Genetics: Published Articles Ahead of Print, published on February 3, 2008 as 10.1534/genetics.107.084269

Selection in the making: A Worldwide Survey of Haplotypic Diversity around a Causative Mutation in Porcine *IGF2*

A. Ojeda[*], L.-S. Huang[§] J. Ren[§], A. Angiolillo[†], I.-C.Cho[**], H. Soto[§§] , C. Lemús-Flores[††],, S.M. Makuza[***], J.M. Folch[*], M. Pérez-Enciso[*,§§§]

[*] Departament de Ciència Animal i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain

[§] Key Laboratory for Animal Biotechnology of Jiangxi province and the Ministry of Agriculture of China, Jiangxi Agricultural University, Nanchang 330045, China.

[†] Dipartimento di Scienze e Tecnologie per l'ambiente e il territorio, Facoltà di Scienze Matematiche, Fisiche e Naturali, Università del Molise, 86090 Pesche, Italy

[**] Division of Livestock, National Institute of Subtropical Agriculture, R.D.A., 175-6, O-Deung Dong, Jeju, 690-150, South Korea**;**

[§§] Facultad de Ciencias Agroalimentarias, Universidad de Costa Rica, 2060 Ciudad Universitaria Rodrigo Facio. San Pedro; Costa Rica

[††] Universidad Autónoma de Nayarit, 63155 Tepic, Mexico;

[***] Department of Animal Science, University of Zimbabwe, MP167, Mount Pleasant, Harare, Zimbabwe*;*

[§§§] Institut Català de Recerca i Estudis Avançats (ICREA), Passeig de Grácia, 08010 Barcelona, Spain

Running title: Selection in the making at porcine IGF2

Keywords: Domestication, Genetic diversity, *IGF2*, Pig, Selection Footprint

Correspondence

Miguel Pérez-Enciso

email: miguel.perez@uab.es

Tel: +34 93 581 4225

Fax: +34 93 581 2110

Departament de Ciència Animal i dels Aliments,

Facultat de Veterinària,

Universitat Autònoma de Barcelona,

08193 Bellaterra, Spain

ABSTRACT

Domestic species allow us to study dramatic evolutionary changes at an accelerated rate due to the effectiveness of modern breeding techniques and the availability of breeds that have undergone distinct selection pressures. We present a worldwide survey of haplotype variability around a known causative mutation in porcine gene *IGF2* that increases lean content. We genotyped 34 SNPs spanning 27 kb in 237 domestic pigs and 162 wild boars. Although, the selective process had wiped out variability for at least 27 kb in the haplotypes carrying the mutation, there was no indication of an overall reduction in genetic variability of international *vs*. European local breeds. There was no evidence either of a reduction in variability caused by domestication. The haplotype structure and a plot of Tajima's *D* against the frequency of the causative mutation across breeds suggested a temporal pattern, where each breed corresponded to a different selective stage. This was observed comparing the haplotype NJ-trees of breeds that have undergone increasing selection pressures for leanness, *e.g.*, European local breeds *vs.* Pietrain. These results anticipate that comparing current domestic breeds will decisively help to recover the genetic history of domestication and contemporary selective processes.

- - - - - -

A major goal of current genetics research, in livestock as in plants or humans, is to identify the polymorphisms responsible for the variability in complex traits, i.e., traits affected by the environment as well as by more than one locus. This endeavor has proved to be difficult. In livestock, despite the large number of chromosome regions associated with phenotypes of economic interest (QTL), very few causative polymorphisms have been convincingly identified. The number of published QTL amount to hundreds in pigs (http://www.animalgenome.org/cgi-bin/QTLdb/SS/summary)(ROTHSCHILD *et al.* 2007), but less than ten causative mutations have been reported so far in this species. A comparable picture exists in all species. In order to accelerate causative gene discovery, traditional QTL studies are usually pursued with gene or genome-wide association studies. A complementary approach is to infer the action of selection at specific loci from their nucleotide variability, the so called selection footprint. To date, several works have shown the usefulness of this approach in humans and in other species, e.g., (CAICEDO *et al.* 2007; DUMONT and AQUADRO 2005; NIELSEN *et al.* 2005; WRIGHT *et al.* 2005). However, different genome wide scans have picked up different regions as affected by selection, see reviews for humans in (NIELSEN *et al.* 2007; THORNTON *et al.* 2007). Possibly, one of the reasons for conflicting results is that disentangling selective from purely demographic forces is very challenging. One of the additional difficulties is that the target of selection can not always be identified, even when the observed nucleotide variability pattern is not explained by demography alone and selection is the most plausible explanation.

Domestic plant and animal species offer underexploited genetic resources which are extremely valuable to disentangle demographic from selective processes. Modern breeding and artificial selection techniques allow us to study dramatic evolutionary changes at an accelerated rate. Domestic species have several advantages over natural or human species: the global phenotypic variability across breeds is often larger than in the wild species, the target of artificial selection is known and can differ between breeds or lines, their demographic history and origin are relatively well documented, the domestication process can be studied with unprecedented accuracy if the wild ancestor is available and, finally, they are easy to sample. So far, however, relatively little is known of livestock fine haplotype structure and of the effects of artificial selection on haplotype variability.

Here, we present the first worldwide study of haplotype variability in a porcine autosomal locus. In order to characterize the footprint of selection in a known selective target, we chose the *IGF2* region. This gene, located in a telomeric position on pig chromosome 2, harbors a paternally expressed mutation that increases muscle growth and leanness (VAN LAERE *et al.* 2003). The causative mutation (intron.3-g.3072G>A) occurs in a CpG island of intron 3 which has a regulatory role; pigs receiving the A allele from their sire have a three fold increase of *IGF2* mRNA in muscle. The mutation has a considerable effect, explaining about 10 - 30 % of the total phenotypic variability for these traits, and has been confirmed in several independent studies (ESTELLE *et al.* 2005; JUNGERIUS *et al.* 2004).

MATERIAL AND METHODS

Samples

A panel of 399 *Sus scrofa* animals comprising 237 domestic or feral pigs and 162 wild boars was genotyped, together with one bearded pig (*Sus barbatus*) and one babirusa (*Babyrousa babyrussa*) as outgroups. The domestic breeds pertained to 40 breeds from Europe, Asia (China, Korea, and Vietnam), the Americas (USA, Mexico, Costa Rica, Bolivia and Argentina) and Africa (Kenya and Zimbabwe) and there were wild boars sampled from Europe, North Africa and Asia (17 countries represented). The complete list of breeds and countries is in Table 1. More details about pig breeds are available in (PORTER 1993) or in the online resource http://www.ansi.okstate.edu/breeds/swine/.

We divided the panel into seven main groups: International breeds, Asian local breeds, European / USA local breeds, African local breeds, Creole and American feral breeds, Asian wild boar, and European / Maghreb wild boar (Table 1). Representatives of two hybrid terminal sire lines are also in the list (*Primus* and *Maximus*). International breeds are those that are primarily used worldwide, representing most of the genetics stock employed by multinational companies. Pietrain and Hampshire are among the leanest animals, followed by Large White, that are primarily used in terminal sire lines. Landrace is valued for its good reproductive and growth abilities, while Duroc is known for a better meat quality. Local Asian breeds were sampled primarily in China. Although local, it is well documented that Chinese pigs were imported into Europe during the 19th century and earlier and contributed significantly to some modern breeds like Large White (GIUFFRA *et al.* 2000; PORTER 1993). Non widely distributed European breeds are within the European local breeds label. They fall broadly in two groups, Mediterranean black pigs and British breeds. The former comprise Southern Italian breeds (Casertana, Sicilian, ...) and Iberian pigs, subdivided into different lines (*e.g.*, *Retinto* – red -, *Lampiño* – black hairless -); Portuguese *Porco Alemtejano* is closely related and genetic interchange between these breeds has occurred frequently during history. It is currently accepted that Mediterranean breeds have not been crossed to Asian animals (ALVES *et al.* 2003), in contrast to British pigs, which were actively crossed to Chinese pigs during the 18th and 19th centuries. The primary purpose was to increase prolificacy and, now paradoxically, fatness.

Polymorphisms genotyped

The SNPs were identified after aligning the 15 published sequences of ~ 28 kb length (VAN LAERE *et al.* 2003). Initially, the goal was to select uniformly spaced SNPs that were identified as tag-SNPs by Haploview (BARRETT *et al.* 2005) plus all coding SNPs. In practice, the SNPs chosen were conditioned by the genotyping platform software in order to genotype simultaneously as many SNPs as possible. Eventually, we were able to genotype 33 SNPs in two plexes using the MassARRAY SNP genotyping system (Sequenom Inc., San Diego, CA), following the manufacturer's instructions. The principles of this method are detailed elsewhere (BUETOW *et al.* 2001). The causative *IGF2* mutation was also genotyped using the pyrosequencing protocol as described (VAN LAERE *et al.* 2003) because it could not be genotyped with MassARRAY technology. The distance between the first and last SNPs was ~ 27 kb, and thus the average spacing was 790 bp, although the maximum gap between consecutive SNPs, 28 and 29, was 5 kb, while the minimum was 15 bp (Table 2).

