Unexpected high polymorphism at FABP4 gene unveils a complex history for pig populations

Ana Ojeda*, Julio Rozas†, Josep M. Folch*, Miguel Pérez-Enciso*††

* Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
† Departament de Genètica, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain
†† Institut Català de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23, 08010 Barcelona, Spain

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. XXXXXX-XXXXXX.

NB: We have now submitted the sequences and we should have accession ids immediately.
Running title: High polymorphism at porcine FABP4

Key words: FABP4, Pig, Selection footprint

Correspondence:
Miguel Pérez-Enciso
Departament de Ciència Animal i dels Aliments
Facultat de Veterinària
Universitat Autònoma de Barcelona
08193 Bellaterra, Spain

email: miguel.perez@uab.es
Fax: +34 93 581 2106
Tel: +34 93 581 4225
Abstract

Fatty acid binding protein 4 (FABP4) plays a key role in fat regulation in mammals and is a strong positional candidate gene for the FAT1 quantitative trait locus located on porcine chromosome 4. DNA resequencing of the FABP4 gene region in 23 pigs from ten breeds and wild boar revealed 134 variants in 6.4 kb, representing a silent nucleotide diversity of $\pi_S = 0.01$, much higher than reported so far in animal domestic species. Moreover, this diversity was highly structured. Also strikingly, the FABP4 phylogenetic tree did not show any geographical or breed origin clustering, with distant breeds sharing similar haplotypes and some of the most heterozygous samples pertaining to highly inbred animals like Iberian Guadyerbas (inbreeding coefficient ~ 0.3) or British Tamworth. In contrast, the cytochrome b (mtDNA) phylogenetic tree was coherent with geographical origin. The estimated age of the most recent common ancestor for the most divergent Iberian or Tamworth haplotypes was much older than domestication. An additional panel of a hundred pigs from eight different breeds and wild boar from Spain, Tunisia, Sardinia and Japan was genotyped for seven selected single nucleotide polymorphisms and shows that high variability at porcine’s FABP4 is the rule rather than the exception. Pig populations, even highly inbred, can maintain high levels of variability for surprisingly long periods of time.

----------

Domestication and modern breeding has shaped dramatically the genomic variability of both animal and plant species that humans currently employ to satisfy their needs. In particular, domestic animal populations are a challenging material to study the footprint of selection because both natural and artificial selection overlap, and because the
species are made up of a complex web of semi isolated populations (the breeds) but that interchange genetic material occasionally. Many breeds, particularly commercial lines, have incorporated genetic material from other lines and the same breed is actually a mosaic of many different sublines. In pigs, for instance, it is well known that European breeds were crossed to Asian individuals to improve their prolificacy and docility. In this sense, the situation in pigs could be different from, e.g., dogs, where breeds consist mainly of closed breeding populations with a high degree of genetic differentiation (Parker et al. 2004).

The classical view considered domestication as a single event resulting in a strong bottleneck, followed by additional bottlenecks caused by breed formation and modern breeding (Clutton-Brock 1999). Although this simple model seems to fit some species like the dog (Lindblad-Toh et al. 2005) or maize (Wright et al. 2005), recent data suggests that domestication has been a much more complex and recurrent phenomenon than previously thought (Bruford et al. 2003). There is now widespread evidence of multiple domestication events in, e.g., pigs (Larson et al. 2005), cattle (Beja-Pereira et al. 2006) or chicken (Liu et al. 2006). Nevertheless, most nucleotide diversity studies in animals have analyzed mtDNA, which is inherited maternally and thus provides a partial story, or microsatellites, that have a very high mutation rate and are sparsely distributed along the genome. Since these phenomena are well known, it would be of interest to study in detail SNP autosomal variability for livestock species to complete our view of domestication and modern breeding.

Although some works have been recently published in dogs (Fondon and Garner 2004), detailed information on autosomal DNA sequence variation in different breeds is
lacking, the major effects of domestication and recent artificial selection on genome variation are still largely unknown. Certainly, regions that contain positional candidate genes are of particular scientific and economic interest. Here we report a comprehensive study of the fatty acid binding protein 4 (FABP4, also called Ap2 or A-FABP) in pigs, a strong positional candidate gene for the quantitative trait locus FAT1, located in swine chromosome 4. FAT1 was the first quantitative trait locus (QTL) reported in any animal domestic species and was detected initially in a wild boar x Large White pig cross (ANDERSSON et al. 1994). This QTL affects fatty acid deposition and growth and has been confirmed in numerous independent experiments involving many different breeds (PÉREZ-ENCISO et al. 2000; WALLING et al. 2000). Yet, its precise molecular nature has remained elusive so far. Recently, we have shown in an Iberian x Landrace cross that the FAT1 region is made up of at least two loci, with the fatty acid binding protein 4 (FABP4) being a very good positional candidate for one of them (MERCADE et al. 2005; MERCADE et al. 2006). In addition, several studies have demonstrated that FABP4 plays a critical role in fatty acid uptake and metabolism. In particular it has been shown that FABP4 is related to diabetes and obesity both in humans and in mice, and that mutations decreasing the level of FABP4 result in lower risk of coronary diseases and of type 2 diabetes (TUNCMAN et al. 2006).

