selected for a diversity of traits. For European breeds, meat content and growth are important targets, whereas Chinese breeds tend to be more prolific and fatter than their European counterparts.

Most of the sample sequences were available in public databases (Ai et al. 2015; Bianco, Soto et al. 2015; Esteve-Codina et al. 2013; Groenen et al. 2012; Molnár et al. 2014; Pérez-Enciso et al. 2016; Rubin et al. 2012) and were downloaded from the short read archive (SRA, http://www.ncbi.nlm.nih.gov/sra). Two additional samples (Iberian pig IBGU1805 and a British Large White LWGB0348) were specifically sequenced for this study and have been submitted to SRA (accessions SRX2787051 and SRX2788443 within study PRJNA255085). The VCF files containing both raw and

imputed SNPs are available at https://bioinformatics.cragenomica.es/numgenomics (under the heading “data”).

NGS Bioinformatics

We downloaded and mapped raw reads against the reference assembly (Sscrofa10.2, Groenen et al. 2012) using BWA mem option (Li and Durbin 2009). We removed PCR duplicates using SAMtools rmdup v0.1.19 (Li et al. 2009) and realigned around indels with GATK IndelRealigner tool (McKenna et al. 2010). We called genotypes with SAMtools mpileup and bcftools call v1.3.0 (Li et al. 2009) for each individual separately. To call a SNP, we set the minimum and maximum depths between 5 × and twice the average sample's depth plus one, the minimum SNP quality was 10 in each sample, with the further requirements of minimum mapping quality and minimum base quality of 20. We also called the homozygous blocks, which are the parts of the sequence that are equal to the reference. Since SAMtools does not filter by default these homozygous blocks by depth, we filtered them fitting the same depth and quality requirements as for the SNP calling procedure using “samtools depth” utility, BEDtools (Quinlan 2014) and custom scripts. In this way, both SNPs and homozygous blocks were filtered by the same criteria.

We then merged individual gVCF files into a multi-individual VCF file, with all the SNPs from the 163 samples. For this purpose, we followed a two-step approach resembling closely that in Pérez-Enciso et al. (2016). In summary, first, we generated a fasta file from the gVCF file for each individual and we generated a multi-individual VCF file using the individual fasta file to identify whether a position is equal to the reference, polymorphic or missing. An alternative approach would have been to call SNPs using all samples simultaneously, but this strategy has been shown to have less power and similar type I error than the one followed here, because joint SNP calling is less sensitive to rare variants than individual calling (Nevado, Ramos-Onsins, and Perez-Enciso 2014). Furthermore, Asian and European samples are highly divergent and multi-sample algorithms are optimized for single population analyses.

Once the multiple sample file was obtained, we discarded the singletons, SNPs in sex chromosomes, and the SNPs with >30% of missing data of the samples in each group (ASDM, ASWB, EUDM, EUWB). If a given SNP was not called in at least ≥30% of

samples in all groups, it was discarded from further analyses. Finally, we imputed the missing genotypes and inferred phases with Beagle 4.0 (Browning and Browning 2013). We annotated SNPs with Ensembl's Variant Effect Predictor (McLaren et al. 2010). This tool also classifies non-synonymous variants as tolerated or deleterious based on their SIFT scores (Sim et al. 2012), which predicts whether an amino acid substitution affects protein function. For each gene, we computed the ratio of deleterious vs. tolerated SNPs. These statistics were computed for each population (ASDM, ASWB, EUDM, EUWB) separately. R (R Development Core Team and R Core Team 2014) was used to obtain a "heatmap" to represent Euclidean distances between samples' genotypes.

Differentiation and disequilibrium metrics

Selection increases differentiation at positively selected loci between a control population and a population where the loci are beneficial, also causing an increase in linkage disequilibrium around selected haplotypes. These two well-known phenomena (e.g., Sabeti et al. 2006) can be captured by either Fst (allele frequency differentiation) or haplotype based tests, such as nSL (Ferrer-Admetlla et al. 2014). Since the pig was independently domesticated in Asia and in Europe (Larson et al. 2005), we computed Fst (Weir – Cockerham estimate, Weir and Cockerham 1984) between wild and domestic populations in each continent separately, Asia and Europe, using VCFtools (Danecek et al. 2011). The nSL metrics is designed to detect the positive selection signal due to an increase in haplotype homozygosity; for this purpose, nSL measures the length of a segment of haplotype homozygosity in terms of number of mutations. We calculated the statistics with the program *nSL* (http://cteg.berkeley.edu/software.html) within the four different populations of interest (ASDM, ASWB, EUDM and EUWB); the statistics was normalized according to derived allele frequency in ten bins of size 0.10. The ancestral allele is needed for the nSL statistics and was inferred from a consensus outgroup allele, as explained in (Bianco, Nevado, et al. 2015). The consensus was obtained from several species: *S. barbatus*, *S. cebifrons*, *S. verrucosus*, *S. celebensis*, and African warthog, (*Phacochoerus africanus*). The divergence between the different *Sus* species is ~4.2 MYA, whereas that of *Sus* with warthog is ca. 10 MYA (Frantz et al. 2016). We removed those SNPs

for which the ancestral allele could not be reliably identified or with more than two alleles. For each gene, we assessed the average recombination rate based in the linkage map by Tortereau et al. (2012). This map was based on four different F2 crosses between European and Chinese breeds; total autosomal length was ~20M. We obtained a smoothed recombination rate using *loess* R package, to minimize the effect of gaps in the recombination map.

Pathway analysis

We downloaded the complete dataset with pig pathways and genes from NCBI Biosystems v.20160202 (Geer et al. 2010). The downloaded file contained 1789 pathways and 7157 genes. The median number of genes per pathway was 47 and ranged from 1 to 1519. The NCBI biosystems database contains records from different source databases, such as KEGG (http://www.genome.jp/kegg/, Kanehisa et al. 2008), REACTOME (http://www.reactome.org/, Matthews et al. 2009) or WikiPathways (http://www.wikipathways.org/, Pico et al. 2008), which are often redundant. For this reason, we filtered the pathways according to their size and redundancy in two steps. First, we removed pathways with <10 and >150 genes (150 corresponds to two SD in the distribution of number of genes per pathway); this was aimed at discarding pathways that were either not informative or too generic and complex. For instance, among the pathways with over 150 genes we find: metabolic pathways, gene expression, metabolism, hemostasis, immune system, and neuronal system. Second, for pathways sharing >50% of their genes, we selected the largest one.

We obtained an empirical P-value for Fst and nSL for each pathway following Dall'Olio et al. (2012). First, an empirical P-value for each SNP was obtained by ranking the statistics (Fst or nSL). Thus, a SNP with Fst (or nSL) ranked as the i-largest out of N SNPs, was assigned a P-value of i/N. Next, we obtained a gene P-value with Fisher' statistics, which combines several independent P-values:

$$x = -2\sum_{j=1}^{S} log(P_j) ,$$

where *S* is the number of SNPs for the gene analyzed (i.e., those within the gene boundaries in Ensembl database) and P_j each associated P-value; since *x* is distributed as a χ^2 with 2*N* d.f., we can obtain a combined P-value for the gene. In a

second step, we repeated the same procedure by combining the P-values of each gene in the pathway to obtain a pathway P-value. The actual significance of this P-value is difficult to interpret, since the null-hypothesis is not clearly defined and therefore we carried out permutations to determine significance. Since each pathway differs in number of genes, we carried out 1000 permutations for random gene-sets of sizes 10-150 genes and differing by increments of 10 genes. In these permutations, dummy pathways were assembled using the P-values of randomly sampled genes, and the actual pathway P-value was compared with the null distribution obtained by permutation. To account for multiple testing, we used the q-value (Benjamini and Hochberg 1995), computed with R-package *qvalue* (Storey et al. 2015), to determine significant pathways using the P-values obtained by permutation.

Critically, Fisher' statistics is based on the premise of independence between P-values, and this is not guaranteed with sequence data given the extreme disequilibrium between nearby SNPs. To avoid this, we pruned the SNP dataset by selecting those positions that minimized linkage disequilibrium using PLINK v.1.9 program (Chang et al. 2015) setting the variant inflation factor (VIF) equal to 2. With this approach, the P-value obtained from Fisher' statistics was independent of the number of SNPs for each gene (Figure S3.2). It should be mentioned that the nSL statistics were computed using all SNPs for which the ancestral allele could be determined, since nSL measures LD in number of SNPs units, but only the values for SNPs in equilibrium were retained to obtain the gene P-value, as for Fst metrics.

To investigate whether significance could be due (partly) explained by Asian introgression in European Domestics, we carried out a semisupervised ADMIXTURE (Alexander, Novembre, and Lange 2009) analysis. We extracted SNPs from all genes pertaining to the given pathway and we run ADMIXTURE with K = 2. We run a semisupervised analyses where all European wild boars were assigned K = 1 and all Asian pigs K = 2, and we let the program compute the fraction of the European domestic genomes due to Asian origin. We did this for each significant pathway and for a random set of pathways with similar number of genes,

Further, we built a co-association network to visualize pathway relationships. Significant pathway to pathway connections were identified using the PCIT network inference algorithm (Reverter and Chan 2008). The PCIT algorithm is a soft-

thresholding method that exploits the twin concepts of Partial Correlation and Mutual Information. In brief, it explores relationships between all possible triplets of nodes (i.e., pathways in our context), in an attempt to determine truly informative correlations between node pairs once the numerical influence of other nodes in the system has been accounted. Clustering was based on ten variables per pathway: the six pathway P-values for Fst (one value per continent) and nSL (one value per population) metrics, and nucleotide diversity in each of the four populations (ASWB, ASDM, EUWB and EUDM), averaged for each gene in the pathway. We estimated Tajima's nucleotide diversity (Tajima 1983, 1989) per gene per population with the methods developed by Ferretti, Raineri, and Ramos-Onsins (2012), which account for missing data, using mstatspop software (Ramos-Onsins, unpublished data, available at http://bioinformatics.cragenomica.es/numgenomics/people/sebas/). We visualized the resulting network using Cytoscape (www.cytoscape.org, (Shannon et al. 2003)). In the visualization scheme, we mapped the pathways (nodes) to a series of attributes to help identify emerging properties. These included number of genes in the pathway, pathway source (KEGG or REACTOME), population with lowest Fst P-value (Asia or Europe), and pathway nucleotide variability.

Results and discussion

Genetics mirrors geography, to an extent

Out of the 163 genomes, we initially identified 71,458,035 autosomal SNPs. After quality filtering, removing the positions with >30% of missing data and discarding singletons, these were reduced to 48,008,185 SNPs. Of those, 31,363,201 (65%) were annotated in dbSNP (https://www.ncbi.nlm.nih.gov/SNP), and the ancestral allele could be determined in 44,417,146 sites. By continent, Asia had a much larger number of private SNPs than Europe, 25,258,008 vs. 5,726,610, as expected from the fact that the species is of Asian origin and that European populations suffered a strong bottleneck, as has been observed in previous studies (Bianco, Nevado, et al. 2015).

A heatmap of genetic distances between all samples and SNPs showed the well-known split between Asia and Europe (Figure S3.3). The European heatmap (Figure 3.1A) shows that the main division is between wild boar and local breeds Iberian, Mangalica and Yucatan vs. International pig breeds Pietrain, Landrace, Large White

and Duroc. Further, all EUWBs were clustered together except the two Near East wild boars, which were grouped in a separate branch. The Yucatan, a miniature pig developed in the USA starting with local Mexican pigs and that still retains an important percentage of ancestry from Iberian pigs (Burgos-Paz et al. 2013) formed a separate group but closer to local pigs than to international breeds. Among those, Duroc was genetically more separate from the rest of international pig breeds.

The picture was somewhat more complex in Asia (Figure 3.1B), although pigs were also grouped by breed. In contrast to Europe, though, we observed a genetic split between North and South wild boars, in agreement with previous results (Ai et al. 2015). Nevertheless, this geographic pattern was not so evident among the domestic pigs, e.g., North Asian breeds (Laiwu, Hetao and Minzhu), which are less separated from those from the South, compared to wild populations.

Pathway statistics

We retrieved 1789 pathways comprising 7157 genes from NCBI database, which were reduced to a final set of 442 pathways with 5713 genes after filtering (Table S3.2) by size (e.g., number of genes) and redundancy. Note that only 25% of pathways but 80% of genes were retained, showing the large redundancy in terms of genes across pathways. Most discarded pathways (676) were very small and contained <10 genes. The distribution of genes per pathway was highly leptokurtic (Figure S3.4).

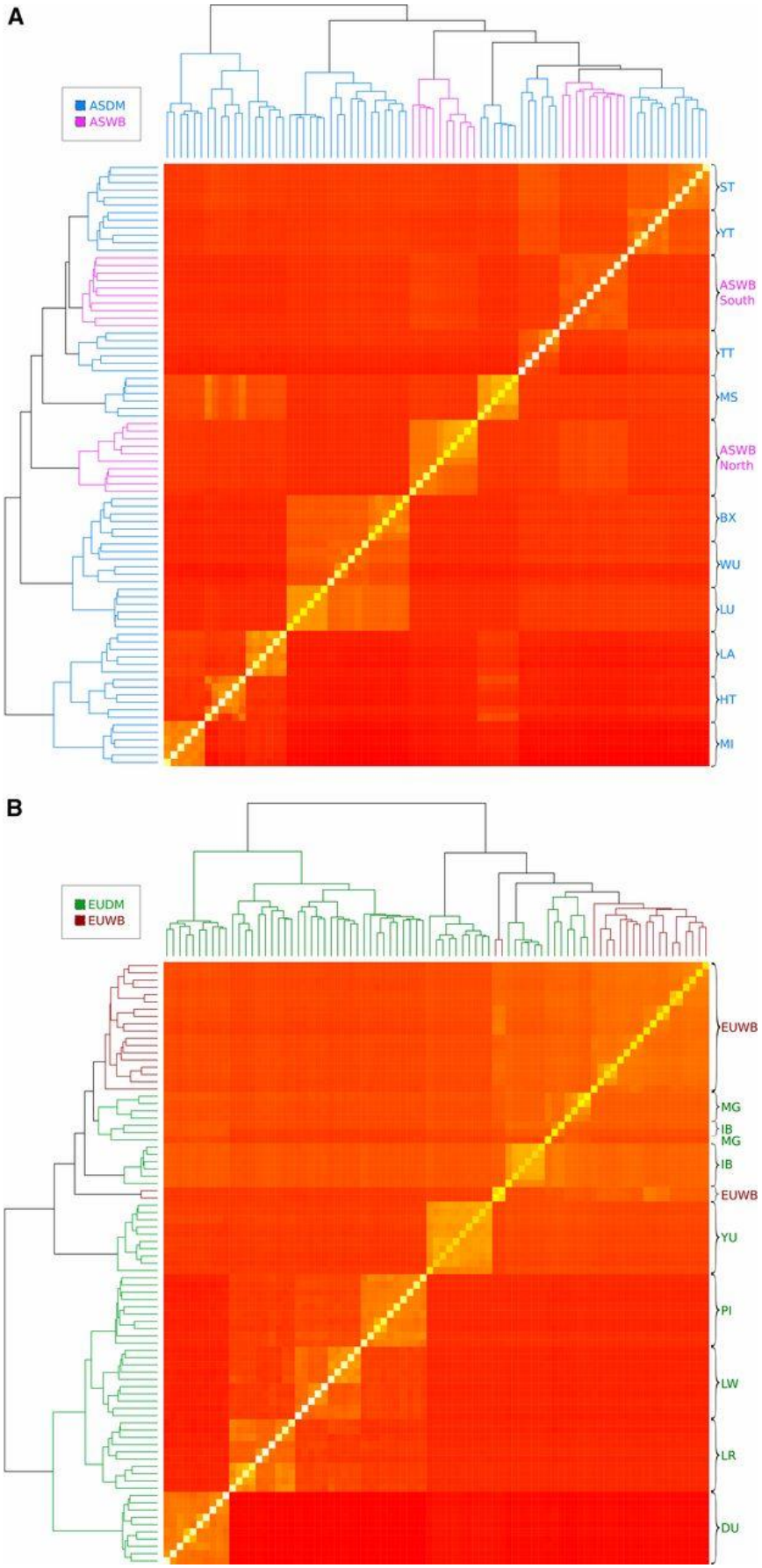
A
ASDM
ASWB
ST
YT
ASWB South
TT
MS
ASWB North
BX
WU
LU
LA
HT
MI
B
EUDM
EUWB
EUWB
MG
IB
MG
IB
EUWB
YU
PI
LW
LR
DU

***Figure 3.1**: A) Heatmap of the European individuals using the molecular relationship matrix, computed using all available autosomal SNPs. B) Heatmap of the Asian pigs. In Europe, breed codes are DU, Duroc; IB, Iberian; LR, Landrace; LW, Large White; MG, Mangalitza; PI, Pietrain; YU, Yucatan minipig. In Asia, breed codes are BX, Bamaxiang; HT, Hetao; LA, Laiwu; LU, Luchuan; MI, Minzhu; MS, Meishan; ST, Sichuan; TT, Tibet; WU, Wuzhishan; YT, Yunnan. Colors are used to differentiate among the populations: ASDM (blue), ASWB (purple), EUDM (green) and EUWB (dark red).*

Differentiation metrics (Fst)

Differentiation (Fst) analysis indicates that allele frequency changes occurred in pathways associated with some important biological processes (Table 3.1). We found more significant pathways in Asia than in Europe; there were 21 pathways significantly differentiated at a q-value < 0.05 in Asia and 12 in Europe, involving a total of 1065 and 576 genes, respectively. Pathways were predominantly continent-specific, but five pathways were differentiated in both continents: integrin cell surface interactions, insulin secretion, pancreatic secretion, ABC transporters, and glutamatergic synapse. Our results are unlikely an artifact caused by differential recombination rate, as we found no correlation between Fst and recombination rate (Figure S3.5). This contrasts with what has been observed in humans (Keinan and Reich 2010).

In Asia, we found six significant pathways related to behavior (serotonergic synapse, dopaminergic synapse, glutamatergic synapse, opioid signaling, long-term depression, and adrenergic signaling in cardiomyocytes). This is remarkable since it has long been recognized that domestication has affected behavior, yet the genetic basis for these changes has not been convincingly identified. The six pathways included a total of 264 genes, which codify for proteins involved in the metabolism of important neurotransmitters like serotonin, dopamine and L-glutamate. Serotonin and dopamine are involved in aggression (serotonin) and reinforcement and reward (dopamine), whereas L-glutamate is the major excitatory neurotransmitter in the central nervous system. Clearly, aggression and reward must have played a role at least during the early stages of domestication and the genetic causes are likely shared between all domestic breeds. Pathway “adrenergic signaling in cardiomyocytes” involves several adrenaline receptors and calcium channels such as ryanodine receptor 2 (RYR2). While RYR2 is primarily expressed in cardiomyocytes, its isoform RYR1 is expressed in skeletal muscle and is well known in pig genetics for being responsible for the pale, soft and exudative syndrome (Fujii et al. 1991).

For the six behavior pathways, Figure 3.2 shows the P-values of all genes that were significant at the 1% nominal level either in Europe, in Asia or both. Although behavior pathways were mainly significant only in Asia (except glutamatergic synapse, Table 3.1), several individual genes were significant in both continents, foremost phospholipase C β 1 (PLCB1), which was significant in both continents and was present in all six behavior pathways. It can be suspected that these pathways were significant only because they contained PLCB1 gene; however, PLCB1 was involved in a total of 35 pathways, and only 14 were significant (q-value < 0.05). Furthermore, we also computed the pathway P-value excluding PLCB1 and we found only a modest decrease in significance (Table S3.3). With our approach, it is unlikely then that a single gene is responsible for significance at the pathway level. Note that this is reasonable under a multi-cause mechanism but may prevent from identifying pathways where a single gene is the main responsible for the rate limiting the whole pathway. In all, PLCB1 plays an important role in the intracellular transduction of many extracellular signals mediated by calcium. It cleaves PIP2 molecule into IP3 and DAG. DAG, together with Ca^{2+} (its secretion is activated ITPR3, also significant in Europe), activates PKC, which plays a central role in activating numerous functions such as transcription, immune response, growth, learning, and smooth muscle contraction. This explains its presence in so many different pathways. Importantly, PLCB1 is expressed in select areas of the brain, including cerebral cortex, hippocampus, amygdala, lateral septum, and olfactory bulb (Koh et al. 2007). In humans, deficiencies in this gene are associated with some kinds of epilepsy (Ngoh et al. 2014). Another interesting and significant gene in both continents was GSK3B (Glycogen synthase kinase-3), which is involved in energy metabolism, neuronal cell development and body pattern formation.

The rest of significantly differentiated pathways comprises pathways related to glucose metabolism (insulin and pancreatic secretion) and development (Wnt signaling, Hippo signaling and axon guidance) in Asia, and recombination or muscle contraction in Europe. Hippo and Wnt signaling pathways are intimately related, and half of all significant genes were shared (Figure S3.6). In contrast, vascular smooth muscle contraction and muscle contraction share only 16 significant out of 85 and 117 genes, respectively. Significant genes in insulin pathway include PLCB1, RYR2 as well as potassium and calcium channels that act on insulin granules and insulin transcription

(KCNN1, KCNN2, KCNMB1). Hippo signaling pathway controls organ size, a fundamental target during domestication and modern breeding. Wnt signaling in turn is one of the most relevant and highly conserved signal transduction pathways, and it has a fundamental role in embryonic development. Hippo and Wnt are tightly interconnected signaling cascades, although their mechanisms differ: Hippo is mainly sensitive to cell density, whereas Wnt responds to concentrations of specific proteins (Irvine 2012). Interestingly, there is also a direct relation between Wnt and insulin pathways, as Wnt signaling increases cell's insulin sensitivity. The three most significant genes in the Wnt pathway were PLCB1 (shared with other pathways, see Figure 3.2 and Figure S3.5), inversin, which contains calmodulin domains and is involved in renal development, and GSK3B, also involved in body pattern formation. GSK3B is also a negative regulator of glucose hormone control.

It is finally worth mentioning two significant pathways involved in recombination, "DNA double strand break repair" and "non-homologous end-joining" (Figure S3.7, which also shows the rest of significant Fst pathways). The issue of the effect of recombination on domestication has been debated for a long time in the literature. Theoretical models have predicted that domestication should increase recombination as rapid selection favors indirectly an increased recombination rate such that Hill-Robertson effect is less limiting for response, and this prediction has been confirmed using chiasma data from the literature (Ross-Ibarra 2004). In a classical paper, Ollivier (1995) also showed that wild boar linkage map was ~33% shorter than domestic pig maps. Nevertheless, other recent studies in sheep, goat and dogs ruled-out changes in recombination rates compared to their wild ancestors (Munoz-Fuentes et al. 2015). Therefore, the significant differentiation found here in this pathway may not be paralleled with changes in recombination rate caused by domestication.

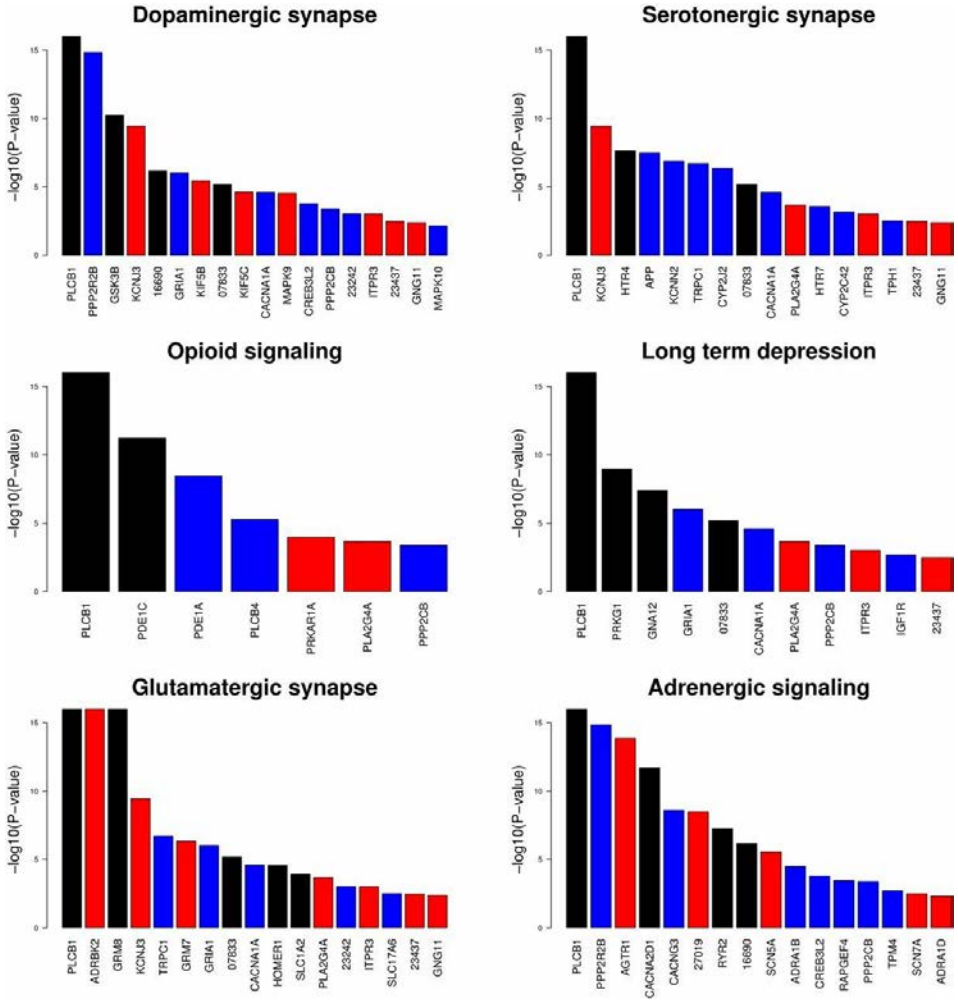

***Figure 3.2**: Gene P-value (-log10) of significant genes at the 1% nominal level in Europe (red bars), in Asia (blue bars) or both continents (black bars) from the significant pathways involved in behavior. When a gene was significant in both continents, the smallest P-value is plotted.*

***Table 3.1**: Significant pathways (q-value < 0.05) obtained in the Fst analysis in Asia and/or Europe.*

Biological process	Pathway name	NCBI ID	Number of genes	P-value* Asia	q-value Asia	P-value* Europe	q-value Europe
Behavior	Opioid Signaling	1337511	46	0.0005	0.010	0.1300	0.686
Behavior	Glutamatergic synapse	213816	75	0.0005	0.010	0.0005	0.018
Behavior	Dopaminergic synapse	469195	88	0.0005	0.010	0.0460	0.495
Behavior	Serotonergic synapse	525344	75	0.0005	0.010	0.1710	0.734
Behavior	Long-term depression	84497	41	0.0005	0.010	0.0230	0.442
Behavior	Adrenergic signaling in cardiomyocytes	908279	104	0.0005	0.010	0.0440	0.495
Biological regulation	Renin secretion	1223594	42	0.0140	0.245	0.0005	0.018
Biological regulation	Phosphatidylinositol signaling system	84464	70	0.0470	0.438	0.0005	0.018
Cell communication	Cell-Cell communication	1336387	61	0.1820	0.570	0.0005	0.018
Cell communication	Integrin cell surface interactions	1337048	51	0.0005	0.010	0.0005	0.018
Cellular process	Assembly of the primary cilium	1336230	123	0.0005	0.010	0.0390	0.479
Cellular process	Hippo signaling pathway	749791	30	0.0005	0.010	0.1590	0.725
Cellular process	Wnt signaling pathway	84473	89	0.0005	0.010	0.2010	0.753
Cellular process	Axon guidance	84477	92	0.0005	0.010	0.0330	0.456

Immune response	Complement cascade	1336947	23	0.1350	0.535	0.0005	0.018
Immune response	Fc-gamma receptor (FCGR) dependent phagocytosis	1336978	36	0.0005	0.010	0.0370	0.467
Immune response	Chagas disease (American trypanosomiasis)	147807	83	0.0005	0.010	0.5210	0.898
Metabolic process	Glycosaminoglycan metabolism	1336589	81	0.0005	0.010	0.2230	0.757
Metabolic process	Phospholipase D signaling pathway	1311111	104	0.1600	0.542	0.0005	0.018
Metabolic process	G alpha (s) signaling events	1337504	88	0.0005	0.010	0.2840	0.810
Metabolic process	Pancreatic secretion	169304	59	0.0005	0.010	0.0005	0.018
Metabolic process	Insulin secretion	777548	60	0.0005	0.010	0.0005	0.018
Muscle contraction	Muscle contraction	1337146	117	0.5250	0.856	0.0005	0.018
Muscle contraction	Vascular smooth muscle contraction	96246	85	0.0440	0.430	0.0005	0.018
Regulation of transcription	Nuclear signaling by ERBB4	1337437	23	0.0005	0.010	0.1310	0.686
Transport	ABC transporters	84452	27	0.0005	0.010	0.0005	0.018

* P-value obtained from permutations.

Linkage disequilibrium metrics

We repeated the same statistical procedure as for Fst with the nSL statistics, which measures linkage disequilibrium instead of differentiation, and that is especially powerful to identify soft sweeps (Ferrer-Admetlla et al. 2014). Each of the four populations, ASDM, ASWB, EUDM and EUWB, was analyzed separately. Overall, there were fewer significant pathwa ys at a q-value < 0.05 with nSL than with Fst (Table 3.1 *vs.* Table 3.2). In particular, we did not find a significant value neither in EUWBs nor in ASWBs, perhaps because there were fewer wild than domestic pigs. In comparison to Fst, concordance between continents with nSL was very high in domestic pigs, as we found the same six out of seven significant pathways in both Asia and Europe. The only exception was pathway "Inflammatory mediator regulation of TRP channels" involved in immune response, which was significant only in Asia. A potential matter of concern with pathway analysis is its definition. As an example, we found that arachidonic acid metabolism pathways annotated by KEGG (NCBI id 84417) and REACTOME (NCBI id 1336691) contained 24 shared genes out of a total of 47 and 39, respectively. Nevertheless, the significant genes (P-value < 0.01) of both pathways were the same. As noted by Mooney et al. (2014), different databases may contain different genes to represent the same biological process; this is a warning to the fact that pathway definition is not an unambiguous concept, and different criteria can legitimately be used to define a given biological process.

Two of the significant pathways are directly linked to reproductive performance ("Ovarian steroidogenesis" and "Steroid hormone biosynthesis"). Importantly, we also identified ovarian steroidogenesis in our previous work (Pérez-Enciso et al. 2016) in a much smaller study on domestication merging Asian and European domestics vs. Asian and European wild boars, and where a completely different analytical approach was employed. The remaining significant pathways were related to lipid metabolism, in particular to linoleic and arachidonic metabolism. Importantly, some of the final products of linoleic metabolism are THF-diols, which are converted into prostaglandins and are involved in sexual behavior of males and ovarian cycle in females. Therefore, most pathways identified with nSL are interrelated and linked to reproduction.

Among the most significant genes in ovarian steroidogenesis, there appears the uncharacterized gene ENSSSCG00000003824. This gene seems to be orthologous

to UGT2B (UDP glycosyltransferase), a cluster of genes involved in the glucuronidation of estrogens. Figure 3.3 shows the significant gene P-values for the significant nSL pathways. It is interesting to remark that a more coherent signal across continents emerge with nSL than with Fst metrics, since the most significant genes with nSL are shared between continents (e.g., compare Figures 3.2 and 3.3).

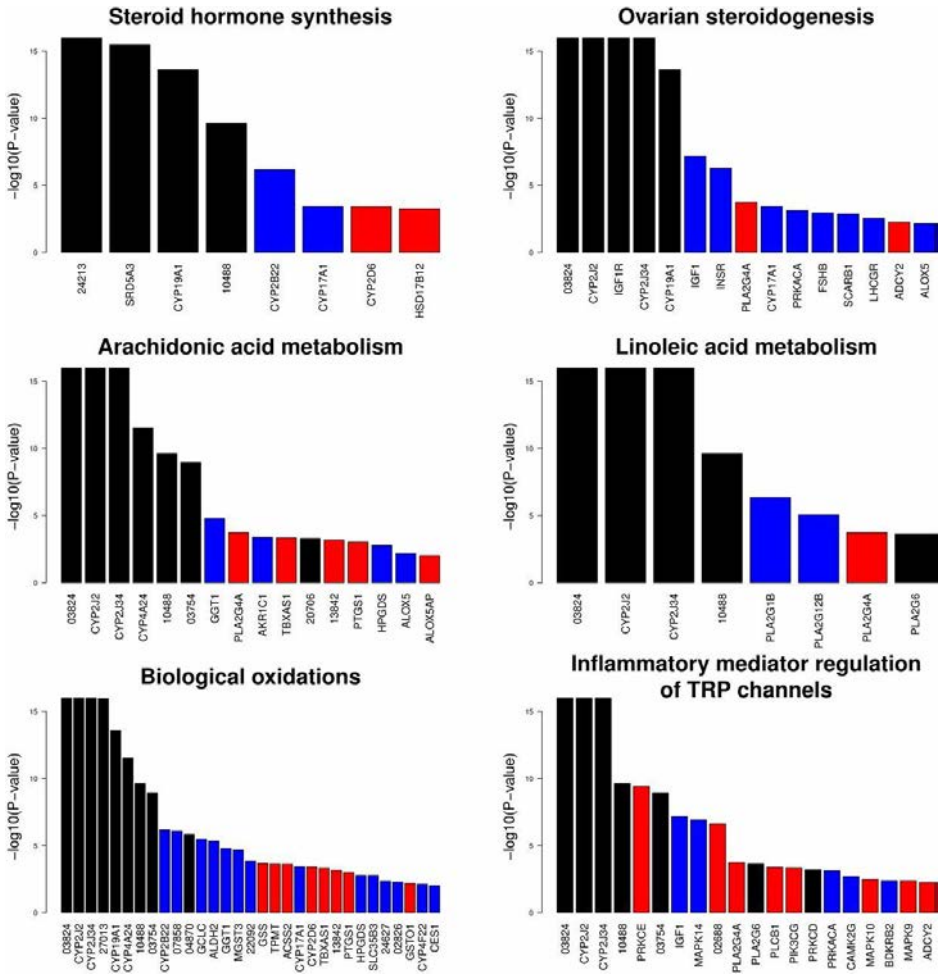

***Figure 3.3**: Significant genes at the 1% nominal level either in Europe, in Asia or both, present in the significant pathways obtained from the nSL analysis.*

***Table 3.2**: Significant pathways (q-value < 0.05) obtained in the nSL analysis for the four populations, according to continent Europe (EU) / Asia (AS) and domestic (DM) / wild (WB) status.*

Biological process	Pathway name	NCBI ID	Number of genes	P-value* ASDM	q-value ASDM	P-value* ASWB	q-value ASWB	P-value* EUDM	q-value EUDM	P-value* EUWB	q-value EUWB
Behaviour / Reproduction	Ovarian steroidogenesis	791446	30	5e-04	0.032	0.012	1	5e-04	0.037	0.524	1
Biological regulation	Biological oxidations	1336802	120	5e-04	0.032	0.148	1	5e-04	0.037	1	1
Hormone synthesis	Steroid hormone biosynthesis	84375	32	5e-04	0.032	0.029	1	5e-04	0.037	0.457	1
Immune response	Inflammatory mediator regulation of TRP channels	948291	73	5e-04	0.032	0.259	1	0.120	1.000	1	1
Lipid metabolic process	Arachidonic acid metabolism	1336691	39	5e-04	0.032	5e-04	0.110	5e-04	0.037	0.998	1
Lipid metabolic process	Arachidonic acid metabolism	84417	47	5e-04	0.032	0.041	1	5e-04	0.037	0.977	1
Lipid metabolic process	Linoleic acid metabolism	84418	24	5e-04	0.032	5e-04	0.110	5e-04	0.037	0.755	1

* P-value obtained from permutations.

Pathways are interrelated

Much as genes do not act in isolation, neither do pathways. To represent this, a co-association network was built with all 35 significant pathways, either with Fst or nSL analysis (Tables 3.1 and 2 merged). The metrics used for the clustering contained the pathway Fst and nSL P-values together with nucleotide variabilities (*see Materials and Methods*). The entire network of pathways contained 83 negative and 129 positive connections. Note that the interpretation of a "positive" or "negative" connection is not straightforward, as is often the case in multivariate methods. The sign would indicate that domestication and/or breeding has exerted similar or opposite changes in the variables used to build the network, conditional on the fact that pathways are significant in at least one analysis. The three most connected pathways, each with 19 connections, were arachidonic acid metabolism (NCBI id 1336691) with 45 genes, glutamatergic synapse with 122 genes, and dopaminergic synapse with 119 genes. When the minimum correlation was set to 0.80 in absolute value (Figure 3.4), four pathways were not sufficiently connected (nuclear signaling by ERBB4, Chagas disease, serotonergic synapse and ABC transporters). In turn, three clusters of highly interconnected pathways are immediately apparently in the network visualization of Figure 3.4. Cluster A contains nine pathways with higher than average nucleotide diversity and show strong positive connections between them. Prominent in this cluster is Axon guidance pathway with 119 genes. Its connections with cell-cell communication, G α signaling events and assembly of the primary cilium via the Insulin secretion pathways suggest this cluster mainly involves extracellular guidance such as growth and hormonal regulation helping axons reach their targets (Dickson 2002).

Most of the corresponding pathways of the other two clusters are negatively related between them. Cluster C is composed by processes related with the sympathetic nervous system, which is activated as response to stress by neurotransmitters such as dopamine (dopaminergic synapse) and glucocorticoids, which induce glutamate release (glutamatergic synapse, Popoli et al. 2012). Several other processes in cluster C are activated in response to the activation of the sympathetic nervous system, for instance increased heart contraction (adrenergic signaling in cardiomyocytes), blood vessels constriction in some parts of the body (vascular smooth muscle contraction and renin secretion), blood vessels dilatation in muscle and muscle contraction

(muscle contraction), and energy obtainment by lipids and carbohydrates degradation (pancreatic secretion). These processes are negatively connected with pathways in cluster B, which contains hormone-controlled processes related with reproduction (steroid hormone biosynthesis, ovarian steroidogenesis, linoleic acid metabolism, and arachidonic acid metabolism), and that are inhibited in stress events, like the pathway "Inflammatory mediator regulation of TRP channels".

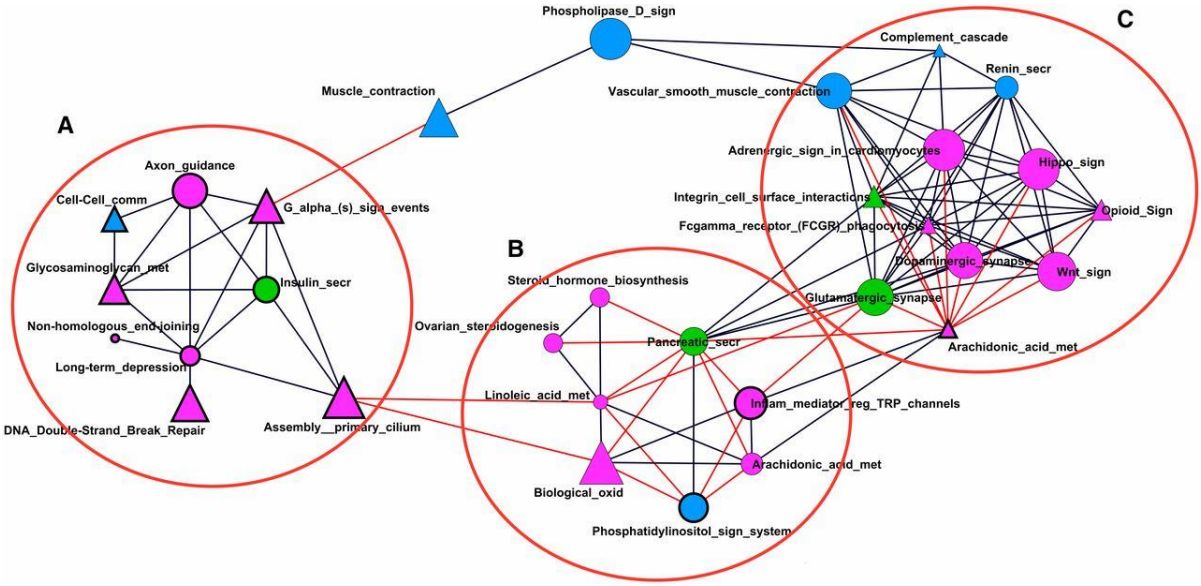

***Figure 3.4**: Co-association network among the 31 significant pathways that are interconnected. Each node represents a pathway that is connected by an edge if partial correlation with another pathway is significant and larger than 0.8 (in absolute value). Node size is proportional to number of genes in the pathway. Node shapes represent pathway source: triangles for REACTOME and circles for KEGG. Colors indicate the population with lowest Fst P-value: pink for Asia, blue for Europe and green for equal significance. Node line width to pathway variability: thin and thick lines for pathways with variability below and above average, respectively. Black and red edges represent positive and negative correlations between pathways, respectively. The three main pathway clusters are identified with letters A, B, C*

Finally, some genes also appeared repeatedly across pathways, which may indicate a central role in some biochemical routes. The list of genes presents in at least 10 of the significant pathways is in Table 3.3. Most of these genes are enzymes involved in general processes, such as phospholipases PLCB1 and PLCB3, involved in the transduction of many signals or PLA2G4A and PLA2G4B, that release arachidonic acid; kinases (PRKCA, PRKCG, PRKACA, MAPK1) involved in development and adenylate cyclases (ADCY2, ADCY3, ADCY4), which are part of the signal

transduction of G proteins, e.g., affecting dopamine. In addition, we also found the receptor ITPR3, which has an important role in taste transduction and is involved in the activation process of PKC that, as explained above, acts on several processes. Some taste receptors have been shown to be affected by domestication (da Silva *et al.* 2014).