Data analysis

Phases were reconstructed with Phase v2.1.1 (LI and STEPHENS 2003) using default options except that the program was run 5 times and the last iteration was 10 times longer, as suggested by the authors. We retained only those phases known with high probability (P > 0.8) for further analyses. Several parameters were estimated with DnaSP v4.10 (ROZAS *et al.* 2003): the mean number of pairwise differences across loci (π_N), Tajimas's *D*, Fu and Li's *D* using babirusa and *S. barbatus* as outgroups to distinguish between ancestral and derived alleles. These indices are reported only for those populations where more than six haplotypes were available. Tajima's D measures the discrepancy between the average polymorphism differences between haplotypes and the scaled number of segregating sites. Under the neutral null model, the expected values of the D statistics are zero. Directional selection causes negative D values while balancing selection, a positive value. Demographic events like admixture also result in positive D's. It should also be considered in interpreting Tajima's D here that an upwards bias is expected because SNP data and not full resequencing are employed (KELLEY *et al.* 2006). Coalescence simulations under the neutral model were also carried out with DnaSP in order to obtain the probability of number of haplotypes conditional on the observed number of segregating sites; 1000 replicates were run with either moderate linkage ($4Ne\ r = 10$) or completely linked sites. The ancestral allele could not be determined for four out of the 34 SNPs and thus some information is lost when applying the Fu and Li's *D* test (Table 1). NJ phylogenetic trees with the p-distance (percentage of differences) were drawn using MEGA3 (KUMAR *et al.* 2004), with standard errors obtained from 1000 bootstrap replicates. The population-scaled recombination rate ($\rho = 4\ Ne\ r$) was estimated using the Hudson's composite likelihood method implemented in LDhat (http://www.stats.ox.ac.uk/~mcvean/LDhat/ (MCVEAN *et al.* 2002)), which assumes a finite-sites mutation model. This we did separately for European wild boar, Asian wild boar, local European breeds, Asian local breeds and International breeds. The Haploview v3.32 program (BARRETT *et al.* 2005) was used to compute disequilibrium measures (r^2 and *D'*) and to identify haplotype blocks.

RESULTS

SNPs genotyped as proxies for complete variability

We firstly investigated how representative were the 34 SNPs of the true relationship between haplotypes (VAN LAERE *et al.* 2003). To study that, we compared the NJ-trees obtained from the 34 SNPs (Figure 1a,b) and the complete sequences published by Van

Laere *et al.* (Figure 1c). Except for sequence AY242110, which pertains to a recombinant Hampshire haplotype that carried the causative mutation, the two topologies were identical. Figure 1 also shows the main phylogenetic clades, **E**, **C**, **J**, **A**, and **M**. The rationale for this arrangement is discussed below (section haplotype phylogenies).

Causative allele frequency

The causative mutation was segregating in 13 breeds and absent in 23 breeds. A noticeable trend emerges simply by comparing the frequency of the selected allele between international breeds (P_A = 0.86) and Asian (P_A = 0.06) or European (P_A = 0.03) local breeds; the mutation was not found in wild boars. We defer the analysis of African and American populations because they are derived populations and its history is more complicated. This pattern suggests that the mutation is recent (after domestication) but that, nonetheless, has spread out across many breeds around the globe. Due to modern selection emphasis on lean content and the introgression of Asian genes into European breeds, its frequency has dramatically increased in international breeds. We found that the derived allele was fixed in two breeds (Hampshire and Pietrain). All these animals shared an identical haplotype, the same found by Van Laere et al. (Figure 1a)(VAN LAERE *et al.* 2003). In Duroc, 89 out of 90 haplotypes were identical, the only heterozygous animal was a Spanish Duroc that differed in 11 positions.

In agreement with previous results (YANG *et al.* 2006), we found the mutation in the Licha Black breed (Shandong province, China). Interestingly, this is one of the leanest breeds from China. (YANG *et al.* 2006) also reported the mutation in the Erhualian breed in the nearby province of Jiangsu but at a lower frequency (5%). We report for the first time the presence of the mutation in the Korean native pig at a rather high frequency (25%) but not in any Asian wild boar. Although the Asian wild boars have not been extensively sampled, these findings, together with the fact that the *A* allele is present in a single haplotype, would suggest that the mutation occurred in Eastern Asia after domestication and has a unique origin. To resolve definitely this question, however, more SNPs on SSC2 and a larger number of Asian samples should be genotyped.

The mutation in Mediterranean local populations was very rare. The *A* allele was absent in the Iberian breed except in Andalusian spotted (*Manchado de Jabugo*), which is a synthetic strain made up of crossing purebred Iberian to Berkshire and Large White (GARCÍA DORY *et al.* 1990). It is also the only line that harbored Asian haplotypes (see below). The *A* allele was also present in Mukota's pig (a Zimbabwean local breed with influence from European and Asian lines), Argentinean feral pigs, Mexican hairless pigs (*Pelón*) and Costa Rican creole pigs.

Nucleotide variability

The number of segregating sites (*S*) and the mean number of pairwise differences across loci (π_N), Tajimas's *D*, and Fu and Li's *D* are in Table 1. Overall, genetic variability was much larger in Asian than European local breeds, both for domestic pigs and wild boar. This is in agreement with analysis of mtDNA that uncovered a bottleneck / expansion demographic process that was stronger in European than Asian pig populations (FANG and ANDERSSON 2006; LARSON *et al.* 2005). But a relevant observation was that, in contrast to what has been currently observed in other species (e.g., in maize, (WRIGHT *et al.* 2005), domestication has not produced a detectable decrease in variability. In fact, the lowest variability was found in European wild boar, in agreement with results at FABP4 gene (OJEDA *et al.* 2006), while some of the most variable breeds are the endangered British Tamworth breed or local Zimbabwean Mukota.

In parallel, selection has wiped out variability almost completely for this region in some breeds, like Pietrain, Hampshire or Duroc, a clear signal of selective sweep. These breeds are typically selected for growth and leanness and are used as sire lines. A single haplotype was found in 34 Pietrain sequences; similarly, all haplotypes except one were identical in 90 Duroc animals. Coalescent simulations showed that it is highly unlikely ($P < 10^{-6}$) to get a single haplotype conditionally on the observed number of segregating sites (S = 11 in Duroc), and thus the neutral model can be rejected. Nucleotide variabilities were comparable in the two main porcine breeds, Landrace and Large White.

We found a large variability in Tajima's or Fu-Li's *D* (Table 1), ranging from highly positive (e.g. Calabrese, Korean wild boar) to strong negative values (Japanese wild boar, Duroc). Thus, in addition to the effects of selection caused by the causative mutation, there must be strong demographic forces affecting nucleotide variability. A genomewide analysis is required to disentangle the two phenomena, though. In order to elucidate the effect of directional selection for the causative mutation, we plotted the frequency of the selected allele against Tajima's *D* (Figure 2). The observed pattern was illuminating. Before the appearance of the mutation ($P_A = 0$), Tajima's *D* was highly variable, likely the result of demographic and / or sampling effects. We assumed that the only non-neutral polymorphism is the intron 3 causative allele. In stark contrast, as the frequency of the *A* allele increases, Tajima's *D* first increases and declines very rapidly after P_A > 0.5. A similar, although more ragged pattern was observed with Fu-Li's *D*. Although the behavior of Tajima's D with genotypic data requires further study specially its dynamics over the selection process, several authors (JENSEN *et al.* 2005; KELLEY *et al.* 2006) report that this statistic is biased upwards and can even take positive values under directional selection and with partial selective sweeps, specially with population structure. Yet in breeds where P_A was close to 1, i.e., when the mutation was almost fixed, Tajima's D taked highly negative values as expected in a clasical selective sweep.

Haplotype phylogenies

A broad view of the porcine genetic landscape before modern selection in China and the Mediterranean (Iberian and Italian Peninsulas) can be seen through the NJ-trees in Figures 3a,b and S1, respectively. In the light of these trees, we selected the five most divergent and frequent haplotypes as clade representatives: '**C**' for causative, is the typical haplotype of Pietrain that carries the derived allele; '**E**' for European, is at high frequencies in European wild boar and local European populations but also found in Asian populations; '**J**' for Japanese was described in a Japanese wild boar (VAN LAERE *et al.* 2003) but also present in Continental Asian populations. Two additional clades were not described previously in (VAN LAERE *et al.* 2003): '**A**' for Asian and '**M**' for Mediterranean, the latter was frequent in European local populations and absent from Asian pigs. All the haplotypes and the NJ-trees are in Figure 1a,b and outlined by colored rectangles. Note that the clade names are rather conventional because Asian populations harbored all haplotypes except **M**, whereas the **M** haplotype was at high frequency in Mediterranean populations abut present also in other populations. All these haplotypes, marked with colored rectangles, are included in the NJ-trees in order to facilitate visual comparisons between populations.

As expected from the fact that the species *S. scrofa* evolved initially in Eastern Asia, this region has maintained higher levels of variability than the European subspecies, as can be intuitively seen by deeper clades at intermediate frequencies and corroborated by higher π_N in Asian *vs.* Mediterranean pigs (Table 1). Chinese pigs harbored all main clades, except **M**, at intermediate frequencies (Figure 3a).

