RAPID COMMUNICATION: A haplotype information theory method reveals genes of evolutionary interest in European vs. Asian pigs[1]

Nicholas J. Hudson,[*] Marina Naval-Sánchez,[†] Laercio Porto-Neto,[†] Miguel Pérez-Enciso,[‡,§] and Antonio Reverter[†,2]

[*]School of Agriculture and Food Sciences, University of Queensland, Gatton, Queensland 4343, Australia [†]CSIRO Agriculture & Food, 306 Carmody Road., Street Lucia, Brisbane, Queensland 4067, Australia [‡]Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB Consortium, Bellaterra 01893, Spain [§]Institut Català de Recerca i Estudis Avançats (ICREA), Carrer de Lluís Companys 23, Barcelona 08010, Spain

ABSTRACT: Asian and European wild boars were independently domesticated ca. 10,000 yr ago. Since the 17th century, Chinese breeds have been imported to Europe to improve the genetics of European animals by introgression of favorable alleles, resulting in a complex mosaic of haplotypes. To interrogate the structure of these haplotypes further, we have run a new haplotype segregation analysis based on information theory, namely compression efficiency (**CE**). We applied the approach to sequence data from individuals from each phylogeographic region (n = 23 from Asia and Europe) including a number of major pig breeds. Our genome-wide CE is able to discriminate the breeds in a manner reflecting phylogeography. Furthermore, 24,956 nonoverlapping sliding windows (each comprising 1,000 consecutive SNP) were quantified for extent of haplotype sharing within and between Asia and Europe. The genome-wide distribution of extent of haplotype sharing was quite different between groups. Unlike European pigs, Asian pigs haplotype sharing approximates a normal distribution. In line with this, we found the European breeds possessed a number of genomic windows of dramatically higher haplotype sharing than the Asian breeds. Our CE analysis of sliding windows captures some of the genomic regions reported to contain signatures of selection in domestic pigs. Prominent among these regions, we highlight the role of a gene encoding the mitochondrial enzyme *LACTB* which has been associated with obesity, and the gene encoding *MYOG* a fundamental transcriptional regulator of myogenesis. The origin of these regions likely reflects either a population bottleneck in European animals, or selective targets on commercial phenotypes reducing allelic diversity in particular genes and/or regulatory regions.

Key words: evolution, genome sequence, pig

J. Anim. Sci. 2018.XX:XX–XX
doi: 10.1093/jas/sky225

INTRODUCTION

Pigs were independently domesticated in Asia and Europe about 10,000 yr ago. Much later, starting in the late 17th century, genetics from Asian breeds were introgressed into European animals to improve certain phenotypes, such as the ability to feed from forage, fatness, and prolificacy (Giuffra et al., 2000; White 2011; Bosse et al., 2015). Moreover, subsequent breed formation through artificial selection on diverse traits occurred in both phylogeographic groups. This combination of historic events has left the genomes of extant populations made up of a complex mosaic of haplotypes. We have recently

[1]In 2015, Miguel Pérez-Enciso visited CSIRO on a Sir Frederick McMaster Visiting Fellowship which laid the foundation for this work.

[2]Corresponding author: tony.reverter-gomez@csiro.au

Received April 2, 2018.
Accepted June 4, 2018.

developed a new method for discriminating populations and assessing haplotype structure (Hudson et al., 2014, 2015, 2017). It exploits the concept of data compression from information theory, highlighting by high compression efficiency (**CE**) those genomic regions that can be efficiently compressed due to regular allele patterns within and between genomes.

Related to disequilibrium and nucleotide diversity, CE combines both features in a single metric. Overall, this yields a very inclusive yet simple analysis that identifies haplotype subsets present in high frequency in one population but not another. Here, we used 23 individuals in each of the two phylogeographic groups. The use of a previously published (Leno-Colorado et al., 2017; Perez-Enciso et al., 2017) *de novo* SNP identification approach from genome sequencing gave us a very high resolution analytical tool with which to scrutinize the population history and genetics of these animals.

MATERIALS AND METHODS

Animal Sequences and SNP Resources

We analyzed complete genome sequence from 46 domestic and wild pigs from Asia and Europe (Leno-Colorado et al., 2017; Pérez-Enciso et al., 2017). In brief, the sampling scheme was designed to be as balanced as possible while simultaneously aiming at capturing variability within continents and within domestic status. Among the 23 sequences from Asian pigs, we have seven wild boars and four from each of Tibetan, Meishan, Hetao, and Bamaxiang. Similarly, among the 23 sequences from European pigs, we have eight wild boars, three Hungarian Mangalica, and four sequences from each of Iberian, Duroc, and Large White.