Impact of deleterious mutations rate

We found 123,571 synonymous and 138,121 non-synonymous SNPs, of which 75,486 were predicted by VEP tool (McLaren et al. 2010) to be tolerated and 62,635, deleterious. In order to investigate whether the significant pathways and their genes have a larger proportion of deleterious vs. tolerated variants than the rest of pathways, we classified the SNPs in three groups: (i) SNPs in non-significant genes of non-significant pathways; (ii) SNPs in non-significant genes (P-value > 0.01) from significant pathways; and (iii) SNPs in significant genes (P-value < 0.01) of significant pathways. Table 3.4 shows the count of predicted deleterious and tolerated SNPs by group according to continent and domestic/wild status populations. In all four populations, there was a systematic trend of decreasing deleterious/tolerated rate with SNPs in significant genes of significant pathways compared to SNPs from non-significant pathways. The χ^2 test was significant in Europe and when all populations were jointly considered ($P < 0.01$) but not in Asia. These results can be interpreted as an increased functional constraint (lower ratio of deleterious mutations) in significant genes from significant pathways than in genes from non-significant pathways.

Previous studies suggest that domestication has resulted in an increased accumulation of deleterious mutations (Cruz, Vilà, and Webster 2008; Renaut and Rieseberg 2015; Pérez-Enciso et al. 2016). In agreement with this, here we observed larger ratio of deleterious *vs.* tolerated in domestics than in wild boars (λ_{ASDM} / λ_{ASWB} and λ_{EUDM} / λ_{EUWB} in Table 3.4). Interestingly, these ratios are higher in significant genes than in random gene SNPs, and also higher in Europe than in Asia. Therefore, these data suggest that potential purifying selection is weaker/less effective in Europe than in Asia, likely because the well-known low effective population size of old world pigs (Groenen et al. 2012). Besides, even if within-population purifying selection was

stronger in significant genes, it was comparatively weaker in domestic than in wild populations.

Table 3.3: ***Genes present in 10 or more significant pathways (Fst).***

Gene symbol	Ensembl Gene ID	Pathways	Significant pathways	P-value (Asia)	P-value (Europe)	Genomic position
PLCB3	ENSSSCG00000013034	35	15	0.128	0.741	2 : 6911684-6927124
PLCB1	ENSSSCG00000007056	35	15	1e-16	5e-04	17 : 19691509-19860912
PRKCA	ENSSSCG00000017268	46	13	0.248	0.887	12 : 13502757-13602445
PRKACA	ENSSSCG00000013771	47	13	0.830	0.583	2 : 65350514-65371241
PLA2G4A	ENSSSCG00000023351	21	11	0.388	2e-04	9 : 140460880-140623439
PRKCG	ENSSSCG00000003256	36	11	0.985	0.157	6 : 52982004-53001741
ITPR3	ENSSSCG00000001518	27	11	0.938	9e-04	7 : 34443056-34510838
ENSSSCG00000000175	ENSSSCG00000000175	35	11	0.579	0.385	5 : 15129052-15145294
ENSSSCG00000023437	ENSSSCG00000023437	25	10	0.435	0.003	13 : 67554829-67604679
ADCY2	ENSSSCG00000017101	32	10	0.024	0.249	16 : 80358753-80624210
MAPK1	ENSSSCG00000010081	74	10	0.696	0.633	14 : 53590167-53614842
ADCY3	ENSSSCG00000008578	32	10	0.999	0.994	3 : 121107128-121201171
ENSSSCG00000007833	ENSSSCG00000007833	39	10	7e-06	0.003	3 : 23052200-23174810
ADCY4	ENSSSCG00000001988	32	10	0.999	0.463	7 : 80227590-80243075

***Table 3.4**: Deleterious and tolerated SNPs grouped according to Europe (EU) / Asia (AS) continent and domestic (DM) / wild (WB) status.*

Continent	Population	SNP type*	Non-significant genes of non-significant pathways	Significant genes of non-significant pathways	Non-significant genes of significant pathways	Significant genes of significant pathways	P-value**
Asia	ASDM	Tolerated	8534	1195	2625	1176	0.756
		Deleterious	6825	954	2144	927	
		λ_{ASDM}	0.800	0.798	0.817	0.788	
	ASWB	Tolerated	6245	843	2029	636	0.033
		Deleterious	4432	587	1320	391	
		λ_{ASWB}	0.710	0.696	0.651	0.615	
	ASDM / ASWB	λ_{ASDM} / λ_{ASWB}	1.127	1.147	1.255	1.282	
Europe	EUDM	Tolerated	5191	716	1675	620	0.023
		Deleterious	3429	483	1200	349	
		λ_{EUDM}	0.661	0.675	0.716	0.563	
	EUWB	Tolerated	2654	416	893	346	0.001
		Deleterious	1153	198	382	103	
		λ_{EUWB}	0.434	0.476	0.428	0.298	
	EUDM / EUWB	λ_{EUDM} / λ_{EUWB}	1.523	1.418	1.675	1.891	
Total	Total	Tolerated	22624	3170	7222	2778	0.003
		Deleterious	15839	2222	5046	1770	
		λ_{Total}	0.700	0.701	0.699	0.637	

* λ corresponds to the ratio of deleterious vs. tolerated SNPs.

** P-value obtained from Chi-square test of the 2x2 table containing non-overlapping SNP sets (non-significant genes from non-significant pathways vs. significant genes from significant pathways).

General discussion

We report a functional analysis of pig domestication and breeding using a large complete sequence dataset that consisted of 40 wild boars and 123 domestic pig genomes from Asia (mainly China) and Europe. Rather than a standard exploratory genome-wide analysis, we focused on an analysis where the unit of study is the pathway. Genes do not function in isolation, but coordinately, and thus metabolic pathways provide a reasonable scaffold to accommodate this fact (e.g., Daub et al. 2013). In a previous study, we observed that the "heritability" of domestic status varied according to pathway and that differences were not due to the number of genes in the pathway, suggesting that pathway can be a meaningful analysis unit (Pérez-Enciso et al. 2016). In fact, one of the main advantages of this approach is that it provides a direct biological interpretation of the analyses, although independent source of information may be required to conclude which tissue and developmental stage the perturbed pathway may act in. In contrast, standard window-based genome wide scans may pinpoint regions devoid of annotations or where the functional relation between significant windows is unknown. We assessed two metrics, differentiation (Fst) and disequilibrium (nSL), and although some of the pathways were connected (Figure 3.4), we found little concordance between the two analyses. Lack of agreement between differentiation and disequilibrium statistics have been reported previously (e.g., Chen et al. 2016; Dall'Olio et al. 2012), and this is likely because of the different timing and persistence of effects caused by selection (Sabeti et al. 2006). In particular, since disequilibrium erodes rapidly, our analysis suggests that reproductive changes (Table 3.2) are among the most recent ones whereas others such as development and behavior (Table 3.1) were earlier targets of domestication and/or breeding. This is coherent with current knowledge, as behavioral changes must have occurred in earlier stages of domestication, as exemplified by the important experiment for tameness in foxes (Kukekova et al. 2012), whereas emphasis in increasing reproductive performance is a more recent target of modern breeding.

Our approach has limitations as well. Foremost, many genes are not assigned to any pathway. In the NCBI Biosystems database v.160202 used here, 7157 genes out of a total of 21,691 annotated genes (Ensembl genes v. 83) were assigned to

at least one pathway. After filtering, we further restricted the analysis to 442 pathways containing 5713 genes; these correspond to a total of 220 Mb or ~8.5% of the whole genome. Another issue is redundancy and the definition of pathway itself, since there are several databases (KEGG, REACTOME, Interactome, etc.) that contain lists of functionally related genes. Definition of the same pathway can actually be quite different between databases (Mooney et al. 2014) as we observed here with the arachidonic metabolism pathways (Table 3.2). Here, as in Daub et al. (2013), we decided to initially consider all available pathways from the NCBI biosystems database, although we set a maximum redundancy between pathways of 50%. But in contrast to Daub et al., who considered the most significant SNP from each gene and removed all gene redundancy between pathways, here we combined all SNPs (after pruning for linkage disequilibrium) from a given gene into a single statistic using Fisher's method, and we allowed a 50% gene redundancy. It is not evident which method is best, but it seems that our approach is more conservative since outlier Fst will be smoothed out unless a general trend across SNPs in the whole gene is maintained. Allowing for gene redundancy in turn allows us to keep the original gene set instead of pathway pruning.

The history of domestication and domestic breeds is in all quite complex. In addition to multiple independent domestication events, as occurred in Asia and in Europe in the case of the pig, local adaptive processes have occurred since different breeds have been selected for different traits. Therefore, it is not surprising that previous works (e.g., Amaral et al. 2011) reported that most of selective signals were breed-specific, although our work demonstrates that shared domestication and breeding signals across breeds can still be detected. These signals are numerous and none of them are strong enough to explain the whole process. Once more, the polygenic model prevails. We further show that Asian and European domestication/breeding processes have both distinctive and shared pathways, and that multiple processes have been involved such as an increase in disequilibrium and in differentiation. Differentiation metrics (Fst) revealed a larger number of signals than disequilibrium (Tables 3.1 *vs.* 3.2), but this may due to the experimental design: we analyzed several breeds jointly and disequilibrium is more rapidly eroded by demographic processes than

differentiation (Sabeti et al. 2006). Nevertheless, we often found genes that were significant in both continents, such as several genes involved in behavior (PLCB1, GSK3B, HTR4, Figure 3.2), while the pathways they belong to were significant in only in one continent. Given that most European breeds have been admixed with Asian pigs (Groenen 2016), it is possible that these shared signals may actually be due to introgression. To verify this, we run a semisupervised ADMIXTURE analyses on all significant (Tables 3.1 and 3.2) and a set of random pathways (Figure S3.8). The average Asian component in EUDM pigs across significant pathways was q = 0.11 (SD = 0.03), which is nearly identical to that observed in a random set of pathways (0.11, SD = 0.02). Similarly, we did not find differences in Asian component between shared significant pathways across continents (q = 0.105) and those that were continent specific (q = 0.102). This suggests that Asian introgression is unlikely to have caused a shared signal between continents, which can be explained because the Asian signature in European breeds seems to be quite heterogeneous, i.e., due to different Asian origins (Bosse et al. 2014; Bianco, Soto, et al. 2015).

In contrast to previous works in humans, which reported an enrichment of pathways related to pathogen response in adaptation (Daub et al. 2013), we did not find a strong over representation of immune – system related pathways. Only three related pathways were detected with Fst (complement cascade, Chagas disease, and FCGR phagocytosis, Table 3.1). This could be due to the fact that domestication was accompanied by stronger selection for traits other than disease resistance such as behavior, reproduction or development. Another explanation is that these disease resistance signals were breed specific and therefore remained undetected in this experimental design.

As in other studies (Cruz, Vilà, and Webster 2008; Renaut and Rieseberg 2015; Pérez-Enciso et al. 2016), we found an increased accumulation of deleterious mutations in domestic animals. We systematically observed a higher proportion of deleterious variants in domestic groups compared to wild boars. This was observed for all genes, regardless they were significant or not. On the other hand, a decreased accumulation of deleterious mutations was observed in significant genes from significant pathways, suggesting, as it is shown in Table 3.3, that these genes perform essential and central tasks in the physiology and

development of the pig. Therefore, these genes seem to be under stronger functional constraint than randomly sampled genes.

Conclusions

We have studied the functional basis of domestication and breeding in the pig. This was possible because a modern equivalent of the wild ancestor is still available for study and was facilitated by the numerous sequences in the public domain. We show that these processes predominantly involved pathways related to behavior, especially in Asia, but also others like insulin, organ size development, recombination and female reproduction. At least in part, these results can be explained by a relaxation of purifying selection associated with the domestication and/or breeding processes. Nevertheless, this purifying selection was stronger in genes and pathways that were significant using Fst than in random genes, likely because these genes play central roles and are highly functionally constrained. Negative selection was also stronger in Asia than in Europe, likely due to larger effective size of Asian population. In all probability, this analysis is conservative since we have focused on SNPs and genes that are consistently differentiated between all domestic breeds pooled together *vs.* wild boar. Focusing on a specific breed may have increased power but also could be prone to false discovery rate, identifying signals that are breed specific rather than domestic specific.

Acknowledgments

We thank Sara Guirao-Rico, Marcel Amills and Sebastián Ramos-Onsins for comments. Part of this work was developed while M.P.-E. was visiting CSIRO Agriculture in St. Lucia campus (Brisbane, Australia), funded by a CSIRO McMaster Visiting Fellowship. J.L.-C is the recipient of a Training Grant for Research Staff (FPI) grant to achieve the PhD, from the Ministry of Economy and Science (MINECO, Spain). Work was also funded by grants AGL2013-41834-R and AGL2016-78709-R from MINECO to M.P.-E. We also acknowledge the support of MINECO for the “Centro de Excelencia Severo Ochoa 2016-2019” award SEV-2015-0533 and of the Centres de Recerca de Catalunya program

(CERCA, Generalitat de Catalunya, Spain) to the Centre for Research in Agricultural Genomics.

References

Ai, Huashui, Xiaodong Fang, Bin Yang, Zhiyong Huang, Hao Chen, Likai Mao, Feng Zhang, et al. 2015. "Adaptation and Possible Ancient Interspecies Introgression in Pigs Identified by Whole-Genome Sequencing." *Nature Genetics* 47 (3): 217–25. https://doi.org/10.1038/ng.3199.

Alexander, David H, John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Amaral, Andreia J, Luca Ferretti, Hendrik-Jan Megens, Richard P M a Crooijmans, Haisheng Nie, Sebastian E. Ramos-Onsins, Miguel Perez-Enciso, Lawrence B Schook, and Martien a M Groenen. 2011. "Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA." Edited by Hans Ellegren. *PLoS ONE* 6 (4). https://doi.org/10.1371/journal.pone.0014782.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society* 57 (1): 289–300. http://www.ams.org/mathscinet-getitem?mr=1325392.

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

Bianco, Erica, H.W. Soto, L. Vargas, and Miguel Pérez-Enciso. 2015. "The Chimerical Genome of Isla Del Coco Feral Pigs (Costa Rica), an Isolated Population since 1793 but with Remarkable Levels of Diversity." *Molecular Ecology* 24: 2364–78. https://doi.org/10.1111/mec.13182.

Bosse, Mirte, Hendrik-Jan Megens, Laurent A. F. Frantz, Ole Madsen, Greger Larson, Yogesh Paudel, Naomi Duijvesteijn, et al. 2014. "Genomic Analysis

Reveals Selection for Asian Genes in European Pigs Following Human-Mediated Introgression." *Nature Communications* 5: 4392. https://doi.org/10.1038/ncomms5392.

Browning, Brian L, and Sharon R Browning. 2013. "Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data." *Genetics* 194 (2): 459–71. https://doi.org/10.1534/genetics.113.150029.

Buitenhuis, Bart, Luc L G Janss, Nina A Poulsen, Lotte B Larsen, Mette K Larsen, and Peter Sørensen. 2014. "Genome-Wide Association and Biological Pathway Analysis for Milk-Fat Composition in Danish Holstein and Danish Jersey Cattle." *BMC Genomics* 15: 1112. https://doi.org/10.1186/1471-2164-15-1112.

Burgos-Paz, William, C a Souza, Hendrik-Jan Megens, Yuliaxis Ramayo-Caldas, M Melo, C Lemús-Flores, E Caal, et al. 2013. "Porcine Colonization of the Americas: A 60k SNP Story." *Heredity* 110 (4): 321–30. https://doi.org/10.1038/hdy.2012.109.

Chang, Christopher C, Carson C Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4: 7. https://doi.org/10.1186/s13742-015-0047-8.

Chen, M, D Pan, H Ren, J Fu, J Li, G Su, A Wang, L Jiang, Q Zhang, and JF Liu. 2016. "Identification of Selective Sweeps Reveals Divergent Selection between Chinese Holstein and Simmental Cattle Populations." *Genetics Selection Evolution* in press.

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

Dall'Olio, Giovanni Marco, Hafid Laayouni, Pierre Luisi, Martin Sikora, Ludovica Montanucci, and Jaume Bertranpetit. 2012. "Distribution of Events of Positive Selection and Population Differentiation in a Metabolic Pathway: The Case of Asparagine N-Glycosylation." *BMC Evolutionary Biology* 12 (98): 1–13. https://doi.org/10.1186/1471-2148-12-98.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Daub, Josephine T., Tamara Hofer, Emilie Cutivet, Isabelle Dupanloup, Lluis Quintana-Murci, Marc Robinson-Rechavi, and Laurent Excoffier. 2013. “Evidence for Polygenic Adaptation to Pathogens in the Human Genome.” *Molecular Biology and Evolution* 30 (7): 1544–58. https://doi.org/10.1093/molbev/mst080.

Diamond, Jared. 2002. “Evolution, Consequences and Future of Plant and Animal Domestication.” *Nature* 418 (6898): 700–707. https://doi.org/10.1038/nature01019.

Dickson, Barry J. 2002. “Molecular Mechanisms of Axon Guidance.” *Science* 298: 1959–64.

Esteve-Codina, Anna, Yogesh Paudel, Luca Ferretti, Emanuele Raineri, Hendrik-Jan Megens, Luis Silió, María C Rodríguez, Martein a M Groenen, Sebastian E Ramos-Onsins, and Miguel Pérez-Enciso. 2013. “Dissecting Structural and Nucleotide Genome-Wide Variation in Inbred Iberian Pigs.” *BMC Genomics* 14 (January): 148. https://doi.org/10.1186/1471-2164-14-148.

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. “On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure.” *Molecular Biology and Evolution* 31 (5): 1275–91. https://doi.org/10.1093/molbev/msu077.

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

Frantz, Laurent A. F., Erik Meijaard, Jaime Gongora, James Haile, Martien A. M. Groenen, and Greger Larson. 2016. “The Evolution of Suidae.” *Annual Review of Animal Biosciences* 4 (1): 61–85. https://doi.org/10.1146/annurev-animal-021815-111155.

Frantz, Laurent A. F., Joshua G. Schraiber, Ole Madsen, Hendrik-Jan Megens, Alex Cagan, Mirte Bosse, Yogesh Paudel, Richard P. M. A. Crooijmans, Greger Larson, and Martien A. M. Groenen. 2015. "Evidence of Long-Term Gene Flow and Selection during Domestication from Analyses of Eurasian Wild and Domestic Pig Genomes." *Nature Genetics* 47 (10): 1141–48. https://doi.org/10.1038/ng.3394.

Fujii, J, K Otsu, F Zorzato, S de Leon, V K Khanna, J E Weiler, P J O'Brien, and D H MacLennan. 1991. "Identification of a Mutation in Porcine Ryanodine Receptor Associated with Malignant Hyperthermia." *Science (New York, N.Y.)* 253 (5018): 448–51.

Geer, Lewis Y., Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H. Bryant. 2010. "The NCBI BioSystems Database." *Nucleic Acids Research* 38 (Database issue): D492-6. https://doi.org/10.1093/nar/gkp858.

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

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

Ha, Ngoc-Thuy, Josef Johann Gross, Annette van Dorland, Jens Tetens, Georg Thaller, Martin Schlather, Rupert Bruckmaier, and Henner Simianer. 2015. "Gene-Based Mapping and Pathway Analysis of Metabolic Traits in Dairy Cows." *Plos One* 10 (3): 1–15. https://doi.org/10.1371/journal.pone.0122325.

Irvine, Kenneth D. 2012. "Integration of Intercellular Signaling through the Hippo Pathway." *Seminars in Cell & Developmental Biology* 23 (7): 812–17. https://doi.org/10.1016/j.semcdb.2012.04.006.

Kanehisa, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika

Hirakawa, Masumi Itoh, Toshiaki Katayama, 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.

Keinan, Alon, and David Reich. 2010. “Human Population Differentiation Is Strongly Correlated with Local Recombination Rate.” *PLoS Genetics* 6 (3): e1000886. https://doi.org/10.1371/journal.pgen.1000886.

Koh, H.-Y., D. Kim, J. Lee, S. Lee, and H.-S. Shin. 2007. “Deficits in Social Behavior and Sensorimotor Gating in Mice Lacking Phospholipase C Beta 1.” *Genes, Brain and Behavior* 0 (0): 070816104557003-??? https://doi.org/10.1111/j.1601-183X.2007.00351.x.

Kukekova, Anna V., Svetlana V. Temnykh, Jennifer L. Johnson, Lyudmila N. Trut, and Gregory M. Acland. 2012. “Genetics of Behavior in the Silver Fox.” *Mammalian Genome* 23 (1–2): 164–77. https://doi.org/10.1007/s00335-011-9373-z.

Larson, Greger, Keith Dobney, Umberto Albarella, Meiying Fang, Elizabeth Matisoo-Smith, Judith Robins, Stewart Lowden, et al. 2005. “Worldwide Phylogeography of Wild Boar Reveals Multiple Centers of Pig Domestication.” *Science* 307 (5715): 1618–21. https://doi.org/10.1126/science.1106927.

Li, Heng, and Richard Durbin. 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, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Gonçalo R. Abecasis, Richard Durbin, 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.

Matthews, Lisa, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, et al. 2009. “Reactome Knowledgebase of Human Biological Pathways and Processes.” *Nucleic Acids Research* 37 (SUPPL. 1): 619–22. https://doi.org/10.1093/nar/gkn863.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

McLaren, William, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. 2010. “Deriving the Consequences of Genomic Variants with the Ensembl API and SNP Effect Predictor.” *Bioinformatics* 26 (16): 2069–70. https://doi.org/10.1093/bioinformatics/btq330.

Molnár, János, Tibor Nagy, Viktor Stéger, Gábor Tóth, Ferenc Marincs, and Endre Barta. 2014. “Genome Sequencing and Analysis of Mangalica , a Fatty Local Pig of Hungary.” *BMC Genomics* 15 (761): 1–12.

Mooney, Michael A., Joel T. Nigg, Shannon K. McWeeney, and Beth Wilmot. 2014. “Functional and Genomic Context in Pathway Analysis of GWAS Data.” *Trends in Genetics* 30 (9): 390–400. https://doi.org/10.1016/j.tig.2014.07.004.

Munoz-Fuentes, Violeta, Marina Marcet-Ortega, Gorka Alkorta-Aranburu, Catharina Linde Forsberg, Jane M. Morrell, Esperanza Manzano-Piedras, Arne Soderberg, et al. 2015. “Strong Artificial Selection in Domestic Mammals Did Not Result in an Increased Recombination Rate.” *Molecular Biology and Evolution* 32 (2): 510–23. https://doi.org/10.1093/molbev/msu322.

Nevado, Bruno, Sebastian E. Ramos-Onsins, and Miguel Perez-Enciso. 2014. “Resequencing Studies of Nonmodel Organisms Using Closely Related Reference Genomes: Optimal Experimental Designs and Bioinformatics Approaches for Population Genomics.” *Molecular Ecology* 23 (7): 1764–79. https://doi.org/10.1111/mec.12693.

Ngoh, Adeline, Amy McTague, Ingrid M Wentzensen, Esther Meyer, Carolyn Applegate, Eric H Kossoff, Denise A Batista, Tao Wang, and Manju A Kurian. 2014. “Severe Infantile Epileptic Encephalopathy Due to Mutations in PLCB1: Expansion of the Genotypic and Phenotypic Disease Spectrum.” *Developmental Medicine and Child Neurology* 56 (11): 1124–28.

https://doi.org/10.1111/dmcn.12450.

Ollivier, L. 1995. “Genetic Differences in Recombination Frequency in the Pig (Sus Scrofa).” *Genome* 38 (5): 1048–51.

Pavlidis, Pavlos, Jeffrey D. Jensen, Wolfgang Stephan, and Alexandros Stamatakis. 2012. “A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans.” *Molecular Biology and Evolution* 29 (10): 3237–48. https://doi.org/10.1093/molbev/mss136.

Pérez-Enciso, M., G. de los Campos, N. Hudson, J. Kijas, and A. Reverter. 2016. “The ‘Heritability’ of Domestication and Its Functional Partitioning in the Pig.” *Heredity* 118: 160–68. https://doi.org/10.1038/hdy.2016.78.

Pico, Alexander R., Thomas Kelder, Martijn P. Van Iersel, Kristina Hanspers, Bruce R. Conklin, and Chris Evelo. 2008. “WikiPathways: Pathway Editing for the People.” *PLoS Biology* 6 (7): 1403–7. https://doi.org/10.1371/journal.pbio.0060184.

Popoli, Maurizio, Zhen Yan, Bruce S McEwen, and Gerard Sanacora. 2012. “The Stressed Synapse: The Impact of Stress and Glucocorticoids on Glutamate Transmission.” *Nature Reviews. Neuroscience* 13 (1): 22–37. https://doi.org/10.1038/nrn3138.

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

R Development Core Team, R, and R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Edited by R Development Core Team. *R Foundation for Statistical Computing*. Vol. 1. R Foundation for Statistical Computing. Vienna: R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7.

Ramírez, Oscar, William Burgos-Paz, Encarna Casas, Maria Ballester, Erica Bianco, Iñigo Olalde, Gabriel Santpere, et al. 2014. “Genome Data from a Sixteenth Century Pig Illuminate Modern Breed Relationships.” *Heredity*

114 (2): 175–84. https://doi.org/10.1038/hdy.2014.81.

Ramos-Onsins, Sebastian E., William Burgos-Paz, Arianna Manunza, and Marcel Amills. 2014. “Mining the Pig Genome to Investigate the Domestication Process.” *Heredity* 113 (6): 471–84. https://doi.org/10.1038/hdy.2014.68.

Renaut, Sebastien, and Loren H. Rieseberg. 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.

Reverter, Antonio, and Eva K F Chan. 2008. “Combining Partial Correlation and an Information Theory Approach to the Reversed Engineering of Gene Co-Expression Networks.” *Bioinformatics* 24 (21): 2491–97. https://doi.org/10.1093/bioinformatics/btn482.

Ross-Ibarra, Jeffrey. 2004. “The Evolution of Recombination under Domestication: A Test of Two Hypotheses.” *The American Naturalist* 163 (1): 105–12. https://doi.org/10.1086/380606.

Rubin, Carl-Johan, Hendrik-Jan Megens, Alvaro Martinez Barrio, Khurram Maqbool, Shumaila Sayyab, Doreen Schwochow, Chao Wang, 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.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. “Positive Natural Selection in the Human Lineage.” *Science* 312 (5780): 1614–20. https://doi.org/10.1126/science.1124309.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504. https://doi.org/10.1101/gr.1239303.

Sim, Ngak-Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. 2012. “SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins.” *Nucleic Acids Research* 40 (Web Server issue): W452-7. https://doi.org/10.1093/nar/gks539.

Storey, John D., ndrew J. Bass, Alan Dabney, and David Robinson. 2015. “Qvalue: Q-Value Estimation for False Discovery Rate Control.” https://www.bioconductor.org/packages/release/bioc/html/qvalue.html.

Tajima, Fumio. 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.

Tajima. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123: 585–95. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1203831/pdf/ge1233585.pdf

Tortereau, Flavie, Bertrand Servin, Laurent A. F. Frantz, Hendrik-Jan Megens, Denis Milan, Gary Rohrer, Ralph Wiedmann, et al. 2012. “A High Density Recombination Map of the Pig Reveals a Correlation between Sex-Specific Recombination and GC Content.” *BMC Genomics* 13 (January): 586. https://doi.org/10.1186/1471-2164-13-586.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. “Analysing Biological Pathways in Genome-Wide Association Studies.” *Nature Reviews. Genetics* 11 (12): 843–54. https://doi.org/10.1038/nrg2884.

Weir, B. S., and C. Clark Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” *Evolution* 38 (6): 1358. https://doi.org/10.2307/2408641.

Zeder, Melinda A. 2015. “Core Questions in Domestication Research.” *Proceedings of the National Academy of Sciences* 112 (11): 3191–98. https://doi.org/10.1073/pnas.1501711112.

Chapter 4

Selection pressure and network topology in wild and domestic pigs

Selection pressure and network topology in wild and domestic pigs

J. Leno-Colorado[1], S. Guirao-Rico[2], L. Silió[3], M.C. Rodríguez[3], M. Pérez-Enciso[1,4], S.E. Ramos-Onsins[1*]

[1] Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Spain

[2] Instituto de Biología Evolutiva (IBE), Passeig Marítim de la Barceloneta 37-49, Barcelona 08003, Spain

[3] Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Ctra. de La Coruña km 7, Madrid, 28040, Spain

[4] ICREA, Carrer de Lluís Companys 23, Barcelona 08010, Spain

* Author for correspondence: Sebastián E. Ramos-Onsins, Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain; sebastian.ramos@cragenomica.es

(In preparation)

Abstract

Domestication is a process of artificial selection driven by humans, which modifies the features of an ancestral species into new phenotypical traits, such as less aggressiveness and greater productivity. The study of markers linked to this evolutionary process may shed light on its biological basis. Here, we investigated the distribution of selective pressures at the genome level to discern the impact of the domestication in the pig genome. To that end, we selected 20 wild boars and 26 domestic pigs, which are composed by two different breeds: Iberian and Large White. These domestic populations are very different between them, Iberian is an autochthonous breed with no evidence of introgression whereas Large White is a commercial breed that has been artificially improved and has admixture with Asian pigs in its genome. To analyze the strength of selection, we designed a new approach which considers the missing information in the data. This method uses four different estimates of the nucleotide variability, based on different parts of the site frequency spectrum. Nevertheless, when we analyzed the selection patterns we did not detected clear signals of domestication. Instead, the effect of demography on the selection patterns is larger than the domestication, Iberian has low variability and its pattern showed the influence of a population reduction, whereas Large White has high variability, probably due to Asian contribution in its genome, and its pattern showed the presence of deleterious and beneficial mutations and the possible influence of a population expansion and/or migration. We conclude that the selection patterns for this data are compatible with the presence of deleterious mutations segregating in all three breeds and the influence of the demography and discuss several hypotheses that explain the apparent lack of domestication signal and the different alternatives to detect the effect of domestication on the genome.

Introduction

Domestic animal histories are evolutionary experiments often lasting several millennia with the result of dramatic phenotypic changes to suit human needs. Domestic species are structured in subpopulations (e.g., breeds) that are, to some degree, genetically isolated from each other and together exhibit a broad

catalog of phenotypes. The demographic history of domestic animals is usually complex, and many events remain unknown or poorly documented. Therefore, the genomes of domesticated species offer a material of utmost interest to study the interplay of demography and accelerated adaptation (Ojeda et al. 2011).

The pig (*Sus scrofa*) is a particularly interesting species and one in which there is abundant genetic tools and sequence data available. It originated in the Southeast Asian region ca. 4 MYA and migrated towards the west (ca. 1.2 MYA (Frantz et al. 2013)), colonizing all climates in Eurasia except the driest. Subsequently, the pig was domesticated out of local wild boars independently in both Asia and Europe ~9,000 years ago. To complicate the history, modern European pig breeds were crossed with Asian pigs during the late 17th century and onwards. In some commercial breeds such as Large White, about 30% of the genome is estimated to be of Asian origin (Bosse, Megens, Madsen, et al. 2014). However, some European local breeds, like the Iberian breed, were spared the genetic contact with Asia and there is no evidence of genetic introgression has been reported (Esteve-Codina et al. 2013). Instead, this breed has suffered a reduction of its effective population size (or population bottleneck) during the last century (Alves et al. 2006). Hence, the differences in both, the demographic history and the artificial selective pressure among pig breeds might have caused differences in the distribution of selective constrains among their genomes.

So far, the evidence of the genetic changes associated with pig domestication and/or artificial breeding are conflictive. While the most accepted view is that regulatory changes have been the predominant target of these changes, several studies underscored the influence of non-synonymous changes (Rubin et al. 2012) and an overall increase of the deleterious mutation rate (Cruz, Vilà, and Webster 2008; Renaut and Rieseberg 2015; Pérez-Enciso et al. 2016; Leno-Colorado et al. 2017) in domestic pigs.

As we and others have shown previously (Leno-Colorado et al. 2017; Ha et al. 2015; Buitenhuis et al. 2014; Daub et al. 2013), selection may impact global pathways rather than individual genes. In this sense, several studies have shown that the strength of the selection is affected by the position of genes in the networks in which they participate. Those genes that are more central in a

network and more connected with other genes are more evolutionary constrained, which makes them less likely to have undergone positive selection compared to peripheral genes (Fraser et al. 2002; Hahn and Kern 2005; Montanucci et al. 2011; Alvarez-Ponce and Fares 2012). Furthermore, the gene evolutionary rate increases as it moves from upstream to downstream, possibly due to the pleiotropy of upstream genes, since they are involved in more functions and processes; therefore, upstream genes should be more conserved (Rausher, Miller, and Tiffin 1999; Riley, Jin, and Gibson 2003; Livingstone and Anderson 2009; Ramsay, Rieseberg, and Ritland 2009).

The main purpose of this work is to detect changes in the distribution of mutational effects produced by the process of domestication at different functional scales (i.e., genes, pathways, genome-wide), using the pig as a model species. Domestication is an intensive selection process that occurred in a short time (<10,000 years) and reduced the genetic variation originally present in the wild population, eventually fixing some of the variants (both selected and non-selected). This process is accompanied by a dramatic change in the environment of the domesticated population, and therefore, in the fitness effects of the variants already present in the ancestral population, for instance, the white color is strongly deleterious in the wild environment but is positively selected in the domesticated one). Consequently, segregating variants that are present in both wild and domesticated organisms may have different frequencies according to their differences in their selective effects. Similarly, new mutations are predicted to have different frequencies between these populations due to either environmental differences and to the strong selective pressures imposed by artificial selection in domestic breeds.

Here, we investigated the impact of domestication process in the distribution of selective pressures across genomes and how this process affected the levels and patterns of variation in domestic and wild pigs. In other words, if the main domestication effects have been produced by new or by segregating mutations, and whether they have affected a large number of positions with weak effect or, alternatively, few positions with strong effect. To that end, we have designed a new approach based on the McDonald-Kreitman test (MKT) (McDonald and Kreitman 1991), that explicitly takes into account the missing information at high-

throughput data. Our method replaces variant counts at different frequencies with four summary statistics of the population mutation rate (θ), to estimate the rate and the strength of adaptation in different segments of the site frequency spectrum. To our knowledge, the comparison of the rate and the strength of adaptation at genomic level between animal domestic species versus its wild counterpart has not been performed so far.

Material and methods

Biological samples

We analyzed the sequence of 46 pig (*Sus scrofa*) genomes (Table S4.1), which includes 20 European wild boars (WB) and 26 domestic pigs, represented by Iberian (IB, n = 6) and Large White (LW, n = 20) breeds. These two domestic breeds were selected because they have very different features; Iberian is a local breed subjected to low artificial selection intensity and with no evidence of introgression, whereas Large White is a breed that has undergone strong artificial selection and frequently introgressed with Asian germplasm (Bosse, Megens, Madsen, et al. 2014; Groenen 2016). In order to polarize the nucleotide changes between the different breeds, we used as outgroup the consensus sequence obtained from several *Sus* species (*S. barbatus*, *S. cebifrons*, *S. verrucosus*, *S. celebensi, around 4.2 MYA divergence*) and the African warthog (*Phacochoerus africanus*, around ~10 MYA divergence), as is detailed in Bianco et al. (2015).

Most of the sequences used in this work were retrieved from public databases (Rubin et al. 2012; Ramírez et al. 2014; Bianco et al. 2015; Frantz et al. 2015; Moon et al. 2015) and were downloaded from the short read archive (SRA, http://www.ncbi.nlm.nih.gov/sra). The VCF file containing all SNPs is available at https://bioinformatics.cragenomica.es/numgenomics. In addition, here we utilized the sequence from 5 wild boars.

Variant calling

The raw reads of each individual were mapped it against the reference pig assembly (Sscrofa10.2, Groenen et al. 2012) using BWA mem option (Li and

Durbin 2009). PCR duplicates were removed using SAMtools rmdup v0.1.19 (Li et al. 2009) and reads were realigned around indels with GATK IndelRealigner tool (McKenna et al. 2010). Genotype calling was performed with SAMtools mpileup and bcftools call v1.3.0 (Li et al. 2009) for each individual separately. We set the minimum and the maximum read depth 5x and twice the average sample's read depth plus one, respectively. Minimum mapping and base quality were established to 20 (P-value=1e-2). Homozygous blocks (regions where both alleles are the same as the reference) were called using 'samtools depth' utility, BEDtools (Quinlan 2014) and custom scripts, following the same criterias as for calling the SNPs, resulting in a gVCF file per individual.

Next, we generated a multi-individual VCF file, merging all individual gVCF files which comprises all the SNPs from the 46 samples. This was done by previously converting every gVCF into a FASTA file and recoding all this information into a new whole sample, multi-gVCF file (Pérez-Enciso et al. 2016)). The main pipeline is available at https://github.com/miguelperezenciso/NGSpipeline.

A principal component analysis (PCA) was performed using the total number of SNPs (coding and non-coding) to analyze the population structure. Missing genotypes (./.) were replaced by the average frequency of the SNP in this position. SNPs with percentage of missing larger than 95% were removed.

Polymorphism and Divergence

Genetic diversity and divergence per gene and population was estimated using mstatspop software (Guirao-Rico et al. 2018, available at https://github.com/cragenomica/mstatspop). We calculated four different estimators of nucleotide variability (θ), which take into account missing data (Ferretti, Raineri, and Ramos-Onsins 2012) and that weigh the frequencies of the SNPs in a different way. The statistics were: Watterson (Watterson 1975), Tajima (Tajima 1983), Fu&Li (Fu and Li 1993) and Fay&Wu's estimators (Fay and Wu 2000). Nucleotide variability was estimated using the following:

$$\hat{\theta} = \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1} i\omega_{i,n_x}\xi_i(x), \quad \frac{1}{L}\sum_{x=1}^{L}\sum_{i=1}^{n_x-1}\omega_{i,n_x} = 1$$

(Equation 1)

Where the weights (w_i) are $w_i = n/(i(n-i)(1+\partial_{i,n-i}))$ for Watterson estimator, $w_i = n/(1+\partial_{i,n-i})$ For Tajima estimator (both for folded spectrum), $w_i = i$ for Fay&Wu estimator and $w_1 = 1$, $w_{i>1} = 0$ for Fu&Li estimator (Achaz 2009). Watterson's estimator weights more the lower frequencies in a harmonic pattern, while Tajima's estimator gives the same weight to all frequencies (using the folded frequency spectrum). On the other hand, Fu&Li's estimator considers only derived singletons and Fay&Wu's estimator increases the weight proportionally to their frequency.

Controlling for the correlation between missing data and variation

Preliminary analyses showed a moderate negative correlation of the levels of variability and divergence with the proportion of missing data for each gene. In order to control for this correlation, we excluded from the analysis all highly variable genes (higher than 99% quantile of the total genes) and also those genes having a ratio of missing data greater than 0.3. The final data set resulting from applying these filters (an average of 13,500 genes, 70% of the total annotated genes) shows no correlation with missing data (Table S4.2).

Proportion of adaptive substitutions

Under the neutral model, observed polymorphisms segregating in a population are most neutral and only a number of positively selected variants are rapidly segregating towards fixation. These positive selected variants are then mainly observed only as fixed variants. Functional positions (i.e., non-synonymous positions) are under relatively strong constriction to variation, and thus, their evolutionary ratios are smaller than the non-functional regions (synonymous positions). In the neutral scenario (see McDonald and Kreitman 1991), both the polymorphism and the divergence (excluding the adaptive fixed variants) are

proportional to the mutation rate (and to the constriction factor in case of non-synonymous positions). That is:

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

(Equation 2)

where θ_s is the synonymous variability, θ_n the non-synonymous variability, K_s the synonymous divergence, K_n the non-synonymous divergence, and α is the proportion of adaptive variants that have been fixed. The proportion of non-synonymous substitutions that are adaptive (α) can be estimated reordering equation 2:

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

(Equation 3)

A higher ratio of non-synonymous/synonymous divergence versus polymorphisms suggests that there has been selection fixing adaptive variants (α > 0), however, the opposite case (α < 0) suggests the effect of deleterious mutations segregating in the population.