As mentioned, one of the large advantages of studying domestic breeds over natural populations is that different breeds / lines may represent different stages of the domestication and artificial selection processes. Moreover, because the target of selection is often well documented, interpretation of the results is also far easier than in natural populations. This temporal pattern can be appreciated by comparing the NJ-trees of British local breeds, Landrace and Large White (Figures 3c,d,f). The set of local British / US breeds would represent the stage just before modern selection for leanness but after introgression of Asian germplasm during the 19th century. They follow a similar pattern to that of Mediterranean local breeds although with some key differences: the **M** haplotype is absent and a more pronounced Asian influence is observed (Figure 3c). The global frequency of the *A* allele (clade **C**) was very low, it was at high frequency only in Chester White, a breed related to Large White. Next, the effect of modern selection can be observed by comparing the clinal NJ-trees of Landrace (Figure 3d), Large White (Figure 3f) and Duroc. The pattern of the Landrace breed is an intermediate stage where the **C** clade represents about 50% of haplotypes, the fact that this clade is rather divergent from the pre-existing clades (clade **E** predominantly) makes it Tajima's *D* become positive (D_T = 1.8). In Large White's NJ-tree (Figure 3f), clade **C** was already predominant at the expense of clade **E** and results in a negative Tajima's *D* overall. Note, however, that the frequency of clade **C** is intermediate (0.5) in the Yorkshire breed, a 'primitive' US Large White representative, Tajima's *D* was positive here. The next most extreme case was the Duroc breed (P_A = 0.99 and D_T = -2.35), i.e., just before fixation. This pattern was not caused by demography alone because we found a highly positive D_T in a subset of this Duroc population for fatty acid binding protein 5 (*FABP5*, Ojeda et al., unpublished). The selective sweep is accomplished in Pietrain and Hampshire, where clade **C** is fixed and genetic variation is removed for at least ~ 27 kb.

There was no haplotype specific to a particular breed, each breed consisted of a mosaic of different haplotypes shared across breeds. This was observed throughout breeds and populations. It suggests that breed effective sizes are larger than what could be suspected a priori. For instance, the Iberian pig is considered as a single breed with different lines (*Lampiño* − hairless −, *Retinto* − red −, *Torbiscal*, *Guadyerbas* ...) that tend to be bred separately. But even highly inbred and isolated lines like the Iberian *Guadyerbas* (inbreeding coefficient > 0.3, (TORO *et al*. 2000) harbored haplotypes in all main clades of the Mediterranean breeds. The Italian NJ-tree was again very similar, and there was no correlation between clade and breed. The two most extreme cases were the British endangered breed Tamworth and Zimbabwe's Mukota. Tamworth is included in the British Rare Breeds Survival Trust protection program (http://www.rbst.org.uk/). Yet, it was one of the most diverse breeds (Table 1), with highly distant haplotypes in clades **E**, **J** and **A** (Figure 3c). The six Mukota haplotypes pertained to three clades, **A**, **C** and **E**, confirming that Mukota has an important Asian influence.

A variety of situations were found in the derived American and African creole / feral populations (Figure S1). We did not find the mutation in Bolivian pigs, collected in the Santa Cruz valley. Costa Rican animals were clearly influenced by modern breeds, as inferred from the high frequency of the causative allele, but also from Mediterranean pigs. Mexican hairless had some Asian influence, evidenced by the presence of the **A** haplotype. This result is in agreement with the report of Asian mtDNA haplotypes for Mexican hairless (LARSON *et al.* 2005). The mutation was segregating even in Argentinean feral pigs. Thus, the genetic structure of derived American populations is complex. No clear pattern emerges and more extensive sampling is required to reconcile history with phylogeny in these populations.

Linkage disequilibrium

Five haplotype blocks were inferred from the algorithm implemented in Haploview (BARRETT *et al.* 2005), spanning 1, 2, 9, 4 and 0.8 kb, respectively (Figure 4a). The third block was the largest and contained the causative mutation. According to the tagger option, 17 SNPs would tag all 34 SNPs with r^2 = 100%, while 10 SNPs would capture all markers with $r^2 > 0.8$. The most associated SNPs with the causative mutation were 14 and 15, r^2 = 0.91 and 0.95, respectively. The linkage disequilibrium pattern looks overall quite complex. In order to simplify the data set, we also analyzed the Landrace and Large White breeds separately, as selecting for the causative mutation in these breeds is more relevant than in local breeds (Figure 4b). As expected, the disequilibrium pattern was simpler, with a single block that nevertheless spanned 14 kb, *i.e.*, only about half of the whole region analyzed. The causative mutation was at high disequilibrium again with SNPs 14 and 15 but also with the contiguous genotyped SNP 22. In summary, the causative SNP was in high disequilibrium with very few of the SNPs analyzed and thus marker assisted selection can be implemented effectively only because the causative SNP has been discovered. There was not relation between physical distance and any disequilibrium measure; this occurred between all pairs of markers (Figure 4c) and in particular between the causative mutation and the rest of SNPs (Figure 4d). Note that the behavior of r^2 was completely different from that of *D'*, as these statistics quantify different aspects of linkage disequilibrium (ARDLIE *et al.* 2002). The index *D'* was one for a large percentage of pairs, denoting the absence of recombination, while r^2 values tended to cluster around zero because the allele frequencies at the SNP pairs were rather different (Figure 4c,d). Although a flat line is the pattern expected if all markers are in the same haplotype block, the picture here was more complex, as many pairs 'escaped' complete disequilibrium. But even for these pairs there was no observable trend between distance and disequilibrium.

DISCUSSION

Selection in the making

Artificial selection in livestock species allows us to study dramatic genetic changes in the making. Different evolutionary stages can be scrutinized by comparing breeds that have undergone very different selection pressures. Knowing that SNP 23 (intron.3-g.3072G>A) is a causative, selected mutation has clearly facilitated the interpretation of the observed pattern of nucleotide and haplotype diversity. Certainly, an ongoing challenge is how to infer selection from the pattern of linkage disequilibrium and nucleotide variability alone, when the target of selection is not known. The fact that artificial selection is much more intense and effective than natural selection, together with the presence of highly structured populations and the availability of the wild ancestor in the pig, should make this task easier than in human or natural populations. Our work provides data that can be used to validate theoretical models for selective sweeps in structured populations.

The criteria of selection in international breeds vary but are primarily leanness, growth and, to a lesser extent, reproductive traits. For most of the remaining breeds, no modern selection and breeding schemes have been set up and thus they tend to be much fatter and of higher meat quality than international breeds. The overall frequencies of the causative allele are in complete agreement with the expectations: high frequencies in international lean breeds and very low frequency in local breeds. The presence of the mutation in Iberian Andalusian spotted is explained by the well known fact that it was created by crossing Iberian to British breeds (ALVES *et al.* 2003; GARCÍA DORY *et al.* 1990); the presence of the derived allele in other local European breeds (Mangalitza and Casertana) is also due to introgression in all likelihood. Some authors (PORTER 1993) have suggested a small influence of Asian breeds in Casertana. But neither in Casertana nor in Mangalitza have Asian mtDNA haplotypes been reported (ALVES *et al.* 2003; ANGIOLILLO *et al.* 2001; LARSON *et al.* 2005), so this would suggest that the introgression of the mutant allele was male mediated. Mexican hairless is thought to be descended from Iberian animals brought by the Spaniards (LEMUS-FLORES *et al.* 2001), but mtDNA analysis (LARSON *et al.* 2005) and the presence of the mutation proofs that interbreeding with Asian germplasm has also occurred. Similarly, the influence of Asian influence in Zimbabwe's Mukota (Figure 3e) is in agreement with mtDNA results (RAMÍREZ *et al.* 2006).

The *D* statistics computed here (Table 1) were obtained with genotypes rather than the complete set of polymorphisms, and thus there is an ascertainment bias that causes an upward bias in Tajimas' *D* (KELLEY *et al.* 2006). Nevertheless, we also expect a positive correlation between sequence and very dense genotyping *D*'s (CARLSON *et al.* 2005; KELLEY *et al.* 2006), which allows us to compare different populations. Directional selection causes an abundance of the haplotype carrying the selected mutation and a relative excess of rare variants that results in a negative Tajima's D. But this is the final stage of a selective sweep. When a mutation occurs and increases in frequency, there will be a moment when the selected haplotype will be at intermediate frequency. At this moment, Tajima's D becomes positive due to an apparent excess of alleles at intermediate frequencies. The pattern observed in Figure 2 is precisely the expected pattern when a selected haplotype replaces the existing ones. The comparison of local British breeds, Landrace, Large White NJ-trees is particularly illuminating. The mutation was at much higher frequency in Large White (P_A = 0.78) than Landrace (P_A = 0.55). Probably, the reason why the mutation has not become fixed in Landrace is that this breed is used in both paternal and maternal lines, and the mutation has detrimental effects in prolificacy because an excess of leanness diminishes reproductive performance in the sow (BUYS *et al.* 2006). In contrast, Large White is mostly used in sire lines in the European market, where the primary interest is to increase growth and leanness. Thus, the indirect selection pressure has been higher in Large White than Landrace. Interestingly, Large White and Landrace are both used as maternal lines in China, and the allele frequencies were similar (YANG *et al.* 2006). Certainly, selective sweeps occur instantaneously (in the evolutionary scale) in natural populations and only the final stage is observed, i.e., here the Pietrain, Hampshire or Duroc populations. The study of a diverse collection of breeds and populations allowed us to carry out a spatial study that mimicked a longitudinal (temporal) process.

Hard or soft sweeps?

There is currently much interest in understanding the effects of directional selection on standing genetic variation, the so called 'soft' sweeps (HERMISSON and PENNINGS 2005; INNAN and KIM 2004; PRZEWORSKI *et al.* 2005; TESHIMA *et al.* 2006). This scenario is particularly relevant here because the selective advantage of a given allele can dramatically change after a domestication process. Simulation results have shown (INNAN and KIM 2004; PRZEWORSKI *et al.* 2005) that the reduction in nucleotide diversity around the selected site can be much smaller than anticipated when the selected allele is already segregating in the population. The reduction will be minimal when the allele is at intermediate frequencies and undistinguishable from neutral variation. Our results and those of (YANG *et al.* 2006) demonstrate that the mutation was segregating in East Asian populations before being selected due to modern emphasis in lean meat, *i.e.*, we would be expecting a soft sweep behavior and a mild reduction in nucleotide diversity. However, the nucleotide pattern is that of a hard sweep, and genetic variability was wiped out for at least 27 kb. The most likely explanation is that the selection process was associated with a strong bottleneck, as probably very few copies of the allele were introgressed in European populations. More extensive sampling of Asian haplotypes harboring the mutation should be carried out to estimate the age of the allele, but the causative mutation seems to be quite recent, after domestication. All in all, the demographic model of modern pig breeding is far more complex than the single population model studied so far in domestication (INNAN and KIM 2004; PRZEWORSKI *et al.* 2005).