Samples were sequenced to an average depth of 11× and aligned against *Sus scrofa* (**SSC**) genome version 10.2 (Groenen et al., 2012). Following previously described approaches (Leno-Colorado et al. 2017; Pérez-Enciso et al. 2017), sequence reads were realigned around indels with GATK IndelRealigner tool (McKenna et al., 2010). Further, SNP calling was performed with samtools/bcftools suite v. 1.2.1 (Li et al., 2009) for each individual separately. Next, all genotype samples were merged in a single file with SNP variants. SNPs with more than 30% missing rate were discarded.

After quality filtering and discarding singletons, we retrieved a total of 24,955,543 autosomal SNPs including 5,978,035 intron variants covering 14,303 genes out of all annotated genes (25,322). Missing SNPs were imputed with Beagle 4 (Browning and Browning, 2013).

Genome-Wide CE and Signatures of Selection

Following previously described approaches (Hudson et al., 2014), CE was computed for each animal from comparing the size in bytes its genotype file before (S_B) and after (S_A) compression as follows:

$$\mathrm{CE} = \frac{S_B - S_A}{S_B}$$

The *gzip* application tool (http://www.gzip.org) was used to compress the files. We first explored the power of CE metrics to cluster individual pig sequences by continent (i.e., Asia vs. Europe) and then applied a sliding window CE approach to detect outlier patterns that may indicate selection. Windows of 1,000 consecutive SNPs were explored for a total of 24,956 windows (i.e., the last one containing 543 SNP only).

Because CE is proportional to regularity, it is therefore related to genetic phenomena that provide regularities both within and across genomes such as runs of homozygosity and linkage disequilibrium (**LD**). Similar to runs of homozygosity and LD, CE does not depend on allele frequencies. However, CE is highly influenced by genome-wide heterozygosity (***Het*** computed from the proportion of heterozygous sites among the 1,000 consecutive SNPs in each window), and therefore CE was adjusted for *Het*, termed **CEh** and computed as follows:

$$\mathrm{CEh} = \frac{\mathrm{CE}}{Het}$$

The CEh sliding window method can be exploited to identify haplotypes segregating between populations of animals. A CEh peak over a genomic region in one population but not another indicates relatively high sharing of haplotypes in the first population, however compositionally complex those haplotypes may be. However, the exact genomic composition at a given compression peak needs assessing on a case-by-case basis, and could be quite different in one genomic location versus another.

This measure of CEh was further *Z*-score normalized into **CEhZ** by subtracting the genome-wide average CEh and dividing by the genome-wide standard deviation.

Candidate windows were visualized with a heat map and hierarchical cluster analysis performed

using the PermutMatrix software (Caraux and Pinloche, 2005) with individual animals in rows, SNPs in columns and genotypes AA, AB, and BB mapped to green, black, and red, respectively.

RESULTS AND DISCUSSION

Genome-Wide CE Adjusted for Heterozygosity (CEh)

Genome-wide CE expressed against genome-wide Het separated the breeds from each other, but a much stronger separation was detected for phylogeographic origin (Fig. 1A). This is in line with what is known on pig phylogeography (Groenen et al., 2012) and with our previous assessment made using a heritability metric (Pérez-Enciso et al., 2017). It implies that the combination of the difference between the two original wild populations, Asia and Europe, plus the independent domestication events dominates the genetics of the modern populations, with breed being a relatively minor contributor. The overall pattern of discrimination is analogous to that made using F_{ST} (allele frequency differentiation) and principal component analysis (Yang et al. 2017). Furthermore, the European pigs have a substantially lower Het than the Asian pigs. This likely reflects the previously described population bottleneck (Groenen et al., 2012; Frantz et al., 2015) that some local European breeds such as Iberian and Mangalica have been exposed to, and which exhibit reduced DNA variability as a consequence (Pérez-Enciso et al., 2017).

Sliding Window CEh and CEhZ

We found that the genome-wide distribution of haplotype sharing as assessed by CEh was dramatically different between European and Asian pigs (Fig. 1B). Across the 24,956 windows, the mean, SD, minimum and maximum CEh values for Asian (European) pigs was 4.49 (13.98), 0.93 (14.47), 1.78 (2.36), and 19.40 (297.89), respectively. Asian pigs sharing distribution is a normal-like symmetric distribution while European pig distribution is close to an exponential-like skewed distribution.