In the case of considering weak deleterious mutations in the model (those detrimental mutations that are not eliminated in the population), we contemplate that a number of non-synonymous detrimental variants are segregating. Their quantity will be higher at lower frequencies and will be lower (or null) deleterious fixations. Following the same notation than in equation 3:

$$\frac{\theta_{in}(1-\beta i)}{\theta_{is}} = \frac{(1-\alpha-\beta min)K_n}{K_s},$$

(Equation 4)

where *i* refers to the frequency at which the calculation of variability is estimated, *β* is the proportion of detrimental mutations, *βmin* is the number of fixed detrimental mutations and $\beta min < \beta i$ at any frequency. Solving for α:

$$\alpha = 1 - \beta min - (1-\beta i)\frac{K_s}{K_n}\frac{\theta_{in}}{\theta_{is}}$$

(Equation 5)

That is, in case of estimating the proportion of adaptive mutations using equation 3, the estimates of α would be underestimated, especially when estimates of variability consider low frequency variants. If we assume that the detrimental variants would never be fixed, a good estimator of α, using equation 3, would be to use estimates of variability based on high frequencies, as they would hardly contain detrimental mutations, which is concordant to arguments used by Messer and Petrov (2013). Analytical equations that estimate the value of α given positive, neutral and negative selection are given in Uricchio, Petrov, and Enard (2019).

If, additionally, we consider weak positively selected variants that are segregating in the population, that means that we consider the possibility to observe adaptive variants segregating in the population. It is expected that the relative frequency of adaptive variants (in relation to neutral) is higher at high frequencies than at lower frequencies, that is:

$$\frac{\theta_{in}(1-\beta i-\gamma i)}{\theta_{is}} = \frac{(1-\alpha-\beta \min-\gamma max)K_n}{K_s},$$

(Equation 6)

Solving for adaptive fixed variants;

$$\alpha+\gamma max = 1-\beta min-(1-\beta i-\gamma i)\frac{K_s}{K_n}\frac{\theta_{in}}{\theta_{is}}$$

(Equation 7)

In that case, the estimation of the proportion of adaptive mutations using equation 3 would also incorrect if we use the estimates of variability at high frequencies. That means that an estimate based on estimates at high frequencies (using equation 3) would underestimate the global proportion of adaptive variants. Note that adaptive variants stabilized at intermediate frequencies are not considered in this approach, which can be an important source of adaptation considering the infinitesimal model.

If we focus on the effects on the polymorphisms, the expression 6 suggest that the ratio of polymorphism non-synonymous versus synonymous would increase by mutations having both positive and negative effects. It is expected that the number of the mutations having negative effects would rapidly decrease at higher

frequencies, while for mutations having positive effects this effect may be the inverse; then higher ratios of non-synonymous versus synonymous variability at higher frequencies may be explained by mutations having positive effects segregating at the population. Furthermore, in case that two populations from the same species have the same divergence ratio and no fixed mutations between them, then:

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

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 8)

In summary, the possible differential effect of selection (positive and negative) at any frequency between populations can be estimated from the ratios of variability synonymous and non-synonymous of the two populations, and the trajectory of this pattern along the frequencies may be used to interpret the effects of the different sources of selection.

Importantly, a number of demographic effects (e.g., bottlenecks) together with mutations having small selective effects may also disturb the ratios of variability. Given that, using this approach, a significant number of variants with small selective effect play an important role in the evolution of the population, and considering that demographic effects changing the population size would also have an effect on mutations with small selective effect, these different processes (selection and demographic changes) may be acting at the same order of magnitude and may be confounded.

In contrast, the effect of linkage between selective (detrimental or adaptive) and neutral variants should not affect the expected estimate of the proportion of adaptive variants, as they would affect in the same proportion both synonymous and non-synonymous positions. On the other hand, the interaction of variants with different selective effects (e.g., negative versus positive) would reduce the effect of selection (Hill and Robertson 1966; Booker and Keightley 2018) and would have a significant consequence on the estimation of adaptive fixed

variants. Uricchio, Petrov, and Enard (2019) estimate analytically the effect of background selection on the fixation probability over a genome.

Bootstrap analysis

Non-parametric bootstrap analysis was performed to estimate the confidence interval for α statistic for each breed and variability estimator. For each breed, synonymous and non-synonymous coding positions were randomly chosen separately (with replacement) until having the total number of positions as in observed data. Then, the α statistic (equation 2) using the four summary statistics of θ was calculated. This process was repeated 100 times to build the distribution of the expected α.

Simulations

We carried out forward simulatins using SLiM (Haller and Messer 2017) in order to explore the expected values of nucleotide diversity, divergence and α under different scenarios and analyzed the obtained results using mstatspop. We simulated different scenarios that consider different selective effects: standard neutral model (SNM), negative selection (NS) and positive selection (PS) occurring from the split of pigs from a recent outgroup to the present time. In addition to this, we also simulated the SNM and NS but adding the effect of PS only to the branch of domesticated pigs, just after the split from wild boars. Specifically, the new and the existing variants that were neutrally segregating at a given proportion of positions, become suddenly beneficial for a particular population until present. Moreover, all these models were also simulated with a reduction or an expansion of the effective population size in the branch corresponding to domestic pigs and adding a migration rate of 0.25 (per individual) from the wild boar branch to the domestic one. (See Table S4.3 for parameter values).

We set a constant effective population of pigs to 10,000 diploid individuals and let then to evolve during 100,000 generations to stabilize the mutations. After that, the population was split in two populations of 10,000 individuals (which roughly corresponds to the divergence between *Sus* and the outgroup) and remained

constant for 200,000 generations. 5,000 generations before present, one of the populations splits in two populations, which would correspond to domestic pigs and wild boars. In those scenarios with a reduction or expansion of the population size in the branch corresponding to domestic pigs, the number of individuals were decreased (or increased) 10 times starting 4,000 generations before present (Figure S4.1). Each scenario was run 100 times in order to have a distribution of the expected values. For each run and population, 100 from the 20,000 simulated sequences were randomly sampled 10 times, ending up with 1,000 samples for each scenario.

The common parameters were: mutation rate of 2.5e-7, recombination rate of 1.17e-8 and a gene length of 10,000 nucleotides (2/3 of positions are non-synonymous and 1/3 synonymous). Additivity (h = 0.5) was assumed for all mutations.

Pathway analysis

We downloaded the complete list of pathways and genes of *Sus scrofa* from KEGG v.20170213 (http://www.genome.jp/kegg/, Kanehisa et al. 2008). The list contained 471 pathways and 5480 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, to discard pathways that were not informative or too generic and complex. The final list contained 171 pathways and 3449 genes.

To analyze the selection pressure of each gene according its position in the pathway, we obtained different topological descriptors. First, we downloaded the XML file of each previously selected pathway from KEGG v.20170213. These files were analyzed with 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 descriptors 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 non-parametric Wilcoxon rank test, due to the extreme leptokurtic distributions.

Results

General description of Variants: Total, Exclusive and Shared Polymorphisms and Fixed Mutations

We first analyzed the population structure of the samples using principal component analyses (PCA), with the aim of visualizing the global differentiation of the studied populations and to identify putative sampling errors and/or to detect recent migrant individuals that may confound the results of the work. PCA analysis (Figure 4.1) shows that samples of same population cluster together and are well separated from each other.

We found a total of 6,684,142 SNPs in autosomal genes, from which 149,440 were present in coding regions. 12.5% and 32.2% of the SNPs in coding regions are shared among three and at least two populations and 31.2%, 2.2 and 34.4% are exclusive (private SNPs) of LW, IB and WB, respectively (Table 4.1). The proportion of private SNPs is in agreement to the known demographic history of the populations; the larger proportion of exclusive SNPs in LW compared to IB and WB is reflecting the introgression of Asian germplasm into LW (Bosse, Megens, Madsen, et al. 2014; Bosse, Megens, Frantz, et al. 2014), whereas the lowest proportion of private SNPs is observed in IB, which suffered a reduction of its effective population size (Esteve-Codina et al. 2013).

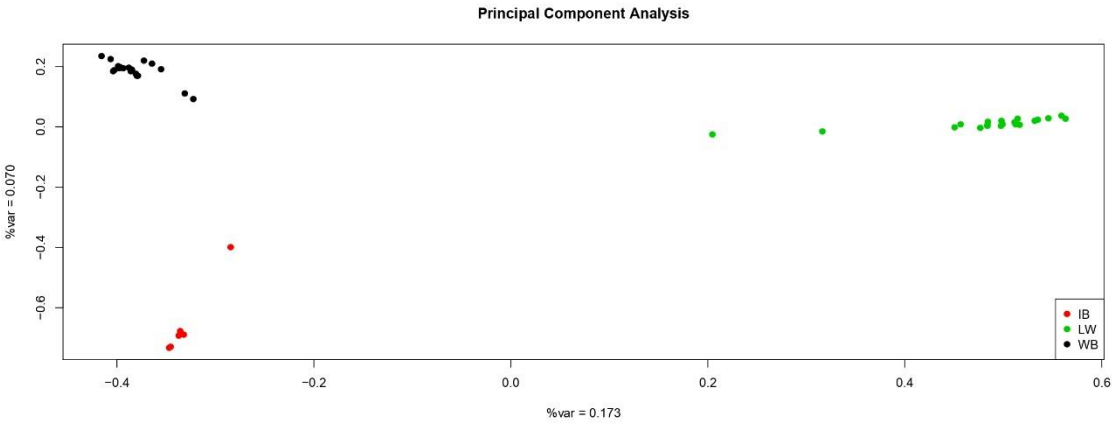

Figure 4.1: *Principal Component Analysis (PCA) based on the whole-genome SNPs of the three pig populations. *IB, Iberian; LW, Large White; WB, Wild boar.*

Focusing on coding regions, were classified variants as polymorphic, fixed (i.e., different allele from outgroup) or ancestral allele (i.e., same allele as outgroup) (Table 4.2). Surprisingly, we did not find any fixed variant completely associated to domestic (IB or LW) versus wild population. Similarly, there are relatively few variants fixed in domestic breeds but polymorphic in the wild boar, suggesting that the fixation of these variants (in IB and LW) present in the original wild boar population (before domestication) or, alternatively, transferred from WB to IB/LW by migration is uncommon. The same is true for fixed variants between any other population pair, indicating a lack of strong population differentiation among populations and therefore, the observed differences in phenotypic traits among populations are hardly explained by fixed coding variants. Only in the case of the IB breed, we found a high proportion of non-exclusive fixed variants (they are polymorphic in the other breeds), which is in agreement with its small population size. Most of new and exclusive variants are polymorphic, pointing to a recent origin (divergence time between populations was not enough to fix them). Remarkably, the ratio of non-synonymous to synonymous variants for each classification (ancestral, fixed, polymorphic) is similar in the three populations and for the four summary statistics, showing a higher proportion of synonymous versus non-synonymous mutations in all cases. This pattern does not suggest the action of differential effects of positive selection between domesticated and wild breeds.

Table 4.1: *Number of SNPs in different functional regions in the three populations of pigs classified as total, shared and exclusive variants. *IB, Iberian; LW, Large White; WB, Wild boar.*

	All SNPs	Shared SNPs between WB, IB and LW	Shared SNPs between WB and IB	Shared SNPs between WB and LW	Shared SNPs between IB and LW	Exclusive SNPs of WB	Exclusive SNPs of IB	Exclusive SNPs of LW
Number of SNPs	24,869,699	7,293,787	666,927	4,017,107	138,378	4,239,052	385,504	8,128,944
Number of SNPs of genes	6,684,142	1,964,562	98,433	1,152,555	48,351	1,138,370	100,550	2,181,321
Number of SNPs of coding regions	149,440	18,611	3,252	25,044	1,177	51,432	3,356	46,568

Table 4.2. *Number of SNPs in coding regions classified according to its allelic status in each population (A: Ancestral allele, F: Fixed allele, P: Polymorphic allele). SNPs that are missing in any of the populations are not considered in this table. *IB, Iberian; LW, Large White; WB, Wild boar.*

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

F: position with a fixed derived variant

P: polymorphic position

A: position with the ancestral variant

The genomic context and the network topology have a limited influence on selective pattern

The heterogeneity in recombination rate, gene density, %GC and amount of CpG islands across the genome can affect the local levels of variability. These effects,

if they are not considered, may mask the effects of different evolutionary forces and hinder the interpretation of the observations. A previous study using the Iberian breed (Esteve-Codina et al. 2013) detected a strong correlation between recombination and variability, although no correlation was observed between variability and gene density, neither with GC content. We did not observe any statistically significant correlation in any of the three studied breeds between α and recombination, neither with gene density (P-value > 0.01), missing rate, %GC and CpG

We also investigated the effect of gene network topology on selective patterns. We found (Figure S4.2) significant large values of betweenness (number of shortest paths going through a vertex) for negative α genes (non-parametric Wilcoxon rank-test, P-value < 0.01) for all populations; tests were significant at LW and WB for in-degree statistic (number of out-going edges). Therefore, the position of the gene in the pathway in which it interacts seems to have an effect on in its selective patterns. Those genes that occupy central positions in a pathway tend to be more evolutionary constrained than the peripherical genes (those that interact more frequently with the environment).

Levels of variability at coding regions

To assess the selective effect of domestication, we explored the pattern of variation at synonymous and non-synonymous positions using different estimators of variability: Watterson, Tajima, Fu&Li and Fay&Wu's estimators, that takes into account missing data and weight the frequencies of the SNPs in a different way (see material and methods). This analysis was performed considering all polymorphism (all) and classifying polymorphism as shared (if a variant in a specific population is shared with other populations) or exclusive of a given population.

i) All polymorphisms: First, we estimated the levels of variability per nucleotide at genome scale, (Figure 4.2 and Table S4.4). We did not observe striking differences in the ratio of non-synonymous versus synonymous regardless of the variability estimator used. The less variable population is IB for any of the estimators and its pattern of variability is also different from the other two

populations; for instance, the variability using the Fu&Li (based on singletons) is high in LW and WB and low in IB. Note than in all three breeds, alleles at high frequencies are proportionally more abundant than at intermediate allele frequencies, suggesting the effect of introgression/admixture or positive selection.

ii) Shared and Exclusive Polymorphisms: Figure 4.2 shows the levels and patterns of variability of the different breeds considering shared and exclusive polymorphisms among breeds. For all three populations, the amount of genome variability at Fu&Li estimates is mainly dominated by exclusive variants, as expected given that singletons are usually very recent variants. Nevertheless, the Iberian breed, a highly inbreeding isolated population, accumulates very low exclusive singletons compared with the other populations, which is also expected in a population that was isolated recently and has a low effective population size (Ramos-Onsins and Rozas 2002). On the other hand, shared polymorphisms prevail in variants at high frequencies (Fay&Wu's estimator). The proportion of exclusive versus shared polymorphisms are similar in Wild boar and Large White populations, while Iberian breed contains mostly shared polymorphisms.

The effect of Domestication inferred from α statistic: The different type and strength of selection (positive and negative, weak and strong) operating differently in domestic and wild populations could be detected by comparing the differential ratio between synonymous and non-synonymous polymorphisms and divergence levels through the calculation of α (see equation 3). Table S4.5 and Figure 4.3 shows the values of genome-wide α estimated using four variability estimators that exploit different information contained in the site frequency spectrum.

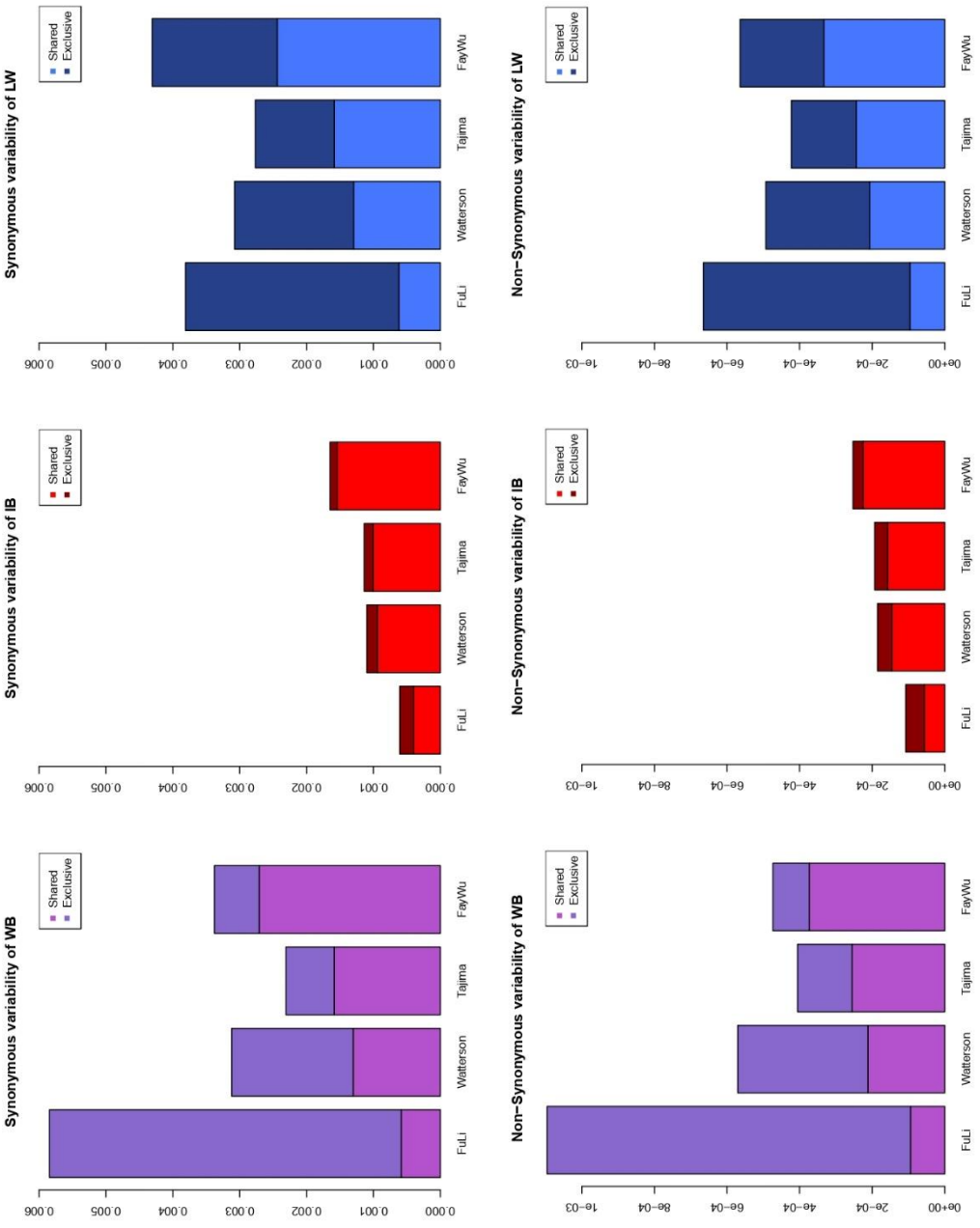

Figure 4.2: *Levels of variability at synonymous and non-synonymous positions for all breeds, variant classification and variability estimators.*

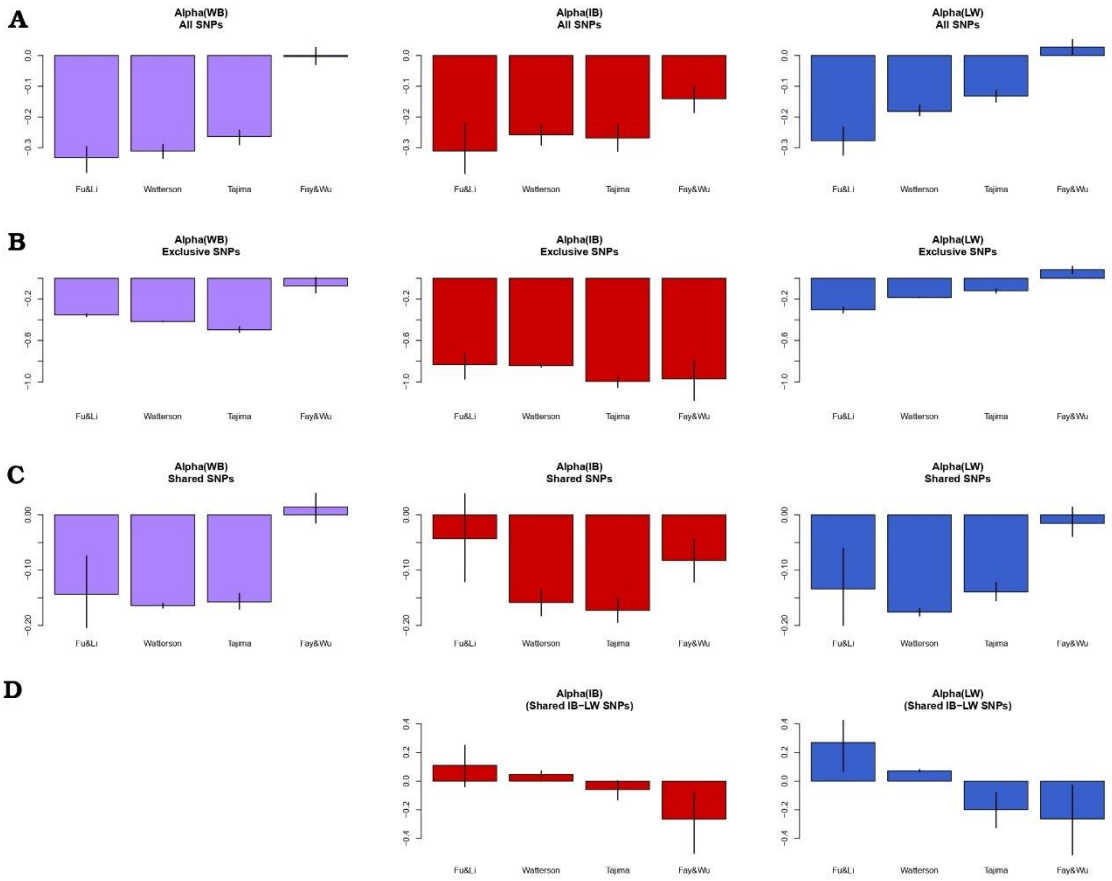

Figure 4.3: *Levels of α for all breeds, variant classification and variability estimators using: A) all SNPs (shared plus exclusive). B) exclusive SNPs of each population. C) shared SNPs between the populations. D) shared SNPs between the domestic breeds. Bootstrap intervals at 95% are indicated by a line at each bar. *IB, Iberian; LW, Large White; WB, Wild boar.*

i) All polymorphisms: We see a general concordance of the *α* patterns estimated at genome-wide scale (from low to high frequencies at each breed: Figure 3A). As expected, the α values are negative when the estimates of variability are based on low frequency variants ($\alpha_{Fu\&Li}$). This may be consequence of a relatively high proportion of deleterious mutations (versus neutral) that are segregating at low frequency. We observe a similar value of α at all three breeds, suggesting a similar proportion of deleterious mutations, irrespective to the domestication process or other demographic events. It is expected a marked decrease of the number of deleterious mutations at higher frequencies (see the tendency curve in Messer and Petrov (2013)). Therefore, we expected that bias in estimating α decreases when using variants at high frequencies (it is assumed

minimum presence of deleterious mutations). We observed higher values of α when using variability estimators based on high frequency variants for all populations, according to expectations that point to a pronounced elimination of deleterious mutations at higher frequencies. Nevertheless, the patterns of α at each breed are very different: wild boar has very low values of α except for $\alpha_{Fay\&Wu}$ (which is zero); domestic Iberian breed has also very negative values for all estimates, including $\alpha_{Fay\&Wu}$ (but this last is significantly less negative); Large White is the only breed that shows a linear increase of the α value, being positive at high frequencies.

The R_βγ ratio (equation 8), summarizes the differences in the ratios of non-synonymous versus synonymous variability between two different breeds, which notoriously have the same divergence ratio and no fixed variants among them. We observe deviations from 1 when using estimators based on high frequency variants, especially when comparing WB and IB (see Figure 4.4A). This indicates a higher proportion of non-synonymous variants at high frequencies (potentially deleterious and/or advantageous) are segregating at Iberian breed, suggesting a strong decay in population size (presence of deleterious variants at high frequencies).

ii) Exclusive and Shared polymorphisms: The distribution of α when using exclusive variants are expected to be more extreme because they are mostly recent variants. A high ratio of non-synonymous variants (versus synonymous) are segregating at low frequency (Figure 4.3B), suggesting that these variants have more deleterious effects than total variants (exclusive plus shared). Nevertheless, the ratio of α at intermediate frequency variants in WB and IB breeds unexpectedly decreased, suggesting a change in the effect of selection for maintaining non-synonymous variants at higher frequency (perhaps by attenuation of the deleterious effect of these mutations, or by positive selection) and insinuating that this pattern is not consequence of domestication but of demographic patterns (e.g., population size decline impacting on the effect of deleterious positions). In contrast, the pattern observed at LW breed is similar to considering total variants, with a linear increase of the value of α at higher variant frequencies. The R_βγ statistic (equation 8, Figure 4.4B) shows the same pattern

than considering total variants but amplified. The excess of non-synonymous variants at IB breed is likely due to the recent breed specific process.

The observed pattern of α at shared variants (Figure 4.3C) shows, in global terms, values closer to zero. Nevertheless, the pattern of α of shared variants differs from total; the value of α considering singletons have smaller negative values than α considering intermediate frequency variants. This pattern may also indicate that selective forces have increased the ratio non-synonymous polymorphisms up to intermediate frequencies. However, the shared α values are in general more moderate (closer to zero), perhaps because non-synonymous shared polymorphisms are functionally more constrained. The R_βγ statistic shows ratios close to one, although generally showing higher non-synonymous ratios for high frequencies at the Iberian breed (Figure 4.4C).

If we consider together the domestic breeds (LW and IB, Figure 4.3D), their shared variants show an inverse pattern in relation to total variants: strong positive values of α at low frequencies and very negative at high frequencies. There is a smaller ratio of non-synonymous low frequency variants versus divergence (perhaps is an active elimination of new variants to conserve differences among breeds) and a higher ratio of non-synonymous variants at high frequencies (perhaps the effect of population structure or alternatively the variants responsible of the initial process of domestication).

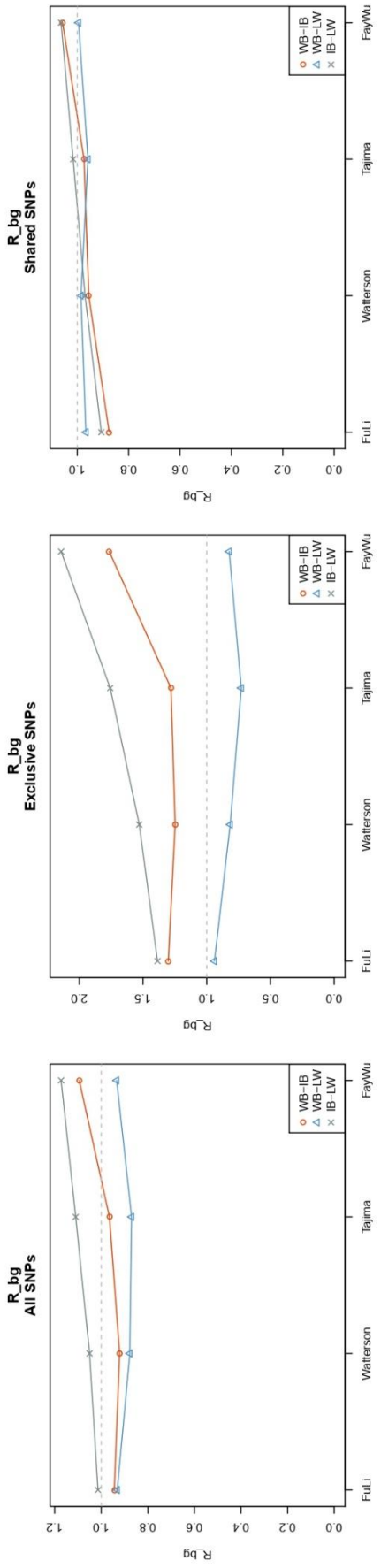

Figure 4.4: *Estimates of R_βγ ratio for all breeds, variant classification and variability estimators. *IB, Iberian, LW, Large White, WB, Wild boar.*

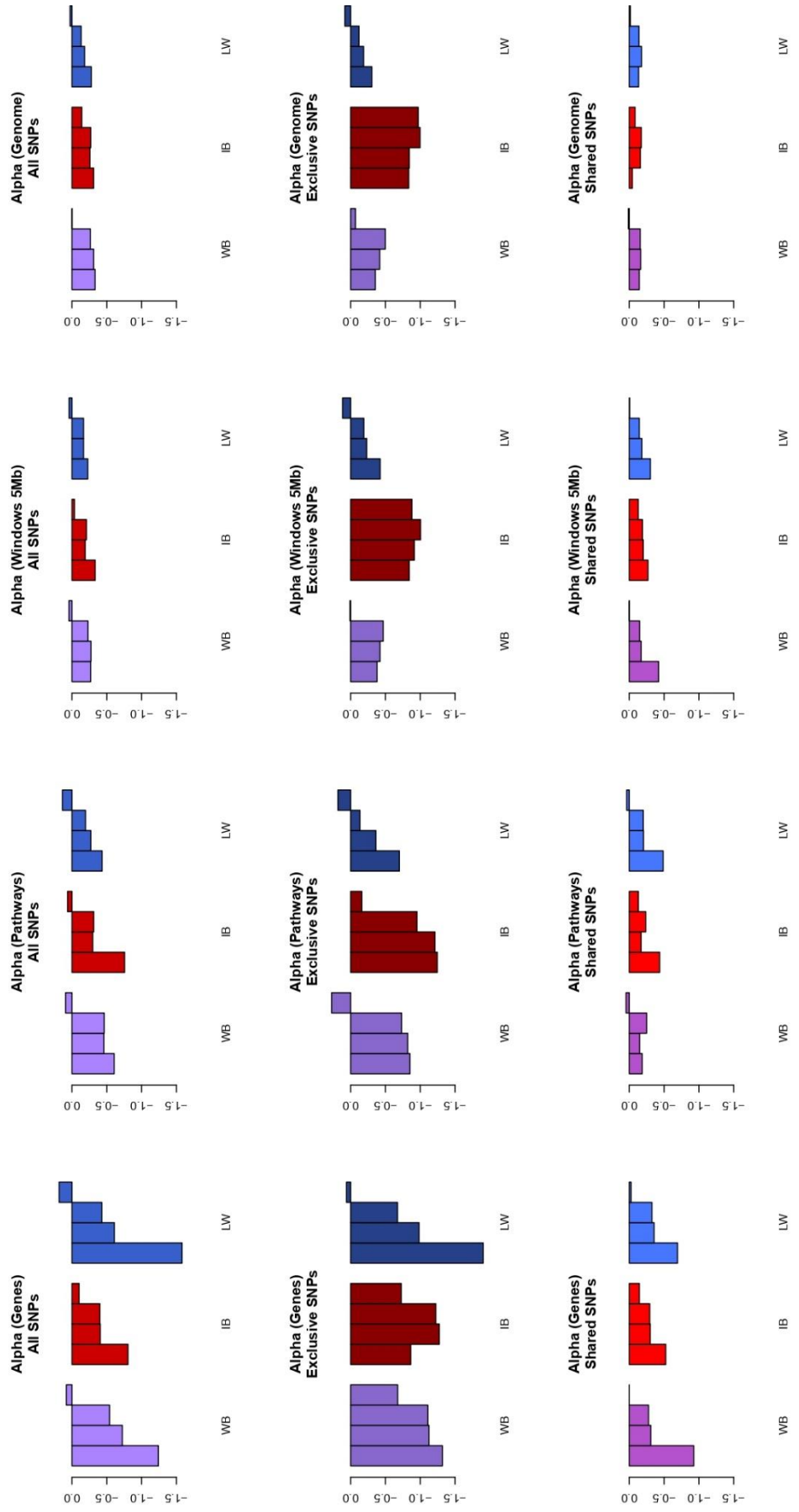

Figure 4.5: *Median values of α for all breeds, variant classification and variability estimators, for each different scale. *IB, Iberian; LW, Large White, WB, Wild boar. In each population of each plot, the order of α is: Fu&Li, Watterson, Tajima and Fay&Wu.*

The distribution of the α estimates at different functional scales of analysis: In addition to the genome-wide analysis, α was estimated for three additional scale levels: i) at gene level, ii) using genes within windows of size 5 Mb, and iii) considering all genes within the same pathway. Figure 4.5 shows the distribution of the median values of α for each scale level based on all, exclusive and shared polymorphisms and for the three populations. We observed that the differences in the value of α are notorious depending on the scale level analysis. The values of α are generally higher at gene scale level (very negative), whereas the α values at genome-wide scale are the closest to zero. The distribution of α values could have a large variance at gene scale since only few variants are used for its estimation. Note that for instance, there is an important number of genes having α = 1 (highest value), because the number of polymorphic non-synonymous variants is zero. On the other hand, if selection is acting at gene scale, the estimation of α using the whole genome data may be biased because the selective effects would be diluted since the proportion of variants located in genes is scarce compared to those located in intergenic regions. In any case, it is possible to identify genes, regions or pathways that are showing extreme α values. There is a moderately high correlation of the values of α between breeds (around 0.7, using Person correlation, and equivalent values considering exclusive or shared variants), suggesting, in general, similar selective effects. In the same way, window regions or gene pathways having high α values are reported as well (Table S4.6 and S4.7).

Simulations under different evolutionary scenarios and comparison with observed results

The different levels and patterns of variability observed among populations may be caused by the domestication process or by other selective and/or demographic events (Eyre-Walker 2002). In order to evaluate the main effects produced by different demographic and selective events, we simulated populations mimicking the process of domestication using SLiM software (Haller and Messer 2017) which also included several demographic events (change in population size and/or migration, see Materials and Methods). We use the α estimate calculated at genome scale to compare the different scenarios and the

observed and simulated data since we believe that this value will account for a summarized effect of selection on its population.

i) The patterns of α based on all variants: The estimates of α differ from the real data when demographic factors are considered (such as population reduction, expansion, migration) but also considering negative selection (see Figures S4.3 and S4). The observed patterns when considering all variants are mostly compatible with a predominant effect of deleterious mutations in all three breeds. Nevertheless, the effect of positive selection, visualized as a positive α value at high frequency variants, may be deeply influenced by demographic factors. For example, population size reduction or migration may increase the estimate of α with no real positive selection acting on the population. Therefore, the positive α value observed at LW population may be produced by recent introgression (a process that has been occurring in this breed). The comparison between the wild and domesticated populations, reflected by the R_βγ ratio, which shows ratios under 1 value for the three estimators or an increase to higher than 1 at high frequency variants, was compatible with combined weak positive and negative selection (Figures S4.5 and S4.6).

ii) The patterns of α based on exclusive variants: The observed patterns of α based on exclusive were differed from those based on total variants except for the LW breed. Under the conditions analyzed in this work, the pattern of α in simulations (Figures S4.7 and S4.8) did not fit the observed patterns with the exception of the LW breed (which is compatible with models including deleterious and beneficial mutations and expansion or migration). Regarding the ratio R_βγ between wild and domestic populations, a number of different models fit the observations, with the exception of the ratio at high frequency variants. This discrepancy may be due to several factors, for example by shorter temporal demographic effects or by selection dependent of frequency (Figures S4.9 and S4.10).

iii) The patterns of α based on shared variants: We have also examined those variants that are shared among populations (domestic versus wild). The α patterns for the observed data did not match the simulated ones for any scenario (Figures S4.11 and S4.12). The patterns that are closest were those considering population reduction, but even then, simulations were not able to explain the

higher values of α at high frequency variants. Considering the ratio R_βγ between wild and domestic populations, the observed patterns were mostly compatible with the SNM (Figures S4.13 and S4.14).

Discussion

Domestication occurred by an adaptive process driven by humans through artificial selection, which modified the features of the ancestral species into new phenotypical traits. Here, we have analyzed the variability and the divergence of several pig breeds (European wild, European domesticated, Commercial domesticated admixed) with the aim to find patterns at genomic level that show the effects of domestication. If the domestication process at pig species is driven by few major gene effects, the genomic signal would be weak, that is, the expected signal would be just the few selective sweeps produced by domestication, Instead, if many variants (with very small effect) are involved in the domestication process, the effect may be as well undetected (as predicted by the infinitesimal model, Fisher 1919). Intermediate scenarios may be possible, and effect of selection may be detected by an excess of the number of functional fixations across the genome.

Signals of domestication at coding regions in pigs

The lack of fixed variants at coding positions in the three breeds indicates that the observed heritable phenotypic differences between breeds are either given by: i) few selective sweeps, ii) selection at non-coding functional regions (not analyzed in this work) or/and iii) by pervasive changes in the frequencies of non-synonymous variants (not achieving fixation). Considering the first hypothesis, it is expected that domestication should fix the adaptive variant(s) for the genes of interest. The studied individuals have no fixed common variants among breeds. A number of soft selective sweeps may explain the surprising absence of fixed coding variants. We have considered the α values for a number of genes where previous studies reported signals of positive selection using other approaches (Groenen 2016). These genes have low or null non-synonymous polymorphisms or divergence (Table S4.8). This pattern is typical from regions close to selective

sweeps, although not necessarily imply that these genes are responsible (there are not signals of fixed variants at coding regions). The only significant values of alpha over zero were at KIT for IB breed, IGF2R and JMJD1C for LW breed and LRRTM3 for WB breed. The second hypothesis implies that the functional regions involved in domestication would not be coding regions. It is possible that many functional non-coding regions (promotors, enhancers, etc.) are affected by selection. This is a sensible hypothesis, but additional information should be considered to perform this analysis. Obtaining this information requires a very accurate analysis of homology and their functionality, which is very complicated at the genome level, especially for non-typical model species (such as Drosophila genus, *M. musculus*, *H. sapiens*). The third hypothesis suggests that a relatively large number of variants with small effect may have changed moderately the frequencies, and the phenotype would have been modified. In this case, depending on the effect size, the number of fixed variants may be significant at genome level, or alternatively, there would be only changes in the frequencies of the variants, not arriving to fixation. In this last case, the effect would be difficult to detect. If we assume that coding regions modified their function by domestication, then a number of coding variants implicated in domestication should be segregating in the populations. These segregating variants may modify the Site Frequency Spectrum, together with deleterious mutations, and should show an excess of non-neutral polymorphisms in relation to divergence. That is, negative α values should also be observed in case of positive selection variants that have not yet fixed.

Irrespective of the exact dating of divergence, focusing the analysis on shared SNPs between populations will be enriched in selective pressures that predate divergence, i.e., domestication. Likewise, while shared variants most likely predate domestication, they may reflect biological constraints on the species but, on the other hand, these shared polymorphisms can be the source of phenotypic variation in a polygenic selective scenario; they may have changed their frequencies (in an infinitesimal scenario) to give different features to each breed. Furthermore, private SNPs (those found segregating only in one of the breeds) may illuminate recent and breed-specific selective signals and differences between domesticated and wild breeds. In both cases, shared and exclusive

polymorphisms are contributing, by definition, to the differences in the site frequency spectrum between functional and non-functional positions and not to fixed divergence. Consequently, increases in the levels of non-synonymous polymorphisms (whatever be their selective effect, positive or negative) will reflect an increased negative value at the α statistic. In all, a differential increase of the ratio of non-synonymous versus synonymous polymorphisms may also indicate an adaptive change to increase non-synonymous polymorphisms up to intermediate frequencies.

Signals of domestication in two domestic pig breeds

The pig populations studied here have very different recent demographic (and selective) histories, which must be considered for an interpretation of the results. Wild boar population is a European-wide sample: the animals were collected from several regions of Europe. The wild boar population experimented a relatively recent population reduction (Groenen 2016), which can affect the patterns of detection of selection (see simulations, Figures S4.3-S4.13). On the other hand, the two domestic breeds analyzed have very different recent histories: while Iberian breed is a local Spanish breed that suffered a strong bottleneck during the 1970's (Esteve-Codina et al. 2013), the Large White breed ancestors were admixed with pigs of Asian origin (Bosse, Megens, Madsen, et al. 2014). Currently, about 30% of LW genome has been estimated to be of Asian origin, while so far, no evidence of introgression has been reported in the IB pig.

As expected, the IB breed had the lowest synonymous and non-synonymous variability among breeds studied because of its small effective population size and because the IB samples come from a very closed population of pigs. However, we expected a higher variability in LW than in WB due to the artificial selection and the Asian introgression, although we detected a very similar variability between them. We observed high variability in WB for the singletons and high in LW for the high-frequency derived alleles. This may be because the wild boar samples come from different regions of Europe, which can increase the number of exclusive SNPs of the wild boars depending of its origin region.

There are not more common patterns within domestic breeds than between wild and domestic breeds. In fact, wild boar and Iberian breeds have more similarities in the patterns of α than Iberian with Large White. A possible explanation may be the different history of each breed: Large White has been hybridized with Asian domestic pigs and thus, has accumulated a large number of differences in relation to European breeds. The domestication process is assumed independent (in genetical terms) in Asian and in European breeds. Thus, not sharing variants may have similar phenotypes (domesticated behavior), and no clear signals of domestication would be shared.