Since modern breeding towards leaner animals must have dramatically affected the metabolism of fat deposition in pigs, it will be particularly illuminating to know the selection footprint left on genes involved in this process. A priori, one would expect a marked decrease in variability and a clear haplotype structure that would distinguish lean breeds from wild boar or autochthonous non-improved breeds. Our data clearly
refute this hypothesis in the case of FABP4 gene and show that high nucleotide diversities can be maintained for long periods of time even in highly inbred animals.

MATERIAL AND METHODS

Sequencing panel

We resequenced a panel of 23 pigs pertaining to the following breeds: Iberian (IB, IB113 and IB114 from retinto strain Villalón, South Spain, and IB415 and IB421 from Guadyerbas strain, a very fat and inbred strain), Porc Negre (PN, a very fat black breed, autochthonous from Balearic islands, n = 2), Tamworth (TW, UK endangered breed, n=1), British Lop (BL, UK endangered breed, n=1), Landrace (LD, hyperprolific line of French origin, n = 4), Large White (LW, LW419 pertains to an old very fat Large White line imported to Spain in 1931 and kept in a closed herd until 1992, when they were slaughtered (RODRIGÁÑEZ et al. 1998), and LW399 from a Spanish company), Duroc (DU, Spain, n=1), Meishan (MS, imported from China to INRA, France, n=1), Vietnamese pot belly (VT, VT104, from Madrid’s Zoo and VT310 from a private farmer in Spain), wild boar (WB) from Central and NE Spain (n=5). One babirusa (BB, Babyrusa babiroussa) from Madrid’s zoo was used as outgroup.

Genotyping panel

The genotyping panel was used to confirm the variability found in the sequenced animals, where a set of SNPs and microsatellites (described below) were genotyped. The panel consisted of wild boar (Spain, n=5, Sardinia, Italy, n=4, Tunisia, n=4 and Japan, n=5), Iberian (Guadyerbas, n=6, Torbiscal, n=5, Villalón, n=2), Sicilian (n=10), Pelón (Mexican hairless, n=8), Piétrain (Spain, France and Germany, n=12), Meishan (INRA, France, n=10), Large White (France, UK, Finland, Denmark and Holland,
n=10), Landrace (France, UK, Finland, Spain, Netherlands, n=10), and Duroc (USA, Canada, Denmark, n=10). This panel comprises the main commercial breeds used worldwide (Piétrain, Duroc, Large and Landrace), two Mediterranean autochthonous breeds (Iberian and Sicilian), one creole breed with likely Iberian origin (Mexican Pelón (LEMUS-FLORES et al. 2001), one Asiatic breed (Meishan), as well as wild boar from its extreme geographical boundaries. This panel should be thus representative of the variability at the species level.

Sequencing

We resequenced about 6.4 kb of FABP4 gene region that comprises all four exons and three introns as well as the 5’ and 3’ flanking regions (Figure 1). The region was amplified in six PCRs and sequenced. Primers and conditions in Suppl. Table. The amplified products were sequenced using the BigDye Terminator v3.1 Ready Reaction Cycle Sequencing Kit in an ABI PRISM 3730 (Applied Biosystems). About 900 bp of cytochrome b was sequenced as described (CLOP et al. 2004). The sequences obtained were analyzed using the SeqScape v2.5 software (Applied Biosystems) with standard filter settings and manually edited and verified.