The classical hitchhiking effect predicts that a selective sweep will reduce genetic variability around the selected target. One of the striking observations from our study is that the footprint of this phenomenon is clearly visible across different international breeds which, in principle, behave as semi - isolated distinct populations. The case of Pietrain and Duroc are particularly noticeable: a single and identical haplotype spanning at least 27 kb was found in all 17 Pietrain animals sampled from Spain, Germany, France and UK and pertaining to several pig breeding companies. As for Duroc, all 90 haplotypes except one (a Spanish Duroc differing in 11 SNPs) were identical at all positions except for the last SNP. Duroc was sampled from Spain, USA, Mexico, Denmark, UK, Hungary and France. Coalescent simulations conditional on the number of segregating sites suggest that it is virtually impossible to get just two haplotypes with $S = 11$ and $n = 90$ in a neutral model. These results are in stark contrast with results at FABP5 on chromosome 4, where there seems to be evidence of balancing selection in Duroc (Ojeda et al. and unpublished results). At the very least in Duroc, the lack of genetic diversity cannot be explained by bottlenecks alone. The picture is more complex in the two most popular international breeds: Large and Landrace. While British, French and Finnish Large White haplotypes were again identical to that in Pietrain, Spanish Landraces were far more variable, the frequency of the derived allele was only 36% and no extreme Tajima's *D* showed up.

Domestication and modern breeding are complex processes

We found no apparent reduction in nucleotide variability after domestication, as can be seen in the diversity indices of Table 1. To contrast the diversity indices π_N, we also estimated the parameter $\rho = 4Ne\ r$, for several populations separately: International breeds ($\hat{\rho}$ = 8.2), Mediterranean local breeds (Italian and Iberian Peninsulas, $\hat{\rho}$ = 8.2), British local breeds ($\hat{\rho}$ = 4.1), European wild boar ($\hat{\rho}$ = 6.1), Asian local breeds ($\hat{\rho}$ = 13.3) and Asian wild boar ($\hat{\rho}$ = 11.2). Although these estimates are subject to large sampling errors, they do suggest that: 1) domestication has not decreased genetic variability in the porcine species, at least for this region, and in agreement with the opinion that domestication – at least in animals – was a complex process and cannot be fully explained by a simple bottleneck (FANG and ANDERSSON 2006; LARSON *et al.* 2005; VILA *et al.* 2005; WONG *et al.* 2004); and 2) despite intense selection in international breeds targeting the *IGF2* region, their effective sizes still seem comparable to local unselected breeds. This is likely the result of two balancing forces: the introgression of Asian genes vs. the selective sweep process. Interestingly, we also reported a very low nucleotide diversity in European wild boar for FABP4 gene as compared to domestic breeds (OJEDA *et al.* 2006). The low nucleotide variability of European wild boar has been also reported for mtDNA in several studies and is consistent with a bottleneck followed by recent expansion that occurred prior to domestication (FANG and ANDERSSON 2006; LARSON *et al.* 2005).

Given that the mutation seems to have a unique origin, it is remarkable how widespread was it across breeds and continents. This proofs that porcine breeds have regularly interchanged genetic material. This regular interchange would also explain the high heterozygosity observed within breeds, which consist of a mosaic of distinct haplotypes at different frequencies (Figure 3, S1). Thus, it is very rare that a particular haplotype is specific of a single breed, however isolated is thought to be. Symmetrically, even highly inbred lines were made up of several clades. The extreme case is the endangered British breed Tamworth: we found it to be one of the most variable breeds, despite the fact of being listed among vulnerable breeds by the Rare Breed Survival Trust (http://www.rbst.org.uk/watch-list/pigs.php), it held haplotypes in clades **J**, **A** and **E**. It is also noteworthy that Large White pig LWES0429 was heterozygous at 9 positions, harboring the two specific Asian haplotypes **A** and **J** (Figure 3f). This individual pertained to a highly inbred, very fat and primitive Large White line imported to Spain in 1931 and kept in a closed herd until 1992, when they were slaughtered (RODRIGÁÑEZ *et al.* 1998). It was homozygous for the wild type causative allele, though. The Iberian line *Guadyerbas* has an average inbred coefficient of ~ 30% and has been isolated for over 50 years now (TORO *et al.* 2000), yet it also harbored haplotypes in all main clades **M** and **E** (pink squares in Figure 3b). All in all, the genetic difference between porcine breeds is very tenuous, and contrasts clearly with other species, e.g., the dog, where breed barriers seem to have been enforced more effectively.

In conclusion, we show that: 1) selection can be observed and analyzed in the making by comparing different breeds that represent distinct stages of the selective process; 2) there is no evidence that, overall, domestication reduced genetic variability in the IGF2 region with respect to current wild ancestor in the pig (although a complete selective sweep is found in some very lean breeds like, e.g., Pietrain) ; and 3) there seems to be considerable gene flow between porcine breeds, with the result of common haplotypes shared across breeds and few (if any) specific haplotypes of a single breed.

Thanks to J. Rozas, S. Ramos-Onsins and J. vand der Made for comments and discussions. We thank the following persons and institutions for samples: L. Silió and M.C. Valdovinos (*Instituto Nacional de Investigaciones Agrarias*, Spain), E. Martínez (Madrid's zoo), N. Okumura (STAFF Institute Japan), G. Rohrer (USDA), Z. Bosze (Agricultural Biotechnology Center, Hungary), M. Cumbreras (*Diputación de Huelva*, Spain), E. von Eckardt (Sweden), M.

Kovac (Slovenia), J. Reixach (Batallé, Spain), O. Hanotte (International Livestock Research Institute, Kenya), J. García-Cascos (*Censyra*, Spain), G. Caja (*Universitat Autònoma Barcelona*, Spain), E. González (Seporsa, Spain), P. Martínez (*Universidad de Santiago de Compostela*, Spain), D. Vidal and C. Gortázar (*Instituto de Investigación en Recursos Cinegéticos*, Spain), S. Casu (*Istituto Zootecnico e Caseario per la Sardegna*, Italy), M. Lukaszewicz (Academy of Sciences, Poland), C. Renard (Institut National Recherche Agronomique, France), E. Grindflek (NORSVIN, Norway), M.A. Revidatti (*Universidad Nacional del Nordeste*, Argentina), A. Clop (Genesis Faraday, GB), R. Silva (Portugal), J.W. Ma (Jiangxi Agricultural University, China), *Instituto de Capacitación del Oriente* (Bolivia), *Veterinarios Sin Fronteras* (Spain), *Junta de Castilla – La Mancha* (Spain), London's zoo, *Copaga* (Spain), *Degesa* (Spain), *Semen Porcino Andalucía* (Spain), Kaliningrad's zoo (Russia). AO is recipient of a PhD grant (*Ministerio de Educación y Ciencia*, Spain). Work funded by projects AGL2004-0103, AECI/5267/06 (Spain), and National Natural Science Foundation of China (grant 30425045). MassARRAY genotyping was subsidized by *Fundación Genoma España* and carried out at the *Cegen* (www.cegen.org) facilities.

LITERATURE CITED

ALVES, E., C. OVILO, M. C. RODRIGUEZ and L. SILIO, 2003 Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. Anim Genet **34:** 319-324.

ANGIOLILLO, A., F. PILLA, D. MATASSINO, A. CLOP and A. SÁNCHEZ, 2001 Genetic characterization of Italian local breeds of pigs by mtDNA analysis, pp. in *EAAP meeting*, Budapest.

ARDLIE, K. G., L. KRUGLYAK and M. SEIELSTAD, 2002 Patterns of linkage disequilibrium in the human genome. Nat Rev Genet **3:** 299-309.

BARRETT, J. C., B. FRY, J. MALLER and M. J. DALY, 2005 Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21:** 263-265.

BUETOW, K. H., M. EDMONSON, R. MACDONALD, R. CLIFFORD, P. YIP *et al.*, 2001 High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc Natl Acad Sci U S A **98:** 581-584.

BUYS, N., G. VAN DEN ABEELE, A. STINCKENS, J. DELEY and M. GEORGES, 2006 Effect of the IGF2-intron3-G3072A mutation on prolificacy on sows, pp. in *World Cong Genet Appl Livest Prod*, Belo Horizonte, Brazil.

CAICEDO, A. L., S. H. WILLIAMSON, R. D. HERNANDEZ, A. BOYKO, A. FLEDEL-ALON *et al.*, 2007 Genome-Wide Patterns of Nucleotide Polymorphism in Domesticated Rice. PLoS Genetics **3:** e163.

CARLSON, C. S., D. J. THOMAS, M. A. EBERLE, J. E. SWANSON, R. J. LIVINGSTON *et al.*, 2005 Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res. **15:** 1553-1565.

DUMONT, V. B., and C. F. AQUADRO, 2005 Multiple Signatures of Positive Selection Downstream of Notch on the X Chromosome in Drosophila melanogaster. Genetics **171:** 639-653.