In conclusion, the observed data for all three breeds shows patterns compatible with the presence of deleterious mutations segregating in all three breeds and no clear signals of positive selection, i.e., the standard neutral model. Nevertheless, when the variants are split into shared and exclusive groups, we observe unexpected patterns that could not be mimicked using simulations with standard demographic scenarios (expansion, reduction, migration). There is a clear effect of deleterious mutations at low variant frequencies and mild or null effect at high frequencies. Additional work and tools for contrasting evolutionary models (e.g., an ABC analysis) that consider the effects of weak beneficial mutations segregating at the population and the effects of standing variants may shed more light to understand the patterns of variation in the domestication process.

Acknowledgments

JL-C is recipient of an FPI grant to achieve the PhD research from Ministry of Economy and Science (MINECO, Spain). Work funded by AGL2016-78709-R grants (MINECO). We also acknowledge the support of MINECO for the 'Centro de Excelencia Severo Ochoa 2016-2019' award SEV-2015-0533.

References

Achaz, Guillaume. 2009. "Frequency Spectrum Neutrality Tests: One for All and All for One." *Genetics* 183: 249–58. https://doi.org/10.1534/genetics.109.104042.

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

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

Bianco, Erica, Bruno Nevado, Sebastian E. Ramos-Onsins, and Miguel Pérez-Enciso. 2015. “A Deep Catalog of Autosomal Single Nucleotide Variation in the Pig.” *PLoS ONE* 10 (3): 1–21. https://doi.org/10.1371/journal.pone.0118867.

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

Bosse, Mirte, Hendrik-Jan Megens, Laurent A. F. Frantz, Ole Madsen, Greger Larson, Yogesh Paudel, Naomi Duijvesteijn, et al. 2014. “Genomic Analysis Reveals Selection for Asian Genes in European Pigs Following Human-Mediated Introgression.” *Nature Communications* 5: 4392. https://doi.org/10.1038/ncomms5392.

Bosse, Mirte, Hendrik-Jan Megens, Ole Madsen, Laurent A. F. Frantz, Yogesh Paudel, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2014. “Untangling the Hybrid Nature of Modern Pig Genomes: A Mosaic Derived from Biogeographically Distinct and Highly Divergent Sus Scrofa Populations.” *Molecular Ecology* 23 (16): 4089–4102. https://doi.org/10.1111/mec.12807.

Buitenhuis, Bart, Luc L G Janss, Nina A Poulsen, Lotte B Larsen, Mette K Larsen, and Peter Sørensen. 2014. “Genome-Wide Association and Biological Pathway Analysis for Milk-Fat Composition in Danish Holstein and Danish Jersey Cattle.” *BMC Genomics* 15: 1112.

https://doi.org/10.1186/1471-2164-15-1112.

Cruz, Fernando, Carles Vilà, and Matthew T. Webster. 2008. “The Legacy of Domestication: Accumulation of Deleterious Mutations in the Dog Genome.” *Molecular Biology and Evolution* 25 (11): 2331–36. https://doi.org/10.1093/molbev/msn177.

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

Daub, Josephine T., Tamara Hofer, Emilie Cutivet, Isabelle Dupanloup, Lluis Quintana-Murci, Marc Robinson-Rechavi, and Laurent Excoffier. 2013. “Evidence for Polygenic Adaptation to Pathogens in the Human Genome.” *Molecular Biology and Evolution* 30 (7): 1544–58. https://doi.org/10.1093/molbev/mst080.

Esteve-Codina, Anna, Yogesh Paudel, Luca Ferretti, Emanuele Raineri, Hendrik-Jan Megens, Luis Silió, María C Rodríguez, Martien A. M. Groenen, Sebastian E. Ramos-Onsins, and Miguel Pérez-Enciso. 2013. “Dissecting Structural and Nucleotide Genome- Wide Variation in Inbred Iberian Pigs.” *BMC Genomics* 14 (148): 1. https://doi.org/10.1186/1471-2164-14-148.

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

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

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

Fisher, R. A. 1919. “XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.” *Transactions of the Royal Society of Edinburgh* 52 (02): 399–433. https://doi.org/10.1017/S0080456800012163.

Frantz, Laurent A. F., Joshua G. Schraiber, Ole Madsen, Hendrik-Jan Megens,

Mirte Bosse, Yogesh Paudel, Gono Semiadi, et al. 2013. “Genome Sequencing Reveals Fine Scale Diversification and Reticulation History during Speciation in Sus.” *Genome Biology* 14 (9): R107. https://doi.org/10.1186/gb-2013-14-9-r107.

Frantz, Laurent A. F., Joshua G. Schraiber, Ole Madsen, Hendrik-Jan Megens, Alex Cagan, Mirte Bosse, Yogesh Paudel, Richard P. M. A. Crooijmans, Greger Larson, and Martien A. M. Groenen. 2015. “Evidence of Long-Term Gene Flow and Selection during Domestication from Analyses of Eurasian Wild and Domestic Pig Genomes.” *Nature Genetics* 47 (10): 1141–48. https://doi.org/10.1038/ng.3394.

Fraser, Hunter B, Aaron E Hirsh, Lars M Steinmetz, Curt Scharfe, and Marcus W Feldman. 2002. “Evolutionary Rate in the Protein Interaction Network.” *Science* 296 (5568): 750–52. https://doi.org/10.1126/science.1068696.

Fu, Yun-Xin, and Wen-Hsiung Li. 1993. “Maximum Likelihood Estimation of Population Parameters.” *Genetics* 134: 1261–70. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205593/pdf/ge13441261.pdf.

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

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

Guirao-Rico, Sara, Oscar Ramirez, Ana Ojeda, Marcel Amills, and Sebastian E. Ramos-Onsins. 2018. “Porcine Y-Chromosome Variation Is Consistent with the Occurrence of Paternal Gene Flow from Non-Asian to Asian Populations.” *Heredity* 120 (1): 63–76. https://doi.org/10.1038/s41437-017-0002-9.

Ha, Ngoc-Thuy, Josef Johann Gross, Annette van Dorland, Jens Tetens, Georg

Thaller, Martin Schlather, Rupert Bruckmaier, and Henner Simianer. 2015. "Gene-Based Mapping and Pathway Analysis of Metabolic Traits in Dairy Cows." *Plos One* 10 (3): 1–15. https://doi.org/10.1371/journal.pone.0122325.

Hahn, Matthew W., and Andrew D. Kern. 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, Benjamin C., and Philipp W. Messer. 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 A Robertson. 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, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, 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.

Leno-Colorado, Jordi, Nick J Hudson, Antonio Reverter, and Miguel Pérez-Enciso. 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, Heng, and Richard Durbin. 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, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Gonçalo R. Abecasis, Richard Durbin, 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.

Livingstone, Kevin, and Stephanie Anderson. 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.

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

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

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

Montanucci, Ludovica, Hafid Laayouni, Giovanni Marco Dall'Olio, and Jaume Bertranpetit. 2011. "Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway across Primates." *Molecular Biology and Evolution* 28 (1): 813–23. https://doi.org/10.1093/molbev/msq259.

Moon, Sunjin, Tae-Hun Kim, Kyung-Tai Lee, Woori Kwak, Taeheon Lee, Si-Woo Lee, Myung-Jick Kim, et al. 2015. "A Genome-Wide Scan for Signatures of Directional Selection in Domesticated Pigs." *BMC Genomics* 16 (1): 1–12. https://doi.org/10.1186/s12864-015-1330-x.

Ojeda, A, S E Ramos-Onsins, D Marletta, L S Huang, J M Folch, and M Pérez-Enciso. 2011. "Evolutionary Study of a Potential Selection Target Region in the Pig." *Heredity* 106 (2): 330–38. https://doi.org/10.1038/hdy.2010.61.

Pérez-Enciso, M., G. de los Campos, N. Hudson, J. Kijas, and A. Reverter. 2016. "The 'Heritability' of Domestication and Its Functional Partitioning in the Pig." *Heredity* 118: 160–68. https://doi.org/10.1038/hdy.2016.78.

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

Ramírez, Oscar, William Burgos-Paz, Encarna Casas, Maria Ballester, Erica Bianco, Iñigo Olalde, Gabriel Santpere, et al. 2014. “Genome Data from a Sixteenth Century Pig Illuminate Modern Breed Relationships.” *Heredity* 114 (2): 175–84. https://doi.org/10.1038/hdy.2014.81.

Ramos-Onsins, Sebastian E., and Julio Rozas. 2002. “Statistical Properties of New Neutrality Tests Against Population Growth.” *Molecular Biology and Evolution* 19 (12): 2092–2100. https://doi.org/10.1093/oxfordjournals.molbev.a004034.

Ramsay, Heather, Loren H Rieseberg, and Kermit Ritland. 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, Mark D, Richard E Miller, and Peter Tiffin. 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, Sebastien, and Loren H. Rieseberg. 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, Rebecca M., Wei Jin, and Greg Gibson. 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.

Rubin, Carl-Johan, Hendrik-Jan Megens, Alvaro Martinez Barrio, Khurram Maqbool, Shumaila Sayyab, Doreen Schwochow, Chao Wang, 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, Fumio. 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.

Uricchio, Lawrence H., Dmitri A. Petrov, and David Enard. 2019. “Exploiting Selection at Linked Sites to Infer the Rate and Strength of Adaptation.” *Nature Ecology & Evolution*, May, 1. https://doi.org/10.1038/s41559-019-0890-6.

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

Chapter 5

VCFcheck: A tool for VCF files diagnostics

VCFcheck: A tool for VCF files diagnostics

J. Leno-Colorado[1,*], M. Pérez-Enciso[1,2]

[1] Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, 08193 Bellaterra, Spain

[2] ICREA, Carrer de Lluís Companys 23, Barcelona 08010, Spain

* Author for correspondence: Jordi Leno-Colorado, Centre for Research in Agricultural Genomics (CRAG), 08193 Bellaterra, Spain; jordi.leno@cragenomica.es

(In preparation)

Abstract

Summary: Next-generation sequencing technologies have become the most powerful tool to identify genetic variants. However, this technology is error-prone due to different factors, such as alignment and base-calling errors. We have developed VCFcheck, a Python standalone and web-based program that allows assessing SNP quality and filtering VCF files in multisample and multipopulation analyses. VCFcheck can perform different analyses in real time from VCF file and represent them graphically, allowing to visually inspect weird patterns, e.g., a dependence of genotype frequencies on read depth or population. Percentage of missing data, reference allele frequency, coverage of samples, principal component analysis and Hardy-Weinberg test, among others, are available.

Availability and implementation: VCFcheck is available at https://github.com/CRAGENOMICA/VCFcheck

Contact: jordi.leno@cragenomica.es

Introduction

Advances in Next-Generation sequencing (NGS) technologies have revolutionized biology across numerous fields. The possibility of sequencing complete genomes from a large number of individuals is indeed a hallmark of modern Genomics. However, the large amount of data generated by NGS makes it also difficult handling and analyzing these data. Furthermore, NGS are error-prone technologies, errors that propagate through the successive bioinformatic steps. An example is the SNP calling process: accurate genotype identification depends on numerous factors such as read depth, base and map qualities. The exact error in a particular analysis is unlikely to be known so the practitioner needs to reach a compromise between a false SNP that is called and a real SNP that is not identified by the pipeline. A typical dilemma is the minimum – and maximum – read depths required to call a SNP, which can be critical to reliably identify heterozygous genotypes (Nielsen et al. 2011).

The VCF (variant call format) is the standard format to represent the mutations found from NGS data, and was developed by the 1000 Genomes project (Danecek et al. 2011; Auton et al. 2015). VCF format stores all the information of

the mutations (reference and alternative alleles, genotype, mapping quality, depth, etc.) and is compatible with most bioinformatic tools.

The need to analyze the mutations in a clear and intuitive way and identify a bad processing of the data or other problems, prompted us to develop a tool with an intuitive graphic interface that produces several SNP statistics.

Methods

The core system of VCFcheck is implemented in Dash (https://plot.ly/products/dash/), a Python framework for building web applications. The program runs as a local server and works on any web browser. The input of VCFcheck is a (g)VCF with one or more samples and an optional file with their corresponding populations. When the files are loaded into VCFcheck, the user will be able to perform several analyses in a simple way in real time and visualize the results.

For this purpose, the uploaded VCF file is converted into a Pandas dataframe (http://pandas.pydata.org) for an easier and faster manipulation of the data. Once the VCF is uploaded, the file is displayed in a table format. The user has the possibility of choosing the type of positions to analyze (all positions, SNPs (biallelic, multiallelic or both), INDELs and ROHs) and filter the genotypes according to three criteria: sample depth, mapping quality by SNP and the percentage of missing data per SNP. When these parameters are established, the user can perform different analyses and visualize them in form of plots. The application can represent the distributions of missing data by SNP and reference allele frequency in order to check whether there is a bias favoring the reference allele. If the VCF contains the coverage of each individual, VCFcheck also can represent its distribution by population or genotype. In the case of a multi-population VCF, it may represent the distribution of missing data by population, perform a principal component analysis (PCA), test whether populations are in Hardy-Weinberg equilibrium or plot the inbreeding coefficient distribution by population. The inbreeding coefficient within populations is calculated as:

$$F = 1 - \frac{H_{observed}}{H_{expected}}$$

where $H_{observed}$ is the actual frequency of heterozygosity in individuals within the population and $H_{expected}$ is the expected heterozygosity within the population assuming Hardy-Weinberg equilibrium. Positive values of F will indicate a deficiency of heterozygotes, whereas negative values indicate an excess (Holsinger and Weir 2009).

Plots representation are interactive, with the possibility of zooming in and out, knowing the value of any point simply by passing over and downloading the plot in PNG format. A filtered VCF file and a summary of the VCF (with warnings of possible biases) can be also exported. An example of the layout with the representation of a multi-individual VCF and the PCA is shown in the Figure 1.

Conclusions

We have developed VCFcheck, an independent stand-alone and web-based tool to process VCF files, obtain different descriptors of SNPs and samples, and perform some descriptive analysis. This application is oriented to end-users to facilitate SNP quality checking and asses possible biases in multi-sample multi-population VCF files.

Acknowledgments

Funding: This study is supported by the Grant for Research Staff (FPI) grant from the Ministry of Economy and Science (MINECO, Spain), of which J. Leno-Colorado is the recipient. Work was also founded by grant AGL2016-78709-R from MINECO to M. Pérez-Enciso.

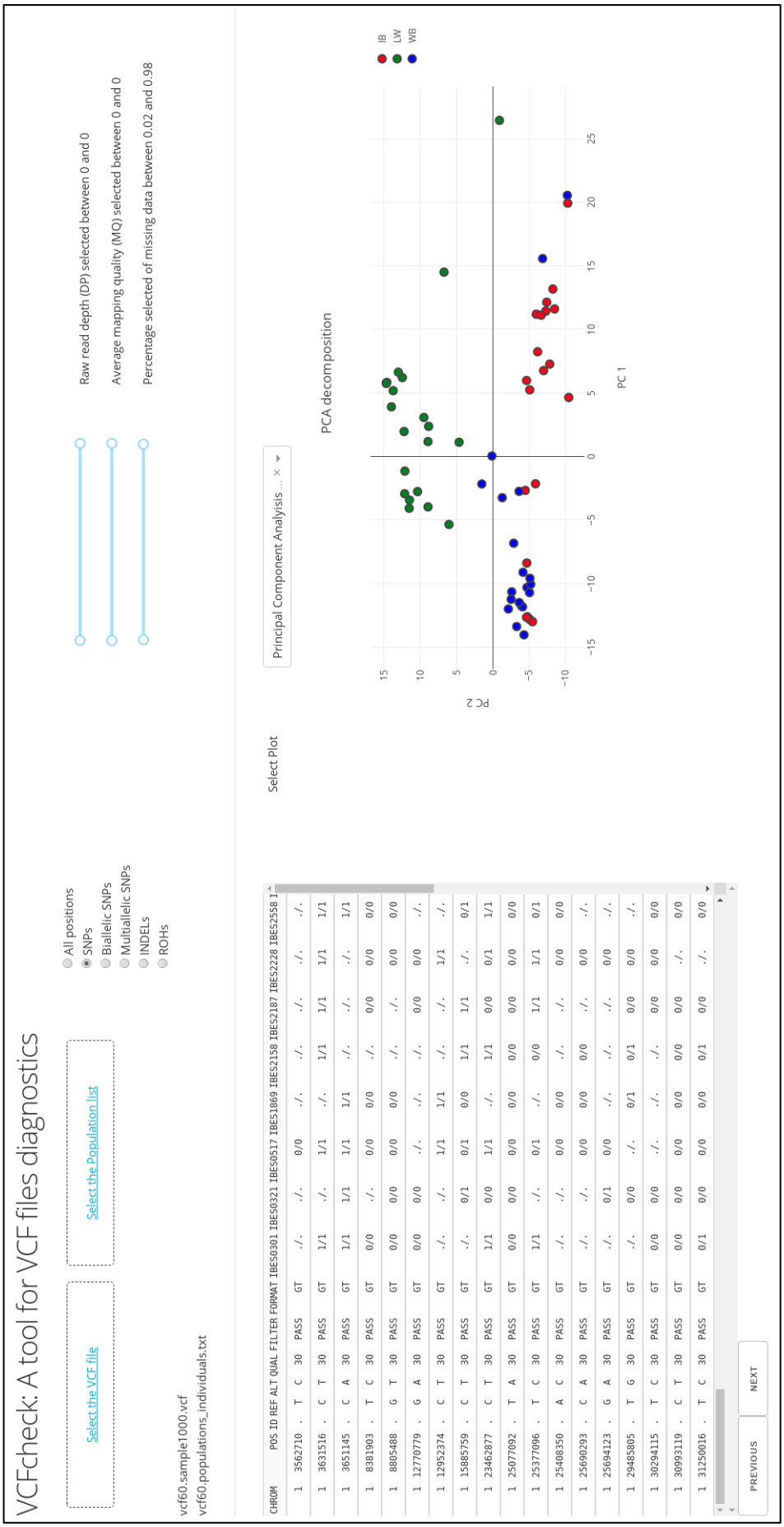

***Figure 5.1:** VCFcheck example layout with a multi-sample VCF and a PCA.*

REFERENCES

Auton, Adam, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Holsinger, Kent E., and Bruce S. Weir. 2009. “Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting FST.” *Nature Reviews Genetics* 10 (9): 639–50. https://doi.org/10.1038/nrg2611.

Nielsen, Rasmus, Joshua S Paul, Anders Albrechtsen, and Yun S Song. 2011. “Genotype and SNP Calling from Next-Generation Sequencing Data.” *Nature Review Genetics* 12 (6): 443–51. https://doi.org/10.1038/nrg2986.

Chapter 6

General Discussion

General Discussion

This thesis informs us about the footprint of selection left by the domestication events in the pig species, and therefore contributes to understand changes that are caused by domestication at the genome level.

The history of pig species has been characterized by different demographic events, including population size changes associated with domestication, and hybridization of domesticated pigs with wild boars. The interplay between demographic history and selection has shaped diversity across pig populations and genomes.

During domestication, animals are removed from its wild environment in order to develop a specific group of traits (aesthetic, physical or at the behavioral level). In the earliest domestication events, animals were likely selected for their behavior (greater docility and less aggressiveness) (Zeder 2015). Later, some traits that are breed-specific were targeted, mainly related to animal production, such as larger size or better reproductive performance.

Domestication at pathway level

As described in the introduction, demographic history of pig is very complex. In addition to the multiple independent domestication events, there were numerous breed-specific selective processes affecting. Previous studies (e.g., Amaral et al. 2011) reported breed-specific selective events, but here we also show (Chapter 3) the existence of selection signals shared across domestic populations.

We performed the analysis of domestication in pigs using the pathway as unit of analysis, comparing domestic breeds vs. wild boars. An advantage of the pathway analysis is a better biological interpretation of results. We made this analysis using two approaches, based on the differentiation and extended haplotype homozygosity, respectively, obtaining results that were partly coincident only. This difference has been observed in other studies (Chen et al. 2016; Dall'Olio et al. 2012) and it is probably due to the time and persistence of

the genetic changes. Because disequilibrium erodes fast, whereas differentiation endures a longer time (Sabeti et al. 2006). Differentiation analysis resulted in pathways related with development and behavior, such as the dopaminergic and serotonergic synapses or the Wnt and Hippo signaling pathways. On the other hand, we identified processes related with the reproduction in the disequilibrium analysis, such as the ovarian steroidogenesis or the arachidonic acid metabolism. These results agree with the demographic history of the pig, being the behavior and development the first selected traits in its domestication, whereas the reproduction is one of the most recent changes in the species and one of the main objectives in modern breeding.

As in other studies (Cruz, Vilà, and Webster 2008; Renaut and Rieseberg 2015; Pérez-Enciso et al. 2016), we observed a larger amount of deleterious mutations in domestic than in wild populations. On the other hand, we observed a decrease of deleterious mutations in genes of significant pathways, which would indicate a more constrained selection in differentiated processes of domestication.

In Chapter 3, we generated a co-association network with the significant pathways obtaining three global processes in which they interact: extracellular guidance, sympathetic nervous system and reproduction. Extracellular guidance is related with the growth and the hormone regulation, whereas the other two processes are negatively connected between them, i.e. when one is activated the other is inhibited, it happens for example in response to stress, which activates the sympathetic nervous system (e.g., muscle and heart contraction and lipids and carbohydrates degradation) and inhibits anabolic processes (e.g., synthesis of lipids and sexual hormones). As these processes are obtained from significant pathways, they will behave differently in wild and domestic pigs.

Different strength of selection in domestic and wild animals?

In Chapter 4, we tried to detect global differences in the coding regions at genomes level in the selection patterns of domestic populations and wild boars. However, when we analyzed the variability and the strength of selection of each population separately, we did not find differences between domestic and wild and no clear signs of domestication at genome level, i.e. there is no more common

patterns between domestic populations than between each of them with wild boars. This may be due to different factors, the selection could have a strong effect on a few genes, in this case, we would observe only signals in the corresponding windows of these genes and not at genome level; or, otherwise, the selection could have affected many variants throughout the genome, but with a very weak effect. Nevertheless, there are other options that must be considered, such as different selective effects in functional non-coding regions, e.g. promotors or enhancers.

In order to detect the real reason of different selection patterns in each population, we performed a series of simulations of different demographic scenarios. In these simulations we compared the standard neutral model, the negative selection and combined models with positive selection; besides we considered demographic factors like reduction, expansion or migration. When comparing real samples with simulations, we could detect the predominant effect of deleterious mutations in all populations. The excess of non-synonymous polymorphism at high frequencies could be caused by different factors: a reduction of the population size during a bottleneck could increase the frequency of these deleterious mutations or, otherwise, genetic hitchhiking caused by positive selection of a mutation could cause an increase in the frequency of linked deleterious alleles (Marsden et al. 2016). However, none of the models fitted the data perfectly well.

Domestication is a relatively recent event (~9,000 years ago), with little time for the fixation of new variants in a population. That is why the nucleotide diversity in the genome predates domestication and explains the general selective processes and constraints of the species. To be able to study the origin of the different variants, we classified the mutations into exclusive of each population and shared among them. With the shared mutations, we focused on the prior selective processes to domestication, and we detected similar patterns in all breeds, some of these patterns seem to be due to a reduction in the effective population size, which is consistent with the demographic history of the pig (Groenen et al. 2012). But this demographic event cannot solely explain the excess of non-synonymous polymorphism at high frequency. On the other hand, when we analyzed the exclusive mutations, obtaining specific patterns for each population, we expected to find differences between domestic and wild animals.

However, in the same way as with the use of all the SNPs, there are no clear selection signatures, we found more differences between domestic breeds than between domestic and wild boars. Large White pigs have a selection pattern that is compatible with the presence of deleterious and beneficial mutations, together with an expansion and/or migration. On the other hand, Iberian pigs and wild boars present a stronger negative selection than Large White, stronger with intermediate frequency variants, which suggest a population reduction in these populations, especially in the Iberian breed.

Overall, we observed a greater effect of demographic history of each population on patterns of selection and variability than the effect of domestication. Iberian pig is an autochthonous Spanish breed that have been suffered a strong bottleneck very recently (Esteve-Codina et al. 2013), Large White is an international commercial breed that have been improved artificially and introgressed with Asian germplasm, for the selection of desired traits, such as high reproductivity or growth (Bosse et al. 2014). As we expected by their demography, Iberian breed presents a low variability compared to Large White or wild boars, due to its small effective population size, while Large White is highly and widely variable, probably due to its hybridization with Asian pigs.

Diffuse or occasional domestication signatures

Analyses based on disequilibrium and on synonymous / non-synonymous rates are different, so in Chapters 3 and 4 we obtain different results. The scale used is also different. In the first work, the analyses of differentiation and extended homozygosity at the pathway level were performed from the statistics obtained in each mutation (nsL and Fst), while in the second work, the variability and selection were analyzed at genome level.

Regarding disequilibrium, we found some significantly differentiated pathways that are related with growth, behavior and reproduction, among others. This is consistent with the known phenotypes and demographic history of domestic pigs. Nevertheless, these are general results that have been obtained from the study of domestication in Europe and Asia separately, and many of the significant pathways come from the Asian analysis, where pigs have a greater genetic

diversity than European ones (Groenen et al. 2012). On the other hand, there are no clear domestication signatures at genome level, but differences in selection patterns due to each breed demographic history are detected, such as a reduction in the population size in Iberian breed or the presence of beneficial mutations and an expansion and/or migration in Large White pigs. In this last study only breeds from Europe were used, a continent in which we found only some significantly differentiated pathways between domestic and wild pigs for the first work. These significant pathways could be the result of domestication, but it is also possible that they are due to breeding processes or because of the effect of Asian alleles that are present in several breeds of those used for the analysis.

A possible explanation of the presence of only weak signals of selection in the patterns is the presence of soft selective sweeps. Soft selective sweeps are events in which an existing beneficial mutation increases its frequency until it becomes fixed or there is more than one mutation present that increases its frequencies competing for fixation (Schrider and Kern 2017). In this way, already existing beneficial mutations are fixed or increase their frequency but without showing a clear signal, possibly observing the selection at the phenotypic level but not at the genotypic level. Nevertheless, there are some studies that question the contribution of soft selective sweeps to the positive selection (e.g., Jensen 2014; Harris, Sackman, and Jensen 2018), due to the lack of evidences of fixations from soft sweeps and the difficulty of distinguishing these events of neutral demographic patterns.

The need for a diagnostic software of VCF files

For the analysis of the large amount of existing genomic data, different bioinformatic tools and software are needed to identify the genetic variants. The problem of these tools is the widespread presence of errors in different steps such as base-calling or alignment. The possibility of errors in the sequence and SNP-calling must be considered since this will affect the subsequent analysis of the data.

Because all our work is based on SNP data and the possibility of errors in the SNP calling process is high, a tool to analyze the quality of the results is highly

needed. In Chapter 5, we developed a stand-alone web application, VCFcheck, which analyses multi-individual multi-population VCF files, obtaining a series of parameters, which are graphically represented in order to study the quality of the samples and the SNP-calling. Furthermore, the VCF file can be filtered, removing those SNPs with low sample depth or high proportion of missing data. In this way, this tool can give us some warnings about the possible biases in the VCF file, such as the correlation between depth and genotype, different proportion of genotypes between population or genotypes are not in Hardy-Weinberg equilibrium, among others.

Perspectives

The studies presented in this thesis have increased our knowledge about the domestication and its effect in the genome, but there are still unresolved questions and future researches can be identified.

First, in order to know the origin of the signs of differentiation or disequilibrium, the analysis could be repeated by separating the European samples in breeds with introgression and breeds without introgression, and to compare each subgroup with European wild boars. We could therefore investigate whether the results obtained in Europe are due to Asian influence and, therefore, are differences between Europe and Asia and not between domestic and wild animals. However, this requires a larger amount of available sequences of autochthonous animals without introgression.

To strengthen the study of selection in different breeds and the effect of domestication, additional samples of available domestic pigs would be also very useful. Since the complex demographic history of pig breeds influences the patterns of selection, it is very difficult to conclude the reason of the obtained signals. As we described above, selection pattern in Large White could be influenced by Asian alleles and the low variability in Iberian breed, due to the recent bottleneck, causes a mild or null positive selection. In addition to that, the number of individuals, especially in Iberian, is very low. For all these reasons, a greater number of domestic breeds could help to disentangle the actual reason for the different signals. The use of other commercial breeds with Asian

introgression and more autochthonous (without introgression) breeds (and individuals) could shed light on the results, detecting patterns due to Asian contribution or to domestication origin.

The estimation of parameters using simulations of evolutionary models, ABC (Approximate Bayesian Computation) or other simulation-based inference methods could be very useful to fit and understand the selection patterns of our observed results. In a recent study by Uricchio, Petrov, and Enard (2019) a new algorithm to infer the rate and strength of selection has been developed, taking into account the weakly beneficial mutations. This method can be of great help for our work, although as we study domestication and did not find fixed sites, we do not know how useful the method will be for our data.

Finally, as we commented previously, detection of possible errors in the data is fundamental to avoid problems that would influence the whole work. It was a problem that we suffered during the analysis of Chapter 4, in which we originally had more samples of Iberian breeds (20), which caused strange results, such as a greater variability in this breed than in the other populations. VCFcheck is a useful diagnostics tool and the continuous development of application functions and implementations will help to detect biases related with the quality and the sample depth.

Chapter 7
Conclusions

Conclusions

1. The analysis of functional processes involved in the domestication and/or breeding processes in pig has affected pathways related with behavior, development, growth and reproduction, among others. Domestication/breeding also seemingly modified processes related with stress and with cellular and hormonal regulation.
2. A higher proportion of deleterious mutations was detected in domestic animals compared with wild boars.
3. Nevertheless, a decreased accumulation of deleterious mutations was found in significant pathways, which indicates a stronger evolutionary constraint in those genes/pathways, likely because they play a central role in the development of the pig.
4. Genes in a central pathway position were more evolutively constrained than peripheral genes.
5. There are no clear signatures of domestication at the genomic level. Instead, the demography of each population seems to have played a major role than selection. Signs of a reduction in the population size are obtained in Iberian breed, while in Large White there is an influence of deleterious and beneficial mutations, together with an expansion and/or migration.
6. Quality control is essential in these large-scale analyses. To help in this task, we have developed a web-based application that allows to assess the SNP quality and filter VCF files.

References

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

Ai, Huashui, Xiaodong Fang, Bin Yang, Zhiyong Huang, Hao Chen, Likai Mao, Feng Zhang, et al. 2015. “Adaptation and Possible Ancient Interspecies Introgression in Pigs Identified by Whole-Genome Sequencing.” *Nature Genetics* 47 (3): 217–25. https://doi.org/10.1038/ng.3199.

Albrechtsen, Anders, Finn Cilius Nielsen, and Rasmus Nielsen. 2010. “Ascertainment Biases in SNP Chips Affect Measures of Population Divergence.” *Molecular Biology and Evolution* 27 (11): 2534–47. https://doi.org/10.1093/molbev/msq148.

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

Amaral, Andreia J, Luca Ferretti, Hendrik-Jan Megens, Richard P M a Crooijmans, Haisheng Nie, Sebastian E. Ramos-Onsins, Miguel Perez-Enciso, Lawrence B Schook, and Martien a M Groenen. 2011. “Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA.” Edited by Hans Ellegren. *PLoS ONE* 6 (4). https://doi.org/10.1371/journal.pone.0014782.

Bianco, Erica, H.W. Soto, L. Vargas, and Miguel Pérez-Enciso. 2015. “The Chimerical Genome of Isla Del Coco Feral Pigs (Costa Rica), an Isolated Population since 1793 but with Remarkable Levels of Diversity.” *Molecular Ecology* 24: 2364–78. https://doi.org/10.1111/mec.13182.

Bosse, Mirte, Hendrik-Jan Megens, Ole Madsen, Laurent A. F. Frantz, Yogesh Paudel, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2014. “Untangling the Hybrid Nature of Modern Pig Genomes: A Mosaic Derived

from Biogeographically Distinct and Highly Divergent Sus Scrofa Populations." *Molecular Ecology* 23 (16): 4089–4102. https://doi.org/10.1111/mec.12807.

Bosse, Mirte, Hendrik Jan Megens, Ole Madsen, Yogesh Paudel, Laurent A. F. Frantz, Lawrence B. Schook, Richard P.M.A. Crooijmans, and Martien A.M. Groenen. 2012. "Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape." *PLoS Genetics* 8 (11). https://doi.org/10.1371/journal.pgen.1003100.

Buitenhuis, Bart, Luc L G Janss, Nina A Poulsen, Lotte B Larsen, Mette K Larsen, and Peter Sørensen. 2014. "Genome-Wide Association and Biological Pathway Analysis for Milk-Fat Composition in Danish Holstein and Danish Jersey Cattle." *BMC Genomics* 15: 1112. https://doi.org/10.1186/1471-2164-15-1112.

Casillas, Sònia, and Antonio Barbadilla. 2017. "Molecular Population Genetics." *Genetics* 205 (3): 1003–35. https://doi.org/10.1534/genetics.116.196493.

Chen, M, D Pan, H Ren, J Fu, J Li, G Su, A Wang, L Jiang, Q Zhang, and JF Liu. 2016. "Identification of Selective Sweeps Reveals Divergent Selection between Chinese Holstein and Simmental Cattle Populations." *Genetics Selection Evolution* in press.

Cornille, Amandine, Pierre Gladieux, Marinus J.M. Smulders, Isabel Roldán-Ruiz, François Laurens, Bruno Le Cam, Anush Nersesyan, et al. 2012. "New Insight into the History of Domesticated Apple: Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties." *PLoS Genetics* 8 (5). https://doi.org/10.1371/journal.pgen.1002703.

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

Dall'Olio, Giovanni Marco, Hafid Laayouni, Pierre Luisi, Martin Sikora, Ludovica Montanucci, and Jaume Bertranpetit. 2012. "Distribution of Events of Positive Selection and Population Differentiation in a Metabolic Pathway:

The Case of Asparagine N-Glycosylation." *BMC Evolutionary Biology* 12 (98): 1–13. https://doi.org/10.1186/1471-2148-12-98.

Daub, Josephine T., Tamara Hofer, Emilie Cutivet, Isabelle Dupanloup, Lluis Quintana-Murci, Marc Robinson-Rechavi, and Laurent Excoffier. 2013. "Evidence for Polygenic Adaptation to Pathogens in the Human Genome." *Molecular Biology and Evolution* 30 (7): 1544–58. https://doi.org/10.1093/molbev/mst080.

Diamond, Jared. 2002. "Evolution, Consequences and Future of Plant and Animal Domestication." *Nature* 418 (6898): 700–707. https://doi.org/10.1038/nature01019.

Dobney, Keith, and Anton Ervynck. 2000. "Interpreting Developmental Stress in Archaeological Pigs: The Chronology of Linear Enamel Hypoplasia." *Journal of Archaeological Science* 27 (7): 597–607. https://doi.org/10.1006/jasc.1999.0477.

Duarte, Carlos M, Nuria Marba, and Marianne Holmer. 2007. "Rapid Domestication of Marine Species." *Science* 316 (5823): 382–83. http://www.sciencemag.org.

Endler, John A. 1986. *Natural Selection in the Wild*. Princeton University Press. https://press.princeton.edu/titles/2354.html.

Eriksson, Jonas, Greger Larson, Ulrika Gunnarsson, Bertrand Bed'hom, Michele Tixier-Boichard, Lina Strömstedt, Dominic Wright, et al. 2008. "Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken." *PLoS Genetics* 4 (2): e1000010. https://doi.org/10.1371/journal.pgen.1000010.

Esteve-Codina, Anna, R. Kofler, H. Himmelbauer, L. Ferretti, A. P. Vivancos, M. A.M. Groenen, J. M. Folch, M. C. Rodríguez, and M. Pérez-Enciso. 2011. "Partial Short-Read Sequencing of a Highly Inbred Iberian Pig and Genomics Inference Thereof." *Heredity* 107 (3): 256–64. https://doi.org/10.1038/hdy.2011.13.

Esteve-Codina, Anna, Yogesh Paudel, Luca Ferretti, Emanuele Raineri, Hendrik-Jan Megens, Luis Silió, María C Rodríguez, Martein a M Groenen,

Sebastian E Ramos-Onsins, and Miguel Pérez-Enciso. 2013. “Dissecting Structural and Nucleotide Genome-Wide Variation in Inbred Iberian Pigs.” *BMC Genomics* 14 (January): 148. https://doi.org/10.1186/1471-2164-14-148.

Eyre-Walker, Adam. 2006. “The Genomic Rate of Adaptive Evolution.” *Trends in Ecology and Evolution* 21 (10): 569–75. https://doi.org/10.1016/j.tree.2006.06.015.

Fang, Meiying, Greger Larson, Helena Soares Ribeiro, Ning Li, and Leif Andersson. 2009. “Contrasting Mode of Evolution at a Coat Color Locus in Wild and Domestic Pigs.” Edited by Gregory S. Barsh. *PLoS Genetics* 5 (1): e1000341. https://doi.org/10.1371/journal.pgen.1000341.

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

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. “On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure.” *Molecular Biology and Evolution* 31 (5): 1275–91. https://doi.org/10.1093/molbev/msu077.

Frantz, Alain C., Giovanna Massei, and Terry Burke. 2012. “Genetic Evidence for Past Hybridisation between Domestic Pigs and English Wild Boars.” *Conservation Genetics* 13 (5): 1355–64. https://doi.org/10.1007/s10592-012-0379-1.

Frantz, Laurent A. F., Ole Madsen, Hendrik-Jan Megens, Joshua G. Schraiber, Yogesh Paudel, Mirte Bosse, Richard P.M.A. Crooijmans, Greger Larson, and Martien A.M. Groenen. 2015. “Evolution of Tibetan Wild Boars.” *Nature Genetics* 47 (3): 188–89. https://doi.org/10.1038/ng.3197.

Frantz, Laurent A. F., Erik Meijaard, Jaime Gongora, James Haile, Martien A. M. Groenen, and Greger Larson. 2016. “The Evolution of Suidae.” *Annual Review of Animal Biosciences* 4 (1): 61–85. https://doi.org/10.1146/annurev-animal-021815-111155.

Frantz, Laurent A. F., Joshua G. Schraiber, Ole Madsen, Hendrik-Jan Megens,

Mirte Bosse, Yogesh Paudel, Gono Semiadi, et al. 2013. “Genome Sequencing Reveals Fine Scale Diversification and Reticulation History during Speciation in Sus.” *Genome Biology* 14 (9): R107. https://doi.org/10.1186/gb-2013-14-9-r107.

Frantz, Laurent A. F., Joshua G. Schraiber, Ole Madsen, Hendrik-Jan Megens, Alex Cagan, Mirte Bosse, Yogesh Paudel, Richard P. M. A. Crooijmans, Greger Larson, and Martien A. M. Groenen. 2015. “Evidence of Long-Term Gene Flow and Selection during Domestication from Analyses of Eurasian Wild and Domestic Pig Genomes.” *Nature Genetics* 47 (10): 1141–48. https://doi.org/10.1038/ng.3394.

Fraser, Hunter B, Aaron E Hirsh, Lars M Steinmetz, Curt Scharfe, and Marcus W Feldman. 2002. “Evolutionary Rate in the Protein Interaction Network.” *Science* 296 (5568): 750–52. https://doi.org/10.1126/science.1068696.

Fritzsche, Peter, Karsten Neumann, Karsten Nasdal, and Rolf Gattermann. 2006. “Differences in Reproductive Success between Laboratory and Wild-Derived Golden Hamsters (Mesocricetus Auratus) as a Consequence of Inbreeding.” *Behavioral Ecology and Sociobiology* 60 (2): 220–26. https://doi.org/10.1007/s00265-006-0159-3.

Fu, Yun-Xin, and Wen-Hsiung Li. 1993. “Maximum Likelihood Estimation of Population Parameters.” *Genetics* 134: 1261–70. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205593/pdf/ge13441261.pdf.

Geer, Lewis Y., Aron Marchler-Bauer, Renata C. Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H. Bryant. 2010. “The NCBI BioSystems Database.” *Nucleic Acids Research* 38 (Database issue): D492-6. https://doi.org/10.1093/nar/gkp858.

Gentry, Anthea, Juliet Clutton-Brock, and Colin P. Groves. 2004. “The Naming of Wild Animal Species and Their Domestic Derivatives.” *Journal of Archaeological Science* 31 (5): 645–51. https://doi.org/10.1016/j.jas.2003.10.006.

Giuffra, E, J M H Kijas, V. Amarger, J-t Jeon, and Leif Andersson. 2000. “The Origin of the Domestic Pig: Independent Domestication and Subsequent

Introgression." *Genetics* 154: 1785–91. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461048/pdf/10747069.pdf.