Genotyping

In order to confirm the sequencing results in a wider sample, we selected seven SNPs: C2634del, A3159G, G4074A, T4091C, T4180A, G4205A, and T6252C (marked with numbers 1-4 in Figure 2). Indel 2634 was associated to fatness in an Iberian x Landrace cross (MERCADE et al. 2006), while SNP 3159 was chosen because is segregating at intermediate frequencies. The last five SNPs 4074 – 6252 were selected because they allow distinguishing clade A from B (explained below, Figure 3). Genotyping of the
indel and C6252T polymorphisms was done as described (MERCADE et al. 2006). For the A3172G polymorphism, a 112 bp-long fragment of intron 1 was amplified using FABP4A3172GF and FABP4A3172GR primers (Suppl. Table). PCRs were performed in a 25µL final volume containing 1.5mM MgCl₂, 200 µM dNTPs, 500 nM of each primer, 45 ng ADN and 0.6 U Taq DNA polymerase (Eco Taq). Thermocycling was 95 ºC for 3 min, 45 cycles of 95 ºC for 1 min, 60 ºC for 1 min, and 72 ºC for 1.5 min, and a final extension of 72 ºC for 5 min. The genotyping of this polymorphism was done in a PSQ HS 96 system (Pyrosequencing AB) using FABP4A3172GSeq2 as sequencing primer (Suppl. Table). The remaining polymorphisms were genotyped by resequencing using FABP4C4111TF and FABP4E1/E2R primers (Suppl. Table). PCRs were performed in a 25 µL final volume containing 1.5 mM MgCl₂, 200 µM dNTPs, 500 nM of each primer, 45 ng and 0.6 U AmpliTaq Gold (Applied Biosystem). Thermocycling was 95 ºC for 10 min, 40 cycles of 94 ºC for 30 s, 60 ºC for 1 min, and 72 ºC for 1.5 min, and a final extension of 72 ºC for 15 min.

In addition, six microsatellites (FABP4, SW35 and SW317 on chromosome 4, SW1517 on chromosome 2, SW1701 on chromosome 7 and S0144 on chromosome 8) were genotyped. PCR reactions were carried out in an GeneAmp PCR System 9700 (Perkin Elmer, Foster City,CA) and analyzed with fluorescent detection in an ABI PRISM 3730 Genetic Analyzer (Applied Biosystems, Foster City, CA). Genotypes were determined using the GeneScan 3.7 analysis software (Applied Biosystems).

**Sequence 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. For estimating \( \pi \) and for tests based in genotype frequencies (Tajima’s \( D \) and Fu-Li’s \( D \)), all polymorphisms were considered. For phylogenetic, recombination and linkage disequilibrium analyses, we considered only phases considered as reliable (P > 0.8) by Phase. Several parameters were estimated with DnaSP v4.10 (ROZAS et al. 2003): nucleotide diversity (\( \pi \)), Tajima’s \( D \) and Fu and Li’s \( D \) using babirusa as outgroup. Phylogenetic sequence analyses were carried out with Kimura’s two parameter model and neighbor-joining (NJ) trees using Mega3, bootstrap was performed to test the phylogenies (KUMAR et al. 2004). For completeness, phylogenetic relationships were also computed with the more parametric Bayesian approach as implemented in MrBayes3 with default parameters (RONQUIST and HUELSENBECK 2003). Recombination rates (\( \rho = 4 Ne r \)) were estimated using coalescence based techniques with LDhat (McVEAN et al. 2002). This method allows fitting a variable recombination rate along the sequence to infer the presence of recombination hotspots (or coldspots). Estimates were obtained with two million iterates and considering three block penalties 5, 20 and 50. Linkage disequilibrium measures were obtained with DnaSP. The Haploview v3.31 program (BARRETT et al. 2005) was used for graphical purposes and to determine the presence and length of haplotype blocks and of potential tag-SNPs.

**Genotype analysis**

For microsatellites, population structure was studied via the Wright’s \( F \) indices, a widely used measure for this purpose. \( F_{ST} \) measures the reduction in heterozygosity in a subpopulation due to drift, while \( F_{IS} \) measures the reduction caused by non random mating within a subpopulation (HARTL and CLARK 1997). Here we report the Weir and Cockerham’s estimates and the Slatkin’s \( R_{ST} \) index, the latter is equivalent to \( F_{ST} \) and is
appropriate when a stepping stone mutation model applies. The Package FSTAT was employed (J. Goudet, http://www.unil.ch/izea/softwares/fstat.html).

RESULTS

Sequence analysis

We detected 134 nucleotide polymorphisms within Sus, including 12 indels and 122 single nucleotide polymorphisms (SNPs, Figure 2). All were located outside the exons except one synonymous mutation in exon 3. This polymorphism level corresponds to one variant every 50 bp and a total and silent nucleotide diversity indices $\pi = 1.17\%$ and $\pi_s = 1.20\%$ respectively (Table 1). Table 1 also presents statistics for some breeds separately (Iberian, Landrace and wild boar). Note that the Iberian population alone had a nucleotide diversity that is half of the whole set of sequences, while Spanish wild boar is the least polymorphic population, followed closely by Landrace. Table 1 also shows Tajima’s $D$ and Fu & Li’s $D$ using babirusa as outgroup. 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. Iberian parameters were quite different from either wild boar, Landrace or in general the whole porcine population studied.