ESTELLE, J., A. MERCADE, J. L. NOGUERA, M. PEREZ-ENCISO, C. OVILO *et al.*, 2005 Effect of the porcine IGF2-intron3-G3072A substitution in an outbred Large White population and in an Iberian x Landrace cross. J Anim Sci **83:** 2723-2728.

FANG, M., and L. ANDERSSON, 2006 Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. Proc Biol Sci **273:** 1803-1810.

GARCÍA DORY, M. A., S. MARTÍNEZ and F. OROZCO, 1990 *Guía de Campo de las Razas Autóctonas Españolas*. Alianza Editorial, Madrid.

GIUFFRA, E., J. M. KIJAS, V. AMARGER, O. CARLBORG, J. T. JEON *et al.*, 2000 The origin of the domestic pig: independent domestication and subsequent introgression. Genetics **154:** 1785-1791.

HERMISSON, J., and P. S. PENNINGS, 2005 Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics **169:** 2335-2352.

INNAN, H., and Y. KIM, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci U S A **101:** 10667-10672.

JENSEN, J. D., Y. KIM, V. B. DUMONT, C. F. AQUADRO and C. D. BUSTAMANTE, 2005 Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. Genetics **170:** 1401-1410.

JUNGERIUS, B. J., A. S. VAN LAERE, M. F. TE PAS, B. A. VAN OOST, L. ANDERSSON *et al.*, 2004 The IGF2-intron3-G3072A substitution explains a major imprinted QTL effect on backfat thickness in a Meishan x European white pig intercross. Genet Res **84:** 95-101.

KELLEY, J. L., J. MADEOY, J. C. CALHOUN, W. SWANSON and J. M. AKEY, 2006 Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. **16:** 980-989.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5:** 150-163.

LARSON, G., K. DOBNEY, U. ALBARELLA, M. FANG, E. MATISOO-SMITH *et al.*, 2005 Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. Science **307:** 1618-1621.

LEMUS-FLORES, C., R. ULLOA-ARVIZU, M. RAMOS-KURI, F. J. ESTRADA and R. A. ALONSO, 2001 Genetic analysis of Mexican hairless pig populations. J Anim Sci **79:** 3021-3026.

LI, N., and M. STEPHENS, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213-2233.

MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231-1241.

NIELSEN, R., C. BUSTAMANTE, A. G. CLARK, S. GLANOWSKI, T. B. SACKTON *et al.*, 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol **3:** e170.

NIELSEN, R., I. HELLMANN, M. HUBISZ, C. BUSTAMANTE and A. G. CLARK, 2007 Recent and ongoing selection in the human genome. Nat Rev Genet **8:** 857-868.

OJEDA, A., J. ROZAS, J. M. FOLCH and M. PEREZ-ENCISO, 2006 Unexpected High Polymorphism at the FABP4 Gene Unveils a Complex History for Pig Populations. Genetics **174:** 2119-2127.

PORTER, V., 1993 *Pigs: A handbook to the breeds of the world*. Helm Information Ltd., Mountfield (East Sussex, UK).

PRZEWORSKI, M., G. COOP and J. D. WALL, 2005 The signature of positive selection on standing genetic variation. Evolution Int J Org Evolution **59:** 2312-2323.

RAMÍREZ, O., A. TOMÁS, A. CLOP, O. GALMANOMITOGUN, S. M. MAKUZA *et al.*, 2006 Microsatellite and chromosome Y sequence analysis of wild boar and autochthonous pig breeds from Asia, Europe, South America and Africa, pp. in *ISAG Meeting*, Porto Seguro, Brazil.

RODRIGÁÑEZ, J., M. TORO, M. RODRÍGUEZ and L. SILIÓ, 1998 Effect of founder allele aurvivaland inbreeding depression on litter size in a closed line of Large White pigs. Anim Sci **67:** 573-582.

ROTHSCHILD, M. F., Z. L. HU and Z. JIANG, 2007 Advances in QTL mapping in pigs. Int J Biol Sci **3:** 192-197.

ROZAS, J., J. C. SANCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496-2497.

TESHIMA, K. M., G. COOP and M. PRZEWORSKI, 2006 How reliable are empirical genomic scans for selective sweeps? Genome Res **16:** 702-712.

THORNTON, K. R., J. D. JENSEN, C. BECQUET and P. ANDOLFATTO, 2007 Progress and prospects in mapping recent selection in the genome. Heredity **98:** 340-348.

TORO, M., J. RODRIGÁÑEZ, L. SILIÓ and M. RODRÍGUEZ, 2000 Genealogical analysis of a closed herd of black hairless Iberian pigs. Conservation Biol **14:** 1843-1851.

VAN LAERE, A. S., M. NGUYEN, M. BRAUNSCHWEIG, C. NEZER, C. COLLETTE *et al.*, 2003 A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature **425:** 832-836.

VILA, C., J. SEDDON and H. ELLEGREN, 2005 Genes of domestic mammals augmented by backcrossing with wild ancestors. Trends Genet **21:** 214-218.

WONG, G. K., B. LIU, J. WANG, Y. ZHANG, X. YANG *et al.*, 2004 A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature **432:** 717-722.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The Effects of Artificial Selection on the Maize Genome. Science **308:** 1310-1314.

YANG, G. C., J. REN, Y. M. GUO, N. S. DING, C. Y. CHEN *et al.*, 2006 Genetic evidence for the origin of an IGF2 quantitative trait nucleotide in Chinese pigs. Anim Genet **37:** 179-180.

Table 1: Main statistics for the populations analyzed.

Breed (symbol)	Countries	N	P_A	S	π_N	D_T	D_{FL}
INTERNATIONAL		218	0.86	31	0.07	-1.28	0.65
Duroc (DU)	DK, ES, FI, GB, US	90	0.99	11	0.01	-2.35	-1.16
Hampshire (HS)	US	10	1.00	1	0.01	-1.11	-1.35
Landrace (LD)	DE, ES, FI, FR, GB	36	0.55	10	0.15	1.80	1.43
LD Spain	ES	14	0.21	11	0.14	0.55	1.57
LD Great Britain	GB	12	1.00	0	-	-	-
LD USA	US	6	0.00	10	0.19	0.94	0.87
Large White (LW)	ES, FI, FR, GB	32	0.78	23	0.15	-0.81	1.08
LW Spain	ES	14	0.36	23	0.31	0.85	1.57
LW Great Britain	GB	14	1.00	0	0.00	-	-
Yorkshire (YK)	US	12	0.50	15	0.23	1.30	1.24
LW + YK	ES, FI, FR, GB, US	44	0.70	21	0.17	-0.07	1.34
Pietrain (PI)	DE, ES, FR	34	1.00	0	0.00	-	-
Synthetic sire lines	GB	4	1.00	0	0.00	-	-
LOCAL ASIA		106	0.06	34	0.32	1.64	1.78
Fengjing (FG)	CN	4	0.00	10	0.20	-	-
Huai (HU)	CN	10	0.00	18	0.29	1.58	1.50
Jiaxin Black (JB)	CN	12	0.00	11	0.11	-0.35	1.56
Jinhua (JH)	CN	8	0.00	12	0.14	-1.15	-0.89
Licha Black (LI)	CN	12	0.25	27	0.34	0.58	0.46
Luchuan (LU)	CN	10	0.00	7	0.10	1.16	1.49
Minzhu (MI)	MI	14	0.00	22	0.35	1.58	1.81

Meishan (MS)	CN (FR, US)	18	0.00	21	0.24	0.21	-0.15
Native pig (NP)	KR	12	0.25	11	0.17	1.15	1.6
I potbelly (VT)	VT (ES, VT)	6	0.00	7	0.10	1.27	1.63
LOCAL EUROPE / USA		244	0.03	32	0.12	-0.69	1.77
Iberian (IB)	ES	76	0.05	17	0.13	0.10	1.61
Iberian excluding Andalusian spotted	ES	70	0.01	13	0.09	0.39	1.53
Porco Alemtejano (PA)	PT	12	0.00	5	0.08	1.24	1.32
Casertana (CA)	IT	20	0.10	18	0.17	-0.08	1.29
Calabrese (CL)	IT	20	0.00	7	0.12	2.31	1.38
Cinta Sienese (CS)	IT	20	0.00	2	0.07	1.44	1.13
Sicilian (SI)	IT	18	0.00	5	0.07	1.14	1.24
Corsica (CO)	FR	4	0.00	4	0.07	-	-
Mangalitza (MG)	HU	18	0.06	14	0.12	-0.52	-0.10
Black Eslavonian (HR)	HR	6	0.00	0	0.00	-	-
Berkshire (BK)	US, GB	14	0.00	25	0.20	-1.34	-0.71
BK USA	US	10	0.00	25	0.25	-1.08	-0.43
British lop (BL)	GB	2	0.00	4	0.00	-	-
Large Black (LB)	GB	6	0.00	0	0.00	-	-
Middle White (MW)	GB	4	0.00	4	0.06	-	-
Old Spot (OS)	GB	6	0.00	6	0.09	-0.06	1.01
Saddle Back (SB)	GB	6	0.00	2	0.04	1.75	1.16
Tamworth (TW)	GB	4	0.00	21	0.39	-	-
Poland China (PL)	US	4	0.00	24	0.37	-	-
Chester White (CW)	US	4	0.75	9	0.14	-	-
LOCAL AMERICA		54	0.22	24	0.17	0.78	1.55
Creole Bolivia (BO)	BO	6	0.00	3	0.06	1.65	1.33
Creole Costa Rica (CR)	CR	28	0.32	21	0.22	0.87	1.34