Goedbloed, D. J., H. J. Megens, P. Van Hooft, J. M. Herrero-Medrano, W. Lutz, P. Alexandri, R. P.M.A. Crooijmans, et al. 2013. "Genome-Wide Single Nucleotide Polymorphism Analysis Reveals Recent Genetic Introgression from Domestic Pigs into Northwest European Wild Boar Populations." *Molecular Ecology* 22 (3): 856–66. https://doi.org/10.1111/j.1365-294X.2012.05670.x.

Goedbloed, Daniel J, Pim van Hooft, Hendrik-Jan Megens, Katharina Langenbeck, Walburga Lutz, Richard PMA Crooijmans, Sip E van Wieren, Ron C Ydenberg, and Herbert HT Prins. 2013. "Reintroductions and Genetic Introgression from Domestic Pigs Have Shaped the Genetic Population Structure of Northwest European Wild Boar." *BMC Genetics* 14 (1): 43. https://doi.org/10.1186/1471-2156-14-43.

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

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

Ha, Ngoc-Thuy, Josef Johann Gross, Annette van Dorland, Jens Tetens, Georg Thaller, Martin Schlather, Rupert Bruckmaier, and Henner Simianer. 2015. "Gene-Based Mapping and Pathway Analysis of Metabolic Traits in Dairy Cows." *Plos One* 10 (3): 1–15. https://doi.org/10.1371/journal.pone.0122325.

Haasl, R. J., and B. A. Payseur. 2011. "Multi-Locus Inference of Population Structure: A Comparison between Single Nucleotide Polymorphisms and Microsatellites." *Heredity* 106 (1): 158–71. https://doi.org/10.1038/hdy.2010.21.

Hahn, Matthew W., and Andrew D. Kern. 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.

Hanotte, O, Bradley D.G., Ochieng J.W., Verjee Y., Hill E.W., and Rege E.O. 2002. “African Pastoralism : Genetic Imprints of Origins and Migrations.” *Science* 296 (April): 336–39.

Harris, Rebecca B., Andrew Sackman, and Jeffrey D. Jensen. 2018. “On the Unfounded Enthusiasm for Soft Selective Sweeps II: Examining Recent Evidence from Humans, Flies, and Viruses.” Edited by Jeffrey Ross-Ibarra. *PLOS Genetics* 14 (12): e1007859. https://doi.org/10.1371/journal.pgen.1007859.

Heerwaarden, Joost van, John Doebley, William H. Briggs, Jeffrey C. Glaubitz, Major M. Goodman, Jose de Jesus Sanchez Gonzalez, and Jeffrey Ross-Ibarra. 2011. “Genetic Signals of Origin, Spread, and Introgression in a Large Sample of Maize Landraces.” *Proceedings of the National Academy of Sciences* 108 (3): 1088–92. https://doi.org/10.1073/pnas.1013011108.

Helyar, S. J., J. Hemmer-Hansen, D. Bekkevold, M. I. Taylor, R. Ogden, M. T. Limborg, A. Cariani, et al. 2011. “Application of SNPs for Population Genetics of Nonmodel Organisms: New Opportunities and Challenges.” *Molecular Ecology Resources* 11 (SUPPL. 1): 123–36. https://doi.org/10.1111/j.1755-0998.2010.02943.x.

Jensen, Jeffrey D. 2014. “On the Unfounded Enthusiasm for Soft Selective Sweeps.” *Nature Communications*. Nature Publishing Group. https://doi.org/10.1038/ncomms6281.

Jordana, J., P. M. Pares, and A. Sanchez. 1995. “Analysis of Genetic Relationships in Horse Breeds.” *Journal of Equine Veterinary Science* 15 (7): 320–28. https://doi.org/10.1016/S0737-0806(06)81738-7.

Kanehisa, Minoru, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, 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.

Kimura, Motoo. 1968. “Evolutionary Rate at the Molecular Level.” *Nature* 217 (5129): 624–26. http://www.ncbi.nlm.nih.gov/pubmed/5637732.

Kimura, Motoo. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Larson, Greger, Umberto Albarella, Keith Dobney, Peter Rowley-Conwy, Jörg Schibler, Anne Tresset, Jean-Denis Vigne, et al. 2007. “Ancient DNA, Pig Domestication, and the Spread of the Neolithic into Europe.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (39): 15276–81. https://doi.org/10.1073/pnas.0703411104.

Larson, Greger, and Joachim Burger. 2013. “A Population Genetics View of Animal Domestication.” *Trends in Genetics* 29 (4): 197–205. https://doi.org/10.1016/j.tig.2013.01.003.

Larson, Greger, Keith Dobney, Umberto Albarella, Meiying Fang, Elizabeth Matisoo-Smith, Judith Robins, Stewart Lowden, et al. 2005. “Worldwide Phylogeography of Wild Boar Reveals Multiple Centers of Pig Domestication.” *Science* 307 (5715): 1618–21. https://doi.org/10.1126/science.1106927.

Larson, Greger, and Dorian Q. Fuller. 2014. “The Evolution of Animal Domestication.” *Annual Review of Ecology, Evolution, and Systematics* 45 (1): 115–36. https://doi.org/10.1146/annurev-ecolsys-110512-135813.

Lewontin, Richard C. 1970. “The Units of Selection.” *Annual Review of Ecology and Systematics* 1 (1): 1–18. https://doi.org/10.1146/annurev.es.01.110170.000245.

Lewontin, Richarc C. 1974. *The Genetic Basis of Evolutionary Change*. New York. https://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/evolution/public/papers/lewontin1974/lewontin1974_chap1andfront.pdf.

Livingstone, Kevin, and Stephanie Anderson. 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.

Ljungqvist, Marcus, Mikael Åkesson, and Bengt Hansson. 2010. “Do

Microsatellites Reflect Genome-Wide Genetic Diversity in Natural Populations? A Comment on Väli et Al. (2008).” *Molecular Ecology* 19 (5): 851–55. https://doi.org/10.1111/j.1365-294X.2010.04522.x.

Lou, D. I., J. A. Hussmann, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer. 2013. “High-Throughput DNA Sequencing Errors Are Reduced by Orders of Magnitude Using Circle Sequencing.” *Proceedings of the National Academy of Sciences* 110 (49): 19872–77. https://doi.org/10.1073/pnas.1319590110.

Luikart, Gordon, Ludovic Gielly, Laurent Excoffier, Jean-denis Vigne, Jean Bouvet, and Pierre Taberlet. 2001. “Multiple Maternal Origins and Weak Phylogeographic Structure in Domestic Goats.” *Proceedings of the National Academy of Sciences* 98 (10): 5927–32.

Marsden, Clare D, Diego Ortega-Del Vecchyo, Dennis P O’Brien, Jeremy F Taylor, Oscar Ramirez, Carles Vilà, Tomas Marques-Bonet, Robert D Schnabel, Robert K Wayne, and Kirk E Lohmueller. 2016. “Bottlenecks and Selective Sweeps during Domestication Have Increased Deleterious Genetic Variation in Dogs.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (1): 152–57. https://doi.org/10.1073/pnas.1512501113.

Marshall, F. B., K. Dobney, T. Denham, and J. M. Capriles. 2014. “Evaluating the Roles of Directed Breeding and Gene Flow in Animal Domestication.” *Proceedings of the National Academy of Sciences* 111 (17): 6153–58. https://doi.org/10.1073/pnas.1312984110.

Matthews, Lisa, Gopal Gopinath, Marc Gillespie, Michael Caudy, David Croft, Bernard de Bono, Phani Garapati, et al. 2009. “Reactome Knowledgebase of Human Biological Pathways and Processes.” *Nucleic Acids Research* 37 (SUPPL. 1): 619–22. https://doi.org/10.1093/nar/gkn863.

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

Megens, Hendrik-Jan, Richard P.M.A. Crooijmans, Magali San Cristobal, Xiao Hui, Ning Li, and Martien A.M. Groenen. 2008. “Biodiversity of Pig Breeds from China and Europe Estimated from Pooled DNA Samples: Differences

in Microsatellite Variation between Two Areas of Domestication." *Genetics Selection Evolution* 40: 103–28. https://doi.org/10.1051/gse:2007039.

Minoche, André E., Juliane C. Dohm, and Heinz Himmelbauer. 2011. "Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems." *Genome Biology* 12 (11): R112. https://doi.org/10.1186/gb-2011-12-11-r112.

Montanucci, Ludovica, Hafid Laayouni, Giovanni Marco Dall'Olio, and Jaume Bertranpetit. 2011. "Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway across Primates." *Molecular Biology and Evolution* 28 (1): 813–23. https://doi.org/10.1093/molbev/msq259.

Myles, Sean, Adam R. Boyko, Christopher L. Owens, Patrick J. Brown, Fabrizio Grassi, Mallikarjuna K. Aradhya, Bernard Prins, et al. 2011. "Genetic Structure and Domestication History of the Grape." *Proceedings of the National Academy of Sciences* 108 (9): 3530–35. https://doi.org/10.1073/pnas.1009363108.

Nielsen, Rasmus. 2005. "Molecular Signatures of Natural Selection." *Annual Review of Genetics* 39 (1): 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420.

Nielsen, Rasmus, Joshua S Paul, Anders Albrechtsen, and Yun S Song. 2011. "Genotype and SNP Calling from Next-Generation Sequencing Data." *Nature Review Genetics* 12 (6): 443–51. https://doi.org/10.1038/nrg2986.

Nuijten, Edwin, Robbert van Treuren, Paul C. Struik, Alfred Mokuwa, Florent Okry, Béla Teeken, and Paul Richards. 2009. "Evidence for the Emergence of New Rice Types of Interspecific Hybrid Origin in West African Farmers' Fields." *PLoS ONE* 4 (10). https://doi.org/10.1371/journal.pone.0007335.

Ohta, Tomoko. 1973. "Slightly Deleterious Mutant Substitutions in Evolution." *Nature* 246 (5428): 96–98. http://www.ncbi.nlm.nih.gov/pubmed/4585855.

Ohta, Tomoko. 1992. "The Nearly Neutral Theory of Molecular Evolution." *Annual Review of Ecology and Systematics* 23: 263–86. https://doi.org/10.1146/annurev.es.23.110192.001403.

Ohta, Tomoko, and John H. Gillespie. 1996. "Development of Neutral and

Nearly Neutral Theories." *Theoretical Population Biology* 49 (7): 128–42. http://ac.els-cdn.com/S0040580996900076/1-s2.0-S0040580996900076-main.pdf?_tid=db129526-08e7-11e7-a54c-00000aab0f6b&acdnat=1489517936_86d51a61f7f0b8d31ee3d70efd4f837c

Olsen, Kenneth M, and Jonathan F Wendel. 2013. "A Bountiful Harvest: Genomic Insights into Crop Domestication Phenotypes." *Annual Review of Plant Biology* 64 (1): 47–70. https://doi.org/10.1146/annurev-arplant-050312-120048.

Ottoni, Claudio, Linus Girdland Flink, Allowen Evin, Christina Geörg, Bea De Cupere, Wim Van Neer, László Bartosiewicz, et al. 2013. "Pig Domestication and Human-Mediated Dispersal in Western Eurasia Revealed through Ancient DNA and Geometric Morphometrics." *Molecular Biology and Evolution* 30 (4): 824–32. https://doi.org/10.1093/molbev/mss261.

Pedrosa, Susana, Metehan Uzun, Juan-José Arranz, Beatriz Gutiérrez-Gil, Fermín San Primitivo, and Yolanda Bayón. 2005. "Evidence of Three Maternal Lineages in near Eastern Sheep Supporting Multiple Domestication Events." *Proceedings of the Royal Society B* 272: 2211–17. https://doi.org/10.1098/rspb.2005.3204.

Pérez-Enciso, M., G. de los Campos, N. Hudson, J. Kijas, and A. Reverter. 2016. "The 'Heritability' of Domestication and Its Functional Partitioning in the Pig." *Heredity* 118: 160–68. https://doi.org/10.1038/hdy.2016.78.

Pierpaoli, M., Z. S. Birò, M. Herrmann, K. Hupe, M. Fernandes, B. Ragni, L. Szemethy, and Ettore Randi. 2003. "Genetic Distinction of Wildcat (Felis Silvestris) Populations in Europe, and Hybridization with Domestic Cats in Hungary." *Molecular Ecology* 12 (10): 2585–98. https://doi.org/10.1046/j.1365-294X.2003.01939.x.

Ponsuksili, S., E. Murani, B. Brand, M. Schwerin, and K. Wimmers. 2011. "Integrating Expression Profiling and Whole-Genome Association for Dissection of Fat Traits in a Porcine Model." *Journal of Lipid Research* 52: 668–78. https://doi.org/10.1194/jlr.m013342.

Ramos-Onsins, Sebastian E., William Burgos-Paz, Arianna Manunza, and

Marcel Amills. 2014. “Mining the Pig Genome to Investigate the Domestication Process.” *Heredity* 113 (6): 471–84. https://doi.org/10.1038/hdy.2014.68.

Ramos, Antonio M., Richard P. M. A. Crooijmans, Nabeel A. Affara, Andreia J. Amaral, Alan L. Archibald, Jonathan E. Beever, Christian Bendixen, et al. 2009. “Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology.” *PLoS ONE* 4 (8): e6524. https://doi.org/10.1371/journal.pone.0006524.

Ramsay, Heather, Loren H Rieseberg, and Kermit Ritland. 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, Mark D, Richard E Miller, and Peter Tiffin. 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, Sebastien, and Loren H. Rieseberg. 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, Rebecca M., Wei Jin, and Greg Gibson. 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.

Rossel, S., F. Marshall, D. O’Connor, M. D. Adams, J. Peters, and T. Pilgram. 2008. “Domestication of the Donkey: Timing, Processes, and Indicators.”

Proceedings of the National Academy of Sciences 105 (10): 3715–20. https://doi.org/10.1073/pnas.0709692105.

Rubin, Carl-Johan, Hendrik-Jan Megens, Alvaro Martinez Barrio, Khurram Maqbool, Shumaila Sayyab, Doreen Schwochow, Chao Wang, 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.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. "Positive Natural Selection in the Human Lineage." *Science* 312 (5780): 1614–20. https://doi.org/10.1126/science.1124309.

Scandura, M., L. Iacolina, B. Crestanello, E. Pecchioli, M. F. Di Benedetto, V. Russo, R. Davoli, M. Apollonio, and G. Bertorelle. 2008. "Ancient vs. Recent Processes as Factors Shaping the Genetic Variation of the European Wild Boar: Are the Effects of the Last Glaciation Still Detectable?" *Molecular Ecology* 17 (7): 1745–62. https://doi.org/10.1111/j.1365-294X.2008.03703.x.

Schrider, Daniel R., and Andrew D. Kern. 2017. "Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome." *Molecular Biology and Evolution* 34 (8): 1863–77. https://doi.org/10.1093/molbev/msx154.

Smith, Bruce D. 2011. "General Patterns of Niche Construction and the Management of 'wild' Plant and Animal Resources by Small-Scale Pre-Industrial Societies." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1566): 836–48. https://doi.org/10.1098/rstb.2010.0253.

Smith, Nick G. C., and Adam Eyre-Walker. 2002. "Adaptive Protein Evolution in Drosophila." *Nature* 415 (6875): 1022–24. https://doi.org/10.1038/4151022a.

Stoneking, Mark, and Johannes Krause. 2011. "Learning about Human Population History from Ancient and Modern Genomes." *Nature Reviews Genetics* 12 (9): 603–14. https://doi.org/10.1038/nrg3029.

Tajima, Fumio. 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.

Tajima, Fumio. 1989. “Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123: 585–95. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1203831/pdf/ge1233585.pdf

Uricchio, Lawrence H., Dmitri A. Petrov, and David Enard. 2019. “Exploiting Selection at Linked Sites to Infer the Rate and Strength of Adaptation.” *Nature Ecology & Evolution*, May, 1. https://doi.org/10.1038/s41559-019-0890-6.

Usha, A. P., S. P. Simpson, and J. L. Williams. 1995. “Probability of Random Sire Exclusion Using Microsatellite Markers for Parentage Verification.” *Animal Genetics* 26 (3): 155–61. https://doi.org/10.1111/j.1365-2052.1995.tb03155.x.

VanRaden, Paul M., Jeffrey R. O’Connell, George R. Wiggans, and Kent A. Weigel. 2011. “Genomic Evaluations with Many More Genotypes.” *Genetics Selection Evolution* 43 (1): 10. https://doi.org/10.1186/1297-9686-43-10.

Verkaar, Edward L.C., Isaäc J. Nijman, Maurice Beeke, Eline Hanekamp, and Johannes A. Lenstra. 2004. “Maternal and Paternal Lineages in Cross-Breeding Bovine Species. Has Wisent a Hybrid Origin?” *Molecular Biology and Evolution* 21 (7): 1165–70. https://doi.org/10.1093/molbev/msh064.

Vignal, Alain, Denis Milan, Magali SanCristobal, and André Eggen. 2002. “A Review on SNP and Other Types of Molecular Markers and Their Use in Animal Genetics.” *Genet. Sel. Evol.* 34: 275–305. https://doi.org/10.1051/gse.

Vigne, Jean Denis. 2011. “The Origins of Animal Domestication and Husbandry: A Major Change in the History of Humanity and the Biosphere.” *Comptes Rendus - Biologies* 334 (3): 171–81. https://doi.org/10.1016/j.crvi.2010.12.009.

Vilà, Carles, Jennifer A. Leonard, Anders Götherström, Stefan Marklund, Kaj

Sandberg, Kerstin Lidén, Robert K. Wayne, and Hans Ellegren. 2001. "Widespread Origins of Domestic Horse Lineages." *Science* 291 (5503): 474–77. https://doi.org/10.1126/science.291.5503.474.

Wang, J. Y., Y. R. Luo, W. X. Fu, X. Lu, J. P. Zhou, X. D. Ding, J. F. Liu, and Q. Zhang. 2012. "Genome-Wide Association Studies for Hematological Traits in Swine." *Animal Genetics* 44: 34–43. https://doi.org/10.1111/j.1365-2052.2012.02366.x.

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

Wealleans, Alexandra L. 2013. "Such as Pigs Eat: The Rise and Fall of the Pannage Pig in the UK." *Journal of the Science of Food and Agriculture* 93 (9): 2076–83. https://doi.org/10.1002/jsfa.6145.

Weir, B. S., and C. Clark Cockerham. 1984. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38 (6): 1358. https://doi.org/10.2307/2408641.

White, Sam. 2011. "From Globalized Pig Breeds to Capitalist Pigs: A Study in Animal Cultures and Evolutionary History." *Environmental History* 16 (1): 94–120. https://doi.org/10.1093/envhis/emq143.

Wilkinson, Samantha, Zen H. Lu, Hendrik Jan Megens, Alan L. Archibald, Chris Haley, Ian J. Jackson, Martien A.M. Groenen, Richard P.M.A. Crooijmans, Rob Ogden, and Pamela Wiener. 2013. "Signatures of Diversifying Selection in European Pig Breeds." *PLoS Genetics* 9 (4). https://doi.org/10.1371/journal.pgen.1003453.

Yang, Bin, Leilei Cui, Miguel Perez-Enciso, Aleksei Traspov, Richard P. M. A. Crooijmans, Natalia Zinovieva, Lawrence B. Schook, et al. 2017. "Genome-Wide SNP Data Unveils the Globalization of Domesticated Pigs." *Genet Sel Evol* 49 (71). https://doi.org/10.1186/s12711-017-0345-y.

Zeder, Melinda A. 2006. *Documenting Domestication: New Genetic and Archaeological Paradigms*. Edited by Bruce D. Smith Melinda A. Zeder, Daniel Bradley, Eve Emshwiller. University of California Press.

https://www.ucpress.edu/book/9780520246386/documenting-domestication.

Zeder, Melinda A. 2012. “The Domestication of Animals.” *Journal of Anthropological Research* 68 (2): 161–90. https://doi.org/10.3998/jar.0521004.0068.201.

Zeder, Melinda A. 2015. “Core Questions in Domestication Research.” *Proceedings of the National Academy of Sciences* 112 (11): 3191–98. https://doi.org/10.1073/pnas.1501711112.

Zhang, D, and GM Hewitt. 2003. “Nuclear DNA Analyses in Genetic Studies of Populations: Practice, Problems and Prospects.” *Molecular Ecology* 12: 563–84. http://www.ncbi.nlm.nih.gov/pubmed/12675814.

Zuckerkandl, Emile, and Linus Pauling. 1965. “Evolutionary Divergence and Convergence in Proteins.” *Evolving Genes and Proteins*, January, 97–166. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6.

Annexes

Supplementary material Chapter 3: "A pathway-centered analysis of pig domestication and breeding in Eurasia"

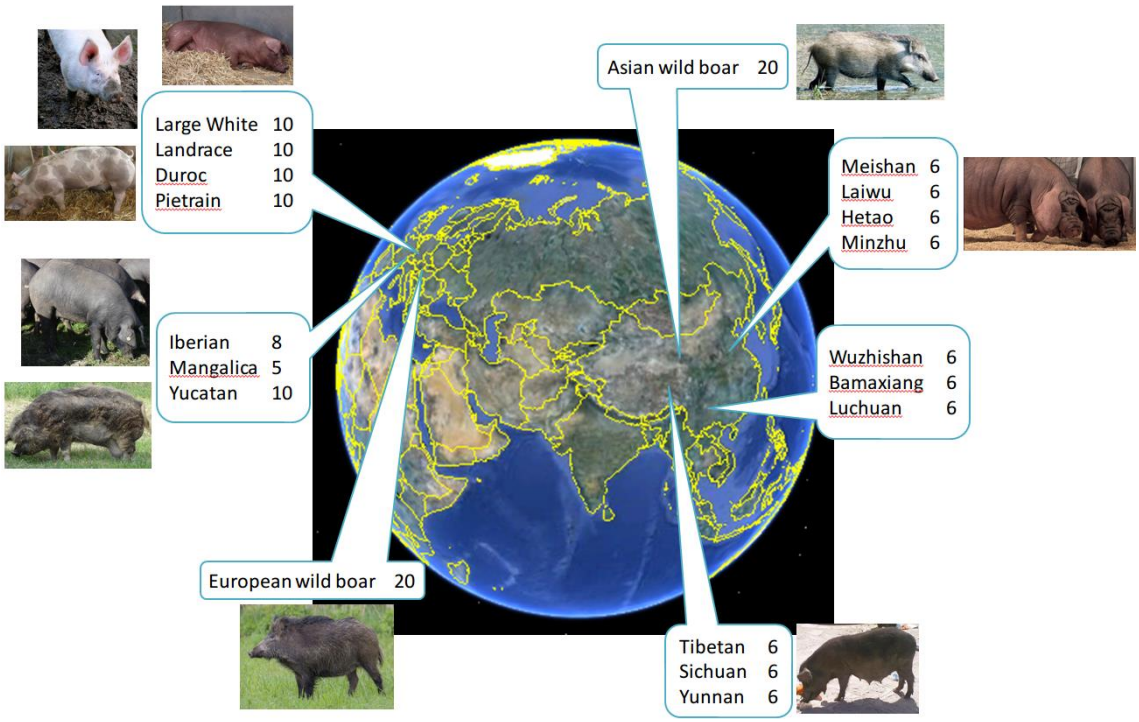

Figure S3.1: *Geographic map of pig samples.*

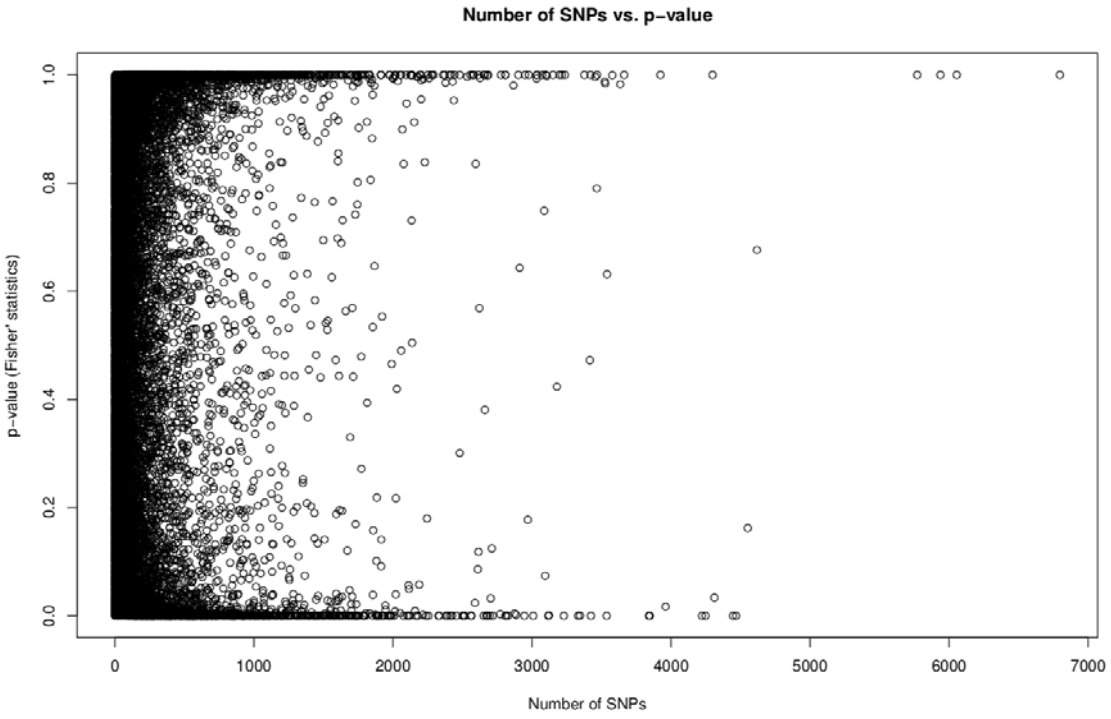

Figure S3.2: *Number of SNPs per gene vs. gene P-value, each dot corresponds to a different gene.*

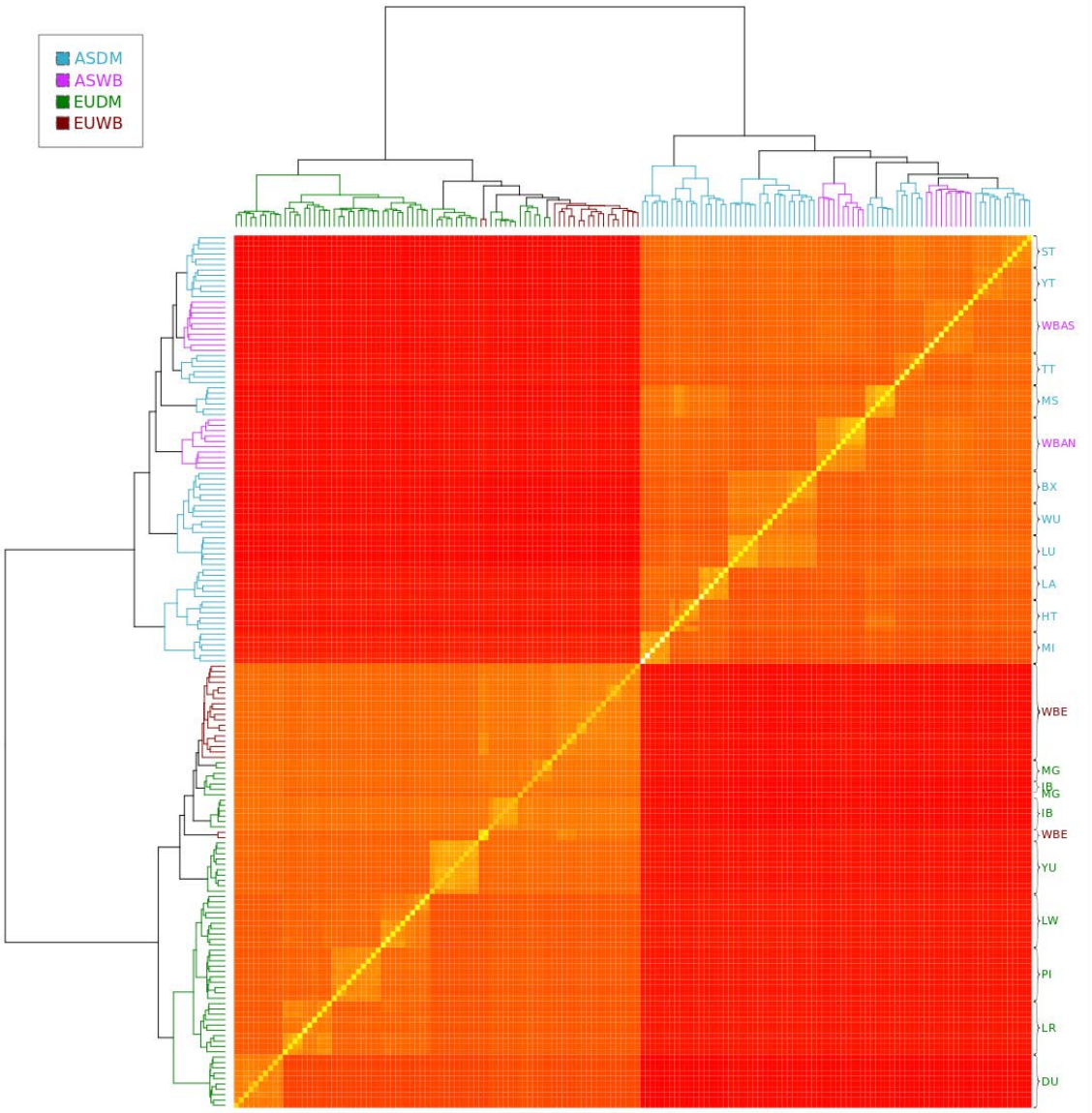

Figure S3.3: *Heatmap of all samples produced using the molecular relationship matrix, computed using all available autosomal SNPs. Colors are used to differentiate among the populations: ASDM (blue), ASWB (purple), EUDM (green) and EUWB (dark red).*

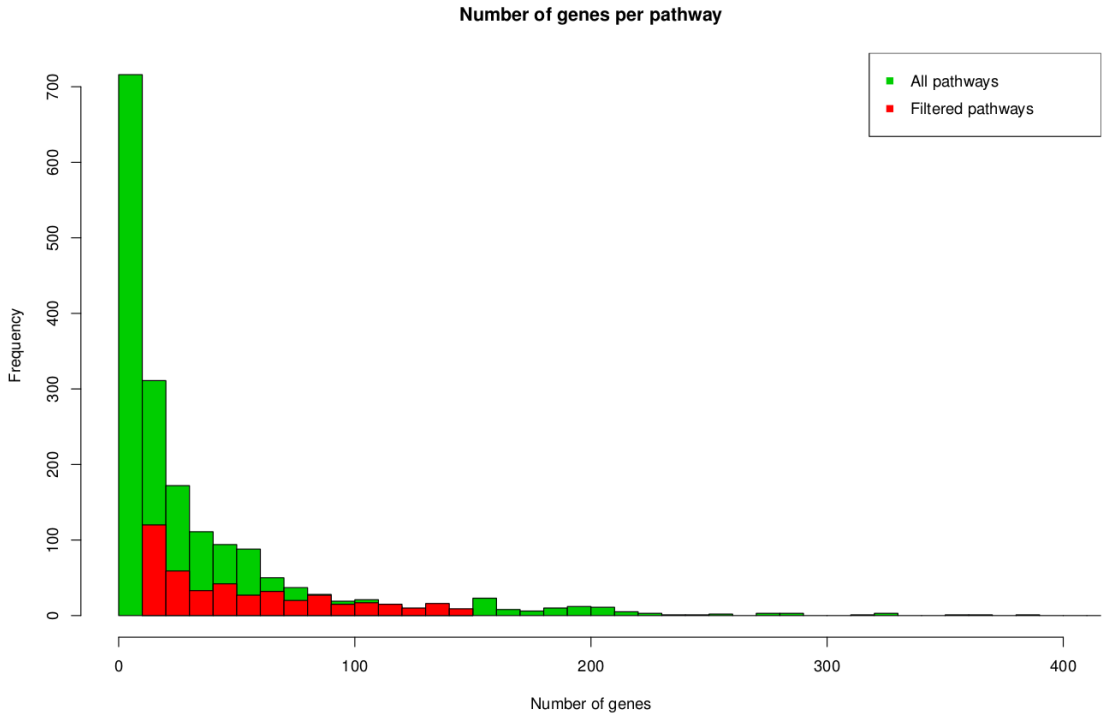

Figure S3.4: *Frequency of the number of genes per pathway using all the 1,789 pathways of Sus scrofa retrieved from NCBI Biosystems database or the 442 filtered pathways.*

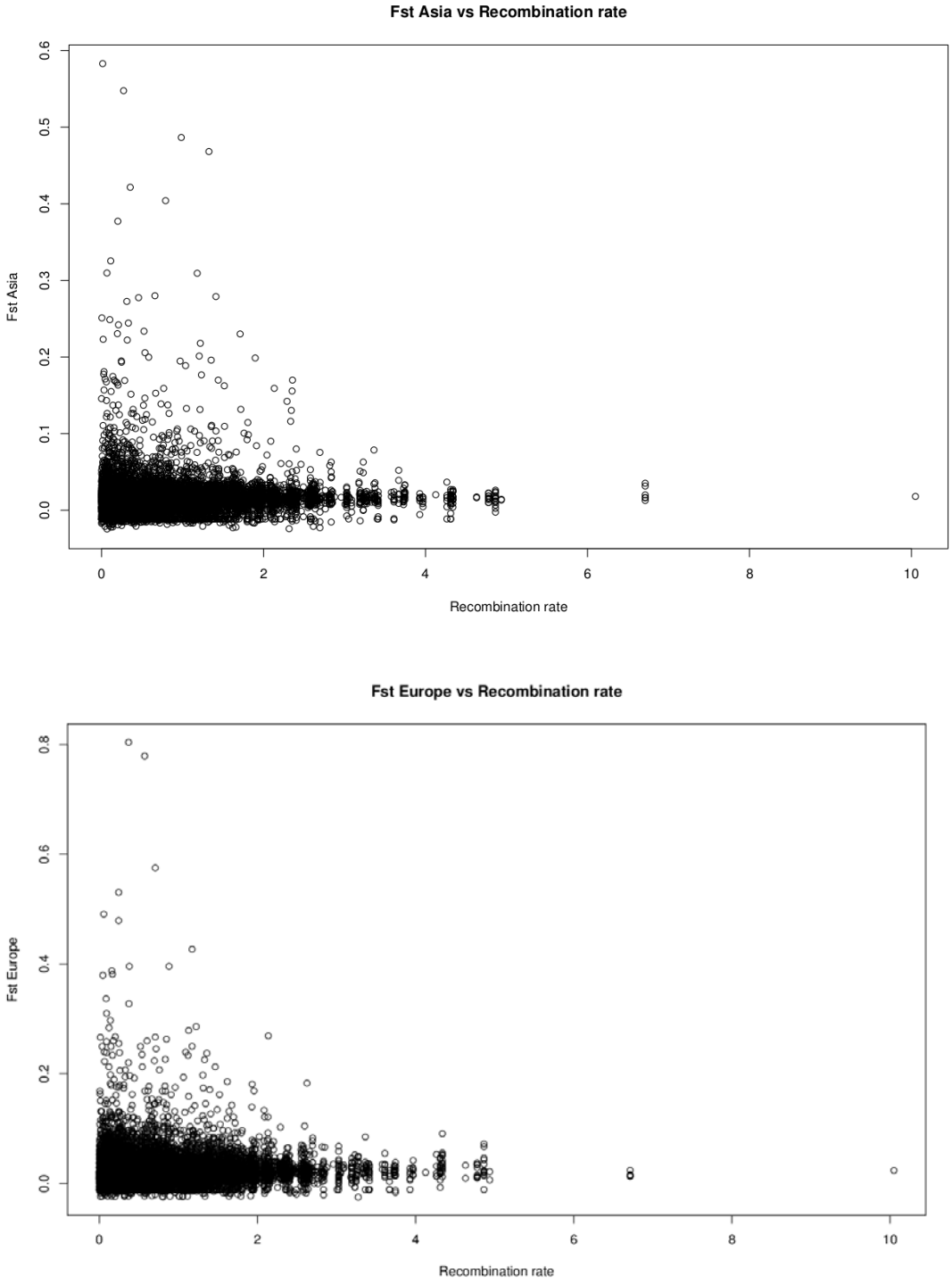

Figure S3.5: *Recombination rate (cM/Mb) vs. Fst per gene in Asia (top) and Europe (bottom).*

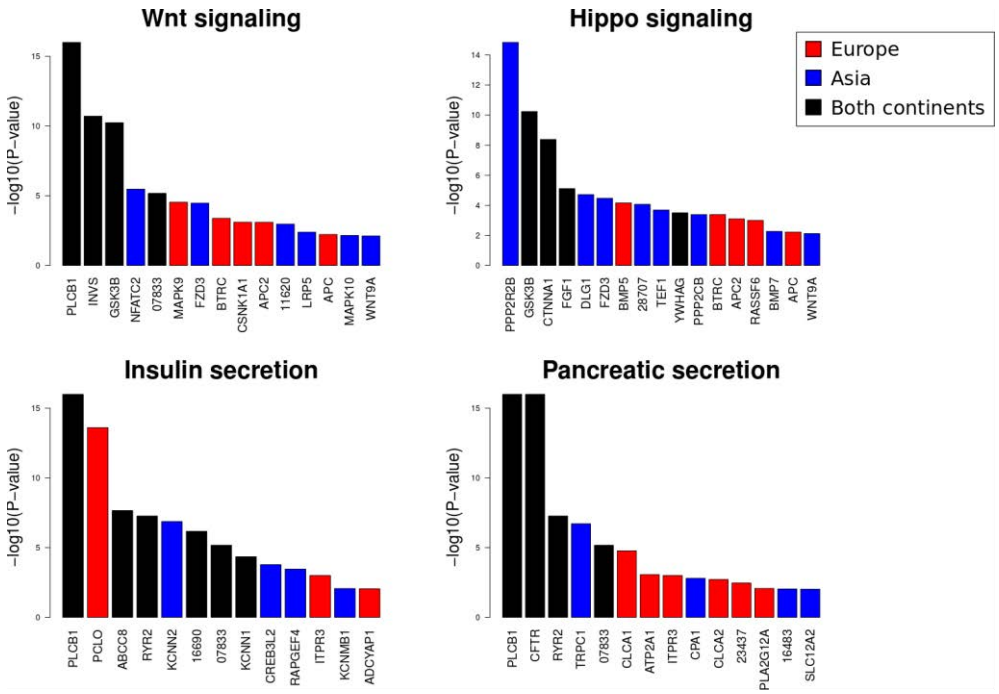

Figure S3.6: *Gene P-value (-log10) of significant genes at the 1% nominal level in Europe (red bars), in Asia (blue bars) or both continents (black bars) from the significant pathways involved in the development of the animal or in the insulin-related pathways. When a gene was significant in both continents, the smallest P-value is plotted. Gene symbols are provided when available; otherwise numbers indicate ensembl ENSSSCG id.*