After reconstructing the gametic phases with Phase (Li and Stephens 2003), we obtained the neighbor-joining (NJ) phylogenetic tree. This tree shows an arrangement in two main distinct clades (named A and B in Figure 3 for convenience) that do not bear a relation to geographical origin: some local breeds like Iberian (Spain), Tamworth (UK) and Meishan (China) harbored haplotypes in both clades. The Bayesian analysis yielded an almost identical clade structure (Supplementary Information). The only Tamworth
animal sequenced, a breed considered endangered by the Rare Breed Survival Trust in Britain, was highly heterozygous, as was the only Meishan individual sequenced, imported from China by INRA (France) and maintained with low breeding numbers for several generations. The Iberian Guadyerbas strain was also highly polymorphic, containing haplotypes in both clades. In contrast to FABP4, the NJ-tree obtained from resequencing of ~660 bp of mitochondrial cytochrome b has a more clear geographical interpretation than that from FABP4 sequence (Figure 4 and Suppl. Figure for the Bayesian result), as breeds of Asiatic origin cluster together, including some Large White where Asiatic haplotypes are common due to interbreeding in the 19th century (GIUFFRA et al. 2000). Similarly, all Iberian pigs and Spanish wild boars had European haplotypes as has been reported previously (ALVES et al. 2003; CLOP et al. 2004).

The estimate of the scaled recombination rate $\rho = 4 Ne r$ obtained with LDhat (MCVEAN et al. 2002) was $\rho = 9.2$ for the whole region, or 1.44 per kb. Interestingly, LDhat pointed out two distinct regions, a high recombination region up to SNP 25 (position 3341) with $\rho = 17$ and an almost negligible recombination region thereafter (Figure 4). An additional hotspot could exist in the vicinity of the microsatellite, around SNPs 71-72, but it was detected only with small block penalty 10 (results not presented). This recombination hotspot is detected by the Haploview’s algorithm as this region is not included in any haplotype block (Figure 5). The haplotype structure of this gene seems complex, containing four small haplotype blocks. Forty four tag SNPs were required to capture all variability, according to the aggressive tagging algorithm with the default options from Haploview (BARRETT et al. 2005), 10 SNPs allowed tagging of 70% of all SNPs. Moreover, there was no clear relation between distance and significance of disequilibrium measures (Figure 6).
Genotypic data

In agreement with sequencing results (Figure 2), almost complete disequilibrium was confirmed in the genotyping panel for the SNPs that separate clades A from B (Figure 3): G4074A, T4091C, T4180A, G4205A, and T6252C. Complete disequilibrium was found between SNPs 4091 – 4205, and only one recombinant haplotype was identified between SNPs 4091 – 4205 and 6252 (in the Sardinian wild boar), and three recombinants between position 4074 and the following SNPs (two in Meishan and one in Japanese wild boar). Table 2 confirms the high variability of Iberian, Meishan and Japanese wild boars for SNPs that separate clade A vs. B. Only Mexican Pelón and Spanish wild boar had almost fixed clade A haplotypes. Overall, the general frequency of clade A haplotypes was 70% (last row in Table 3), comparable to that found in the sequencing panel (Figures 2 and 3).

Table 3 shows the F-indices obtained from the microsatellites. We reasoned that microsatellite at FABP4 should have fixation indices ($F_{ST}$) different from other microsatellites as is unlikely that the whole porcine genome is as variable as reported here for FABP4. However, this expectation was not confirmed and we found that FABP4’s microsatellite had $F_{ST}$’s comparable and not significantly different from other microsatellites (Table 3). The number of alleles was also similar to those found for other microsatellites. All $F_{ST}$’s were significantly different from zero for all markers, which supports that porcine breeds are highly structured.