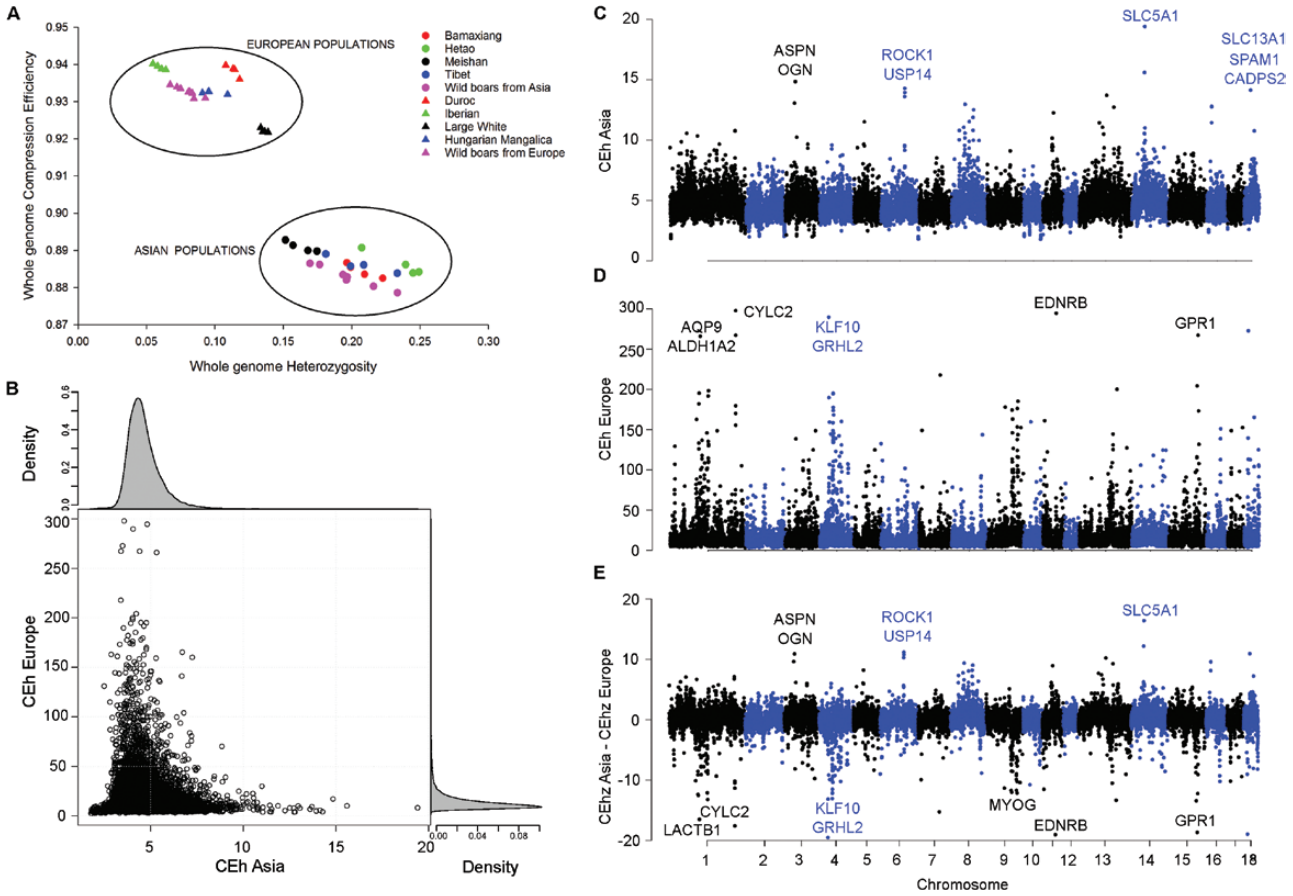

Figure 1. Compression efficiency (CE) of pig genomes: (A) Scatter plot of whole-genome sequence compression efficiency over heterozygosity for 46 genomes from 10 pig breeds in two phylogeographic regions, Asia and Europe; (B) Density distribution and relationship between CE corrected for heterozygosity (CEh) in Asian (*x*-axis) and European (*y*-axis) pig breeds; (C) Distribution of CEh along the genome for Asian pigs breeds listing candidate genes; (D) Distribution of CEh along the genome for European pigs breeds listing candidate genes; (E) Distribution along the genome for the difference between normalized CEh (CEhZ) in Asian minus European pig breeds listing candidate genes.