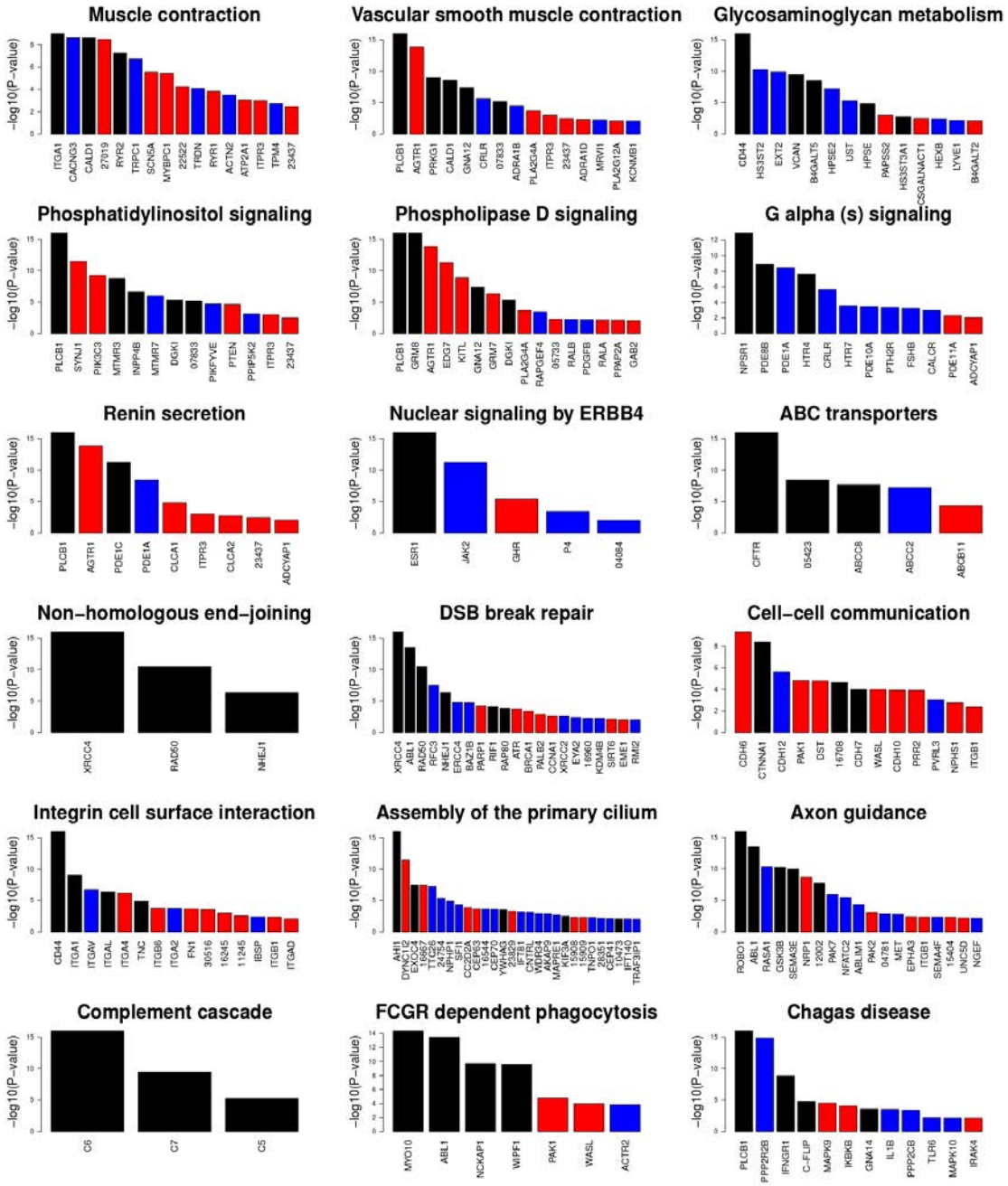

Figure S3.7: *Gene P-value (-log10) of significant genes at the 1% nominal level in Europe (red bars), in Asia (blue bars) or both continents (black bars) from the significant pathways in terms of Fst not shown in Fig S5 and Fig 3. When a gene was significant in both continents, the smallest P-value is plotted. Gene symbols are provided when available; otherwise numbers indicate ensembl ENSSSCG id.*

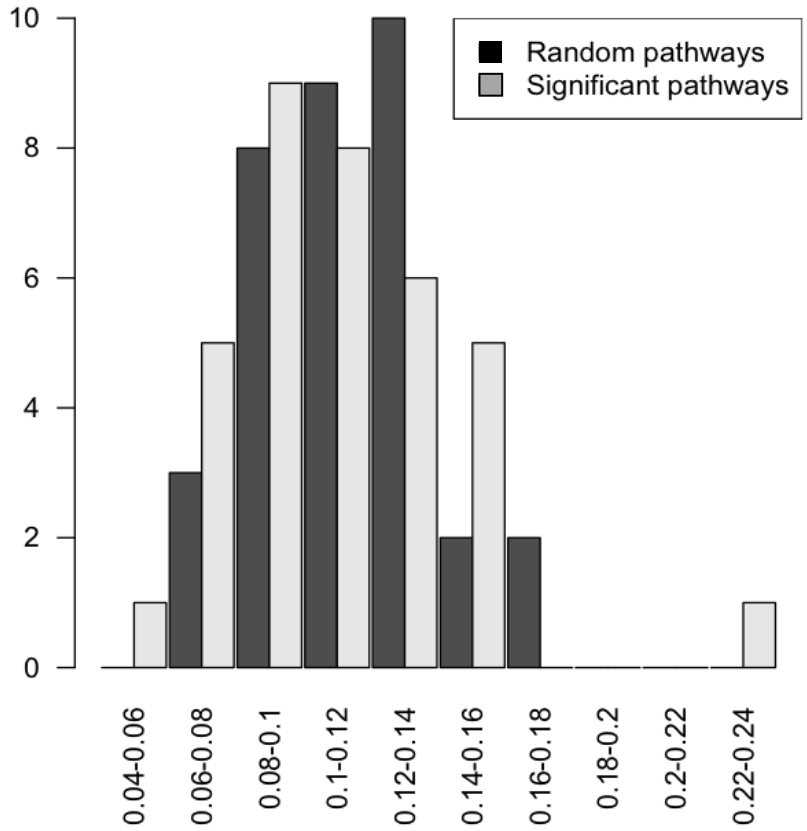

Figure S3.8: *Distribution of inferred Asian contribution across the significant pathways from Tables 1 and 2 (in grey) and a random sample of non-significant pathways (black).*

Table S3.1: *Sample list.*

Continent	Status	Breed	Origin	Sample	Accession	Depth*
Europe	Domestic	Duroc	International	DUNA3503	SRX703503	14.6
Europe	Domestic	Duroc	International	DUNA3504	SRX703504	14.0
Europe	Domestic	Duroc	International	DUNA3507	SRX703507	13.2
Europe	Domestic	Duroc	International	DUNA3508	SRX703508	13.1
Europe	Domestic	Duroc	International	DUNA3509	SRX703509	13.9
Europe	Domestic	Duroc	International	DUNA3512	SRX703512	14.2
Europe	Domestic	Duroc	International	DUNA3534	SRX703534	13.4
Europe	Domestic	Duroc	International	DUNA3536	SRX703536	12.9
Europe	Domestic	Duroc	International	DUNA3538	SRX703538	13.3
Europe	Domestic	Duroc	International	DUNA3539	SRX703539	12.7
Europe	Domestic	Iberian	Spain	IBGM0327	SRR1513307	13.0
Europe	Domestic	Iberian	Spain	IBGU1330	SRR765849	7.2
Europe	Domestic	Iberian	Spain	IBGU1802	SRX245748	12.4
Europe	Domestic	Iberian	Spain	IBGU1803	SAMN05362554	13.0
Europe	Domestic	Iberian	Spain	IBGU1804	SRR1917381	14.5
Europe	Domestic	Iberian	Spain	IBGU1805	SRX278843 **	12.4
Europe	Domestic	Iberian	Spain	IBNI01U07	ERP011076	13.4
Europe	Domestic	Iberian	Spain	IBRE01F51	ERP011076	14.3

Europe	Domestic	Mangalica	Hungary	MGHU6909	SRX476909	14.1
Europe	Domestic	Mangalica	Hungary	MGHU6910	SRX476910	8.7
Europe	Domestic	Mangalica	Hungary	MGHU6911	SRX476911	14.5
Europe	Domestic	Mangalica	Hungary	MGHU01F18	ERP011076	9.7
Europe	Domestic	Mangalica	Hungary	MGHU01F20	ERP011076	11.0
Europe	Domestic	Yucatan	Mexico	YUUS1489	SRR1873293	14.1
Europe	Domestic	Yucatan	Mexico	YUUS3553	SRX703553	12.9
Europe	Domestic	Yucatan	Mexico	YUUS3554	SRX703554	13.5
Europe	Domestic	Yucatan	Mexico	YUUS3555	SRX703555	11.7
Europe	Domestic	Yucatan	Mexico	YUUS3557	SRX703557	12.9
Europe	Domestic	Yucatan	Mexico	YUUS3558	SRX703558	12.1
Europe	Domestic	Yucatan	Mexico	YUUS3559	SRX703559	11.7
Europe	Domestic	Yucatan	Mexico	YUUS3560	SRX703560	12.0
Europe	Domestic	Yucatan	Mexico	YUUS3561	SRX703561	13.1
Europe	Domestic	Yucatan	Mexico	YUUS3565	SRX703565	12.9
Europe	Domestic	Landrace	International	LR24F01	ERX149144	13.7
Europe	Domestic	Landrace	International	LRNA3540	SRX703540	12.8
Europe	Domestic	Landrace	International	LRNA3541	SRX703541	12.9
Europe	Domestic	Landrace	International	LRNA3542	SRX703542	12.2
Europe	Domestic	Landrace	International	LRNA3543	SRX703543	11.8
Europe	Domestic	Landrace	International	LRNA3547	SRX703547	9.6

Europe	Domestic	Landrace	International	LRNA3549	SRX703549	10.6
Europe	Domestic	Landrace	International	LRNA3550	SRX703550	10.9
Europe	Domestic	Landrace	International	LRNA3551	SRX703551	11.8
Europe	Domestic	Landrace	International	LRNA3552	SRX703552	12.2
Europe	Domestic	Large White	International	LW22F04	ERX149151	10.1
Europe	Domestic	Large White	International	LW22F07	ERX149153	11.6
Europe	Domestic	Large White	International	LW22M07	ERX149155	10.4
Europe	Domestic	Large White	International	LWGB0348	SRX2787051 **	12.0
Europe	Domestic	Large White	International	LWNA3584	SRX703584	11.8
Europe	Domestic	Large White	International	LWNA3594	SRX703594	11.8
Europe	Domestic	Large White	International	LWNA3595	SRX703595	12.8
Europe	Domestic	Large White	International	LWNA3596	SRX703596	12.0
Europe	Domestic	Large White	International	LWNA3597	SRX703597	12.5
Europe	Domestic	Large White	International	LWNA3599	SRX703599	12.3
Europe	Domestic	Pietrain	International	PI21F02	ERX149167	10.6
Europe	Domestic	Pietrain	International	PI21F06	ERX149168	10.5
Europe	Domestic	Pietrain	International	PI21M21	ERX149171	11.2
Europe	Domestic	Pietrain	International	PINA4919	ERX954919	11.4
Europe	Domestic	Pietrain	International	PINA4921	ERX954921	11.5
Europe	Domestic	Pietrain	International	PINA4923	ERX954923	12.2
Europe	Domestic	Pietrain	International	PINA4925	ERX954925	9.7

Europe	Domestic	Pietrain	International	PINA4928	ERX954928	10.3
Europe	Domestic	Pietrain	International	PINA4929	ERX954929	10.4
Europe	Domestic	Pietrain	International	PINA4930	ERX954930	10.1
Europe	Wild	European Wild Boar	Switzerland	WBCH26M09	ERX149181	14.4
Europe	Wild	European Wild Boar	Spain	WBES0494	SAMN05362552	12.6
Europe	Wild	European Wild Boar	Spain	WBES0717	SRR1513306	13.0
Europe	Wild	European Wild Boar	France	WBFR25U11	ERX149180	9.4
Europe	Wild	European Wild Boar	Netherlands	WBNL21M03	ERX149177	11.4
Europe	Wild	European Wild Boar	Netherlands	WBNL21F04	ERP011076	15.3
Europe	Wild	European Wild Boar	Netherlands	WBNL22M02	ERP011076	16.6
Europe	Wild	European Wild Boar	Netherlands	WBNL22M03	ERP011076	14.4
Europe	Wild	European Wild Boar	Tunisia	WBTN0965	SAMN05362553	12.4
Europe	Wild	European Wild Boar	Greece	WBGR31F05	ERP011076	14.2
Europe	Wild	European Wild Boar	Greece	WBGR31M09	ERP011076	11.2
Europe	Wild	European Wild Boar	Greece	WBGR32F07	ERP011076	10.7
Europe	Wild	European Wild Boar	Greece	WBGR32U05	ERP011076	10.2
Europe	Wild	European Wild Boar	Italy	WBIT44U06	ERP011076	13.2
Europe	Wild	European Wild Boar	Italy	WBIT44U07	ERP011076	11.8
Europe	Wild	European Wild Boar	Italy	WBIT28F31	ERP011076	17.3
Europe	Wild	European Wild Boar	Italy	WBIT28M39	ERP011076	12.6
Europe	Wild	European Wild Boar	Italy	WBIT42M09	ERP011076	13.6

Europe	Wild	European Wild Boar	Near East	WBNE33U04	ERP011076	12.4
Europe	Wild	European Wild Boar	Near East	WBNE33U05	ERP011076	11.4
Asia	Domestic	Bamaxiang	South China	BXCN5762	SRS465762	13.1
Asia	Domestic	Bamaxiang	South China	BXCN5763	SRS465763	14.2
Asia	Domestic	Bamaxiang	South China	BXCN5764	SRS465764	12.6
Asia	Domestic	Bamaxiang	South China	BXCN5765	SRS465765	13.6
Asia	Domestic	Bamaxiang	South China	BXCN5766	SRS465766	13.1
Asia	Domestic	Bamaxiang	South China	BXCN5767	SRS465767	13.5
Asia	Domestic	Hetao	North China	HTCN5750	SRS465750	13.1
Asia	Domestic	Hetao	North China	HTCN5751	SRS465751	12.0
Asia	Domestic	Hetao	North China	HTCN5752	SRS465752	9.7
Asia	Domestic	Hetao	North China	HTCN5753	SRS465753	11.9
Asia	Domestic	Hetao	North China	HTCN5754	SRS465754	12.6
Asia	Domestic	Hetao	North China	HTCN5755	SRS465755	12.2
Asia	Domestic	Laiwu	North China	LACN5768	SRS465768	12.9
Asia	Domestic	Laiwu	North China	LACN5769	SRS465769	12.9
Asia	Domestic	Laiwu	North China	LACN5770	SRS465770	13.7
Asia	Domestic	Laiwu	North China	LACN5771	SRS465771	13.1
Asia	Domestic	Laiwu	North China	LACN5772	SRS465772	12.8
Asia	Domestic	Laiwu	North China	LACN5773	SRS465773	12.8
Asia	Domestic	Luchuan	South China	LUCN5722	SRS465722	13.9

Asia	Domestic	Luchuan	South China	LUCN5723	SRS465723	13.2
Asia	Domestic	Luchuan	South China	LUCN5724	SRS465724	13.1
Asia	Domestic	Luchuan	South China	LUCN5725	SRS465725	12.8
Asia	Domestic	Luchuan	South China	LUCN5726	SRS465726	13.9
Asia	Domestic	Luchuan	South China	LUCN5727	SRS465727	13.2
Asia	Domestic	Minzhu	North China	MICN5756	SRS465756	12.3
Asia	Domestic	Minzhu	North China	MICN5757	SRS465757	12.8
Asia	Domestic	Minzhu	North China	MICN5758	SRS465758	13.1
Asia	Domestic	Minzhu	North China	MICN5759	SRS465759	11.4
Asia	Domestic	Minzhu	North China	MICN5760	SRS465760	13.4
Asia	Domestic	Minzhu	North China	MICN5761	SRS465761	13.4
Asia	Domestic	Meishan	South China	MS20U10	ERX149162	9.1
Asia	Domestic	Meishan	South China	MS20U11	ERX149163	9.1
Asia	Domestic	Meishan	South China	MS21M07	ERX149164	8.8
Asia	Domestic	Meishan	South China	MS21M14	ERX149165	10.1
Asia	Domestic	Meishan	South China	MS21M05	ERP011076	11.5
Asia	Domestic	Meishan	South China	MS21M08	ERP011076	9.9
Asia	Domestic	Sichuan	Tibet	STCN5740	SRS465740	13.2
Asia	Domestic	Sichuan	Tibet	STCN5741	SRS465741	13.2
Asia	Domestic	Sichuan	Tibet	STCN5742	SRS465742	12.8
Asia	Domestic	Sichuan	Tibet	STCN5743	SRS465743	13.0

Asia	Domestic	Sichuan	Tibet	STCN5744	SRS465744	13.7
Asia	Domestic	Sichuan	Tibet	STCN5745	SRS465745	13.2
Asia	Domestic	Tibetan	Tibet	TTCN5728	SRS465728	13.3
Asia	Domestic	Tibetan	Tibet	TTCN5729	SRS465729	12.4
Asia	Domestic	Tibetan	Tibet	TTCN5730	SRS465730	9.5
Asia	Domestic	Tibetan	Tibet	TTCN5731	SRS465731	10.6
Asia	Domestic	Tibetan	Tibet	TTCN5732	SRS465732	11.3
Asia	Domestic	Tibetan	Tibet	TTCN5733	SRS465733	13.6
Asia	Domestic	Wuzhishan	South China	WUCN5708	SRS465708	12.7
Asia	Domestic	Wuzhishan	South China	WUCN5709	SRS465709	12.7
Asia	Domestic	Wuzhishan	South China	WUCN5710	SRS465710	12.7
Asia	Domestic	Wuzhishan	South China	WUCN5711	SRS465711	12.6
Asia	Domestic	Wuzhishan	South China	WUCN5712	SRS465712	12.9
Asia	Domestic	Wuzhishan	South China	WUCN5713	SRS465713	13.2
Asia	Domestic	Yunnan	Tibet	YTCN5734	SRS465734	13.9
Asia	Domestic	Yunnan	Tibet	YTCN5735	SRS465735	10.5
Asia	Domestic	Yunnan	Tibet	YTCN5736	SRS465736	13.6
Asia	Domestic	Yunnan	Tibet	YTCN5737	SRS465737	14.1
Asia	Domestic	Yunnan	Tibet	YTCN5738	SRS465738	13.3
Asia	Domestic	Yunnan	Tibet	YTCN5739	SRS465739	12.7
Asia	Wild	South Asian Wild Boar	South China	WBCN5716	SRS465716	13.3

Asia	Wild	South Asian Wild Boar	South China	WBCN5717	SRS465717	13.7
Asia	Wild	South Asian Wild Boar	South China	WBCN5718	SRS465718	8.8
Asia	Wild	South Asian Wild Boar	South China	WBCN5719	SRS465719	9.0
Asia	Wild	South Asian Wild Boar	South China	WBCN5720	SRS465720	9.4
Asia	Wild	South Asian Wild Boar	South China	WBCN5721	SRS465721	9.6
Asia	Wild	South Asian Wild Boar	South China	WBCNS29U04	ERX149182	5.3
Asia	Wild	South Asian Wild Boar	South China	WBCNS29U12	ERX149183	10.1
Asia	Wild	South Asian Wild Boar	South China	WBCNS29U14	ERP011076	9.6
Asia	Wild	South Asian Wild Boar	South China	WBCNS29U16	ERP011076	13.6
Asia	Wild	North Asian Wild Boar	North China	WBCNN30U01	ERX149184	5.2
Asia	Wild	North Asian Wild Boar	North China	WBCNN30U08	ERX149185	9.7
Asia	Wild	North Asian Wild Boar	North China	WBCNN30U09	ERP011076	14.3
Asia	Wild	North Asian Wild Boar	Korea	WBKR3566	SRX703566	11.6
Asia	Wild	North Asian Wild Boar	Korea	WBKR3568	SRX703568	11.8
Asia	Wild	North Asian Wild Boar	Korea	WBKR3569	SRX703569	12.2
Asia	Wild	North Asian Wild Boar	Korea	WBKR3570	SRX703570	11.6
Asia	Wild	North Asian Wild Boar	Korea	WBKR3571	SRX703571	11.9
Asia	Wild	North Asian Wild Boar	Korea	WBKR3572	SRX703572	13.5
Asia	Wild	North Asian Wild Boar	Russia	WBRU1064	SAMN05362551	6.9

* Depth is genome average number of reads per position when filtered by quality.
** New sample sequence data provided in this study (PRJNA255085)

Table S3.2: *Filtered pathway list.*

NCBI ID	Source	Number of genes	Pathway name
84361	KEGG	55	Glycolysis / Gluconeogenesis
84362	KEGG	30	Citrate cycle (TCA cycle)
84363	KEGG	24	Pentose phosphate pathway
84364	KEGG	26	Pentose and glucuronate interconversions
84365	KEGG	30	Fructose and mannose metabolism
84366	KEGG	27	Galactose metabolism
84368	KEGG	13	Fatty acid biosynthesis
84369	KEGG	22	Fatty acid elongation
84372	KEGG	21	Steroid biosynthesis
84374	KEGG	10	Ubiquinone and other terpenoid-quinone biosynthesis
84375	KEGG	52	Steroid hormone biosynthesis
84377	KEGG	135	Oxidative phosphorylation
84378	KEGG	19	Arginine biosynthesis
84381	KEGG	101	Pyrimidine metabolism
84384	KEGG	36	Glycine, serine and threonine metabolism
84387	KEGG	49	Valine, leucine and isoleucine degradation
84388	KEGG	52	Lysine degradation
84389	KEGG	44	Arginine and proline metabolism
84390	KEGG	22	Histidine metabolism
84391	KEGG	33	Tyrosine metabolism
84392	KEGG	18	Phenylalanine metabolism
84394	KEGG	41	Tryptophan metabolism
84396	KEGG	29	beta-Alanine metabolism
84397	KEGG	11	Taurine and hypotaurine metabolism
84398	KEGG	16	Selenocompound metabolism
84401	KEGG	53	Glutathione metabolism
84402	KEGG	49	Starch and sucrose metabolism
84403	KEGG	49	N-Glycan biosynthesis
84404	KEGG	17	Other glycan degradation
84405	KEGG	28	Mucin type O-Glycan biosynthesis
84407	KEGG	19	Glycosaminoglycan degradation
84408	KEGG	21	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
84410	KEGG	24	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
84412	KEGG	58	Glycerolipid metabolism
84414	KEGG	26	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
84415	KEGG	86	Glycerophospholipid metabolism
84416	KEGG	43	Ether lipid metabolism
84417	KEGG	62	Arachidonic acid metabolism
84418	KEGG	32	Linoleic acid metabolism

84421	KEGG	26	Glycosphingolipid biosynthesis - lacto and neolacto series
84422	KEGG	15	Glycosphingolipid biosynthesis - globo series
84423	KEGG	14	Glycosphingolipid biosynthesis - ganglio series
84424	KEGG	37	Pyruvate metabolism
84426	KEGG	26	Glyoxylate and dicarboxylate metabolism
84428	KEGG	30	Propanoate metabolism
84430	KEGG	24	Butanoate metabolism
84431	KEGG	18	One carbon pool by folate
84436	KEGG	26	Nicotinate and nicotinamide metabolism
84437	KEGG	19	Pantothenate and CoA biosynthesis
84438	KEGG	13	Folate biosynthesis
84439	KEGG	54	Retinol metabolism
84440	KEGG	34	Porphyrin and chlorophyll metabolism
84442	KEGG	15	Nitrogen metabolism
84443	KEGG	17	Sulfur metabolism
84447	KEGG	46	Aminoacyl-tRNA biosynthesis
84450	KEGG	31	Drug metabolism - other enzymes
84451	KEGG	19	Biosynthesis of unsaturated fatty acids
84452	KEGG	43	ABC transporters
84453	KEGG	143	Ribosome
84455	KEGG	42	Basal transcription factors
84457	KEGG	68	PPAR signaling pathway
84458	KEGG	44	Nucleotide excision repair
84459	KEGG	29	Homologous recombination
84461	KEGG	86	ErbB signaling pathway
84464	KEGG	96	Phosphatidylinositol signaling system
84466	KEGG	115	Cell cycle
84467	KEGG	75	p53 signaling pathway
84468	KEGG	134	Ubiquitin mediated proteolysis
84469	KEGG	36	SNARE interactions in vesicular transport
84470	KEGG	38	Regulation of autophagy
84471	KEGG	60	mTOR signaling pathway
84472	KEGG	90	Apoptosis
84473	KEGG	129	Wnt signaling pathway
84474	KEGG	44	Notch signaling pathway
84476	KEGG	78	TGF-beta signaling pathway
84477	KEGG	119	Axon guidance
84478	KEGG	58	VEGF signaling pathway
84480	KEGG	77	ECM-receptor interaction
84481	KEGG	145	Cell adhesion molecules (CAMs)
84482	KEGG	71	Adherens junction
84483	KEGG	134	Tight junction
84484	KEGG	89	Gap junction
84485	KEGG	73	Complement and coagulation cascades
84486	KEGG	67	Antigen processing and presentation
84487	KEGG	24	Renin-angiotensin system

84488	KEGG	101	Toll-like receptor signaling pathway
84490	KEGG	83	Hematopoietic cell lineage
84491	KEGG	108	Natural killer cell mediated cytotoxicity
84492	KEGG	106	T cell receptor signaling pathway
84493	KEGG	73	B cell receptor signaling pathway
84494	KEGG	64	Fc epsilon RI signaling pathway
84495	KEGG	119	Leukocyte transendothelial migration
84496	KEGG	67	Long-term potentiation
84497	KEGG	59	Long-term depression
84499	KEGG	65	Taste transduction
84501	KEGG	131	Insulin signaling pathway
84502	KEGG	88	GnRH signaling pathway
84503	KEGG	94	Melanogenesis
84504	KEGG	71	Adipocytokine signaling pathway
84505	KEGG	49	Type II diabetes mellitus
84507	KEGG	24	Maturity onset diabetes of the young
84509	KEGG	150	Parkinson's disease
84510	KEGG	55	Amyotrophic lateral sclerosis (ALS)
84513	KEGG	68	Colorectal cancer
84514	KEGG	64	Renal cell carcinoma
84518	KEGG	85	Prostate cancer
84519	KEGG	26	Thyroid cancer
84520	KEGG	52	Basal cell carcinoma
84521	KEGG	69	Melanoma
84522	KEGG	38	Bladder cancer
84523	KEGG	69	Chronic myeloid leukemia
84524	KEGG	54	Acute myeloid leukemia
84525	KEGG	88	Small cell lung cancer
84527	KEGG	23	Asthma
84528	KEGG	50	Autoimmune thyroid disease
84529	KEGG	101	Systemic lupus erythematosus
84532	KEGG	37	Primary immunodeficiency
92797	KEGG	32	Base excision repair
92862	KEGG	20	Terpenoid backbone biosynthesis
92864	KEGG	13	Non-homologous end-joining
92865	KEGG	26	Circadian rhythm
93350	KEGG	73	Cardiac muscle contraction
96246	KEGG	117	Vascular smooth muscle contraction
98759	KEGG	119	Lysosome
101116	KEGG	36	Alanine, aspartate and glutamate metabolism
101117	KEGG	50	Amino sugar and nucleotide sugar metabolism
101119	KEGG	118	Neurotrophin signaling pathway
101120	KEGG	33	Prion diseases
104469	KEGG	43	Cysteine and methionine metabolism
114044	KEGG	27	Dorso-ventral axis formation
114045	KEGG	82	Fc gamma R-mediated phagocytosis

117257	KEGG	73	RNA degradation
117258	KEGG	68	RIG-I-like receptor signaling pathway
119280	KEGG	87	Progesterone-mediated oocyte maturation
121477	KEGG	83	Dilated cardiomyopathy
122187	KEGG	50	NOD-like receptor signaling pathway
125120	KEGG	131	Spliceosome
125121	KEGG	61	Cytosolic DNA-sensing pathway
125122	KEGG	62	Viral myocarditis
126899	KEGG	110	Oocyte meiosis
128755	KEGG	43	Intestinal immune network for IgA production
130624	KEGG	41	Aldosterone-regulated sodium reabsorption
131215	KEGG	82	Peroxisome
131382	KEGG	25	Protein export
143693	KEGG	42	Vasopressin-regulated water reabsorption
144173	KEGG	22	Proximal tubule bicarbonate reclamation
144174	KEGG	64	Leishmaniasis
147584	KEGG	26	Collecting duct acid secretion
147807	KEGG	107	Chagas disease (American trypanosomiasis)
149780	KEGG	76	Bacterial invasion of epithelial cells
149781	KEGG	25	Phototransduction
152663	KEGG	53	Malaria
153374	KEGG	74	Salivary secretion
153902	KEGG	148	Phagosome
154407	KEGG	30	Other types of O-glycan biosynthesis
167313	KEGG	99	Amoebiasis
169304	KEGG	87	Pancreatic secretion
169640	KEGG	115	Toxoplasmosis
170717	KEGG	44	Carbohydrate digestion and absorption
172824	KEGG	46	Staphylococcus aureus infection
172825	KEGG	81	Protein digestion and absorption
173971	KEGG	127	Hepatitis C
193142	KEGG	133	Osteoclast differentiation
193143	KEGG	68	Bile secretion
193315	KEGG	84	mRNA surveillance pathway
194378	KEGG	34	Fat digestion and absorption
194379	KEGG	34	African trypanosomiasis
199370	KEGG	82	Ribosome biogenesis in eukaryotes
199554	KEGG	22	Vitamin digestion and absorption
200307	KEGG	87	Rheumatoid arthritis
212235	KEGG	42	Mineral absorption
213302	KEGG	132	Measles
213303	KEGG	41	Endocrine and other factor-regulated calcium reabsorption
213816	KEGG	122	Glutamatergic synapse
217714	KEGG	112	Cholinergic synapse
218109	KEGG	73	Pertussis
373896	KEGG	59	Synaptic vesicle cycle

375169	KEGG	81	Salmonella infection
377245	KEGG	51	Fanconi anemia pathway
446220	KEGG	10	Acylglycerol degradation
446225	KEGG	10	Pentose phosphate pathway (Pentose phosphate cycle)
446253	KEGG	10	Cytochrome bc1 complex
446264	KEGG	10	NADH dehydrogenase (ubiquinone) Fe-S protein/flavoprotein complex, mitochondria
446266	KEGG	10	Glycolysis, core module involving three-carbon compounds
446270	KEGG	10	Exosome, eukaryotes
446275	KEGG	11	COPII complex
469195	KEGG	119	Dopaminergic synapse
469196	KEGG	64	Legionellosis
471064	KEGG	10	Survival motor neuron (SMN) complex
525344	KEGG	111	Serotonergic synapse
546272	KEGG	45	Cocaine addiction
552645	KEGG	10	Holo-TFIIH complex
552646	KEGG	16	BRCA1-associated genome surveillance complex (BASC)
552647	KEGG	10	BER complex
552651	KEGG	10	Spliceosome, U1-snRNP
552652	KEGG	34	Spliceosome, U4/U6.U5 tri-snRNP
552653	KEGG	37	Spliceosome, 35S U5-snRNP
552654	KEGG	12	ECS complex
552660	KEGG	12	HRD1/SEL1 ERAD complex
583021	KEGG	10	ESCRT-III complex
583275	KEGG	33	Nicotine addiction
585587	KEGG	143	Alcoholism
620367	KEGG	23	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP
620368	KEGG	13	Guanine ribonucleotide biosynthesis IMP => GDP,GTP
634542	KEGG	90	NF-kappa B signaling pathway
673234	KEGG	61	Chemical carcinogenesis
695240	KEGG	101	HIF-1 signaling pathway
749791	KEGG	139	Hippo signaling pathway
777548	KEGG	84	Insulin secretion
791445	KEGG	63	Biosynthesis of amino acids
791446	KEGG	51	Ovarian steroidogenesis
799191	KEGG	96	Estrogen signaling pathway
812268	KEGG	109	TNF signaling pathway
814206	KEGG	67	Prolactin signaling pathway
816324	KEGG	105	Carbon metabolism
835450	KEGG	71	Thyroid hormone synthesis
842783	KEGG	61	Inflammatory bowel disease (IBD)
869518	KEGG	46	Fatty acid metabolism
908279	KEGG	141	Adrenergic signaling in cardiomyocytes
921549	KEGG	133	FoxO signaling pathway

946620	KEGG	113	Thyroid hormone signaling pathway
948291	KEGG	109	Inflammatory mediator regulation of TRP channels
953739	KEGG	124	Platelet activation
989938	KEGG	117	AMPK signaling pathway
1026258	KEGG	132	Signaling pathways regulating pluripotency of stem cells
1060724	KEGG	62	Central carbon metabolism in cancer
1060725	KEGG	97	Choline metabolism in cancer
1085110	KEGG	12	Hedgehog signaling
1085111	KEGG	21	BMP signaling
1085113	KEGG	10	Activin signaling
1085116	KEGG	20	JAK-STAT signaling
1085120	KEGG	22	MAPK (JNK) signaling
1085125	KEGG	10	Cell cycle - G2/M transition
1085126	KEGG	12	cGMP signaling
1146435	KEGG	95	Glucagon signaling pathway
1146436	KEGG	113	Sphingolipid signaling pathway
1223593	KEGG	57	Regulation of lipolysis in adipocytes
1223594	KEGG	67	Renin secretion
1273514	KEGG	76	Aldosterone synthesis and secretion
1273515	KEGG	113	Insulin resistance
1311111	KEGG	140	Phospholipase D signaling pathway
1320226	KEGG	103	AGE-RAGE signaling pathway in diabetic complications
1320227	KEGG	92	Longevity regulating pathway - mammal
1336227	REACTOME	79	Mitochondrial translation
1336230	REACTOME	138	Assembly of the primary cilium
1336239	REACTOME	11	Notch-HLH transcription pathway
1336240	REACTOME	47	Nuclear Receptor transcription pathway
1336247	REACTOME	76	RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription
1336261	REACTOME	98	RNA Polymerase II Transcription
1336274	REACTOME	130	Processing of Capped Intron-Containing Pre-mRNA
1336294	REACTOME	48	Deadenylation-dependent mRNA decay
1336310	REACTOME	132	Translation
1336327	REACTOME	63	Epigenetic regulation of gene expression
1336350	REACTOME	50	DNA Damage Bypass
1336358	REACTOME	121	DNA Double-Strand Break Repair
1336373	REACTOME	105	Nucleotide Excision Repair
1336386	REACTOME	30	Fanconi Anemia Pathway
1336387	REACTOME	75	Cell-Cell communication
1336393	REACTOME	10	Cell-extracellular matrix interactions
1336404	REACTOME	10	Depolarization of the Presynaptic Terminal Triggers the Opening of Calcium Channels
1336405	REACTOME	45	Neurotransmitter Release Cycle
1336423	REACTOME	114	Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell

1336424	REACTOME	10	Acetylcholine Binding And Downstream Events
1336425	REACTOME	10	Activation of Nicotinic Acetylcholine Receptors
1336426	REACTOME	10	Presynaptic nicotinic acetylcholine receptors
1336451	REACTOME	10	GABA A receptor activation
1336458	REACTOME	87	Potassium Channels
1336480	REACTOME	25	Detoxification of Reactive Oxygen Species
1336481	REACTOME	27	Cellular response to heat stress
1336486	REACTOME	129	Cellular Senescence
1336492	REACTOME	61	Macroautophagy
1336500	REACTOME	10	Protein folding
1336501	REACTOME	10	Chaperonin-mediated protein folding
1336502	REACTOME	10	Association of TriC/CCT with target proteins during biosynthesis
1336504	REACTOME	27	Gamma carboxylation, hypusine formation and arylsulfatase activation
1336511	REACTOME	10	The activation of arylsulfatases
1336515	REACTOME	10	Attachment of GPI anchor to uPAR
1336517	REACTOME	60	Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide, LLO) and transfer to a nascent protein
1336525	REACTOME	15	N-glycan trimming in the ER and Calnexin/Calreticulin cycle
1336528	REACTOME	132	Transport to the Golgi and subsequent modification
1336537	REACTOME	10	Reactions specific to the complex N-glycan synthesis pathway
1336539	REACTOME	51	O-linked glycosylation
1336542	REACTOME	10	Termination of O-glycan biosynthesis
1336543	REACTOME	79	SUMOylation
1336553	REACTOME	48	Peptide hormone metabolism
1336556	REACTOME	10	Metabolism of Angiotensinogen to Angiotensins
1336562	REACTOME	15	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
1336563	REACTOME	16	Unfolded Protein Response (UPR)
1336565	REACTOME	10	IRE1alpha activates chaperones
1336569	REACTOME	23	Surfactant metabolism
1336576	REACTOME	66	Glucose metabolism
1336587	REACTOME	11	Pentose phosphate pathway (hexose monophosphate shunt)
1336589	REACTOME	94	Glycosaminoglycan metabolism
1336606	REACTOME	43	Inositol phosphate metabolism
1336611	REACTOME	10	Synthesis of IP2, IP, and Ins in the cytosol
1336613	REACTOME	43	Lipid digestion, mobilization, and transport
1336614	REACTOME	10	Digestion of dietary lipid
1336621	REACTOME	86	Fatty acid, triacylglycerol, and ketone body metabolism
1336640	REACTOME	24	Peroxisomal lipid metabolism
1336649	REACTOME	28	Regulation of cholesterol biosynthesis by SREBP (SREBF)
1336651	REACTOME	30	Bile acid and bile salt metabolism

1336654	REACTOME	10	Synthesis of bile acids and bile salts via 24-hydroxycholesterol
1336657	REACTOME	25	Metabolism of steroid hormones
1336663	REACTOME	123	Phospholipid metabolism
1336686	REACTOME	10	Synthesis of PIPs at the early endosome membrane
1336691	REACTOME	45	Arachidonic acid metabolism
1336692	REACTOME	10	Synthesis of Prostaglandins (PG) and Thromboxanes (TX)
1336701	REACTOME	68	Sphingolipid metabolism
1336705	REACTOME	61	Integration of energy metabolism
1336720	REACTOME	10	eNOS activation
1336725	REACTOME	46	Pyruvate metabolism and Citric Acid (TCA) cycle
1336737	REACTOME	76	Metabolism of nucleotides
1336748	REACTOME	128	Metabolism of vitamins and cofactors
1336767	REACTOME	21	Amino acid synthesis and interconversion (transamination)
1336769	REACTOME	19	Branched-chain amino acid catabolism
1336771	REACTOME	38	Histidine, lysine, phenylalanine, tyrosine, proline and tryptophan catabolism
1336784	REACTOME	17	Amine-derived hormones
1336789	REACTOME	13	Glyoxylate metabolism and glycine degradation
1336791	REACTOME	21	Sulfur amino acid metabolism
1336799	REACTOME	20	Metabolism of porphyrins
1336802	REACTOME	149	Biological oxidations
1336806	REACTOME	10	Xenobiotics
1336826	REACTOME	10	Glutathione synthesis and recycling
1336841	REACTOME	10	Reversible hydration of carbon dioxide
1336845	REACTOME	45	Platelet homeostasis
1336856	REACTOME	13	Signal amplification
1336859	REACTOME	14	Thrombin signalling through proteinase activated receptors (PARs)
1336860	REACTOME	27	GPVI-mediated activation cascade
1336861	REACTOME	32	Platelet Aggregation (Plug Formation)
1336868	REACTOME	85	Response to elevated platelet cytosolic Ca2+
1336871	REACTOME	35	Formation of Fibrin Clot (Clotting Cascade)
1336876	REACTOME	91	Cell surface interactions at the vascular wall
1336879	REACTOME	94	Factors involved in megakaryocyte development and platelet production
1336883	REACTOME	44	TCR signaling
1336888	REACTOME	51	Costimulation by the CD28 family
1336894	REACTOME	144	Signaling by the B Cell Receptor (BCR)
1336908	REACTOME	32	MHC class II antigen presentation
1336909	REACTOME	51	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
1336910	REACTOME	14	Rap1 signalling
1336912	REACTOME	96	Toll-Like Receptors Cascades
1336916	REACTOME	10	IRAK1 recruits IKK complex
1336924	REACTOME	10	ERKs are inactivated

1336936	REACTOME	10	IRAK1 recruits IKK complex upon TLR7/8 or 9 stimulation
1336938	REACTOME	10	TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling
1336947	REACTOME	25	Complement cascade
1336948	REACTOME	10	Initial triggering of complement
1336956	REACTOME	25	Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways
1336958	REACTOME	10	Inflammasomes
1336959	REACTOME	10	The NLRP3 inflammasome
1336961	REACTOME	57	RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways
1336978	REACTOME	40	Fcgamma receptor (FCGR) dependent phagocytosis
1336985	REACTOME	11	MAP2K and MAPK activation
1336986	REACTOME	26	Negative regulation of MAPK pathway
1336998	REACTOME	24	FCERI mediated Ca+2 mobilization
1337001	REACTOME	91	C-type lectin receptors (CLRs)
1337008	REACTOME	35	Interferon Signaling
1337017	REACTOME	30	Interleukin-1 signaling
1337021	REACTOME	12	Regulation of signaling by CBL
1337022	REACTOME	24	Interleukin-6 family signaling
1337026	REACTOME	12	Growth hormone receptor signaling
1337033	REACTOME	50	Collagen formation
1337039	REACTOME	22	Elastic fibre formation
1337041	REACTOME	16	Laminin interactions
1337042	REACTOME	17	Non-integrin membrane-ECM interactions
1337043	REACTOME	10	Syndecan interactions
1337044	REACTOME	21	ECM proteoglycans
1337045	REACTOME	57	Degradation of the extracellular matrix
1337048	REACTOME	61	Integrin cell surface interactions
1337051	REACTOME	57	Semaphorin interactions
1337055	REACTOME	10	SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion
1337061	REACTOME	12	Netrin-1 signaling
1337065	REACTOME	43	L1CAM interactions
1337070	REACTOME	62	EPH-Ephrin signaling
1337083	REACTOME	13	LGI-ADAM interactions
1337115	REACTOME	84	Programmed Cell Death
1337131	REACTOME	10	Apoptotic factor-mediated response
1337132	REACTOME	10	Cytochrome c-mediated apoptotic response
1337134	REACTOME	10	Activation of caspases through apoptosome-mediated cleavage
1337146	REACTOME	129	Muscle contraction
1337158	REACTOME	132	Cell Cycle Checkpoints
1337195	REACTOME	91	Mitotic G2-G2/M phases
1337207	REACTOME	68	Mitotic Prophase
1337215	REACTOME	107	Mitotic Prometaphase
1337224	REACTOME	17	Mitotic Telophase/Cytokinesis