DISCUSSION
FABP4 is an extremely variable gene, even in highly inbred lines: Nucleotide diversities found for this gene are higher than values reported in any domestic species so far, except in histocompatibility genes, e.g., is the double of the average rate found in chicken $\pi = 0.65\%$ (BERLIN and ELLEGREN 2004). Such high nucleotide diversities are typically found in species with high effective sizes like Drosophila, which have nucleotide diversities $\sim 1.0\%$ ($D. melanogaster$) – $1.8\%$ ($D. simulans$ and $D. pseudoobscura$) (MORIYAMA and POWELL 1996). But the results are even more surprising when we consider the fact that the NJ tree does not show a geographical arrangement (Figures 2 and 3) and that highly inbred populations like Tamworth or Iberian Guadyerbas turned out to be highly polymorphic for this gene. The case of Iberian Guadyerbas strain is particularly striking. This is a fat, black hairless and highly inbred line (average $F \sim 0.35$). It is representative of some of the old types of Iberian pigs and has been kept in isolation since 1945 (TORO et al. 2000). Yet, it contains haplotypes in all clades of Figure 3. This high variability was confirmed by genotyping a larger and more diverse sample of pigs (Table 2). In parallel to results in Figure 3, no clear geographical pattern could be observed in Table 2, although the clade B was more frequent in Asian animals. It is interesting to note that porcine breeds do not seem overall less polymorphic than their wild ancestors, as has also been observed in chicken (WONG et al. 2004).

The average pairwise distance between babirusa and pig sequences was $0.051 \pm 0.006$ using a Kimura’s two parameter model. Using an estimated divergence time of 12 – 26 My (THOMPSEN et al. 1996), this results in an estimated substitution rate of $\lambda = 0.051 / 2 \times (12-26) \times 10^6 \sim 1 - 2 \times 10^{-9} / \text{bp} / \text{yr}$, in agreement with usual results in mammals (GRAUR and LI 2000). The average distance within Sus was 0.01. In the only Tamworth
sequenced, the nucleotide distance between both haplotypes was 0.028, and that between the most distant Iberian haplotypes, 0.033; in both cases they are close to the maximum distance within *Sus*, 0.036. These extreme Iberian and Tamworth lineages must have diverged several (~ 7-15) My ago. It can be concluded that a large part of the variability observed is certainly much older than domestication. Thus, it is not surprising that all lineages are found both in European and in Asian populations since FABP4 polymorphisms predate domestication by far. The high and structured nucleotide diversity found between and within breeds strongly supports that domestication was not a single demographic event but rather the result of multiple processes, in agreement with previous results (Vila *et al.* 2005).

**Sequence, mtDNA and microsatellites tell different stories:** Most studies of genetic diversity are carried out with either mitochondrial DNA or microsatellites. But the results can be discordant depending on the markers used. Note that the phylogeny of autosomal FABP4 and maternally inherited cytochrome b are rather different (Figs. 3 and 4), emphasizing distinct male and female driven variabilities as observed in other species, e.g., the dog (Sundqvist *et al.* 2006). It is also important to notice that the FABP4 microsatellite is uncorrelated to the clade structure seen in Figure 3. Ten alleles were identified within the sequencing panel (Figure 2), although no specific variant could be ascribed to any of the clades in Figure 3. In fact, the average correlation between the microsatellite and the 134 polymorphisms was 0.03, and 0.02 with SNP in position 6252, that separates clade A vs. B. Also the $F_{ST}$ of FABP4’s microsatellite was comparable to the rest of microsatellites studied (Table 3), when one would expect to be different as is unlikely that the high variability of FABP4 gene is to be found throughout the porcine’s genome.
**Linkage disequilibrium:** An assessment of linkage disequilibrium is crucial to exploit current fine mapping techniques based on association analysis. The current view in animal domestic species is that disequilibrium occurs over long ranges, even across cM for dairy cattle (FARNIR et al. 2000), sheep (McRAE et al. 2002) or pig (JUNGERIUS et al. 2005; NSENGIMANA et al. 2004). In a recent study of resequencing, JUNGERIUS et al.(2005) observed that significant disequilibrium extended for about 50 kb in three pig breeds (Landrace, Large White and Meishan) implying that disequilibrium mapping could be based on a relatively sparse set of SNPs. The picture found here is more complex. First, the haplotype blocks were smaller and second, a non negligible fraction of the gene was not ascribed to any block (Figure 5). The Haploview algorithm suggests that one tag-SNP would be required every ~ 150 bp, impractical for any genome wide association analysis, although 10 SNPs would tag 94 SNPs. These results suggest that the density of tagging SNPs can vary dramatically between species or at least that long range disequilibrium should not be taken for granted in animal domestic species.