Consequently, in European pigs, there are a number of very extreme regions of high haplotype sharing that is not observed in the Asian pigs (Fig. 1C and D). This is likely a consequence of the bottleneck the European population is known to have gone through. Note also that the variance of CEh is much higher in European than in Asian pigs, and this is likely due to Asian introgression into Europe (Pérez-Enciso, 2014). Signatures of selection may also be partially responsible for this pattern but, as previously observed, the signals that are shared between breeds tend to be weak (Pérez-Enciso et al., 2017). We have ruled out genome-wide patterns of differential recombination frequency as being likely on the grounds that this process would be expected to influence European and Asian pig genomes similarly.

Some of the genomic regions reported to contain signatures of selection in domestic pigs by Rubin et al. (2012) were also captured by our CEh sliding window analyses. These include *PLAG1*, a loci on porcine chromosome SSC4 associated with stature, growth, and puberty in other domestic animals and humans, which ranked 57 out of 14,303 genes (or top 0.40%) by CEh in European breeds; and *MC1R*, a well-known loci on SSC6 associated with coat color, which ranked 34 (top 0.24%) also by CEh in European breeds. Similarly, LCORL, a loci on SSC8 consistently associated with body size in domestic animals, was ranked 170 (top 1.19%) in Asian breeds, while *SERPINA6* on SSC7, also known as corticosteroid-binding globulin (*CBG*) and known to affect fat deposition and muscle content in pigs (Guyonnet-Dupérat et al., 2006; Esteve et al., 2011), was ranked at the bottom 98.60% according to CEh in Asian breeds.

Within Asian breeds (Fig. 1C), windows of high haplotype sharing included *ASPN* and *OGN* on SSC3 both reported to affect growth and carcass traits in a Meishan × Piétrain intercross (Stratil et al., 2006). Similarly, the gene *SLC5A1* on SSC14 was captured by the window associated with the highest CEh value in Asian pigs and corresponds to a putative copy number variant region associated with fatness in three pig breeds (Fowler et al. 2013).

Within European breeds (Fig. 1D), the window with the highest CEh value contains porcine cylicin II (*CYLC2*) on SSC1 encodes a protein of the sperm cytoskeleton and is likely to play a role in spermiogenesis and fertilization (Rousseaux-Prévost et al. 2003). The window with the second largest CEh value was located on SSC11 and captures the gene *EDNRB* reported to contain a signature of diversifying selection affecting breed standard criteria, such as coat color and ear morphology in European pig breeds (Wilkinson et al. 2013).

Among the top five most extreme windows in the Asian vs. European breeds contrast based on CEhZ (Fig. 1E), were regions containing the gene encoding the mitochondrial enzyme *LACTB* (Fig. 2) which has been associated with obesity in mice (Chen et al., 2008) but no role as of yet documented in pigs, and the gene encoding *MYOG* a fundamental transcriptional regulator of myogenesis (Wright et al., 1989), recently shown to be differentially expressed among pigs with extreme intra-muscular fat content (Lim et al. 2017). We hypothesize that these two genes may play a role in driving carcass composition, muscularity, and metabolism in modern European pigs.

In Fig.2, it is worth noting that, after hierarchical cluster analysis of rows (animals), the upper

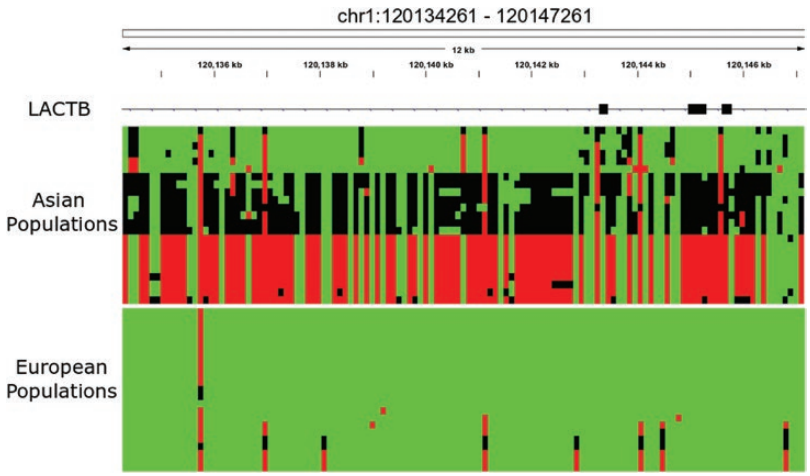

Figure 2. Heatmap of the allele composition (green = AA, black = AB, red = BB) for 127 consecutive intronic SNP (arranged in columns) over the *LACTB* region (chr1:120,134,261 – 120,147,261) in the Asian (top panel) and European populations (bottom panel) with 23 pigs each (arranged in rows).