Feral pig (FP)	AR	6	0.33	9	0.17	1.30	1.69
Hairless (HL)	MX	14	0.07	16	0.16	-0.79	0.44
LOCAL AFRICA		14	0.07	19	0.13	-0.56	0.21
Kenyan (KE)	KE	8	0.00	5	0.05	-0.75	1.43
Mukota (MU)	ZW	6	0.17	13	0.22	0.54	0.57
WILD BOAR (WB) ASIA		30	0.00	19	0.15	-1.17	-0.18
China	CN	6	0.00	17	0.29	1.18	1.39
Japan	JP	12	0.00	18	0.11	-2.16	-1.23
Korea	KR	12	0.00	4	0.07	2.28	1.23
WB EUROPE + MAGHREB	BE, ES, FI, FR, GR, IT, MA, PO, PT, RO, RU, SE, SI, TN	132	0.00	27	0.04	-1.83	1.53
WB Belgium	BE	10	0.00	7	0.09	0.63	0.89
WB Spain	ES	36	0.00	4	0.04	0.08	1.05
WB Italy	IT	6	0.00	3	0.04	-1.23	-0.7
WB Romania	RO	10	0.00	3	0.05	1.44	1.15
WB Sweden	SE	10	0.00	4	0.08	-0.13	0.28
WB Slovenia	SI	8	0.00	4	0.33	0.71	1.87
WB Tunisia	TN	12	0.00	2	0.02	-1.62	-1.12

In some instances, e.g., Landrace, results are presented for the whole breed and by country when more than three individuals were genotyped. For that reason, not all countries for the whole breed are listed separately. A country in parenthesis means the origin of the animals, e.g., Meishan is a Chinese breed but we obtained the samples from France (INRA) and the US. **N**, number of haplotypes (i.e., twice the number of animals); country codes are the ISO two letter abbreviations (http://www.iso.org/iso/en/prods-services/iso3166ma/02iso-3166-code-lists/list-en1.html); Maghreb refers to NW African countries, here we sampled animals from Tunisia and Morocco; $\boldsymbol{P_A}$, frequency of the derived allele (*i.e.*, the causative mutation); ***S***, number of segregating sites; $\boldsymbol{\pi_N}$, mean number of pairwise differences across SNPs; $\boldsymbol{D_T}$, Tajima's *D*; $\boldsymbol{D_{FL}}$, Fu-Li's *D* index; *D* indices are reported when *N* > 6.

Table 2: SNP characteristics

Site	Gene	Region	Position[1]	Ancestral / Derived Allele[2]	Ancestral Allele Frequency
1	*TH*	Exon 14,3’UTR	262	T / C	0.84
2		Intergenic	1236	G / A	0.99
3		Intergenic	1856	G / T	0.57
4		Intergenic	1930	(G,A)	0.57
5	*IGF2*	Exon 1, 5’UTR	3889	G / A	0.17
6	*IGF2*	Intron 1	4706	G / A	0.94
7	*IGF2*	Intron 1	5039	C / T	0.95
8	*IGF2*	Intron 1	5581	G / C	0.83
9	*IGF2*	Intron 1	6755	G / A	0.79
10	*IGF2*	Intron 1	8213	A / C	0.71
11	*IGF2*	Intron 1	8283	G / A	0.60
12	*IGF2*	Intron 1	8592	G / A	0.94
13	*IGF2*	Intron 1	8911	C / T	0.63
14	*IGF2*	Intron 1	9012	T / G	0.29
15	*IGF2*	Intron 1	9568	(T, G)	0.70
16	*IGF2*	Intron 1	9583	C / T	0.93
17	*IGF2*	Intron 1	9986	C / T	0.92
18	*IGF2*	Intron 1	10320	T / A	0.08
19	*IGF2*	Intron 1	10714	C/ T / A	0.77 / 0.16
20	*IGF2*	Exon 2, 5’UTR	11423	G / A	0.93
21	*IGF2*	Intron 2	12501	A / G	0.09

22	*IGF2*	Intron 3	14902	C / A	0.44
23*	*IGF2*	Intron 3	16144	G / A	0.70
24	*IGF2*	Intron 3	17224	T / C	0.74
25	*IGF2*	Intron 3	18013	C / T	0.84
26	*IGF2*	Intron 3	18089	C / T	0.85
27	*IGF2*	Intron 3	18164	G / A	0.15
28	*IGF2*	Intron 3	18194	(T,G)	0.92
29	*IGF2*	Intron 3	18558	(A,C)	0.85
30	*IGF2*	Intron 6	23484	A / G	0.93
31	*IGF2*	Intron 6	23683	C / T	0.96
32	*IGF2*	Intron 6	23846	A / G	0.70
33	*IGF2*	Intron 6	24652	C / T	0.74
34	*IGF2*	Intron 8	27304	(T, C)	0.61

[1] Positions with respect to European wild boar accession AY242112.

[2] Status derived using *S. barbatus* as outgroup (Figure 1); parentheses when allele status could not be determined.

* Causative SNP

Figure captions

Figure 1. — a) Alignment showing the haplotypes from the 34 SNPs genotyped here pertaining to the 15 sequences published (VAN LAERE *et al.* 2003). The last three haplotypes (JXBLACK06741, IB01591 and *Sus barbatus*) were found here; they represent clades **A** (Jianxi Black), clade **M** (Iberian pig) and an outgroup (*S. barbatus*) and have been added for comparison. b) NJ-tree using the p-distance computed using the haplotypes. c) NJ-tree obtained with the full sequence (VAN LAERE *et al.* 2003). The letters in black circles refer to the main clades **E** (European), **C** (causative), **J** (Japanese), **A** (Asian), and **M** (Mediterranean). The representative haplotypes are shown with colored rectangles and used in all phylogenetic trees to facilitate comparisons (Figures 3 and S1).

Figure 2. — Relationship between Tajima's *D* and frequency of the causative mutation (P_A).

Figure 3. — NJ-trees for different populations using the p-distance. Encircled letters point at the clade representative (Figure 2). Different colors correspond to different breeds. The individual code is composed of 4 letters, 5 numbers and a G or an A. The first two letters are the breed abbreviation as in Table 1, the second two letters is the country code, the first four numbers are an internal code, the last number is the haplotype 1 or 2, and the final letter is the causative allele status. The color codes are: **China**, Licha Black (pink), Meishan (red), Fengjing (yellow), Jinhua (blue), Huai (light blue), Minzhu (dark green), Jiaxin Black (black), wild boar (grey), Luchuan (green); **Iberian Peninsula**, *Retinto* (red); Andalusian spotted (blue), *Torbiscal* (yellow),

Guadyerbas (pink), *Negro Lampiño* (black), *Porco Alemtejano* (green); **Great Britain**, Chester White (white), Berkshire (red), Middle white (light blue), Old Spot (green), British Lop (blue), Tamworth (yellow), Saddle Back (orange), Large Black (black). The black circles denote the clades defined in Figure 1 with representative haplotypes marked with colored rectangles: **C**, causative (red rectangle); **E**, European (blue); **A**, Asian (black); **M**, Mediterranean (green), and **J**, Japanese (magenta). All five representative haplotypes are drawn in all NJ-trees to facilitate comparisons.

Figure 4. — Haplotype structure and relationship between distance and disequilibrium measures. a) r^2 plot between pairs of loci and all individuals, haplotype blocks are underlined, the arrow points at the causative SNP; b) r^2 plot between pairs of loci for Landrace and Large White animals; c) relationship between distance and r^2 (full circles) and D' (open circles) for all pairs of loci; d) relationship between distance and r^2 and D' between the causative SNP and the rest of loci.

Supporting information

Figure S1. NJ-trees for different populations. Codes as in Figure 3. Color codes are: European wild boar, each color represents a different country. In Italy, Casertana (black), Sicilian (yellow), Calabrese (green), Cinta Sienese (blue).

FIG1

a)

```
                         10        20  *      30
AY242112EWB      CGTGAGCGGCAGCGTCCACGGAGCCCATAACATT  E
AY242098P208     T.GA.....AG.TTG......CAT.......GC.  C
AY242103LW3      T.GA.....AG.TTG......CAT.......GC.
AY242108LRJ      T.GA.....AG.TTG......CAT.......GC.
AY242111H205     T.GA.....AG.TTG......CAT.......GC.
AY242110H254     T......C.AG.TTG......CAT.......GC.
AY242099M220     T.GA.....AG.TTG......C.T.......GC.
AY242109JWB      T.G.GAT.A.GAT..TTTAAAC.T.TGGC.TGCC  J
AY242107LW1224   T......C.A........T.....T..G....CC
AY242106LW1461   T......C.A........T.....T..G......
AY242105LW197    ..................................
AY242104LW209    T......C.A........T........G....CC
AY242102LW33361  TAGA.....AG..........C.T.T.G......
AY242101LW419    ...........................G......
AY242100LW463    T.G.GAT.A.GA...TT.AAAC.T.T.G.G..C.
JXBLACK06741G    T.G.G....AG..........C.T.TGGCG..C.  A
IBES01591G       T.......AAG............T........C.  M
SUS BARBATUS     T.G?G....AG..T?..T..AC.T..G??...C?
```

