

Genomic Analysis To Identify Signatures of Artificial Selection and Loci Associated with Important Economic Traits in Duroc Pigs

Yunlong Ma,¹ Saixian Zhang,¹ Kaili Zhang, Chengchi Fang, Shengsong Xie, Xiaoyong Du, Xinyun Li, Debin Ni,² and Shuhong Zhao²

Key Laboratory of Agricultural Animal Genetics, Breeding, and Reproduction of the Ministry of Education & Key Laboratory of Swine Genetics and Breeding of the Ministry of Agriculture, Huazhong Agricultural University, Wuhan 430070, P. R. China

ABSTRACT Identifying genetic basis of domestication and improvement in livestock contributes to our understanding of the role of artificial selection in shaping the genome. Here we used whole-genome sequencing and the genotyping by sequencing approach to detect artificial selection signatures and identify the associated SNPs of two economic traits in Duroc pigs. A total of 38 candidate selection regions were detected by combining the fixation index and the Composite Likelihood Ratio methods. Further genome-wide association study revealed seven associated SNPs that were related with intramuscular fat content and feed conversion ratio traits, respectively. Enrichment analysis suggested that the artificial selection regions harbored genes, such as *MSTN*, *SOD2*, *MC5R* and *CD83*, which are responsible for economic traits including lean muscle mass, fertility and immunization. Overall, this study found a series of candidate genes putatively associated with the breeding improvement of Duroc pigs and the polygenic basis of adaptive evolution, which can provide important references and fundamental information for future breeding programs.

KEYWORDS

Duroc pig
Artificial selection signatures
Genome-wide association study
Economic traits

INTRODUCTION

Duroc, an older breed of domestic pig, was developed in America and formed the basis for many mixed-breed commercial boars after a long period of artificial selection. It is predominantly used as the terminal sires in pig industry, and is well known for its superior performance in growth, feed conversion efficiency, carcass and meat quality traits (Briggs and Briggs 1969). The artificial selection affecting these economic traits has left detectable selection signatures within the genome of modern Duroc pigs (Edea *et al.* 2017). Although the identification of selection signatures has been studied for decades

on the basis of microsatellite and SNP arrays, there is now unprecedented opportunity for progress on fine mapping with the advent of large genome sequencing data sets on pig variation (Abbott 2012; Ai *et al.* 2015; Sabeti *et al.* 2002).

In general, most of the economic traits in commercial animals are quantitative and controlled by many genes with small effects (Khatkar *et al.* 2004). This quantitative genetics view is supported by most of the recent gene mapping researches, including the identification of selection signatures and genome-wide association studies (GWAS) (Eusebi *et al.* 2017; Ma *et al.* 2018; Pritchard *et al.* 2010; Ros-Freixedes *et al.* 2014). Therefore, genomic selection has become the main method in genetic improvement in important economic traits (Hayes *et al.* 2009). However, for traits such as meat quality and feed efficiency that are difficult and expensive to measurement, genomic selection is still limited by the sample size of the reference population. So, based on gene mapping, genetic improvement through marker-assisted selection is still an important alternative to improve these traits. Simultaneously, the development of genome selection methods, combined with the information of GWASs, will help further improve the accuracy of genetic evaluation (Zhang *et al.* 2014).

To address the growing demand for high quality pork and minimize pig breeding costs, the understanding of genomic architecture underlying pork quality and feed efficiency is valued by animal scientists

Copyright © 2018 Ma *et al.*

doi: <https://doi.org/10.1534/g3.118.200665>

Manuscript received August 12, 2018; accepted for publication September 10, 2018; published Early Online September 20, 2018.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material available at Figshare: <https://doi.org/10.25387/g3.7077125>.

¹These authors contributed equally to this work

²Correspondence and requests for materials should be addressed to: College of Animal Sciences & Technology, Huazhong Agricultural University, No. 1, Shizishan Street, Hongshan District, Wuhan 430070, People's Republic of China Email: (shzhao@mail.hzau.edu.cn; nicdebin@mail.hzau.edu.cn)

and breeders. Intramuscular fat (IMF) content is one of the most important factors affecting pork quality. Based on marker-assisted selection, there is increasing interest in mapping some of the causal genes of IMF for its genetic improvement (Ros-Freixedes *et al.* 2014). Feed conversion ratio (FCR) is an important indicator for feed efficiency and is usually included in the selection index for genetic improvement in pig breeding (Do *et al.* 2014). Although a series of QTL associated with interesting economic traits were reported based on traditional QTL mapping, it is still scarce to investigate the genetic basis of IMF and FCR using genome-wide association study in pigs (Do *et al.* 2014; Hu *et al.* 2016; Ros-Freixedes *et al.* 2014; Sato *et al.* 2016).

In this study, we used paired-end Illumina sequencing to resequence the genomes of four purebred Durocs and used the genotyping by sequencing approach to characterize the genomes of 282 individuals from the same population. To identify the signatures of artificial selection, sequences from 23 purebred Durocs, 21 Asian wild boars (AWB), and 25 European wild boars (EWB) were downloaded from the most advanced publically available database. Two different statistics, the Composite of Likelihood Ratio (CLR) and the fixation index (F_{ST}), were applied to detect selection signatures. In addition, we also carried out GWAS to identify the key genes related with IMF and FCR that were changed during the breeding improvement of Durocs. We found strong signatures of selection left in Durocs during breed formation exemplified by several striking selective sweeps overlapping with some major QTLs. The results would provide useful information for those who are interested in further understanding the genetic basis of important commercial traits, and facilitate future breeding of Durocs to improve these traits through genomic selection.

MATERIAL AND METHODS

Ethics statement

All research involving animals was conducted under protocols (No. 5 proclaim of the Standing Committee of Hubei People's Congress) approved by the Standing Committee of Hubei People's Congress and the ethics committee of Huazhong Agricultural University in P. R. China. In addition, all experiments were performed in accordance with approved relevant guidelines and regulations.