1337241	REACTOME	63	Chromosome Maintenance
1337252	REACTOME	10	Packaging Of Telomere Ends
1337254	REACTOME	37	ABC-family proteins mediated transport
1337258	REACTOME	90	Transport of inorganic cations/anions and amino acids/oligopeptides
1337270	REACTOME	85	Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds
1337286	REACTOME	31	Transport of vitamins, nucleosides, and related molecules
1337291	REACTOME	32	Aquaporin-mediated transport
1337295	REACTOME	41	Iron uptake and transport
1337297	REACTOME	150	Ion channel transport
1337299	REACTOME	20	Ligand-gated ion channel transport
1337300	REACTOME	91	Stimuli-sensing channels
1337320	REACTOME	14	Translocation of GLUT4 to the plasma membrane
1337321	REACTOME	29	Binding and Uptake of Ligands by Scavenger Receptors
1337323	REACTOME	10	Scavenging by Class A Receptors
1337332	REACTOME	18	EGFR downregulation
1337352	REACTOME	45	Phospholipase C-mediated cascade; FGFR2
1337355	REACTOME	33	Negative regulation of FGFR2 signaling
1337378	REACTOME	81	PI3K Cascade
1337391	REACTOME	73	p75 NTR receptor-mediated signalling
1337414	REACTOME	11	Retrograde neurotrophin signalling
1337422	REACTOME	18	VEGFR2 mediated vascular permeability
1337426	REACTOME	12	Regulation of KIT signaling
1337437	REACTOME	26	Nuclear signaling by ERBB4
1337441	REACTOME	14	RAF-independent MAPK1/3 activation
1337446	REACTOME	106	Rho GTPase cycle
1337449	REACTOME	20	RHO GTPases activate PAKs
1337450	REACTOME	33	RHO GTPases activate PKNs
1337454	REACTOME	12	RHO GTPases activate IQGAPs
1337459	REACTOME	11	RHO GTPases Activate NADPH Oxidases
1337460	REACTOME	22	Signaling by BMP
1337461	REACTOME	53	Signaling by TGF-beta Receptor Complex
1337465	REACTOME	18	Signaling by NOTCH
1337471	REACTOME	10	Signaling by NOTCH3
1337477	REACTOME	59	Chemokine receptors bind chemokines
1337481	REACTOME	10	Relaxin receptors
1337482	REACTOME	35	Amine ligand-binding receptors
1337489	REACTOME	13	Eicosanoid ligand-binding receptors
1337492	REACTOME	15	Nucleotide-like (purinergic) receptors
1337499	REACTOME	32	Class B/2 (Secretin family receptors)
1337502	REACTOME	20	Class C/3 (Metabotropic glutamate/pheromone receptors)
1337504	REACTOME	108	G alpha (s) signalling events
1337507	REACTOME	140	G alpha (q) signalling events
1337511	REACTOME	57	Opioid Signalling

1337517	REACTOME	10	Adenylate cyclase activating pathway
1337522	REACTOME	21	WNT ligand biogenesis and trafficking
1337523	REACTOME	24	Degradation of beta-catenin by the destruction complex
1337527	REACTOME	10	WNT mediated activation of DVL
1337529	REACTOME	44	Formation of the beta-catenin:TCF transactivating complex
1337530	REACTOME	31	Deactivation of the beta-catenin transactivating complex
1337531	REACTOME	10	Degradation of DVL
1337533	REACTOME	11	Regulation of FZD by ubiquitination
1337539	REACTOME	10	WNT5A-dependent internalization of FZD2, FZD5 and ROR2
1337541	REACTOME	15	Signaling by Hippo
1337546	REACTOME	12	Signaling by Activin
1337547	REACTOME	70	Visual phototransduction
1337553	REACTOME	40	Signaling by Retinoic Acid
1337556	REACTOME	110	Signaling by Hedgehog
1337565	REACTOME	42	Death Receptor Signalling

Table S3.3: *P-value of significant pathways that contain PLCB1, including and excluding the gene.*

Biological process	Pathway name	NCBI ID	P-Value with PLCB1	P-Value without PLCB1
Behavior	Opioid Signaling	1337511	0.00	1.65e-09
Behavior	Glutamatergic synapse	213816	0.00	0.00
Behavior	Dopaminergic synapse	469195	0.00	0.00
Behavior	Serotonergic synapse	525344	0.00	0.00
Behavior	Long-term depression	84497	0.00	1.11e-09
Behavior	Adrenergic signaling in cardiomyocytes	908279	0.00	5.44e-15
Biological regulation	Phosphatidylinositol signaling system	84464	1.67e-15	4.43e-07
Biological regulation	Renin secretion	1223594	2.22e-16	1.45e-06
Cellular process	Wnt signaling pathway	84473	0.00	7.14e-10
Immune response	Chagas disease (American trypanosomiasis)	147807	0.00	6.52e-09
Immune response	Inflammatory mediator regulation of TRP channels	948291	3.24e-14	4.38e-06
Metabolic process	Insulin secretion	777548	0.00	5.12e-11
Metabolic process	Pancreatic secretion	169304	0.00	1.50e-12
Metabolic process	Phospholipase D signaling pathway	1311111	1.99e-12	1.36e-05
Muscle contraction	Vascular smooth muscle contraction	96246	2.22e-16	5.01e-08

Supplementary material Chapter 4: "Selection pressure and network topology in wild and domestic pigs"

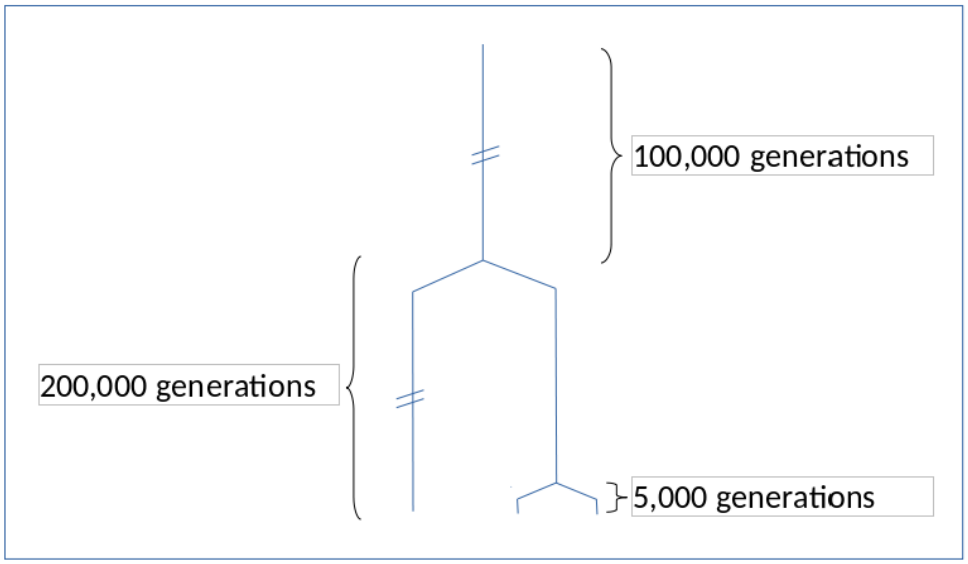

***Figure S4.1**: Graphical representation of simulated demographic model for three pig population*

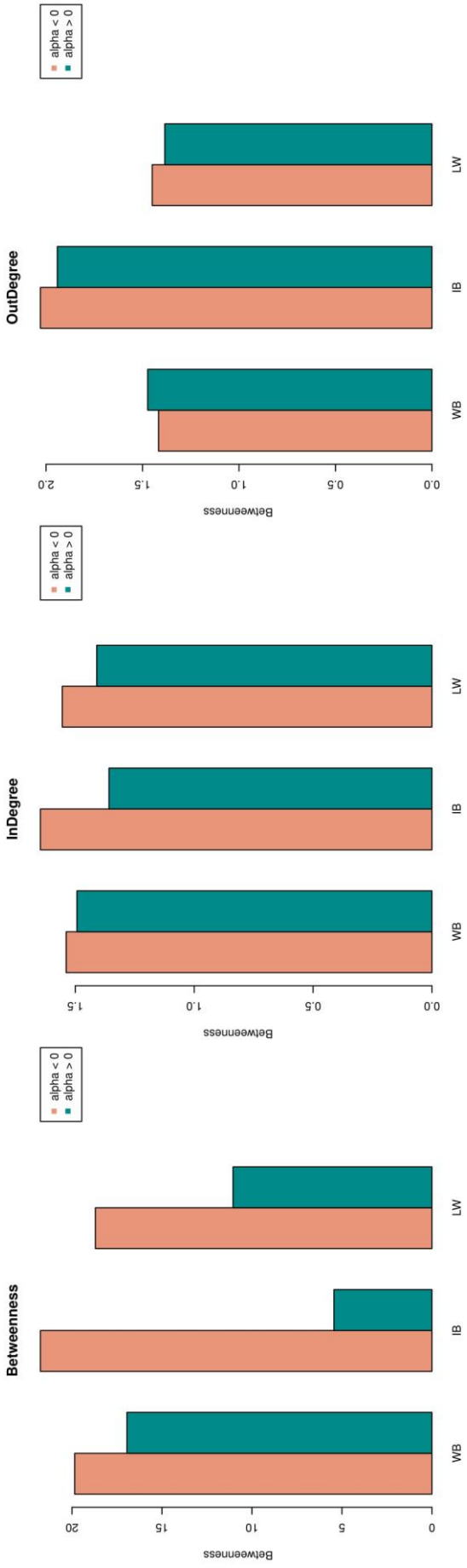

Figure S4.2. *Means of network topology features (betweenness, in-degree and out-degree) of genes within pathways, grouping genes according to positive and negative values of α.*IB, Iberian; LW, Large White; WB, Wild boar.*

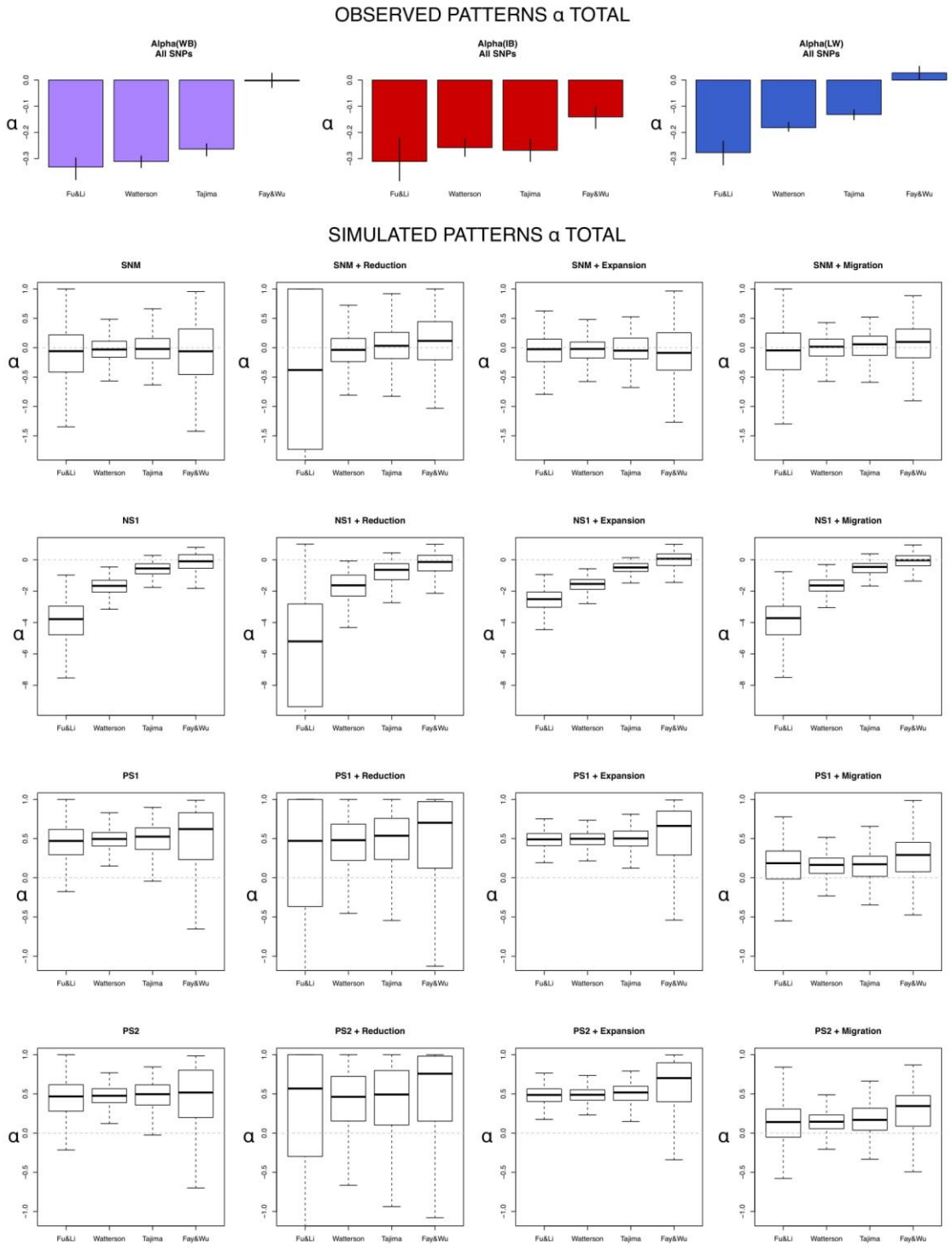

Figure S4.3. *Levels of α calculated from total polymorphisms using different variability estimators for each simulated not-combined scenario. *IB; Iberian; LW, large white; WB, Wild boar. SNM, standard neutral model; NS1; negative selection 1; PS1, positive selection 1, PS2, positive selection 2.*

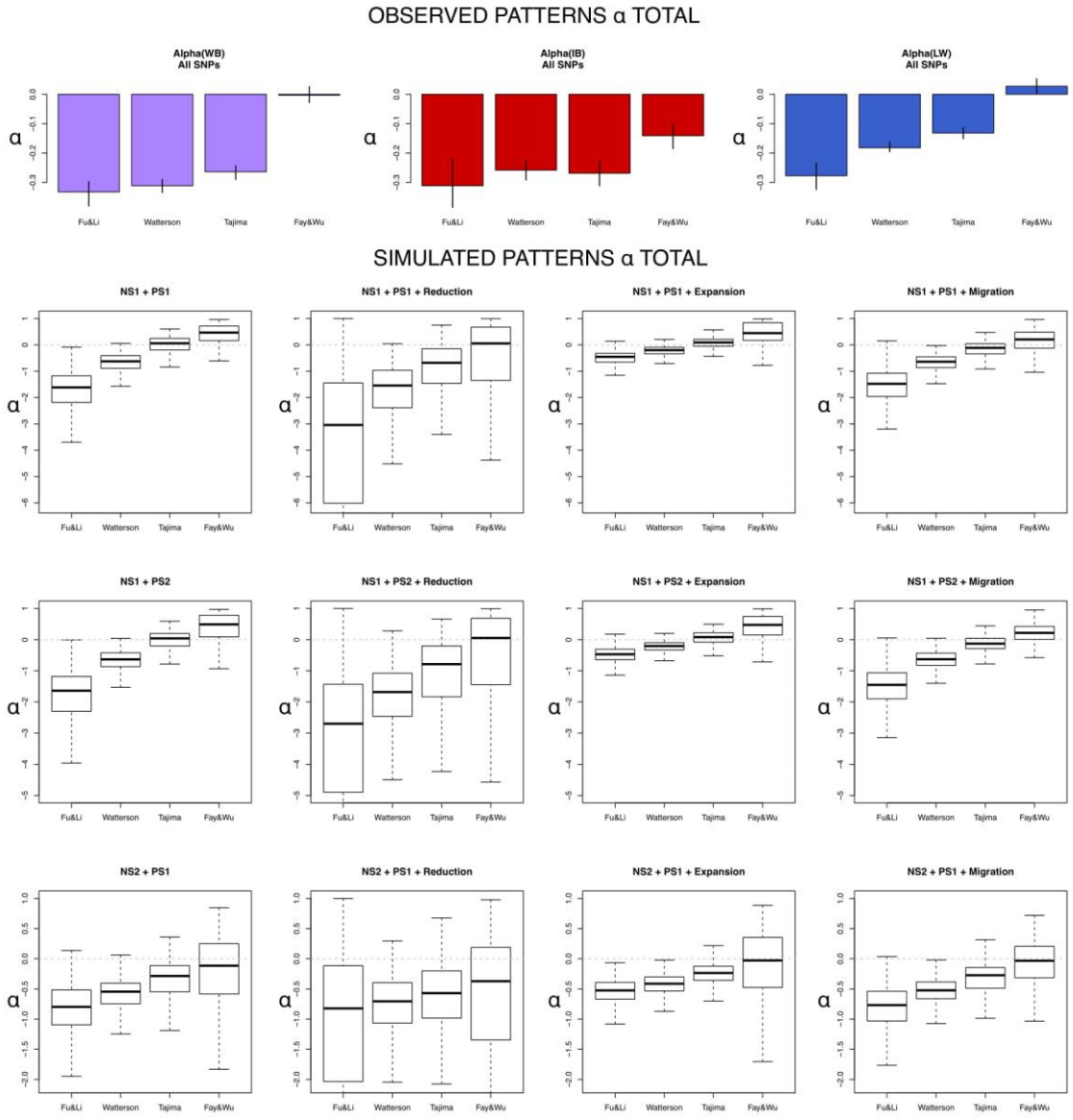

Figure S4.4. *Levels of α calculated from total polymorphisms using differente variability estimator for each simulated combined scenario (negative selection with positive selection). *IB; Iberian; LW, large white; WB, Wild boar. NS1; negative selection 1; NS2, negative selection 2; PS1, positive selection 1, PS2, positive selection 2.*

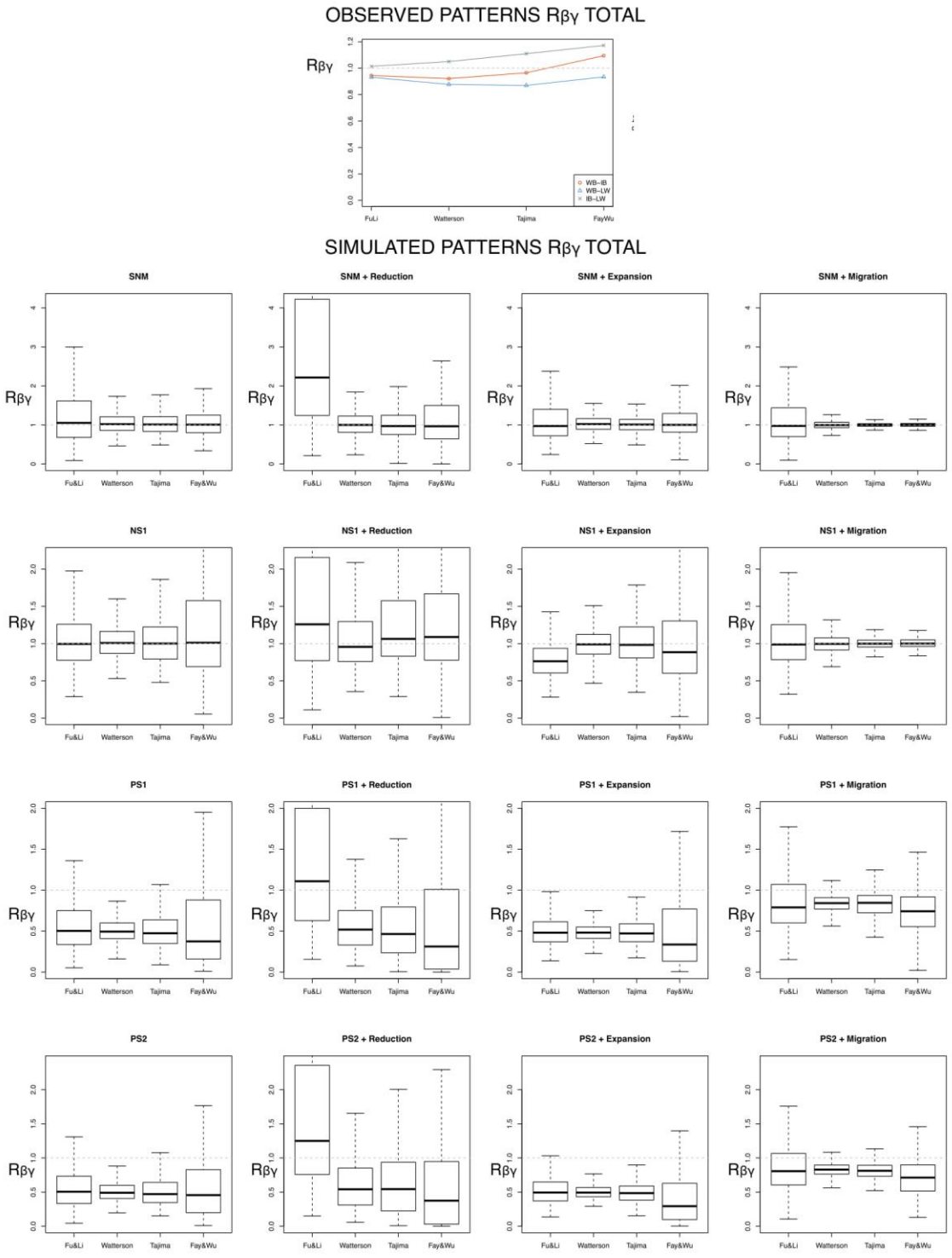

Figure S4.5. *Levels of R_βγ ration calculated from total polymorphisms using differente variability estimator for each simulated non-combined scenario. *IB; Iberian; LW, large white; WB, Wild boar. SNM, standard neutral model; NS1; negative selection 1; PS1, positive selection 1, PS2, positive selection 2.*

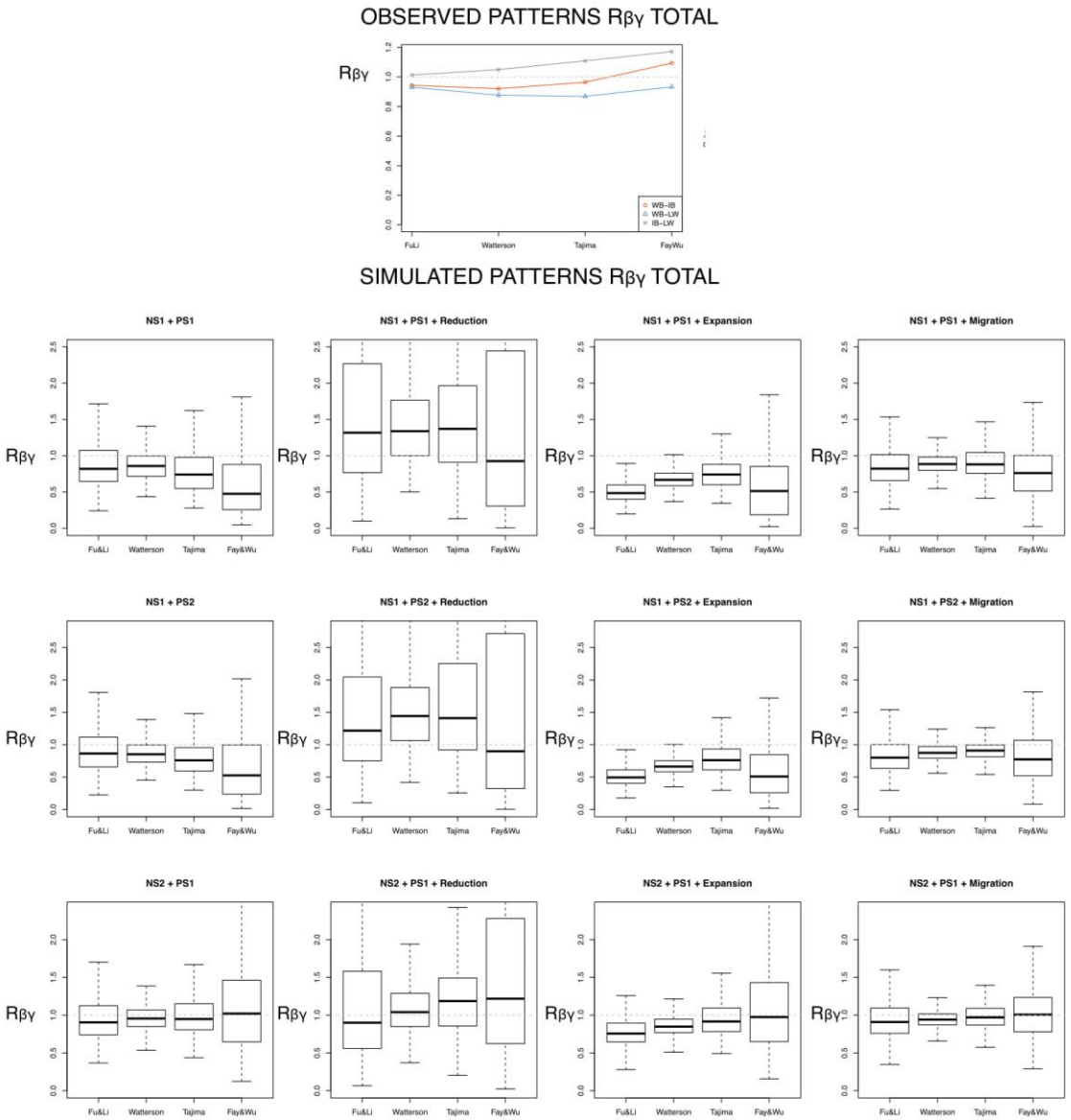

Figure S4.6. *Levels of R_βγ ration calculated from total polymorphisms using different variability estimator for each simulated combined scenario (negative selection with positive selection). *IB; Iberian; LW, large white; WB, Wild boar. NS1; negative selection 1; NS2, negative selection 2; PS1, positive selection 1, PS2, positive selection 2.*

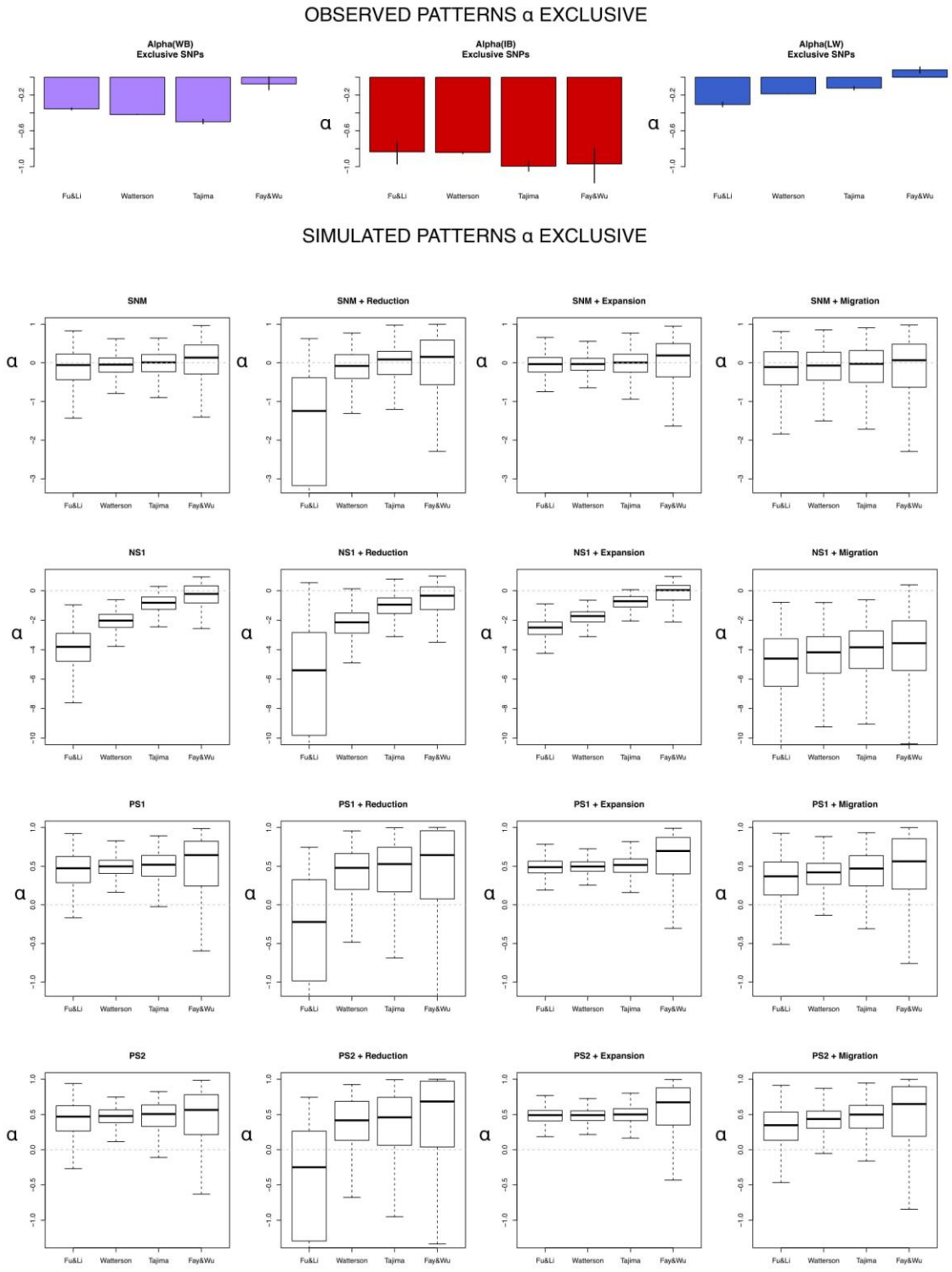

Figure S4.7. *Levels of α calculated from exclusive polymorphisms using different variability estimator for each simulated non-combined scenario. *IB; Iberian; LW, large white; WB, Wild boar. SNM, standard neutral model; NS1; negative selection 1; PS1, positive selection 1, PS2, positive selection 2.*

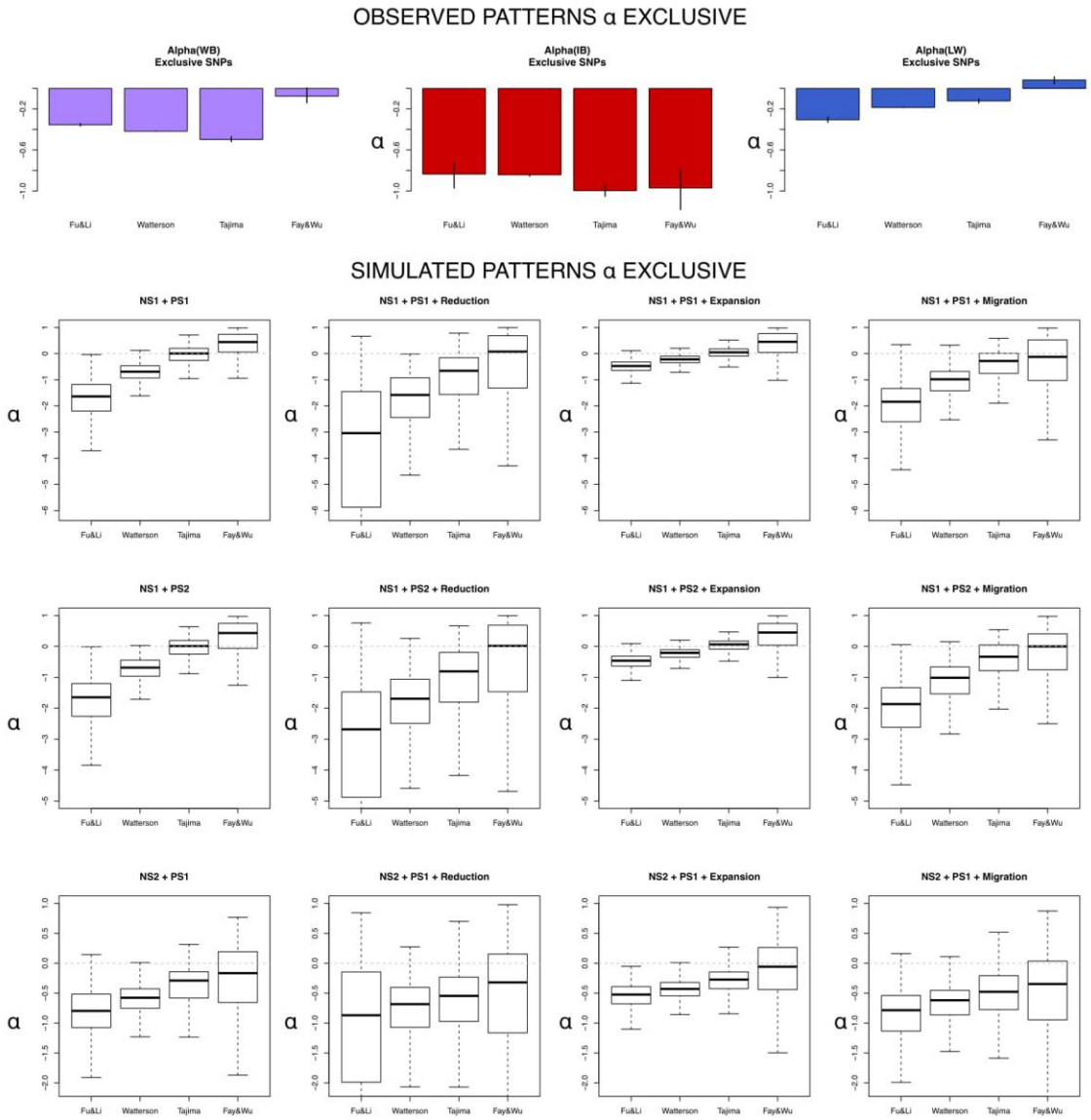

Figure S4.8. *Levels of α calculated from exclusive polymorphisms using different variability estimator for each simulated combined scenario (negative selection with positive selection). *IB; Iberian; LW, large white; WB, Wild boar. NS1; negative selection 1; NS2, negative selection 2; PS1, positive selection 1, PS2, positive selection 2.*

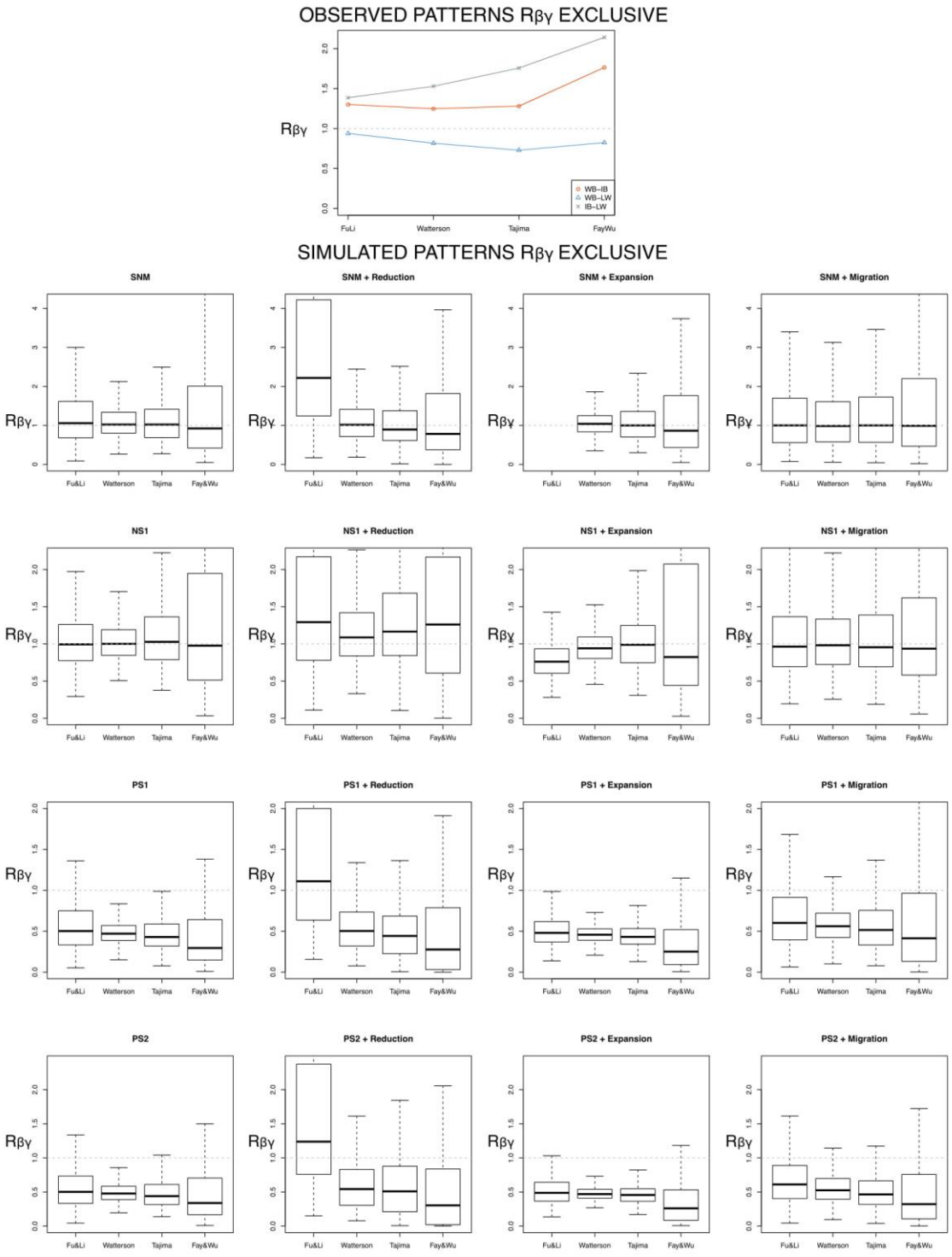

Figure S4.9. *Levels of R_βγ ration calculated from exclusive polymorphisms using different variability estimator for each simulated non-combined scenario. *IB; Iberian; LW, large white; WB, Wild boar. SNM, standard neutral model; NS1; negative selection 1; PS1, positive selection 1, PS2, positive selection 2.*

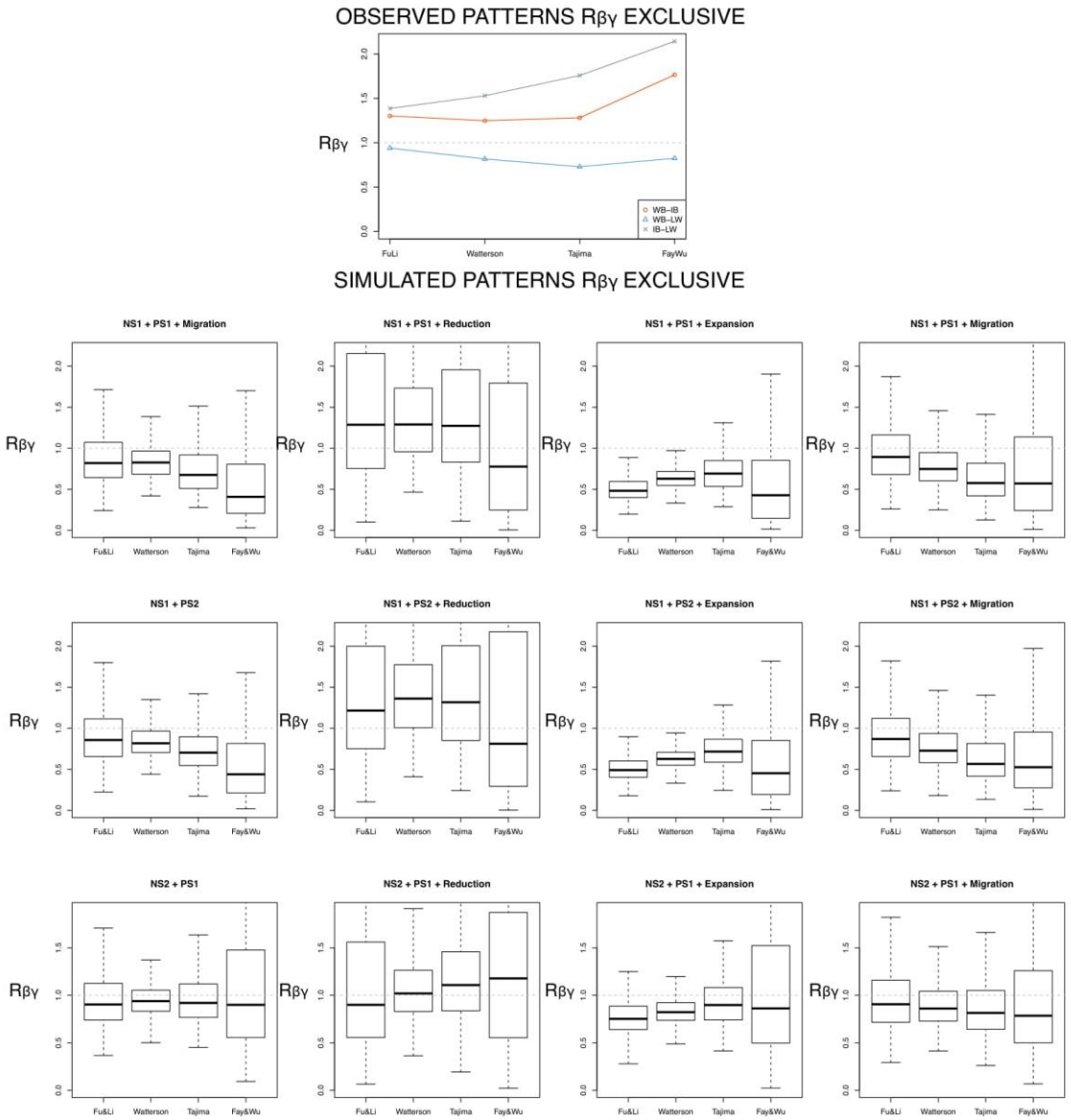

Figure S4.10. *Levels of R_βγ ration calculated from exclusive polymorphisms using different variability estimator for each simulated combined scenario (negative selection with positive selection). *IB; Iberian; LW, large white; WB, Wild boar. NS1; negative selection 1; NS2, negative selection 2; PS1, positive selection 1, PS2, positive selection 2.*