**Evidence for selection:** The pattern and level of DNA polymorphism distribution reported at FABP4 region is clearly unexpected for domestic breeds. The results are even more surprising when one considers that a highly significant QTL (P < 10^{-10}) affecting fatness maps very close to the FABP4 region (MERCADe et al. 2005). At first glance, a selective sweep promoted by a hitchhiking event acting on a beneficial mutation (i.e., directional selection) or a bottleneck event caused by the domestication might generate the observed unusual haplotype structure with two characteristic clades (clades A and B, Figures 2 and 3). Indeed, selection for a favorable mutation affecting fat levels might have generated the current genetic structure via hitchhiking (QUESADA
et al. 2003; ROZAS et al. 2001); the hitchhiking event might be incomplete (not all variation was removed due to recombination) or ongoing. The low nucleotide variability levels and the significant negative values of Tajima’s $D$ and Fu and Li’s $D$ on clade A would support the directional selection hypothesis (Tajima’s $D$ of -2.05, $P < 0.05$; Fu-Li’s $D$ of -2.40, $P < 0.05$).

The analysis just presented, however, does not consider that some breeds share haplotypes from the most distant clades and that there does not seem to exist intermediate haplotypes between clade A and B (Figures 2 and 3). When we analyze each population individually (those with $n > 4$, Table 1), only the Landrace population seems to be under directional selection. Since the number of substitutions in clade A is not highly reduced, the putative selective event should be old, predating the time of domestication. Balancing selection, in contrast, seems to fit well the data in the Iberian population. The segregating haplotypes were very old, the most recent common ancestor of the most distant Iberian haplotypes being $\sim 7 - 15$ My old. Although these ages are based on a simplistic application of the molecular clock and thus these estimates are likely biased (HO and LARSON 2006), no plausible range of substitution or mutation rates can explain that the high divergence observed could appear after domestication. Another indirect evidence of balancing selection can be obtained from the estimate of effective size. Under equilibrium, $\theta = 4Ne\mu = E(\pi) = 0.0059$ (Table 1); therefore using as $\mu \sim \lambda \sim 10^{-9}$ (see above), then $Ne \sim 10^6$. This estimate should nevertheless be taken with caution because domestication and bottlenecks causes that populations were not equilibrium. But even assuming some overestimation due to a violation of assumptions, this effective size is very high considering the small number of known breeding animals of the Iberian breed with a few elite herds providing most of
boars; this breed, in addition, suffered a strong bottleneck after 1960, when foreign breeds began to be imported and replaced the traditional breed (DOBAO et al. 1988), not to mention the particularly highly inbred Guadyerbas line (TORO et al. 2000). This suggests that some sort of balancing selection has maintained the observed levels of variability. The SNP genotype data on the genotyping panel does suggest that high variability is the rule rather than the exception (Table 2); for instance, if we take SNP T6252C as marker for clade A vs. B, we find that only the Mexican Pelón and Spanish wild boars do not show variation, while Iberian, Meishan and Japanese wild boars were the most variable populations. Nevertheless, the fact that only intronic and synonymous variability was found suggests that there are restrictions to the protein sequence and that balancing selection, if it exists, may not targetting the gene product itself.

Finally, admixture could also have played a role also in the gene structure observed. For instance, note that SNP frequencies in Sardinian and Tunisian wild boars were similar (Table 2). It is known that wild boar was introduced only after the seventh millennium BC in Sardinia (LARSON et al. 2005) and it is highly unlikely that Tunisian wild boar is crossed to their domestic counterparts. Then, the similarity between lines would suggest that there might have been interchanges between North Africa and the Mediterranean populations at some point in History, may be similar to that reported recently in cattle (BEJA-PEREIRA et al. 2006). Occasionally, male wild boar may have mated female domestic pigs, distorting the history of mitochondrial vs. autosomal loci. But this genetic flow is unlikely to be large (FANG et al. 2006). Further, other intercrosses seem highly unlikely, like that between Iberian Guadyerbas and Chinese Meishan which nevertheless share some long haplotypes, or between Japanese and Occidental wild boars. Nevertheless, genetic interchange between populations can not satisfactorily
explain the global high nucleotide diversity. Therefore, the observed pattern of nucleotide variation would have been generated by a complex interplay of directional or balancing selection together with a number of bottlenecks events.

**Conclusion:** This is, to our knowledge, the first comprehensive analysis of an autosomal porcine gene where both the wild ancestor and different domestic breeds have been analyzed. At present, we still do not know how general our results are, but they do challenge several currently held views about genetic variability in domestic animals: (i) domestic animals can maintain surprisingly high levels of variability for long periods of time; (ii) there is not necessarily a correlation between genetic and geographic differentiation; (iii) the presence of uninterrupted long haplotype blocks is not guaranteed.