Animals, genome sequencing, quality checking and filtering

In this study, genomic DNA was extracted from ear tissues of each of 286 purebred Duroc pigs using a standard phenol-chloroform method. Among them, four individuals were selected to sequence the whole genomes. Paired-end sequencing libraries with an insert size around 350-bp were constructed for each sample. 2×150-bp paired-end sequencing was performed on an Illumina HiSeq X-ten platform at the BGI-Huada Genomics Institute in Shenzhen. The rest of individuals were genotyped by sequencing (GBS) on an Illumina HiSeq 2000 platform in Buckler Lab for Maize Genetics and Diversity at Cornell University (Elshire *et al.* 2011). In addition, we also downloaded the whole genome sequence data of 69 individuals from the EMBL-EBI database (<https://www.ebi.ac.uk/>), including 23 Duroc pigs, 21 Asian wild boars and 25 European wild boars (Supplementary Table S1). Quality control of sequence data applied the following criteria: reads with (i) > 10 bp aligned to the adapter with up to 10% mismatches, (ii) up to 10% unidentified nucleotides (N), (iii) > 50% bases having a phred quality less than 5 were removed, and (iv) Duplicate reads generated by PCR amplification in the library construction process were also removed.

Read alignment and SNP calling

Clean reads passing quality control filter were aligned to the *Sus scrofa* reference genome (Sscrofa11.1) using the Burrows-Wheeler Aligner (BWA) software (Li and Durbin 2009). The reference genome sequence was indexed and the command 'mem -t 8' was used to find the suffix array (SA) coordinates of good hits for each individual read. SAMtools was then used to convert the SA coordinates into the best alignments in BAM format (Li *et al.* 2009). After alignment, 'CreateSequenceDictionary', 'SortSam' and 'MarkDuplicates' of Picard were separately used to indexing, sorting and removing potential PCR duplications (Li *et al.* 2009). The BAM files were indexed by Samtools. The 'HaplotypeCaller', 'SelectVariants' and 'VariantFiltration' of GATK with default parameters and SAMtools 'mpileup' module with the parameters as '-q 1 -ugf' were used to call SNPs. Finally, the GATK 'VariantFiltration' module was used to exclude SNP calling errors in according to the following criteria (Mckenna *et al.* 2010). For GBS individuals, joint genotyping was performed by the GATK 'GenotypeGVCFs' module.

High quality SNPs with (i) coverage depth ≥ 4 , (ii) RMS mapping quality ≥ 20 , (iii) the distance of adjacent SNPs ≥ 5 bp, (iv) the missing ratio of samples within each population < 50%, and (v) the minor allele frequency (MAF) ≥ 0.01 were kept for further analysis. The SNPs that were used for GWAS need further apply quality control of Hardy-Weinberg equilibrium ($P < 10^{-6}$). To decrease the influence from genotype imputation, only SNPs with the missing ratio of samples within each population < 20% were used to detect selection signatures. Beagle software was used to impute the missing genotypes and infer haplotypes with default settings (Browning and Browning 2016).

SNP annotation

We functionally annotated Single nucleotide variants (SNVs) with the gene-based annotation modules of ANNOVAR (Wang *et al.* 2010). Using ENSEMBL genes, we investigated where the SNVs are located in the regions of gene components.

Linkage disequilibrium

We calculated the correlation coefficient (r^2) for every pair of SNPs to measure the LD level in Duroc pigs and two wild boar populations using PLINK (Purcell *et al.* 2007). To visualize the LD decay in this analysis, the r^2 values for 1000-bp distance bins were averaged and the corresponding figure was drawn by R script.

Identification of artificial selection signatures

Identification of selection signatures is an important field of population genetics. To detect the selection signals accurately and objectively, a larger sample size is able to better reflect genomic patterns shaped by selection in the particular population. Therefore, all 27 whole-genome sequencing individuals were taken together to detect positive selection signatures referring to the result of principal components analysis (Supplementary Fig. S1). To identify artificial selection signatures, AWB and EWB were defined as the reference population, respectively.

In this analysis, two methods were employed to search for the evidence of artificial selection in two steps. The first step is positive selection detection, which was performed using the composite likelihood ratio (CLR) (Nielsen *et al.* 2005). The CLR method calculates the likelihood ratio of selection signals by comparing the spatial distribution of allele frequencies in an observed window to the frequency spectrum of the whole genome. SweepFinder was employed to calculate the CLR with a grid size of 25 kb resulting in a total of 90,588 CLR scores across

the genome. In the second step, we identified the underlying artificial selection signals from the above detected positive selection signals by calculating the F_{ST} statistic for pairwise sites between the observed populations and the reference population (Weir and Cockerham 1984). The unbiased F_{ST} estimate proposed by Weir and Cockerham was used to measure the population differentiation, with values ranging from 0 (no differentiation) to 1 (complete differentiation). To produce comparable CLR and F_{ST} test results, single site scores for F_{ST} were averaged in non-overlapping windows of 25 kb resulting in a total of 90,322 and 90,411 F_{ST} scores across the genome when AWB and EWB were defined as the reference population, respectively. The empirical P-values were generated by genome wide ranking of F_{ST} and CLR values (Qanbari *et al.* 2014; Simianer *et al.* 2014). Finally, the windows that the scores of the statistics fell into the 98th percentile were considered significant in CLR and F_{ST} methods, respectively. Note that the artificial selection signatures in this study were defined in the genomic region in which both CLR statistical value and the F_{ST} statistical value were greater than the cut-off value at the genome level.

Genome-wide Association studies

In this analysis, all 282 castrated Duroc boars were treated similarly. All individuals were in good health and had the same body weight at the beginning of the experiment. The feed conversion ratio was calculated during the growth period from 30Kg to 100Kg, and at the end of the measurement, the intramuscular fat of each pig was determined by B ultrasound (Supplementary Table S2).

After quality control assessment, autosomes of 282 individuals genotyped by sequencing were used in two association studies for IMF and FCR traits, respectively. GEMMA was used to fit the model $y = u + Ss + Xb + W\alpha + e$, where y is the phenotypes vector, u is the intercept, S is a design matrix of the fixed effects, s is the batches effect, b is the SNP effect, X is a design matrix for allele dosages for the imputed SNPs, W is an incidence matrix linking α to y , α is the additive genetic effect $\sim N(0, G\sigma_a^2)$, where σ_a^2 is the additive genetic variance and G is the realized genomic relationship matrix that was estimated using genotype information, and e is the random residual term. To avoid double fitting of SNP effects efficiently, the test SNP and the other SNPs in the same chromosome were removed in constructing the G matrix each time based on the standardized relatedness matrix in GEMMA (Zhou and Stephens 2012).

Since Bonferroni correction is overly conservative especially when genetic data has high linkage disequilibrium, it may cause false negative results (Duggal *et al.* 2008). Therefore, a less conservative significance threshold of 1.03×10^{-5} (0.05/4,853) based on the SimpleM method was used to account for multiple tests in this analysis. A total of 4,853 independent tests were identified here that was in turn inferred by the number of principal components accounting for a 99% of the variance of the SNP matrix (Gao *et al.* 2008).