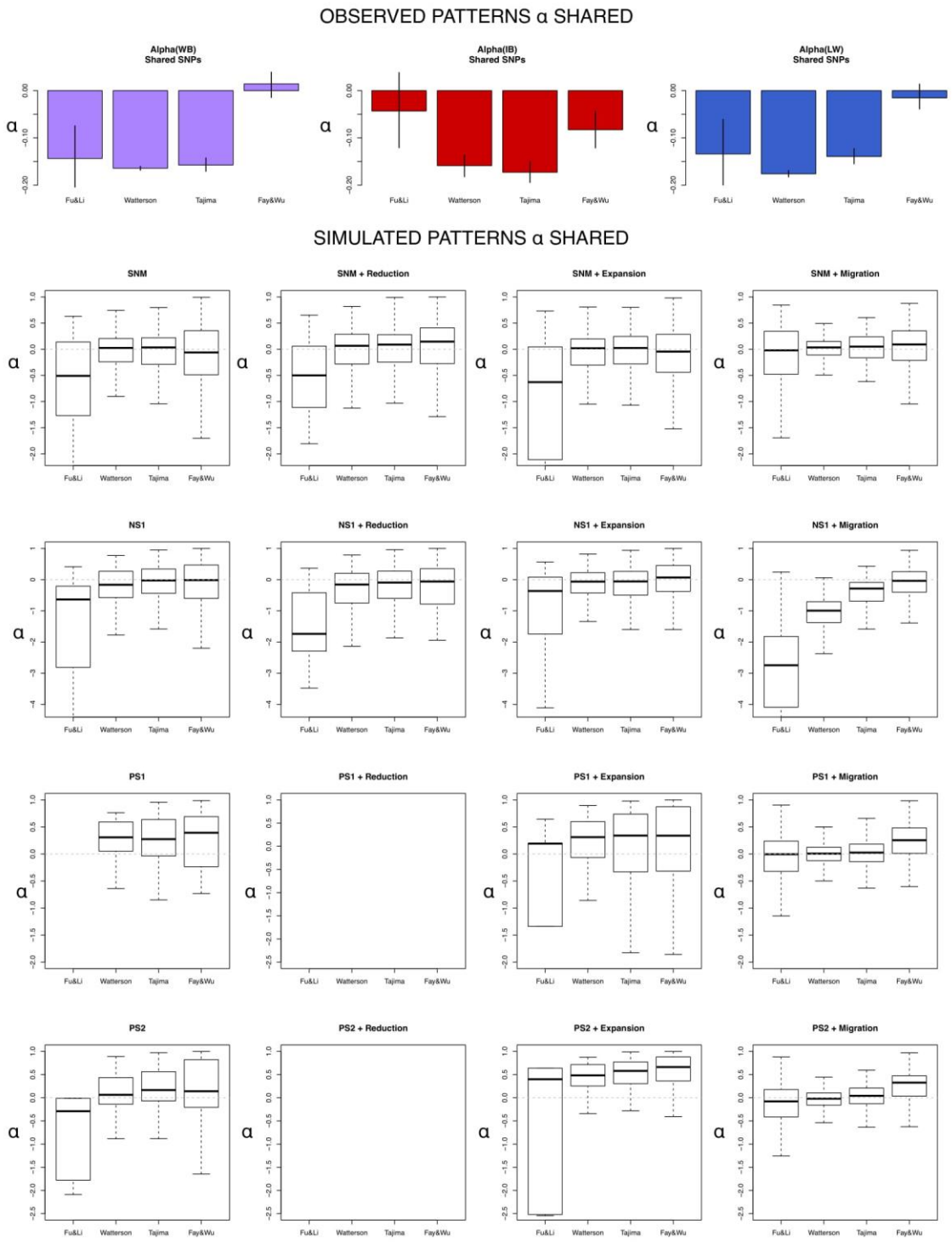

Figure S4.11. *Levels of α calculated from shared polymorphisms using different variability estimator for each simulated non-combined scenario. *IB; Iberian; LW, large white; WB, Wild boar. SNM, standard neutral model; NS1; negative selection 1; PS1, positive selection 1, PS2, positive selection 2.*

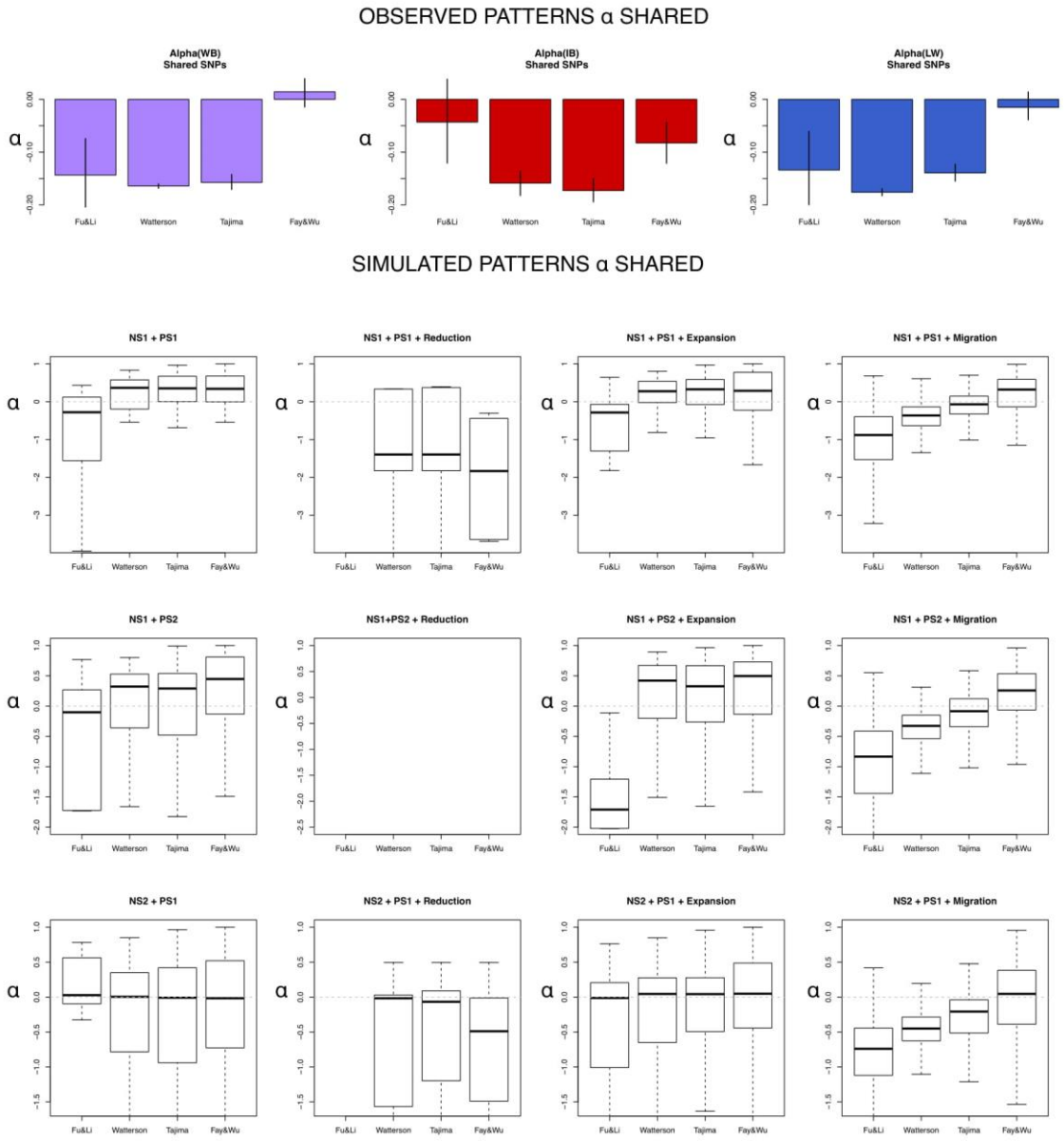

Figure S4.12. *Levels of α calculated from shared polymorphisms using different variability estimator for each simulated combined scenario (negative selection with positive selection). *IB; Iberian; LW, large white; WB, Wild boar. NS1; negative selection 1; NS2, negative selection 2; PS1, positive selection 1, PS2, positive selection 2.*

Figure S4.13. *Levels of R_βγ ration calculated from shared polymorphisms using different variability estimator for each simulated non-combined scenario. *IB; Iberian; LW, large white; WB, Wild boar. SNM, standard neutral model; NS1; negative selection 1; PS1, positive selection 1, PS2, positive selection 2.*

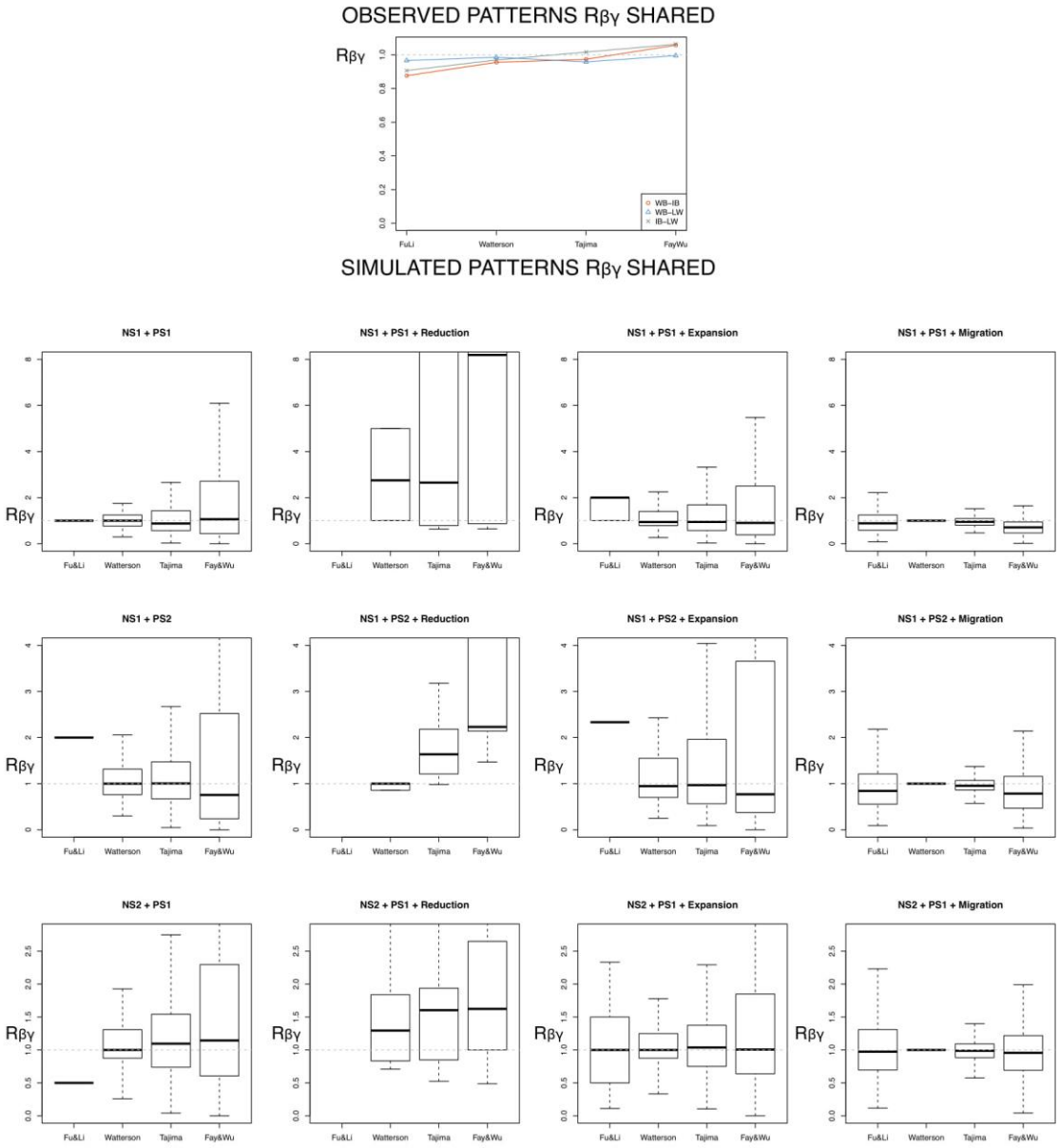

Figure S4.14. *Levels of R_βγ ration calculated from shared polymorphisms using differente variability estimator for each simulated combined scenario (negative selection with positive selection). *IB; Iberian; LW, large white; WB, Wild boar. NS1; negative selection 1; NS2, negative selection 2; PS1, positive selection 1, PS2, positive selection 2.*

Table S4.1: *List of pig samples.*

Status	Breed	Origin	Sample	Accession	Depth
Domestic	Iberian	Spain	IBGM0327	SRR1513307	13,0
Domestic	Iberian	Spain	IBGU1330	SRR765849	7,2
Domestic	Iberian	Spain	IBGU1802	SRX245748	12,4
Domestic	Iberian	Spain	IBGU1803	SRR3745079	13,0
Domestic	Iberian	Spain	IBGU1804	SRR1917381	14,5
Domestic	Iberian	Spain	IBGU1805	SRR5515065	12,4
Domestic	Large White	International	LW22F02	ERR173186	10,0
Domestic	Large White	International	LW22F03	ERR173187	10,1
Domestic	Large White	International	LW22F04	ERR173188	10,1
Domestic	Large White	International	LW22F06	ERR173189	9,3
Domestic	Large White	International	LW22F07	ERR173190	11,6
Domestic	Large White	International	LW22M07	ERR173192	10,4
Domestic	Large White	International	LW36F01	ERR173193	9,8
Domestic	Large White	International	LW36F04	ERR173196	9,4
Domestic	Large White	International	LWGB0348	SRR5513124	12,0
Domestic	Large White	International	LWNA3577	SRR1581108, SRR1581107	9,9
Domestic	Large White	International	LWNA3579	SRR1581111, SRR1581110	10,1
Domestic	Large White	International	LWNA3582	SRR1581121, SRR1581119	9,0
Domestic	Large White	International	LWNA3584	SRR1581128, SRR1581127	11,8
Domestic	Large White	International	LWNA3594	SRR1581137	11,8
Domestic	Large White	International	LWNA3595	SRR1581139	12,8
Domestic	Large White	International	LWNA3596	SRR1581138	12,0
Domestic	Large White	International	LWNA3597	SRR1581140	12,5
Domestic	Large White	International	LWNA3599	SRR1581141	12,3
Domestic	Large White	International	LWNA37F01	ERR977060	17,9
Domestic	Large White	International	LWNA38M02	ERR977062	19,6
Wild	Wild Boar	Switzerland	WBCH26M09	ERR173218	14,4
Wild	Wild Boar	Spain	WBES0231		11,2
Wild	Wild Boar	Spain	WBES0252		12,2
Wild	Wild Boar	Spain	WBES0288		5,2
Wild	Wild Boar	Spain	WBES0291		11,5
Wild	Wild Boar	Spain	WBES0297		5,4
Wild	Wild Boar	Spain	WBES0494	SRR3745077	12,6
Wild	Wild Boar	Spain	WBES0717	SRR1513306	13,0
Wild	Wild Boar	France	WBFR25U11	ERR173217	9,4
Wild	Wild Boar	Netherlands	WBNL21M03	ERR173214	11,4
Wild	Wild Boar	Netherlands	WBNL21F04	ERR977317	15,3
Wild	Wild Boar	Netherlands	WBNL22M02	ERR977342	16,6
Wild	Wild Boar	Tunisia	WBTN0965	SRR3745078	12,4

Wild	Wild Boar	Greece	WBGR32F07	ERR977364	10,7
Wild	Wild Boar	Greece	WBGR32U05	ERR977367	10,2
Wild	Wild Boar	Italy	WBIT44U06	ERR977380	13,2
Wild	Wild Boar	Italy	WBIT44U07	ERR977383	11,8
Wild	Wild Boar	Italy	WBIT28F31	ERR977355	17,3
Wild	Wild Boar	Italy	WBIT28M39	ERR977356	12,6
Wild	Wild Boar	Italy	WBIT42M09	ERR977377	13,6

Table S4.2. *Pearson correlation (and P-value within the parenthesis) between the nucleotide variability and divergence with the ratio of missing data using a filtered dataset of genes. This dataset is composed by the genes with a proportion of missing less than 0.3 and with variability and divergence values lower than the 99% quantile of the total genes. *IB, Iberian; LW, Large White; WB, Wild boar.*

		Missing	
		Synonymous	**Non-Synonymous**
IB	Fu&Li	-0.0023 (0.785)	-0.0142 (0.101)
	Watterson	-0.0115 (0.180)	-0.0204 (0.017)
	Tajima	-0.0087 (0.310)	-0.0178 (0.038)
	Fay&Wu	-0.0163 (0.058)	-0.0205 (0.017)
	Divergence	-0.0123 (0.154)	-0.0523 (1.2e-09)
LW	Fu&Li	-0.0585 (1.1e-10)	-0.0089 (0.327)
	Watterson	-0.0228 (0.012)	-0.0131 (0.148)
	Tajima	0.0109 (0.228)	-0.0110 (0.222)
	Fay&Wu	-0.0048 (0.593)	-0.0248 (0.006)
	Divergence	0.0125 (0.167)	-0.0203 (0.024)
WB	Fu&Li	-0.1851 (<2.2e-16)	-0.1251 (<2.2e-16)
	Watterson	-0.1440 (<2.2e-16)	-0.0970 (<2.2e-16)
	Tajima	-0.0600 (2.4e-12)	-0.0459 (8.4e-08)
	Fay&Wu	-0.0457 (1.0e-07)	-0.0536 (4.1e-10)
	Divergence	-0.0369 (1.7e-05)	-0.0515 (1.9e-09)

Table S4.3. *Models of selection simulated with SLiM and its parameter values. SNM, standard neutral model; NS1, negative selection 1; NS2, negative selection 2; PS1, positive selection 1; PS2, positive selection 2; PS3, positive selection 3.*

Model	Fitness neutral mutations	Fitness deleterious mutations	Fitness beneficial mutations	Proportion mutations (deleterious/ neutral/ beneficial)	Mutation rate	Recombination rate	Length	umber of individuals
SNM	0	-1	0	1/9/0	2,50e-07	1,17e-08	10,000bp	10,000
SNM + PS1	0	-1	0 0.00005 (domestic)	1/9/1				
SNM + PS2	0	-1	0 0.005 (domestic)	1/9/1				
SNM + PS3	0	-1	0 0.1 (domestic)	1/9/0.01				
NS1	0	-0,01	0	1/9/0				
NS2 + PS1	0	-0,0005	0 0.00005 (domestic)	1/9/1				
NS1 + PS1	0	-0,01	0 0.00005 (domestic)	1/9/1				
NS1 + PS2	0	-0,01	0 0.005 (domestic)	1/9/1				
NS1 + PS3	0	-0,01	0 0.1 (domestic)	1/9/0.01				
PS2	0	-0,1	0,005	10/0/1				

Table S3.4. *Levels of variability at synonymous and non-synonymous for all breeds, variant classification and variability estimators at genome level. *IB, Iberian; LW, Large White; WB, Wild boar. Ps, synonymous polymorphism; Pn, non-synonymous polymorphism; Ds, synonymous divergence; Dn, non-synonymous divergence.*

		WB				IB				LW			
		Ps	**Pn**	**Ds**	**Dn**	**Ps**	**Pn**	**Ds**	**Dn**	**Ps**	**Pn**	**Ds**	**Dn**
All SNPs	FuLi	0,00585	0,00110	0,00878	0,00123	0,00061	0,00011	0,00859	0,00116	0,00382	0,00067	0,00854	0,00117
	Watterson	0,00312	0,00057	0,00869	0,00121	0,00110	0,00019	0,00855	0,00114	0,00308	0,00049	0,00846	0,00115
	Tajima	0,00231	0,00041	0,00866	0,00120	0,00114	0,00019	0,00854	0,00114	0,00277	0,00042	0,00840	0,00113
	FayWu	0,00338	0,00047	0,00862	0,00121	0,00165	0,00025	0,00851	0,00114	0,00431	0,00056	0,00835	0,00112
Shared SNPs	FuLi	0,00059	0,00009			0,00040	0,00006			0,00062	0,00010		
	Watterson	0,00130	0,00021			0,00094	0,00015			0,00130	0,00021		
	Tajima	0,00159	0,00026			0,00101	0,00016			0,00159	0,00024		
	FayWu	0,00271	0,00037			0,00154	0,00022			0,00244	0,00033		
Private SNPs	FuLi	0,00526	0,00100			0,00021	0,00005			0,00319	0,00057		
	Watterson	0,00182	0,00036			0,00016	0,00004			0,00178	0,00029		
	Tajima	0,00072	0,00015			0,00013	0,00004			0,00118	0,00018		
	FayWu	0,00067	0,00010			0,00011	0,00003			0,00187	0,00023		

Table S4.5. *Levels of α for all breeds, ariant classification in three different populations of pigs. *IB, Iberian; LW, Large White; WB, Wild boar.*

		WB	IB	LW
All SNPs	FuLi	-0,332	-0,311	-0,277
	Watterson	-0,311	-0,258	-0,182
	Tajima	-0,264	-0,269	-0,132
	FayWu	-0,003	-0,141	0,028
Exclusive SNPs	FuLi	-0,354	-0,835	-0,305
	Watterson	-0,417	-0,843	-0,186
	Tajima	-0,498	-0,997	-0,121
	FayWu	-0,075	-0,971	0,083
Shared SNPs	FuLi	-0,144	-0,043	-0,134
	Watterson	-0,164	-0,159	-0,176
	Tajima	-0,157	-0,173	-0,139
	FayWu	0,014	-0,083	-0,015
Shared IB-LW SNPs	FuLi		0,110	0,269
	Watterson		0,048	0,072
	Tajima		-0,059	-0,199
	FayWu		-0,266	-0,263

Table S4.6. *Pathways with the highest $\alpha_{Fay\&Wu}$ values in the three populations of pigs. *IB, Iberian; LW, Large White; WB, Wild boar.*

WB		IB		LW	
Pathway	**$\alpha_{Fay\&Wu}$**	**Pathway**	**$\alpha_{Fay\&Wu}$**	**Pathway**	**$\alpha_{Fay\&Wu}$**
Maturity onset diabetes of the young	0,99230	Maturity onset diabetes of the young	1,00000	Phototransduction	0,91997
Protein export	0,90743	Nitrogen metabolism	1,00000	Folate biosynthesis	0,84830
Phototransduction	0,90399	Protein export	1,00000	Spliceosome	0,81299
Apelin signaling pathway	0,82168	Basal transcription factors	0,96922	Thyroid cancer	0,81254
Basal transcription factors	0,81404	Hypertrophic cardiomyopathy	0,96097	mRNA surveillance pathway	0,78237
Carbohydrate digestion and absorption	0,78207	Beta-Alanine metabolism	0,86393	Morphine addiction	0,75427
GnRH signaling pathway	0,77080	Carbohydrate digestion and absorption	0,85869	Antifolate resistance	0,72263
Circadian rhythm	0,74982	Mitophagy-animal	0,85324	Basal transcription factors	0,71137
Mineral absorption	0,73443	MAPK (JNK) signaling	0,83623	Choline metabolism in cancer	0,60802
Amyotrophic lateral sclerosis (ALS)	0,72434	Viral myocarditis	0,80584	Arginine and proline metabolism	0,58645

Table S4.7. *Genomic windows of 5Mb with the highest $\alpha_{Fay\&Wu}$ values in the three different populations of pig. *IB, Iberian; LW, Large White; WB, Wild boar*

WB		IB		LW	
Chromosome : Initial - Final Positions	**$\alpha_{Fay\&Wu}$**	**Chromosome : Initial - Final Positions**	**$\alpha_{Fay\&Wu}$**	**Chromosome : Initial - Final Positions**	**$\alpha_{Fay\&Wu}$**
15:10000001-15000000	1,00000	9:150000001-153670197	1,00000	16:10000001-15000000	1,00000
1:170000001-175000000	1,00000	8:20000001-25000000	1,00000	15:45000001-50000000	0,99930
15:45000001-50000000	0,99734	7:110000001-115000000	1,00000	11:30000001-35000000	0,98920
11:65000001-70000000	0,98494	4:55000001-60000000	1,00000	15:100000001-105000000	0,97879
1:215000001-220000000	0,96279	2:95000001-100000000	1,00000	1:170000001-175000000	0,95827
15:100000001-105000000	0,96238	2:115000001-120000000	1,00000	14:125000001-130000000	0,95233
10:1-5000000	0,95963	16:85000001-86898991	1,00000	1:65000001-70000000	0,95093
16:15000001-20000000	0,95849	16:15000001-20000000	1,00000	14:150000001-153851969	0,94730
11:30000001-35000000	0,95648	15:80000001-85000000	1,00000	13:85000001-90000000	0,94534
13:190000001-195000000	0,94697	15:5000001-10000000	1,00000	15:10000001-15000000	0,92277
1:50000001-55000000	0,94079	15:45000001-50000000	1,00000	7:110000001-115000000	0,91288
11:55000001-60000000	0,92279	15:30000001-35000000	1,00000	11:55000001-60000000	0,90170
1:190000001-195000000	0,91743	15:155000001-157681621	1,00000	9:85000001-90000000	0,89437
16:65000001-70000000	0,91263	15:10000001-15000000	1,00000	13:65000001-70000000	0,89346
5:110000001-111506441	0,91231	15:100000001-105000000	1,00000	1:145000001-150000000	0,86119
13:45000001-50000000	0,91105	13:65000001-70000000	1,00000	15:95000001-100000000	0,86061
16:80000001-85000000	0,90642	11:60000001-65000000	1,00000	2:130000001-135000000	0,82900

16:10000001-15000000	0,90341	11:30000001-35000000	1,00000	18:20000001-25000000	0,82542
7:110000001-115000000	0,90178	10:1-5000000	1,00000	3:140000001-144787322	0,82330
6:5000001-10000000	0,89058	16:1-5000000	0,99712	16:15000001-20000000	0,80676
13:65000001-70000000	0,87955	14:125000001-130000000	0,99439	15:50000001-55000000	0,80520
4:65000001-70000000	0,85453	1:240000001-245000000	0,99190	14:85000001-90000000	0,80366
4:25000001-30000000	0,84057	16:65000001-70000000	0,99184	13:190000001-195000000	0,79835
9:85000001-90000000	0,82182	10:60000001-65000000	0,98889	1:315000001-315321322	0,78167
16:1-5000000	0,81841	3:140000001-144787322	0,97741	1:130000001-135000000	0,77286
6:20000001-25000000	0,81645	16:40000001-45000000	0,96543	6:95000001-100000000	0,77138
2:35000001-40000000	0,80054	1:215000001-220000000	0,95732	18:60000001-61220071	0,76378
1:195000001-200000000	0,79284	1:40000001-45000000	0,95056	13:40000001-45000000	0,76357
12:30000001-35000000	0,78398	8:35000001-40000000	0,94947	5:85000001-90000000	0,74643
10:35000001-40000000	0,78328	11:80000001-85000000	0,94127	6:30000001-35000000	0,72954

Table S4.8. *Estimates of variability and α in genes that were previously reported to be under positive selection using other approaches. *IB, Iberian; LW, Large White; WB, Wild boar*

Gene	Function	IB													
		$\theta_{Fu\&Li}$		$\theta_{Watterson}$		θ_{Tajima}		$\theta_{Fay\&Wu}$		Divergence		$\alpha_{Fu\&Li}$	$\alpha_{Watterson}$	α_{Tajima}	$\alpha_{Fay\&Wu}$
		Syn	nSyn	Syn	nSyn	Syn	nSyn	Syn	nSyn	Syn	nSyn				
NR6A1	Body size	0,0000	0,0000	0,0000	0,0004	0,0000	0,0003	0,0000	0,0023	0,0127	0,0011	NA	-Inf	-Inf	-Inf
PLAG1	Body size	0,0000	0,0000	0,0000	0,0004	0,0000	0,0006	0,0000	0,0003	0,0035	0,0004	NA	-Inf	-Inf	-Inf
LCORL	Body size	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	NA	NA	NA	NA
NR6A1	Body size	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0149	0,0000	NA	NA	NA	NA
KIT	Coat color	0,0000	0,0005	0,0055	0,0003	NA	NA	0,0071	0,0002	0,0113	0,0007	-Inf	0,0270	NA	0,6009
EDNRB	Coat color (Asian origin)	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0038	0,0000	NA	NA	NA	NA
LRRTM4	Behavior	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0035	0,0000	NA	NA	NA	NA
LRRTM3	Behavior	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0026	0,0000	NA	NA	NA	NA
LRRTM1	Behavior	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PPP1R1B	Behavior	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0029	NA	NA	NA	NA
LEMD3	Ear morphology	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	NA	NA	NA	NA
IGF2R	Lean growth	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0126	0,0006	NA	NA	NA	NA
JMJD1C	Fertility	0,0006	0,0002	0,0004	0,0002	0,0003	0,0002	0,0000	0,0000	0,0014	0,0001	-2,4615	-6,1567	-7,7022	-10,5845
OSTN	Body composition	0,0000	0,0000	0,0000	0,0012	0,0000	0,0007	0,0000	0,0060	0,0000	0,0030	NA	NA	NA	NA
AHR	Litter size	0,0034	0,0019	0,0011	0,0006	0,0006	0,0003	0,0001	0,0000	0,0173	0,0020	-3,6565	-3,8136	-4,1221	-4,7429

Gene	Function	LW													
		$\theta_{Fu\&Li}$		$\theta_{Watterson}$		θ_{Tajima}		$\theta_{Fay\&Wu}$		Divergence		$\alpha_{Fu\&Li}$	$\alpha_{Watterson}$	α_{Tajima}	$\alpha_{Fay\&Wu}$
		Syn	nSyn	Syn	nSyn	Syn	nSyn	Syn	nSyn	Syn	nSyn				
NR6A1	Body size	0,0042	0,0013	0,0010	0,0006	0,0002	0,0004	0,0000	0,0001	0,0128	0,0015	-1,5686	-4,0377	-12,9888	-68,0563
PLAG1	Body size	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0035	0,0000	NA	NA	NA	NA
LCORL	Body size	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	NA	NA	NA	NA
NR6A1	Body size	0,0000	0,0043	0,0000	0,0010	0,0000	0,0002	0,0000	0,0000	0,0149	0,0001	-Inf	-Inf	-Inf	-Inf
KIT	Coat color	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
EDNRB	Coat color (Asian origin)	0,0000	0,0012	0,0009	0,0003	0,0004	0,0001	0,0000	0,0000	0,0040	0,0000	-Inf	-40,5910	-19,5556	-8,4866
LRRTM4	Behavior	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0035	0,0000	NA	NA	NA	NA
LRRTM3	Behavior	0,0000	0,0000	0,0006	0,0000	0,0009	0,0000	0,0034	0,0000	0,0021	0,0000	NA	NA	NA	NA
LRRTM1	Behavior	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PPP1R1B	Behavior	0,0000	0,0000	0,0000	0,0007	0,0000	0,0008	0,0000	0,0043	0,0000	0,0025	NA	NA	NA	NA
LEMD3	Ear morphology	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	NA	NA	NA	NA
IGF2R	Lean growth	0,0042	0,0006	0,0054	0,0003	0,0036	0,0001	0,0105	0,0000	0,0131	0,0007	-1,7871	-0,1626	0,2498	0,9838
JMJD1C	Fertility	0,0076	0,0011	0,0042	0,0007	0,0024	0,0005	0,0014	0,0001	0,0024	0,0003	-0,1230	-0,3366	-0,7260	0,2899
OSTN	Body composition	0,0000	0,0033	0,0000	0,0016	0,0000	0,0003	0,0000	0,0065	0,0000	0,0033	NA	NA	NA	NA
AHR	Litter size	0,0000	0,0000	0,0039	0,0013	0,0059	0,0017	0,0018	0,0022	0,0037	0,0019	NA	0,3556	0,4274	-1,4077

Gene	Function	WB													
		$\theta_{Fu\&Li}$		$\theta_{Watterson}$		θ_{Tajima}		$\theta_{Fay\&Wu}$		Divergence		$\alpha_{Fu\&Li}$	$\alpha_{Watterson}$	α_{Tajima}	$\alpha_{Fay\&Wu}$
		Syn	nSyn	Syn	nSyn	Syn	nSyn	Syn	nSyn	Syn	nSyn				
NR6A1	Body size	0,0211	0,0013	0,0060	0,0006	0,0022	0,0003	0,0002	0,0000	0,0139	0,0001	-4,8007	-8,8221	-10,9035	-7,6995
PLAG1	Body size	0,0000	0,0000	0,0000	0,0003	0,0000	0,0005	0,0000	0,0004	0,0035	0,0004	NA	-Inf	-Inf	-Inf
LCORL	Body size	0,0107	0,0016	0,0026	0,0004	0,0006	0,0001	0,0000	0,0000	0,0003	0,0000	-0,0687	-0,0519	0,0000	0,0678
NR6A1	Body size	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0149	0,0000	NA	NA	NA	NA
KIT	Coat color	0,0031	0,0000	0,0048	0,0004	0,0042	0,0004	0,0113	0,0009	0,0106	0,0006	1,0000	-0,5965	-0,5629	-0,3554
EDNRB	Coat color (Asian origin)	0,0038	0,0000	0,0009	0,0000	0,0002	0,0000	0,0000	0,0000	0,0039	0,0000	NA	NA	NA	NA
LRRTM4	Behavior	0,0176	0,0000	0,0042	0,0000	0,0009	0,0000	0,0000	0,0000	0,0040	0,0000	NA	NA	NA	NA
LRRTM3	Behavior	0,0000	0,0008	0,0006	0,0004	0,0001	0,0001	0,0051	0,0000	0,0026	0,0001	-Inf	-22,1883	-35,2253	0,9498
LRRTM1	Behavior	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0044	0,0000	NA	NA	NA	NA
PPP1R1B	Behavior	0,0000	0,0029	0,0000	0,0007	0,0000	0,0002	0,0000	0,0000	0,0000	0,0030	NA	NA	NA	NA
LEMD3	Ear morphology	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	NA	NA	NA	NA
IGF2R	Lean growth	0,0042	0,0006	0,0011	0,0002	0,0003	0,0000	0,0000	0,0000	0,0128	0,0007	-1,9225	-2,0012	-2,2248	-2,5313
JMJD1C	Fertility	NA	NA	0,0036	0,0009	0,0024	0,0008	0,0007	0,0003	NA	NA	NA	-0,3150	-0,7923	-1,0365
OSTN	Body compositio	0,0000	0,0000	0,0025	0,0016	0,0019	0,0034	0,0002	0,0036	0,0011	0,0034	NA	0,8037	0,4583	-4,1313
AHR	Litter size	0,0000	0,0009	0,0048	0,0009	0,0024	0,0003	NA	NA	0,0163	0,0018	-Inf	-0,6319	-0,2191	NA

Acknowledgements

Después de 4 años, este apartado es probablemente una de las cosas más difíciles que estoy haciendo ya que como todos los que me conocen saben, no soy muy dado yo a estas cosas. Así que nadie se queje, que suficiente me está costando ya no poner un simple “Gracias a todos”.

En primer lugar, me gustaría agradecer a Miguel y Sebas por darme la oportunidad de realizar esta tesis sin conocerme previamente, ya que gracias a eso he aprendido mucho tanto a nivel científico como personal.

Durante todo este tiempo han pasado muchos compañeros por el CRAG de los que me llevo muy buenos recuerdos. Aunque al principio les costó hacerme socializar, al final lo consiguieron y por ello quiero agradecer a Erica, Manu, Antonia, Dani, Taina, Anna, Arianna y Rayner que, siendo el nuevo, intentaron de todas las maneras que bajara a los cafés, ¡solo les costó 3 meses! De ellos, sobre todo quiero agradecer a Erica porque, entre los videos de gatos y otras fricadas, me ayudó muchísimo en el trabajo, con paciencia consiguió que aprendiera a programar y que entendiese mi trabajo (y a Miguel jeje). Por supuesto, también a Manu y a Antonia, Manu por sus “consejos” memorables (es ordenado hasta en el baño) y el gran aprendizaje con la cultureta general, y Antonia porque su alegría se extendía a todo el grupo mejorando así cualquier jornada dura, en ellos encontré dos buenos amigos. Uno de los mejores recuerdos me lo llevo de Taina, quien me ha ayudado en muchísimas cosas, sobre todo personales, hemos compartido vivencias, recetas (y calorías¬¬), cafés, y hasta he aprendido un idioma nuevo (taino). Dani, por otro lado, ha sido una fuente inagotable de sabiduría friki, condimentos y curiosidades raras, con el que he descubierto que la cara angulosa te hace más atractivo.

Después de mí, llegaron el resto de predocs Marta, Lourdes, Lino, Emilio, María, Elies, Laura, Ioanna y Dailu. Con Marta, aunque me riñera muchas veces por hablar o molestarle, he encontrado a una muy buena amiga con la que compartir muchos momentos y darnos consejos sobre gym y restaurantes pijos. Lourdes a ti gracias por tu tiempo, ¡es broma!, en Lourdes encontré a alguien muy parecida a mí (pero con muy mala leche, eso sí) que ha llegado a ser una gran amiga, nos gustan las mismas cosas, los cotilleos (junto con Taina), criticar y, solo de vez en cuando, ir a tomar alguna cervecilla. Emilio es el estudiante con alma de director, siempre intentando investigar cosas nuevas, ha sido apodado de varias

maneras: el presidente, el IP de animal, el futuro Riechmann, etc. A ti Emilio solo te puedo dar un consejo: “Aprende a medir la pasta, que menudos tanques te haces”. A Laura, la cantante del grupo, aunque solo haya venido a una o dos cosas de las que organizamos, agradecer su ayuda en los momentos que he necesitado. Elies, al principio dabas un poco de miedo por el carácter, pero bueno, al final se te coge hasta un poco de aprecio. En Ioanna he encontrado a alguien con quien mejorar mi inglés (o intentarlo) de forma muy amena, y sobre todo hablar de Juego de Tronos y todas las teorías posibles. María, siendo una de las últimas en llegar, ha sido de las que más se ha involucrado en todo, incluso en el grupo AllYouCanEat (nombrado en su honor), hemos compartido muchos momentos junto a Dani, Lino y Joan en el cine, la cueva, barbacoas, etc. Hablando de ese grupo no puede faltar el agradecimiento a Joan, mi gran mentor en la programación dura que, aunque de primeras parezca muy tímido de segundas también lo es, pero una vez coge confianza puedes hablar con él de todo, ya sea de trabajo, de cine o de fricadas, hasta que consigues que beba whisky que descubres que tú aun eres un niño al su lado. Por último, (pensabas que me olvidaba eh) el pessssado, Lino, que te voy a decir que no sepas ya, en ti he encontrado un gran amigo que sé que nunca me fallará, compañero de fiesta, cerveceo, barbacoa, charlas, cine y un sinfín de cosas más. Ahora te tocará buscarte a otro para que te ayude con la programación, aunque un Bioinformático como tú seguro que podrá hacerlo solo, además necesitarás un nuevo compañero para lo que tú y yo sabemos, y entre tú y yo, Emilio no sirve para eso. Dale también las gracias a Mónica por sus chipaguazus y sopas paraguayas, además de aguantarme más de un día molestando por ahí.

A Tania quiero darle las gracias por todo lo que me ha ayudado durante estos años con todos los papeleos evitándome muchos dolores de cabeza, y por supuesto, al resto de gente con las que he compartido estos años como son Betlem, Anna, Marcel, Álex y Josep María, entre otros.

Agradecer a mis padres por haberme ayudado a llegar hasta aquí por todos los medios posibles. Y, por último, muchísimas gracias a mis personas favoritas, Vanessa y Lucía, por aguantarme y apoyarme durante todo este tiempo, y sobre todo por permitirme pensar en todo menos trabajo al llegar a casa.