panel where the Asian breeds are displayed, shows three distinct groups of animals. The top two groups where genotypes AA predominate (green cells) are made up of a mixture of wild boards and Meishan pigs. However, the bottom group where BB predominates (red cells) is where Hetao and Tibetan breeds are found.

In conclusion, in previous studies we have found genome-wide CE to bear a relationship to *FST* in human, cattle, sheep, and chicken populations (Hudson et al., 2014, 2015, 2017). A sliding window version of the analysis also detects regions of high co-sharing of haplotypes that established genetic methods may overlook. This could be because it is a hypothesis-free screen that makes no *a priori* prediction about haplotype structure, such as Runs of Homozygosity, and because it makes no assumption of Hardy–Weinberg equilibrium so no genomic regions are omitted. These properties have been corroborated in the current study where CE information-based method have been applied to whole-genome sequence data from pigs allowing us to retrieve the pattern expected from the known demographic history of the population. Similarly, when CE is applied by windows, the most extreme regions should be enriched in genes that depart from the average demographic history and therefore could be selection candidates.

LITERATURE CITED

Bosse, M., M. S. Lopes, O. Madsen, H. J. Megens, R. P. Crooijmans, L. A. Frantz, B. Harlizius, J. W. Bastiaansen, and M. A. Groenen. 2015. Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. Proc. Biol. Sci. 282:20152019. doi:10.1098/rspb.2015.2019

Browning, B. L., and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 194:459–471. doi:10.1534/genetics.113.150029

Caraux, G., and S. Pinloche. 2005. Permutmatrix: a graphical environment to arrange gene expression profiles in optimal linear order. Bioinformatics. 21:1280–1281. doi:10.1093/bioinformatics/bti141

Chen, Y., J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, et al. 2008. Variations in DNA elucidate molecular networks that cause disease. Nature. 452:429–435. doi:10.1038/nature06757

Esteve, A., A. Ojeda, L. S. Huang, J. M. Folch, and M. Pérez-Enciso. 2011. Nucleotide variability of the porcine SERPINA6 gene and the origin of a putative causal mutation associated with meat quality. Anim. Genet. 42:235–241. doi:10.1111/j.1365-2052.2010.02138.x

Fowler, K. E., R. Pong-Wong, J. Bauer, E. J. Clemente, C. P. Reitter, N. A. Affara, S. Waite, G. A. Walling, and D. K. Griffin. 2013. Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. BMC Genomics. 14:784. doi:10.1186/1471-2164-14-784

Frantz, L. A., J. G. Schraiber, O. Madsen, H. J. Megens, A. Cagan, M. Bosse, Y. Paudel, R. P. Crooijmans, G. Larson, and M. A. Groenen. 2015. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nat. Genet. 47:1141–1148. doi:10.1038/ng.3394

Giuffra, E., J. M. Kijas, V. Amarger, O. Carlborg, J. T. Jeon, and L. Andersson. 2000. The origin of the domestic pig: independent domestication and subsequent introgression. Genetics. 154:1785–1791. https://www.ncbi.nlm.nih.gov/pubmed/10747069

Groenen, M. A., A. L. Archibald, H. Uenishi, C. K. Tuggle, Y. Takeuchi, M. F. Rothschild, C. Rogel-Gaillard, C. Park, D. Milan, H. J. Megens, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 491:393–398. doi:10.1038/nature11622

Guyonnet-Dupérat, V., N. Geverink, G. S. Plastow, G. Evans, O. Ousova, C. Croisetière, A. Foury, E. Richard, P. Mormède, and M. P. Moisan. 2006. Functional implication of an Arg307Gly substitution in corticosteroid-binding globulin, a candidate gene for a quantitative trait locus associated with cortisol variability and obesity in pig. Genetics. 173:2143–2149. doi:10.1534/genetics.105.053983