Enrichment analysis for artificial selection signatures

Enrichment analysis was carried out for exploring the potential biological functions of genes located in putative artificial selection regions. It is involved all the selected genes in the 200 kb window around the significant signatures, which was determined by the linkage disequilibrium decay (Supplementary Fig. S2). Genes located in putatively selected regions were identified using the BioMart program (<http://www.biomart.org/>, Kasprzyk 2011), and then an enrichment analysis, which included the terms cellular component (CC), molecular function (MF), biological process (BP), and pathway analysis was performed for the identified genes using DAVID 6.7 (<http://david.abcc.ncifcrf.gov/>) (Huang *et al.* 2009).

Data availability

The Illumina sequence reads are available in the NCBI Sequence Read Archive under the accession SRP158574. Supplemental files, the GBS data and phenotype data can be downloaded from Figshare (<https://figshare.com/s/e45cc6d717cd498d5013>). Supplemental material available at Figshare: <https://doi.org/10.25387/g3.7077125>.

RESULTS AND DISCUSSION

Genome resequencing and genetic variation

After quality control, a total of 127.98 Gb of sequence data were generated on the basis of our four Duroc pigs, with the average sequencing depth of approximately 13-fold and the average genome coverage of 99.03%. As shown in Table 1, we identified 9,245,511 SNPs with an average density of 3.76 SNPs/kb. Then, we compared these SNPs that we found with those from the SNPs database that was built using 23 downloaded Duroc pigs. About 88% of the variants (8,205,625 SNPs) in our SNP data set were found in the SNPs database, whereas more than 11% (1,039,886 SNPs) of the variants that we identified were absent from the SNPs database (Supplementary Fig. S3). These novel SNPs substantially expand the database of Duroc genetic variants. Therefore, combining all 27 Duroc individuals together not only increased the sample size, but also significantly increased the SNPs density of the genome. The previous study indicates that a high marker density has positive effects on the identification of selection signatures (Ma *et al.* 2015).

Combined, a total of 14,827,549 SNPs with an average density of 6.03 SNPs/kb were detected using 27 sequencing Duroc individuals. Among them, 8,863,149 were located in intergenic regions, 5,496,375 were in intronic regions and 111,759 were in exonic regions. Similarly, 45.49 Gb of sequence data were generated by GBS, with the population sequencing depth of approximately 17-fold and about 0.56% of the bases in the reference genome being covered by at least one reads. A total of 651,425 SNPs were identified in the 282 pig genomes, including 144 intergenic SNPs, 192 intronic SNPs and 16,332 exonic SNPs (Table 1). The SNPs identified by GBS were not evenly distributed across the genome, with an average density of 0.27 SNPs/kb (Supplementary Fig. S4). Similar genome annotation information of EWB and AWB herein is illustrated in Supplementary Table S3.

Genome-wide artificial selection signatures

To detect positive selection, the CLR scores were calculated using an identical grid size across the genome. We focused the analyses on windows for which the scores of the statistics fell in the top 2nd percentile. As shown in Supplementary Fig. S5, the threshold value is almost greater than the largest CLR scores in the wild populations. In general, we expected a limited amount of artificial selection for commercial traits occurred in the wild populations. Out of 90,588 sliding windows, 1,811 windows were identified as the potential selection signatures in the Duroc population. Then, the windows within a 200kb fragment around the potential selection signatures are merged and the genomic regions were defined as candidate selection regions (CSRs). Correspondingly, a total of 70 fragments, spanning lengths of 81.40 Mb and covering 3.3% of the genome, were identified as CRSs in Duroc pigs (Figure 1A).

To further identify the artificial selection signatures, F_{ST} statistics were used to detect the differentiation between Duroc and wild pigs. We calculated F_{ST} per site and averaged them in non-overlapping 25 kb windows across the genome. The windows that the scores of the statistics fell in the top 2nd percentile were considered as significant. Out of 90,322 sliding windows, 1,806 windows were identified as the potential selection signatures and fall into 758 CRSs when AWB was treated

■ **Table 1** Summary and annotation of SNPs in Duroc pigs

| Category | | WGS | GBS |
|-----------------------------|-------------------|------------------|-------------------|
| Sample size | 27 | 4 | 23 |
| Average depth (X) | 12.10 | 12.79 | 11.98 |
| Average genome coverage (%) | 96.56 | 99.03 | 96.13 |
| Average Mapping rate (%) | 99.67 | 99.66 | 99.67 |
| High-quality base (Gb) | 817.35 | 127.98 | 689.37 |
| Q20 (%) | 97.36 | 95.23 | 97.73 |
| Q30 (%) | 91.18 | 89.67 | 91.44 |
| Number of total SNP | 14,827,549 | 9,245,511 | 13,787,663 |
| Upstream | 92,785 | 56,487 | 85,634 |
| UTR5 | 28,436 | 18,070 | 26,101 |
| Exonic | | | |
| Stopgain | 545 | 265 | 504 |
| Stoploss | 117 | 65 | 114 |
| Synonymous | 66,285 | 41,379 | 61,367 |
| Nonsynonymous | 44,660 | 26,352 | 41,243 |
| Unknown | 152 | 33 | 144 |
| Splicing | 614 | 397 | 565 |
| Intronic | 5,496,375 | 3,459,359 | 5,108,598 |
| UTR3 | 133,194 | 82,605 | 124,775 |
| UTR5/UTR3 | 702 | 454 | 679 |
| Downstream | 98,207 | 59,700 | 91,145 |
| Upstream/downstream | 2,328 | 1,455 | 2,164 |
| Intergenic | 8,863,149 | 5,498,890 | 8,244,630 |

¹the population sequencing depth.

as reference population. Similarly, Out of 90,411 sliding windows, 1,804 windows were identified as the potential selection signatures and fall into 684 CRSs when EWB was treated as reference population (Figure 1B).