**Acknowledgments**
We thank Anna Mercadé and Jordi Estellé for help with genotyping, and Maribel Merchán for help with sequence editing. We are extremely grateful to all persons and institutions who generously provided samples. These are: L. Silió, M.C. Rodríguez (INIA, Madrid), C. Talavera, E. Martínez (Madrid’s zoo), M. Cumbreras (Diputación de Huelva), J. Jaume (Institut de Biologia Animal de Balears), Erik v Eckhardt (Sweden), N. Okumura (STAFF Institute, Japan), C. Lemús (UAN, Mexico), R. Alonso (UNAM, Mexico), J. Reixach (Batallé, Spain), E. González (Seporsa, Spain), O. Vidal (Universitat de Girona), A. Angiolillo, S. Casu (Italy), C. Renard, J.P. Bidanel (INRA, France), P. Conde (Kubus, Spain), J.L. Noguera (IRTA, Spain), J.A. Barbosa (Cooperativa Guissona), Dehesón del Encinar (Junta de Castilla-La Mancha), and Copaga. We bought semen from Semen Cardona, Semen Porcino Andalucía (Spain) and Deerpark pigs (N. Ireland). We also thank one referee for the insightful comments. AO holds a PhD fellowship from Ministerio de Educación y Ciencia (MEC), Spain. Work funded by grants AFG2004-0103/GAN and BFU2004-02253 (MEC, Spain).

**LITERATURE CITED**


THOMPSEN, P., B. HOYHEIM and K. CHRISTENSEN, 1996 Recent fusion events during evolution of pig chromosomes 3 and 6 identified by comparison with the babirusa karyotype. Cytogenet Cell Genet. 73: 203-208.


Table 1

Polymorphism statistics in the sequencing panel

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>π %</th>
<th>πS %</th>
<th>DTajima</th>
<th>DFu-Li</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>47</td>
<td>1.17</td>
<td>1.20</td>
<td>-0.90*</td>
<td>-2.00**</td>
</tr>
<tr>
<td>Iberian</td>
<td>8</td>
<td>0.59</td>
<td>0.62</td>
<td>0.97*</td>
<td>1.01*</td>
</tr>
<tr>
<td>Landrace</td>
<td>8</td>
<td>0.21</td>
<td>0.22</td>
<td>-1.61**</td>
<td>-0.15</td>
</tr>
<tr>
<td>Wild boar</td>
<td>10</td>
<td>0.12</td>
<td>0.13</td>
<td>-0.67</td>
<td>-0.75</td>
</tr>
</tbody>
</table>

π and πS refer to total and silent nucleotide diversities, respectively; Fu-Li’s $D$ is obtained using babirusa as outgroup, $n$ refer to the number of sequences. Significance P-values were obtained with coalescence simulations setting $\rho = 9.2$ (the estimated value with LDhat): #, P < 0.10; *, P < 0.05; **, P < 0.01.
Table 2

Frequencies of the reference allele (Y16039) for a selected set of polymorphisms in the genotyping panel

<table>
<thead>
<tr>
<th>Position</th>
<th>2634</th>
<th>3159</th>
<th>4074</th>
<th>4091</th>
<th>6252</th>
<th>4205</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>-</td>
<td>G</td>
<td>A</td>
<td>TAG</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Alternative</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>CTA</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>Babirusa</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>TTG</td>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>WB Spain</td>
<td>0.70 0.40 1.00 1.00 1.00 5</td>
</tr>
<tr>
<td>WB Japan</td>
<td>0.40 0.10 0.10 0.40 0.40 5</td>
</tr>
<tr>
<td>WB Sardinia</td>
<td>0.83 0.62 0.75 0.75 0.88 4</td>
</tr>
<tr>
<td>WB Tunisia</td>
<td>0.83 0.50 0.83 0.83 0.75 4</td>
</tr>
<tr>
<td>IB Guadyerbas</td>
<td>0.33 0.42 0.42 0.42 0.42 6</td>
</tr>
<tr>
<td>IB Torbiscal</td>
<td>0.40 0.40 0.38 0.38 0.40 5</td>
</tr>
<tr>
<td>Mexican</td>
<td>0.94 0.57 0.93 0.93 0.93 8</td>
</tr>
<tr>
<td>Sicilian</td>
<td>0.85 0.30 0.85 0.85 0.85 10</td>
</tr>
<tr>
<td>Meishan</td>
<td>0.35 0.40 0.40 0.60 0.60 10</td>
</tr>
<tr>
<td>Duroc</td>
<td>0.70 0.61 0.70 0.70 0.70 10</td>
</tr>
<tr>
<td>Landrace</td>
<td>0.83 0.70 0.85 0.85 0.85 10</td>
</tr>
<tr>
<td>Large White</td>
<td>0.80 0.30 0.80 0.80 0.85 10</td>
</tr>
<tr>
<td>Piétrain</td>
<td>0.71 0.04 0.71 0.71 0.71 12</td>
</tr>
<tr>
<td>All</td>
<td>0.68 0.38 0.68 0.72 0.73 99</td>
</tr>
</tbody>
</table>
SNPs in positions 4091 – 4205 are in complete disequilibrium; $n$, number of individuals who were successfully genotyped for at least two SNPs.
Table 3