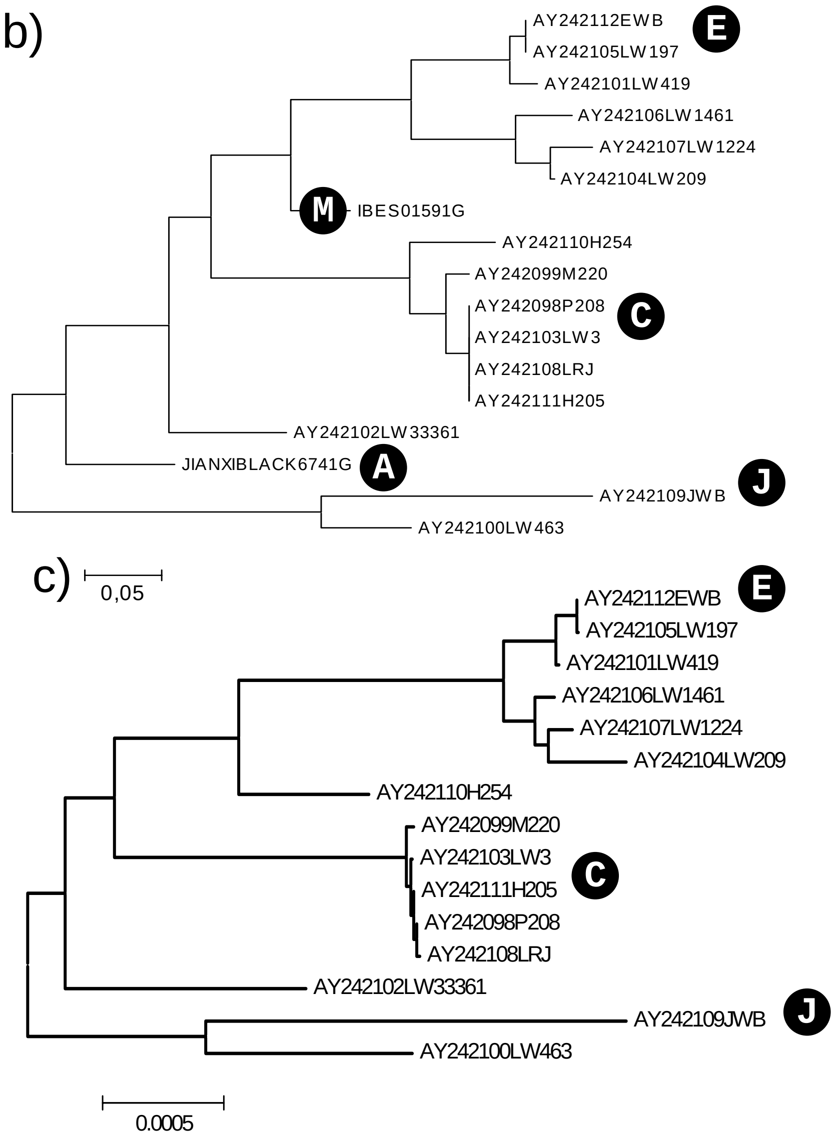


FIG2

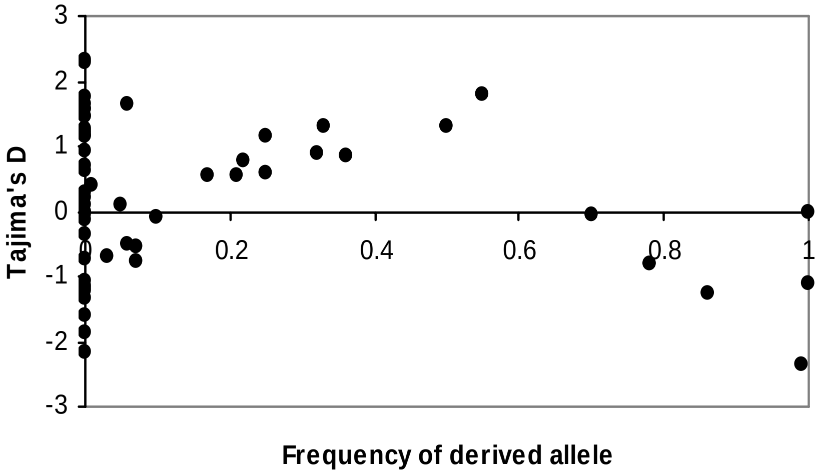


FIG3

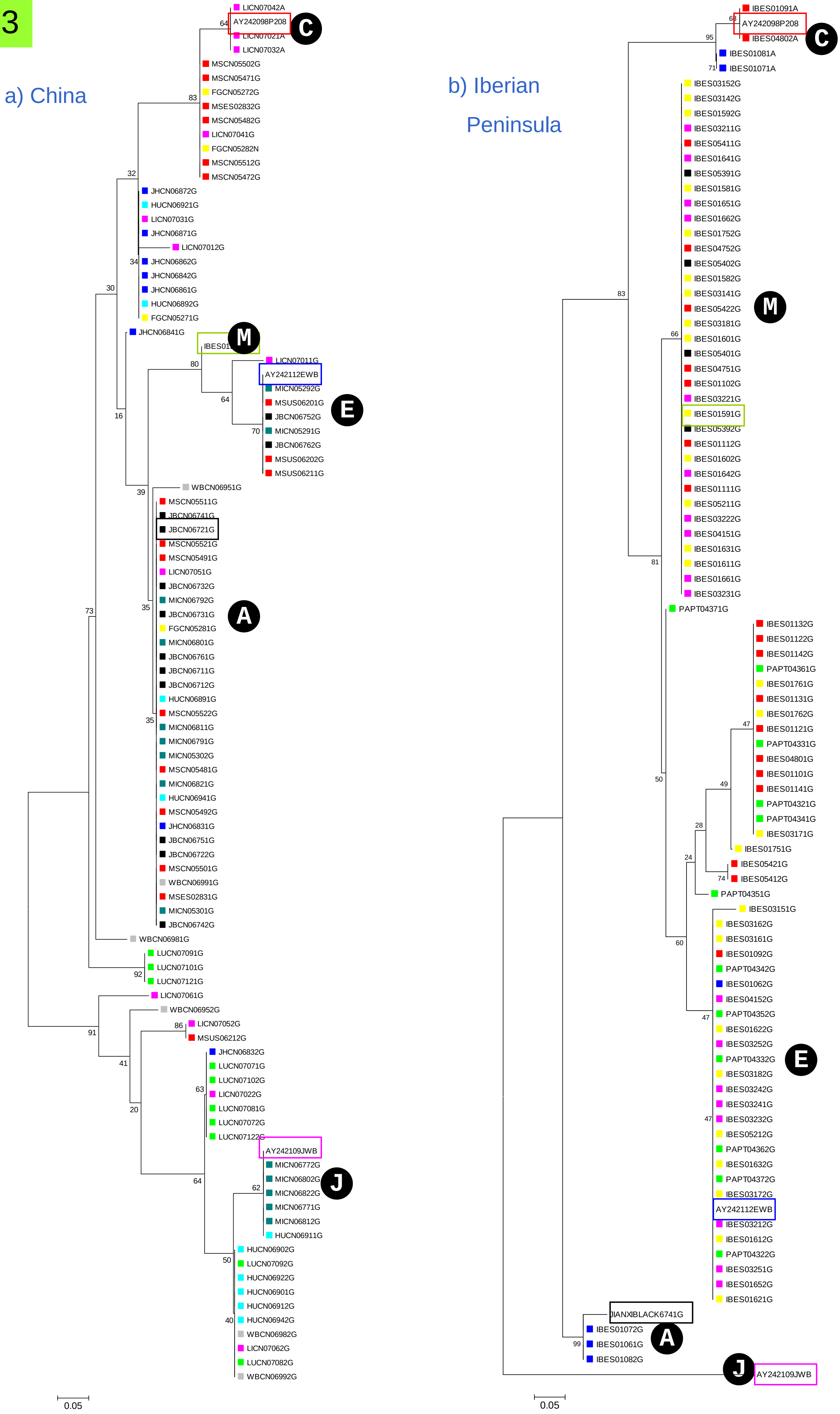
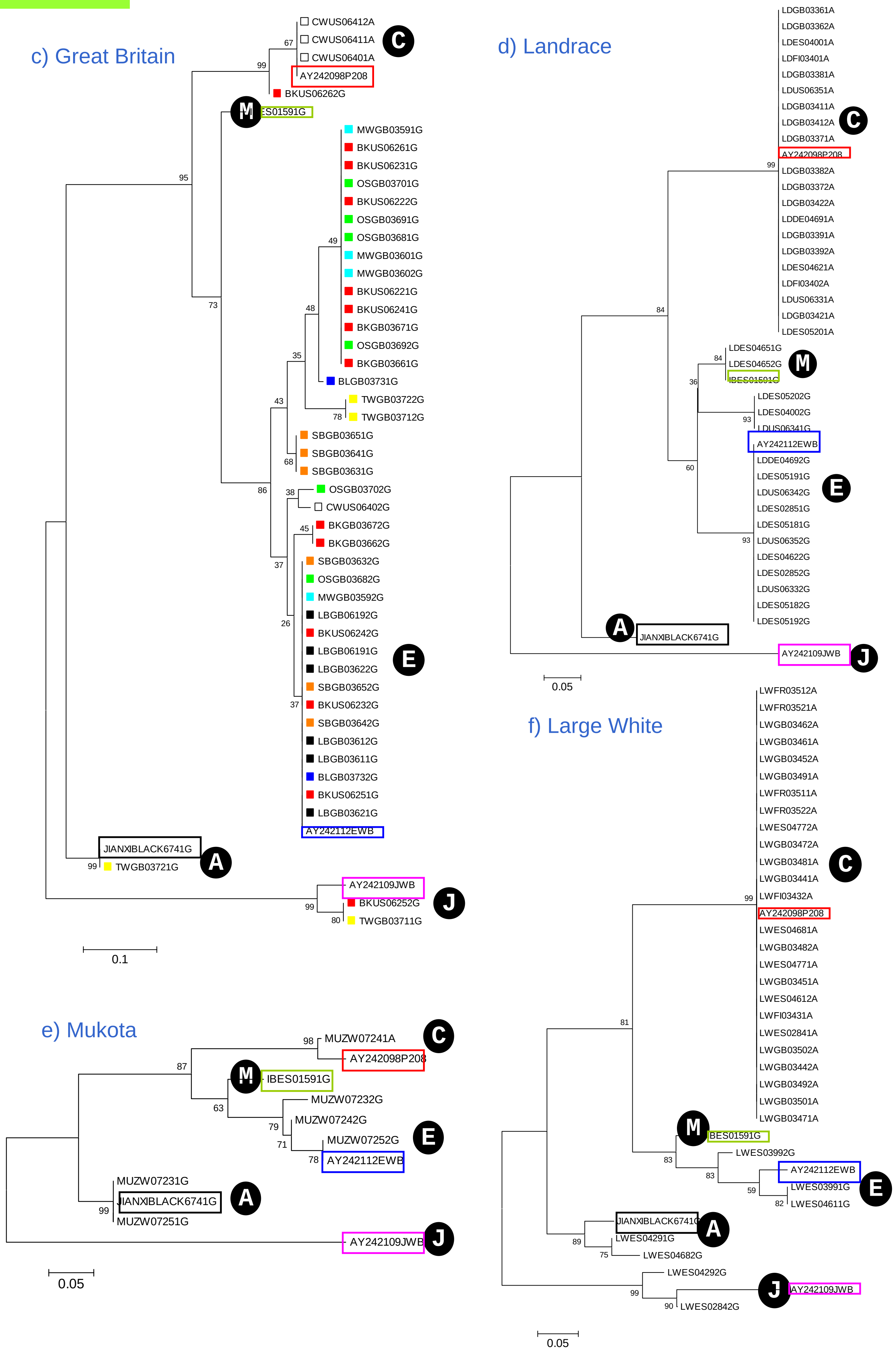