In this analysis, the overlapping candidate selection regions detected by CLR and F_{ST} were defined as candidate artificial selection regions

(CASR). Collectively, a total of 38 CASRs, spanning lengths of 40.48 Mb and covering 1.62% of the genome, were identified in Duroc pigs through compared with the wild populations (Figure 1C, Supplementary Table S4). In general, the adaptive evolution of important traits, due to human-driven or natural select, would leave a number of selection signatures in genomic regions, where should

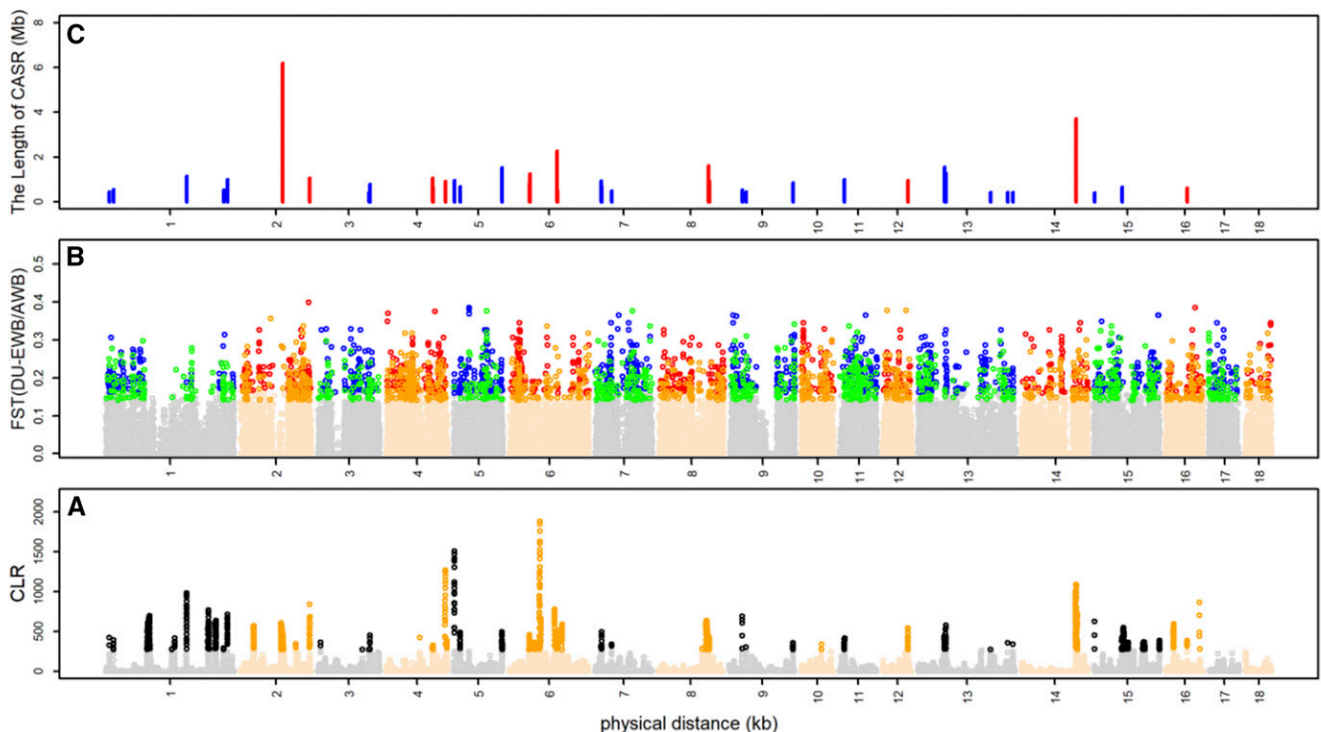


Figure 1 Summary of the genomic regions underlying artificial selection in Duroc pigs. (A) Manhattan plot based on CLR tests. (B) Manhattan plot based on F_{ST} tests, the points in blue and red (black and orange) represent scores when EWB (AWB) was treated as reference population. (C) The lines illustrate the positions and lengths of candidate artificial selection regions (CASR).

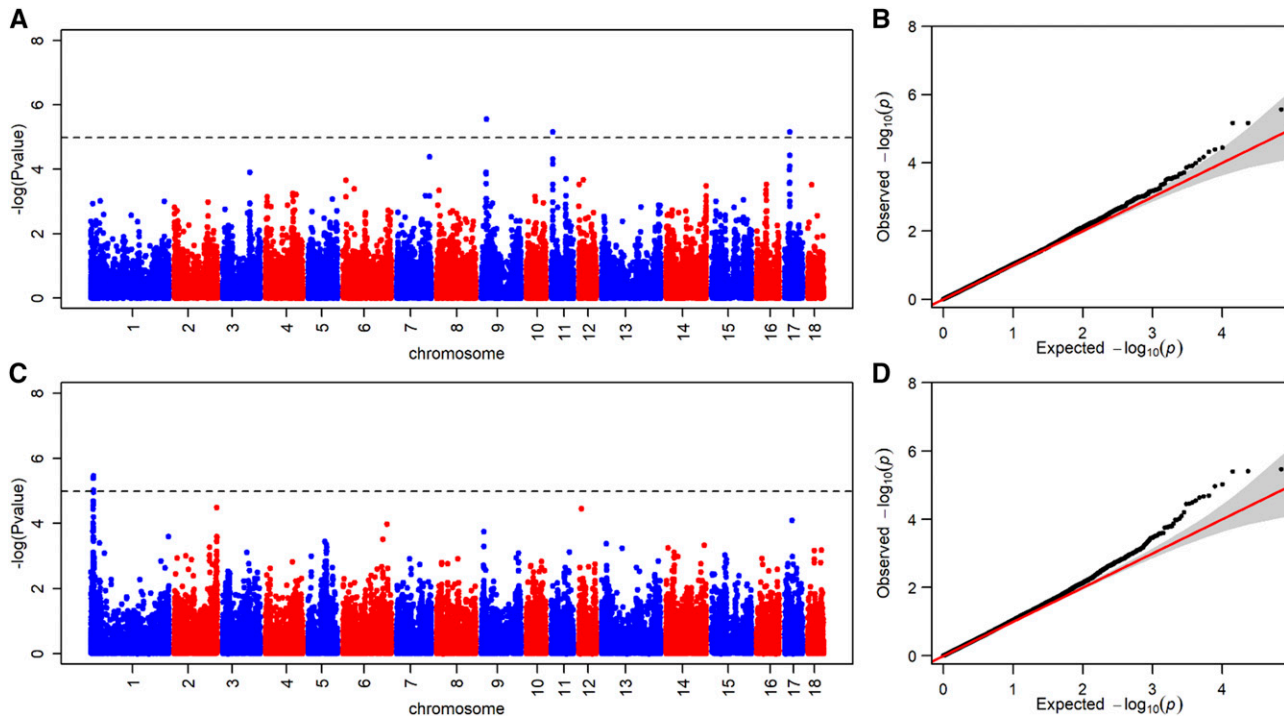


Figure 2 Visualization of the signals revealed by genome-wide association analyses for IMF and FCR traits. (A, C) Manhattan plots present the association of the imputed SNPs with the IMF and FCR traits in 282 Duroc individuals, respectively. (B, D) The corresponding quantile-quantile plots.