F indices obtained for the microsatellites typed in the genotyping panel

<table>
<thead>
<tr>
<th>Marker</th>
<th>Chr</th>
<th># alleles</th>
<th>$\hat{F}_{ST}$</th>
<th>$\hat{F}_{IS}$</th>
<th>$R_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SW1517</td>
<td>2</td>
<td>17</td>
<td>0.15±0.03</td>
<td>0.07±0.05</td>
<td>0.29</td>
</tr>
<tr>
<td>SW317</td>
<td>4</td>
<td>13</td>
<td>0.09±0.03</td>
<td>0.19±0.10</td>
<td>0.09</td>
</tr>
<tr>
<td>FABP4</td>
<td>4</td>
<td>15</td>
<td>0.12±0.04</td>
<td>0.11±0.07</td>
<td>0.16</td>
</tr>
<tr>
<td>SW35</td>
<td>4</td>
<td>6</td>
<td>0.29±0.07</td>
<td>0.29±0.10</td>
<td>0.45</td>
</tr>
<tr>
<td>SW1701</td>
<td>7</td>
<td>18</td>
<td>0.15±0.04</td>
<td>0.09±0.05</td>
<td>0.07</td>
</tr>
<tr>
<td>S0144</td>
<td>8</td>
<td>9</td>
<td>0.22±0.05</td>
<td>0.09±0.09</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Figure 1. Scheme of FABP4 gene showing the region sequenced, from ~ 1300 – 7730 bp. The four exons (E1-E4) span positions 2379-2451, 5098-5269, 6128-6227, 6702-6749 bp, a microsatellite (M) is in position 4951, and a SINE (S) is located at 5617-5914 bp. All positions relative to accession Y16039.

Figure 2. Polymorphisms found by resequencing; haplotypes are arranged as in the NJ tree in Figure 2 and thus note that haplotypes from the same individual are not necessarily contiguous (the last number in the sequence name refers to the haplotype, 1 or 2). First sequence is reference sequence (GenBank accession Y16039) while the second is the outgroup (*Babyrousa babyrussa*). Indels are denoted with ‘-’, lower case indicates that phase could not be reliably determined and ‘?’ , that nucleotide could not be ascertained because the individual was heterozygous for indels or the microsatellite. The microsatellite is located between SNPs 67 and 68 and is indicated by the symbol ‘M’, the allele code (0-20) represents the difference in number of repeats relative to babirusa. The only mutation is an exon is marked with E (position 6094). The bottom line numbers the SNPs genotyped in the genotyping panel (Table 3). These were in positions 2634 (1), 3159 (2), 4074 (3), 4091 (3), 4180 (3), 4205 (3) and 6252 (4).

Figure 3. Neighbor-Joining tree for the FABP4 gene. The last number (1 or 2) indicates the haplotype, thus note that Tamworth (TW372) or Meishan (MS405) has two highly divergent haplotypes, as have Guadyerbas IB415, Landrace LD400 or Large White LW399. Breeds with haplotypes in different clades are marked with symbols. Numbers at selected branches are the bootstrap values after 1000 resampling draws.
Figure 4. Neighbor - Joining tree obtained from cytochrome b sequence, letters refer to haplotypes described by Giuffra et al. (Giuffra et al. 2000): E1 and E3 are European haplotypes and A1 and A2, Asiatic. All non shown sequences (REST) fall within the E1 clade. Babirusa sequences (BB) obtained from GenBank are added with accession numbers shown.

Figure 5. Estimation of blocks with different recombination rates obtained from LDhat (McVean et al. 2002) and block penalty 20.

Figure 6. Linkage disequilibrium ($r^2$) plot obtained with Haploview 3.31 (Barrett et al. 2005). Triangles with solid lines represent the five blocks identified with Gabriel et al’s algorithm. Arrows marks the span of the high recombination hotspot.

Figure 7. Relation between physical distance and significance of linkage disequilibrium. The correlation coefficient between distance and P-value was 0.09.
Figure 4:
Figure 5
Figure 6
Figure 7