a) China
LICN07042A
64
AY242098P208
C
LICN07021A
LICN07032A
MSCN05502G
MSCN05471G
FGCN05272G
83
MSES02832G
MSCN05482G
LICN07041G
FGCN05282N
MSCN05512G
32
MSCN05472G
JHCN06872G
HUCN06921G
LICN07031G
JHCN06871G
LICN07012G
34
JHCN06862G
JHCN06842G
30
JHCN06861G
HUCN06892G
FGCN05271G
JHCN06841G
IBES01
M
80
LICN07011G
AY242112EWB
64
MICN05292G
16
MSUS06201G
E
JBCN06752G
70
MICN05291G
JBCN06762G
MSUS06202G
MSUS06211G
WBCN06951G
39
MSCN05511G
JBCN06741G
JBCN06721G
MSCN05521G
MSCN05491G
LICN07051G
JBCN06732G
MICN06792G
73
35
JBCN06731G
A
FGCN05281G
MICN06801G
JBCN06761G
JBCN06711G
JBCN06712G
HUCN06891G
35
MSCN05522G
MICN06811G
MICN06791G
MICN05302G
MSCN05481G
MICN06821G
HUCN06941G
MSCN05492G
JHCN06831G
JBCN06751G
JBCN06722G
MSCN05501G
WBCN06991G
MSES02831G
MICN05301G
JBCN06742G
WBCN06981G
LUCN07091G
92
LUCN07101G
LUCN07121G
LICN07061G
91
WBCN06952G
86
LICN07052G
41
MSUS06212G
JHCN06832G
LUCN07071G
LUCN07102G
63
LICN07022G
20
LUCN07081G
LUCN07072G
LUCN07122G
AY242109JWB
MICN06772G
64
62
MICN06802G
J
MICN06822G
MICN06771G
MICN06812G
HUCN06911G
HUCN06902G
50
LUCN07092G
HUCN06922G
HUCN06901G
HUCN06912G
40
HUCN06942G
WBCN06982G
LICN07062G
LUCN07082G
WBCN06992G
0.05
b) Iberian
Peninsula
IBES01091A
63
AY242098P208
95
IBES04802A
C
IBES01081A
71
IBES01071A
IBES03152G
IBES03142G
IBES01592G
IBES03211G
IBES05411G
IBES01641G
IBES05391G
IBES01581G
IBES01651G
IBES01662G
IBES01752G
IBES04752G
IBES05402G
IBES01582G
83
IBES03141G
M
IBES05422G
IBES03181G
66
IBES01601G
IBES05401G
IBES04751G
IBES01102G
IBES03221G
IBES01591G
IBES05392G
IBES01112G
IBES01602G
IBES01642G
IBES01111G
IBES05211G
IBES03222G
IBES04151G
IBES01631G
81
IBES01611G
IBES01661G
IBES03231G
PAPT04371G
IBES01132G
IBES01122G
IBES01142G
PAPT04361G
IBES01761G
IBES01131G
IBES01762G
47
IBES01121G
PAPT04331G
IBES04801G
50
IBES01101G
49
IBES01141G
PAPT04321G
PAPT04341G
IBES03171G
28
IBES01751G
24
IBES05421G
74
IBES05412G
PAPT04351G
IBES03151G
IBES03162G
60
IBES03161G
IBES01092G
PAPT04342G
IBES01062G
IBES04152G
47
PAPT04352G
IBES01622G
IBES03252G
PAPT04332G
E
IBES03182G
IBES03242G
IBES03241G
47
IBES03232G
IBES05212G
PAPT04362G
IBES01632G
PAPT04372G
IBES03172G
AY242112EWB
IBES03212G
IBES01612G
PAPT04322G
IBES03251G
IBES01652G
IBES01621G
JIANXIBLACK6741G
IBES01072G
A
99
IBES01061G
IBES01082G
J
AY242109JWB
0.05

FIG3contd



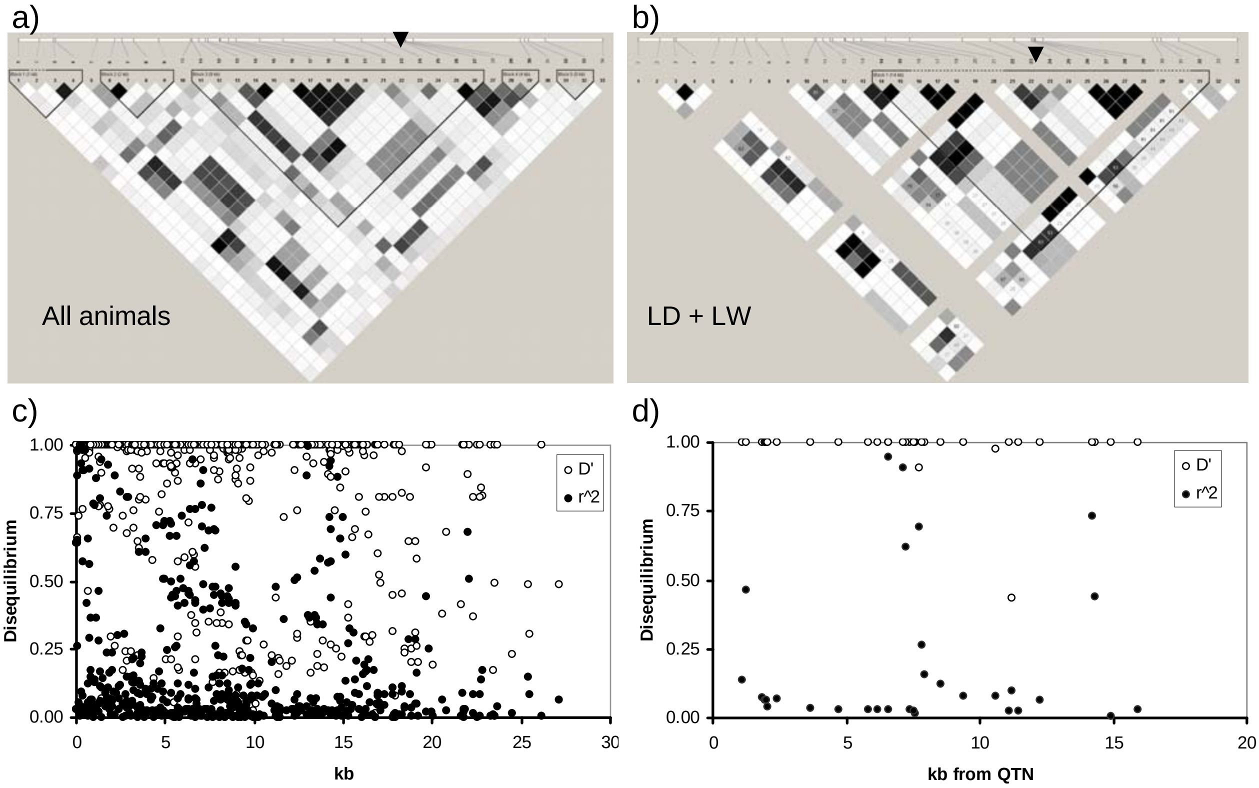
c) Great Britain

d) Landrace

e) Mukota

f) Large White

FIG4


a)
b)
All animals
LD + LW
c)
Disequilibrium
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
30
kb
○ D'
● r^2
d)
Disequilibrium
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
kb from QTN
○ D'
● r^2