harbor the corresponding causal genes. Therefore, the candidate selection regions should be major enriched in genic regions. To test this hypothesis, we identified a total of 246,868 SNPs in the 38 candidate artificial selection regions, of which 143,978 (57.918%) were intergenic, 97,287 (39.136%) were intronic and 1,924 (0.78%) were exonic. Comparing the distribution of SNPs between candidate artificial selection regions and the whole genome, it suggested that there was no predictable pattern in where those selection regions were located (Supplementary Table. S5). This seems to indicate that selection not only plays a role in the gene regions, but also in other regulatory elements of the genome.

Candidate selection regions harboring loci associated with two economic traits

Genome-wide association studies have already been proven effective to reveal the underlying genetics of economic traits (Meyer *et al.* 2016; Visscher *et al.* 2017). If the underlying genetic basis of those domestication traits has been improved by recent artificial selection, the genomic regions that harbored the QTNs revealed through GWAS should overlap with the candidate selection regions. In this analysis, we performed mixed-model association analysis using GBS data of 35,303 high-quality markers. Although the small sample size limited our power, we identified seven associated loci with a less conservative significance threshold ($-\log_{10}(p\text{-value}) \geq 4.98$), including four SNPs for IMF trait and three SNP for FCR trait (Figure 2). Comparing seven associated loci with the candidate selection region, we found that a series of selection signatures were located around the associated loci for both traits (Supplementary Table S6).

Because of concerns that the sample size for GWAS is too small to make meaningful scientific conclusion, we investigated the results of GWAS for IMF and FCR traits using a random sampling of 250 individuals, and the process is repeated 10 times. Among them, the similar

results of GWAS for IMF and FCR traits have been replicated with smaller sample size, respectively (Supplementary Fig. S6, S7). In addition, we further check genotype-class frequencies and phenotypic means for seven significant SNPs. As shown in Supplementary Table S7, we can clearly observe the trend of phenotypic changes with three genotypes. The results suggested that the SNPs passing the less conservative significance threshold are still promising in this study.

For FCR trait, an associated SNP was close to the 16.375-16.425 Mb selection region on SSC9 and the QTL of ‘time spent feeding’ was reported to be overlapping with this region. For IMF trait, 4 associated SNPs were found close to the 7.1-9.0 Mb selection regions on SSC1. After scanning the pig QTLdb (<https://www.animalgenome.org>), we found that these four SNPs were also overlapping with ‘the drip loss’ and ‘the stearic acid content’, respectively (Hu *et al.* 2016).

We highlighted the 7.00-10.00 Mb regions around the most significant SNP on SSC1 (Figure 3), where a series of CLR and F_{ST} scores exceeded the significance threshold and several harboring genes, such as *IGF2R*, *TMEM181*, *SOD2* and *TAGAP*, were responsible for sex determination, growth, muscle and bone development (Supplementary Table. S8). Among them, one possible positional candidate gene, insulin like growth factor 2 receptor (*IGF2R*), serves *IGF2* turnover in IGF signal mediated process (Ludwig *et al.* 1996) and *IGF2* is known as a major gene that influences the meat quality of pigs (Clark *et al.* 2014).

Go terms, pathways and candidate genes for artificial selection

To further investigate the genetic basis of improved economic traits, a total of 371 genes overlapping with all 38 candidate artificial selection regions in Duroc pigs were found and the corresponding orthologous genes from human were used to perform an enrichment analysis by DAVID 6.7 (<https://david.ncifcrf.gov/>) (Huang *et al.* 2009). The results indicated that genes related with a number of terms previously