Hudson, N. J., R. J. Hawken, R. Okimoto, R. L. Sapp, and A. Reverter. 2017. Data compression can discriminate broilers by selection line, detect haplotypes, and estimate genetic potential for complex phenotypes. Poult. Sci. 96:3031–3038. doi:10.3382/ps/pex151

Hudson, N. J., L. R. Porto-Neto, J. Kijas, S. McWilliam, R. J. Taft, and A. Reverter. 2014. Information compression exploits patterns of genome composition to discriminate populations and highlight regions of evolutionary interest. BMC Bioinformatics. 15:66. doi:10.1186/1471-2105-15-66

Hudson, N. J., L. Porto-Neto, J. W. Kijas, and A. Reverter. 2015. Compression distance can discriminate animals by genetic profile, build relationship matrices and estimate breeding values. Genet. Sel. Evol. 47:78. doi:10.1186/s12711-015-0158-9

Leno-Colorado, J., N. J. Hudson, A. Reverter, and M. Pérez-Enciso. 2017. A pathway-centered analysis of pig domestication and breeding in Eurasia. G3 (Bethesda). 7:2171–2184. doi:10.1534/g3.117.042671

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and samtools. Bioinformatics. 25:2078–2079. doi:10.1093/bioinformatics/btp352

Lim, K. S., K. T. Lee, J. E. Park, W. H. Chung, G. W. Jang, B. H. Choi, K. C. Hong, and T. H. Kim. 2017. Identification of differentially expressed genes in longissimus muscle of pigs with high and low intramuscular fat content using RNA sequencing. Anim. Genet. 48:166–174. doi:10.1111/age.12518

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. doi:10.1101/gr.107524.110

Pérez-Enciso, M. 2014. Genomic relationships computed from either next-generation sequence or array SNP data. J.

Anim. Breed. Genet. 131:85–96. doi:10.1111/jbg.12074
Pérez-Enciso, M., G. de Los Campos, N. Hudson, J. Kijas, and A. Reverter. 2017. The 'heritability' of domestication and its functional partitioning in the pig. Heredity (Edinb). 118:160–168. doi:10.1038/hdy.2016.78
Rousseaux-Prévost, R., C. Lécuyer, H. Drobecq, C. Sergheraert, J. L. Dacheux, and J. Rousseaux. 2003. Characterization of boar sperm cytoskeletal cylicin II as an actin-binding protein. Biochem. Biophys. Res. Commun. 303:182–189. doi:10.1016/S0006-291X(03)00317-6
Rubin, C. J., H. J. Megens, A. Martinez Barrio, K. Maqbool, S. Sayyab, D. Schwochow, C. Wang, Ö. Carlborg, P. Jern, C. B. Jørgensen, et al. 2012. Strong signatures of selection in the domestic pig genome. Proc. Natl. Acad. Sci. U. S. A. 109:19529–19536. doi:10.1073/pnas.1217149109
Stratil, A., M. Van Poucke, H. Bartenschlager, A. Knoll, M. Yerle, L. J. Peelman, M. Kopecný, and H. Geldermann. 2006. Porcine OGN and ASPN: mapping, polymorphisms and use for quantitative trait loci identification for growth and carcass traits in a Meishan x Piétrain intercross. Anim. Genet. 37:415–418. doi:10.1111/j.1365-2052.2006.01480.x
White, S. 2011. From globalized pig breeds to capitalist pigs: a study in animal cultures and evolutionary history. Environ. Hist. 16:94–120. doi:10.1093/envhis/emq143
Wilkinson, S., Z. H. Lu, H. J. Megens, A. L. Archibald, C. Haley, I. J. Jackson, M. A. Groenen, R. P. Crooijmans, R. Ogden, and P. Wiener. 2013. Signatures of diversifying selection in European pig breeds. PLoS Genet. 9:e1003453. doi:10.1371/journal.pgen.1003453
Wright, W. E., D. A. Sassoon, and V. K. Lin. 1989. Myogenin, a factor regulating myogenesis, has a domain homologous to MyoD. Cell. 56:607–617.doi:10.1016/0092-8674(89)90583–7
Yang, B., L. Cui, M. Perez-Enciso, A. Traspov, R. P. M. A. Crooijmans, N. Zinovieva, L. B. Schook, A. Archibald, K. Gatphayak, C. Knorr, et al. 2017. Genome-wide SNP data unveils the globalization of domesticated pigs. Genet. Sel. Evol. 49:71. doi:10.1186/s12711-017-0345-y