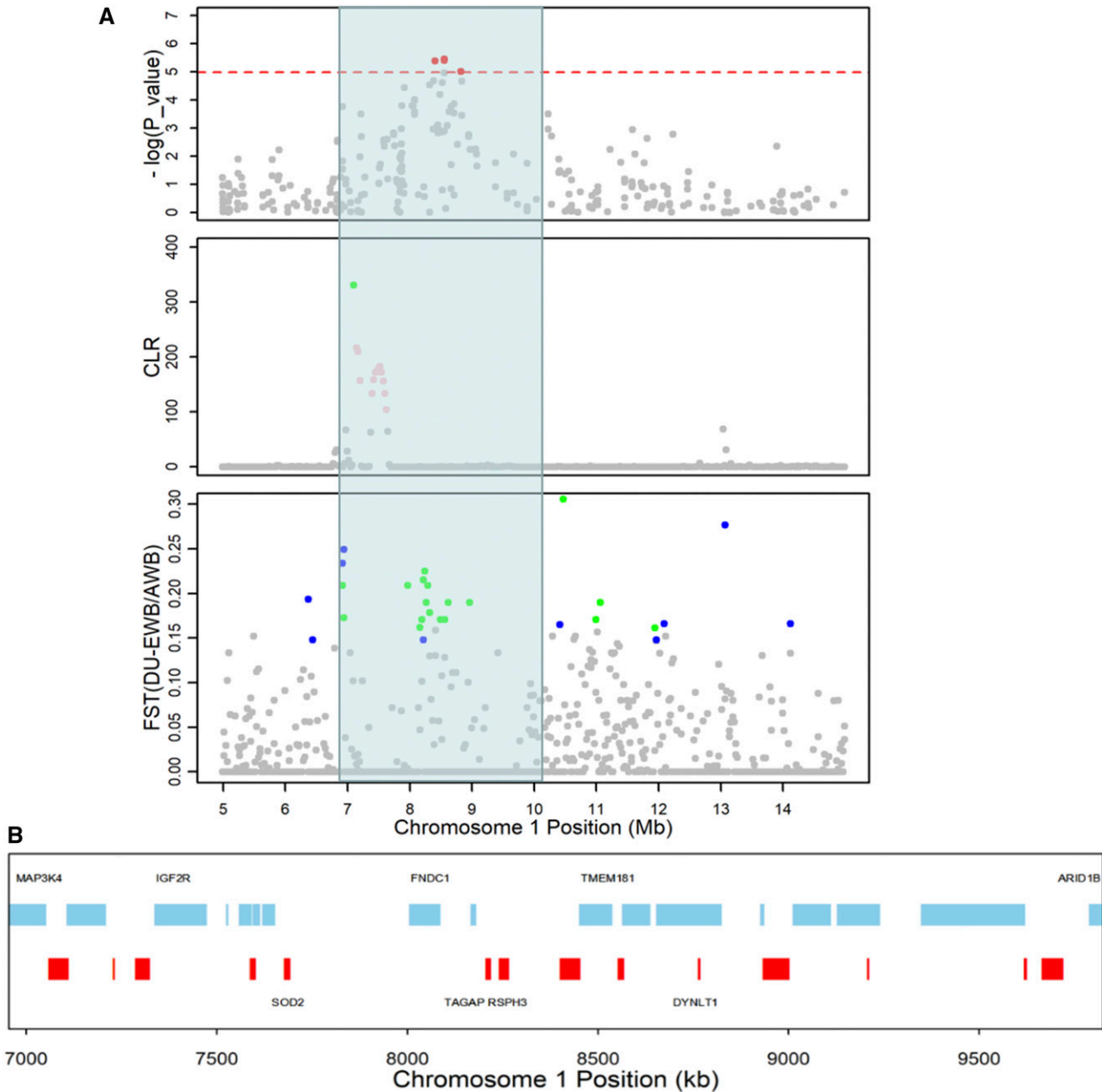


Figure 3 Comparison of GWAS and candidate artificial selection regions (CASR) on SSC1 in Duroc. (A) The top Manhattan plot shows GWAS results for IMF trait. The middle plot shows the distribution of CLR scores calculated in sliding windows. The bottom plot is the distribution of F_{ST} scores when AWB (blue) and EWB (green) were treated as reference population, respectively. The red line is the significance threshold. (B) The positions of 31 gene models in this region are indicated: skyblue, minus-strand genes; red, positive-strand genes. The detailed gene list can be found in Supplementary Table S8.

implicated in breed improvement were present within or close to these sweep regions.

Among them, a set of genes overlapping artificial selection regions were enriched for the ability of immunity, which mainly included ‘GO:0042267~natural killer cell mediated cytotoxicity’ (6 genes), ‘GO:0019882~antigen processing and presentation’ (6 genes) and ‘hsa04650:Natural killer cell mediated cytotoxicity’ (6 genes) (Supplementary Table S9). Additionally, we also found one significant candidate artificial selection regions that overlapped with *CD83* gene, which contributes to T lymphocyte proliferation (Pinho *et al.* 2014) (Table 2). Due to the high density and selection intensity of

rearing animals in the breeding farm, the artificial selection effect is undoubtedly to play an indispensable role on selecting innate immunity genes indirectly. Six genes with strong selection signatures are significantly enriched in ‘GO:0043651~linoleic acid metabolic process’ and the *LPIN2* gene that plays a significant role in fat deposition was found to fall into the 102-104.275 Mb artificial selection regions on SSC6 in Duroc, suggesting that these genes may have been selected during the breeding of lean pigs (Table S2, S9).

In addition to those significant Go and Pathway terms, a set of artificial selection signatures with extreme *P*-value coincide with a cluster of genes involved in meat quality, growth, fertility and so

■ **Table 2** Some candidate genes overlap with the potential regions of artificial selection in Duroc pigs

| Chr. | Pos. (Mb) ¹ | P-value. (method) ² | Gene | Gene function |
|------|------------------------|--------------------------------|--------|--|
| 1 | 7679352..7691724 | 0.01;0.001(AWB); 0.007(EWB) | SOD2 | Fertility (Kwiatkowska <i>et al.</i> 2017) |
| 2 | 51102757..51128346 | 0.048;0.008(EWB) | WNT9A | Chondrogenesis (Später <i>et al.</i> 2006) |
| 2 | 88122788..88255282 | 0.005;<0.001(AWB) | HOMER1 | Muscle development (Hao <i>et al.</i> 2017) |
| 2 | 88021572..88098522 | 0.007;<0.001(AWB) | JMY | Development of porcine embryos (Lin <i>et al.</i> 2015) |
| 2 | 88391280..88507344 | 0.004;<0.001(AWB) | CMYA5 | Carcass trait and meat quality (Xu <i>et al.</i> 2011) |
| 6 | 96322276..96323795 | 0.003 | MC5R | Back fat thickness, lipid metabolism, exocrine function, proinflammatory activity (Switonski <i>et al.</i> 2013) |
| 6 | 103997590..104055816 | 0.009;0.007(AWB) | NDC80 | Fertility (Wei <i>et al.</i> 2014) |
| 6 | 103578544..103728152 | 0.008;0.007(AWB) | LPIN2 | Back-fat thickness (He <i>et al.</i> 2009) |
| 7 | 10359211..10387002 | 0.017;0.005(AWB); 0.006(EWB) | CD83 | Enhances T lymphocyte proliferation (Pinho <i>et al.</i> 2014) |
| 7 | 75161831..75168634 | 0.027;0.001(AWB); 0.004(EWB) | FITM1 | Fat-deposition-related traits (li <i>et al.</i> 2010) |
| 7 | 75241058..75255701 | 0.027;0.001(AWB); 0.004(EWB) | DHRS4 | Meat quality (Hwang <i>et al.</i> 2017) |
| 9 | 35187561..35200085 | 0.032;<0.001(AWB);0.006(EWB) | CASP1 | Fertility (Ashworth <i>et al.</i> 2010) |
| 12 | 50454063..50491164 | 0.036;0.007(AWB) | SPNS2 | Hearing (Chen <i>et al.</i> 2014) |
| 15 | 94623526..94628440 | 0.046, 0.009(EWB) | MSTN | Lean muscle mass (Baati <i>et al.</i> 2017) |

¹This column presents the position of candidate genes which overlap with or close to the potential regions of artificial selection.

²This column presents the genome-wide P-values of sweep statistics.

on (Table 2). An interesting selection signature was located on SSC15 (104.325-104.8 Mb) and close to the *MSTN* gene, which was associated with lean muscle mass and played an important role in the process of muscle development (Baati *et al.* 2017). As a commercial pig breed of economic importance, leanness has consistently been considered as an objective trait of Durocs breeding. Correspondingly, growth rates and meat quality would also be considered and designed in the breeding program. In this study, *WNT9A* gene associated with chondrogenesis and *CMYA5* gene associated with carcass length and meat quality were overlapping with the strong selection signatures. *MC5R* gene, an established sweep in previous studies, was found close to the 97.625-98.725 Mb regions on SSC6, which was not only associated with back fat thickness but also played an important role in proinflammatory activity (Switonski *et al.* 2013). The *HOMER1* gene that is overlapped with the 86.55-92.75 Mb artificial selection regions on SSC2 is associated with muscle differentiation and calcium homeostasis (Hao *et al.* 2017). Note that a series of genes, including *SOD2*, *JMY*, *NDC80* and *CASP1*, overlapping with artificial selection regions in this analysis were associated with the reproductive traits. In comparing with its wild ancestors, litter size, fetal weight and the other fertility traits would change for adapting human-driven selection. These results indicated that the improvement of fertility and other complex traits may due to the polygenic basis, rather than be caused by only a few critical loci.

CONCLUSIONS

In summary, this study detected signatures of artificial selection and identified a number of loci associated with some important economic traits, like IMF and FCR, which provided an important resource in future Durocs breeding programs. SNP annotation implied that there was no predictable distribution in where those artificial selection regions were located in genome. Enrichment analyses suggested that the polygenic basis may be a reasonable explanation for the phenomenon, that many different genes associated with the economic traits have often been selected during the improvement and breed formation. The application of combining sweep analysis and genome-wide association analysis are effective in mapping commercial important genes, especially to the data with small sample size caused by the expensive measurement traits.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 31601916), the NSFC-CGIAR project (31361140365), the

Fundamental Research Funds for the Central Universities (No. 2662015QD018), the National Swine Industry Technology System (No. CARS-35) and SRF (No. 2017046). Yunlong Ma, Debin Ni and Shuhong Zhao conceived and designed the experiments; Shengsong Xie and Xinyun Li performed the experiments; Yunlong Ma, Saixian Zhang, Kaili Zhang, Xiaoyong Du and Chengchi Fang analyzed the data; Shuhong Zhao contributed reagents/materials/analysis tools; Yunlong Ma wrote the paper. All authors contributed to revision and review of the manuscript.

LITERATURE CITED

- Abbott, A., 2012 Pig geneticists go the whole hog. *Nature* 491: 315–316. <https://doi.org/10.1038/491315a>
- Ai, H. S., X. D. Fang, B. Yang, Z. Y. Huang, H. Chen *et al.*, 2015 Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat. Genet.* 47: 217–225. <https://doi.org/10.1038/ng.3199>
- Ashworth, M. D., J. W. Ross, D. R. Stein, F. J. White, U. W. DeSilva *et al.*, 2010 Endometrial caspase 1 and interleukin-18 expression during the estrous cycle and peri-implantation period of porcine pregnancy and response to early exogenous estrogen administration. *Reprod. Biol. Endocrinol.* 8: 33. <https://doi.org/10.1186/1477-7827-8-33>
- Baati, N., C. Feillet-Coudray, G. Fouret, B. Vernus, B. Goustard *et al.*, 2017 Myostatin deficiency is associated with lipidomic abnormalities in skeletal muscles. *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids* 1862: 1044–1055. <https://doi.org/10.1016/j.bbalip.2017.06.017>
- Briggs, H. M., and D. M. Briggs, 1969 *Modern breeds of livestock*, Macmillan, London.
- Browning, B. L., and S. R. Browning, 2016 Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 98: 116–126. <https://doi.org/10.1016/j.ajhg.2015.11.020>
- Chen, J., N. Ingham, J. Kelly, S. Jadeja, D. Goulding *et al.*, 2014 Spinster Homolog 2 (Spns2) Deficiency Causes Early Onset Progressive Hearing Loss. *PLoS Genet.* 10: e1004688. <https://doi.org/10.1371/journal.pgen.1004688>
- Clark, D. L., B. M. Bohrer, M. A. Tavarez, D. D. Boler, J. E. Beever *et al.*, 2014 Effects of the porcine IGF2 intron 3-G3072A mutation on carcass cutability, meat quality, and bacon processing. *J. Anim. Sci.* 92: 5778–5788. <https://doi.org/10.2527/jas.2014-8283>
- Do, D. N., T. Ostensen, A. B. Strathe, T. Mark, J. Jensen *et al.*, 2014 Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. *BMC Genet.* 15: 27. <https://doi.org/10.1186/1471-2156-15-27>

- Duggal, P., E. M. Gillanders, T. N. Holmes, and J. E. Bailey-Wilson, 2008 Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics* 9: 516. <https://doi.org/10.1186/1471-2164-9-516>
- Edea, Z., J. K. Hong, J. H. Jung, D. W. Kim, Y. M. Kim *et al.*, 2017 Detecting selection signatures between Duroc and Duroc synthetic pig populations using high-density SNP chip. *Anim. Genet.* 48: 473–477. <https://doi.org/10.1111/age.12559>
- Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS One* 6: e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Eusebi, P. G., R. Gonzálezprez, R. Quintanilla, J. Tibau, T. F. Cardoso *et al.*, 2017 A genome-wide association analysis for carcass traits in a commercial Duroc pig population. *Anim. Genet.* 48: 466–469. <https://doi.org/10.1111/age.12545>
- Gao, X. Y., J. Stamier, and E. R. Martin, 2008 A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* 32: 361–369. <https://doi.org/10.1002/gepi.20310>
- Hao, X. J., G. Plastow, C. Y. Zhang, S. T. Xu, Z. Q. Hu *et al.*, 2017 Genome-wide association study identifies candidate genes for piglet splay leg syndrome in different populations. *BMC Genet.* 18: 64. <https://doi.org/10.1186/s12863-017-0532-4>
- Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009 Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92: 433–443. <https://doi.org/10.3168/jds.2008-1646>
- He, X. P., X. W. Xu, and B. Liu, 2009 Molecular characterization, chromosomal localization and association analysis with back-fat thickness of porcine LPIN2 and LPIN3. *Mol. Biol. Rep.* 36: 1819–1824. <https://doi.org/10.1007/s11033-008-9385-2>
- Hu, Z. L., C. A. Park, and J. M. Reecy, 2016 Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res.* 44: D827–D833. <https://doi.org/10.1093/nar/gkv1233>
- Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4: 44–57. <https://doi.org/10.1038/nprot.2008.211>
- Hwang, J. H., S. M. An, S. G. Kwon, D. H. Park, T. W. Kim *et al.*, 2017 Associations of the Polymorphisms in DHRS4, SERPING1, and APOR Genes with Postmortem pH in Berkshire Pigs. *Anim. Biotechnol.* 28: 288–293. <https://doi.org/10.1080/10495398.2017.1279171>
- Khatkar, M. S., P. C. Thomson, I. Tammen, and H. W. Raadsma, 2004 Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genet. Sel. Evol.* 36: 163–190. <https://doi.org/10.1186/1297-9686-36-2-163>
- Kwiatkowska, J., B. Wąsowska, and P. Gilun, 2017 Expression of hypoxia inducible factor 1 α and antioxidant enzymes: Superoxide dismutases-1 and -2 in ischemic porcine endometrium. *Reprod. Biol.* 17: 289–293. <https://doi.org/10.1016/j.repbio.2017.05.013>
- Li, D. Z., J. X. He, M. G. Lei, D. Q. Xu, S. W. Jiang *et al.*, 2010 Polymorphism in exon 2 of pig FIT1 gene and its association with fat-deposition-related traits. *Yi Chuan* 32: 375–380. <https://doi.org/10.3724/SP.J.1005.2010.00375>
- Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., B. Handsaker, A. Wysoker, J. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Lin, Z. L., X. S. Cui, S. Namgoong, and N. H. Kim, 2015 Junction-mediating and regulatory protein (JMY) is essential for early porcine embryonic development. *J. Reprod. Dev.* 61: 361–367. <https://doi.org/10.1262/jrd.2015-021>
- Ludwig, T., J. Eggenschwiler, P. Fisher, A. J. D. Ercole, M. L. Davenport *et al.*, 1996 Mouse mutants lacking the type 2 IGF receptor (IGF2R) are rescued from perinatal lethality in Igf2 and Igf1r null backgrounds. *Dev. Biol.* 177: 517–535. <https://doi.org/10.1006/dbio.1996.0182>
- Ma, Y., X. Ding, S. Qanbari, S. Weigend, Q. Zhang *et al.*, 2015 Properties of different selection signature statistics and a new strategy for combining them. *Heredity* 115: 426–436. <https://doi.org/10.1038/hdy.2015.42>
- Ma, Y. L., L. T. Gu, L. B. Yang, C. H. Sun, S. S. Xie *et al.*, 2018 Identifying artificial selection signals in the chicken genome. *PLoS One* 13: e0196215. <https://doi.org/10.1371/journal.pone.0196215>
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Meyer, R. S., J. Y. Choi, M. Sanches, A. Plessis, J. M. Flowers *et al.*, 2016 Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat. Genet.* 48: 1083–1088. <https://doi.org/10.1038/ng.3633>
- Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. *Genome Res.* 15: 1566–1575. <https://doi.org/10.1101/gr.4252305>
- Pinho, M. P., I. K. Migliori, E. A. Flatow, and J. A. Barbuto, 2014 Dendritic cell membrane CD83 enhances immune responses by boosting intracellular calcium release in T lymphocytes. *J. Leukoc. Biol.* 95: 755–762. <https://doi.org/10.1189/jlb.0413239>
- Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Curr. Biol.* 20: R208–R215. <https://doi.org/10.1016/j.cub.2009.11.055>
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira *et al.*, 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81: 559–575. <https://doi.org/10.1086/519795>
- Qanbari, S., H. Pausch, S. Jansen, M. Somel, T. M. Strom *et al.*, 2014 Classic Selection Sweeps Revealed by Massive Sequencing in Cattle. *PLoS Genet.* 10: e1004148. <https://doi.org/10.1371/journal.pgen.1004148>
- Ros-Freixedes, R., S. Gol, R. N. Pena, M. Tor, J. C. M. Dekkers *et al.*, 2014 Genome-wide association study for intramuscular fat content and composition in Duroc pigs, pp. in *World Congress of Genetics Applied To Livestock Production Wcgalp*.
- Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837. <https://doi.org/10.1038/nature01140>
- Sato, S., Y. Uemoto, T. Kikuchi, S. Egawa, K. Kohira *et al.*, 2016 SNP- and haplotype-based genome-wide association studies for growth, carcass, and meat quality traits in a Duroc multigenerational population. *BMC Genet.* 17: 60. <https://doi.org/10.1186/s12863-016-0368-3>
- Simianer, H., Y. Ma, and S. Qanbari, 2014 Statistical Problems in Livestock Population Genomics, pp. in *World Congress on Genetics Applied To Livestock Production*.
- Später, D., T. P. Hill, R. J. O’Sullivan, M. Gruber, D. A. Conner *et al.*, 2006 Wnt9a signaling is required for joint integrity and regulation of Ihh during chondrogenesis. *Development* 133: 3039–3049. <https://doi.org/10.1242/dev.02471>
- Switonski, M., M. Mankowska, and S. Salamon, 2013 Family of melanocortin receptor (MCR) genes in mammals—mutations, polymorphisms and phenotypic effects. *J. Appl. Genet.* 54: 461–472. <https://doi.org/10.1007/s13353-013-0163-z>
- Visser, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy *et al.*, 2017 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101: 5–22. <https://doi.org/10.1016/j.ajhg.2017.06.005>
- Wang, K., M. Y. Li, and H. Hakonarson, 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38: e164. <https://doi.org/10.1093/nar/gkq603>
- Wei, X. M., C. H. Gao, J. Luo, W. Zhang, S. H. Qi *et al.*, 2014 Hec1 inhibition alters spindle morphology and chromosome alignment in

- porcine oocytes. *Mol. Biol. Rep.* 41: 5089–5095. <https://doi.org/10.1007/s11033-014-3374-4>
- Weir, B. S., and C. C. Cockerham, 1984 Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38: 1358. <https://doi.org/10.1111/j.1558-5646.1984.tb05657.x>
- Xu, X. L., X. W. Xu, Q. Yin, L. Sun, B. Liu *et al.*, 2011 The molecular characterization and associations of porcine cardiomyopathy associated 5 (CMYA5) gene with carcass trait and meat quality. *Mol. Biol. Rep.* 38: 2085–2090. <https://doi.org/10.1007/s11033-010-0334-5>
- Zhang, Z., J. He, H. Zhang, P. Gao, M. Erbe *et al.*, 2014 Results of Genome Wide Association Studies Improve the Accuracy of Genomic Selection, pp. in *World Congress on Genetics Applied To Livestock Production*.
- Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44: 821–824. <https://doi.org/10.1038/ng.2310>

Communicating editor: D. J. de Koning