ZOOLOGICAL RESEARCH

OrthReg: a tool to predict *cis*-regulatory elements based on cross-species orthologous sequence conservation

DEAR EDITOR,

Cis-regulatory elements play an important role in the development of traits and disease in organisms (Ma et al., 2020; Woolfe et al., 2005) and their annotation could facilitate genetic studies. The Encyclopedia of DNA Elements (ENCODE) (Davis et al., 2018) and Functional Annotation of Animal Genomes (FAANG) (FAANG Consortium et al., 2015) offer pioneering data on regulatory elements in several species. Currently, however, regulatory element annotation data remain limited for most organisms. In this study, we developed a tool (OrthReg) for annotating conserved orthologous *cis*-regulatory elements in targeted genomes using an annotated reference genome. Cross-species validation of this annotation tool using human and mouse ENCODE data confirmed the robustness of this strategy. To explore the efficiency of the tool, we annotated the pig genome and identified more than 28 million regulatory annotation records using the reference human ENCODE data. With this regulatory annotation, some putative regulatory non-coding variants were identified within domestication sweeps in European and East Asian pigs. Thus, this tool can utilize data produced by ENCODE, FAANG, and similar projects, and can be easily extended to customized experimental data. The extensive application of this tool will help to identify informative single nucleotide polymorphisms (SNPs) in post-genome-wide association studies and resequencing analysis of organisms with limited regulatory annotation data.

In recent years, the application of next-generation sequencing technologies has identified a vast number of candidate variants related to different traits and diseases in organisms (Axelsson et al., 2013; Ma et al., 2019; Rubin et al., 2012; Visscher et al., 2017; Yang et al., 2018, 2019). More than 90% of these variants lie within non-coding sequences and are enriched in various signals associated with transcriptional regulation (Maurano et al., 2012). However, the identification and annotation of these variants in regulatory non-coding elements in genomes remain a challenging task.

To identify regulatory elements in genomes, the ENCODE and FAANG projects were launched to annotate genomes in several organisms. However, regulatory element annotation is still very limited for other organisms (except for humans and mice), which has hindered genetic studies from exploring the role of regulatory variants in the development of diverse traits. The “phylogenetic footprinting” of *cis*-regulatory elements suggests that it may be possible to identify regulatory sequences across species (Gumucio et al., 1992). Many different types of non-coding regulatory elements are conserved under strong evolutionary constraints in a variety of organisms (Bejerano et al., 2004; Cooper et al., 2005; Drake et al., 2006; Siepel et al., 2005), offering an opportunity to identify putative regulatory elements in species by comparative genomics.

In this study, we developed OrthReg for annotating regulatory elements in targeted genomes based on orthologous sequence conservation of regulatory elements in reference genomes (see Supplementary Notes for details). OrthReg can generate abundant regulatory element annotation for targeted organisms by utilizing data provided by ENCODE, FAANG, and similar projects.

We first assessed the reliability of OrthReg using annotation information from the mouse and human genomes in ENCODE. To estimate OrthReg performance, human ENOCDE data were used as a reference to predict the

Received: 26 April 2020; Accepted: 01 June 2020; Online: 11 June 2020

Foundation items: This work was supported by the Chinese Academy of Sciences (XDA24010107), Ministry of Agriculture of China (2016ZX08009003-006), National Natural Science Foundation of China (31621062), Funding for Open Access Charge: Ministry of Agriculture of China (2016ZX08009003-006), and Animal Branch of the Germplasm Bank of Wild Species, Chinese Academy of Sciences (Large Research Infrastructure Funding)

DOI: 10.24272/j.issn.2095-8137.2020.099

regulatory elements in the mouse genome, with the predicted results then validated by the mouse ENCODE data. To avoid the influence of data from different experimental assays and tissues, 64 matched datasets representing the same ChIP-seq experiments and tissues were selected in both species, resulting in a coverage of 12 tissues and 11 ChIP-seq assays (Supplementary Table S1). Generally, OrthReg showed good predictive performance for most ChIP-seq signals and tissue types. On average, the precision of the predicted regulatory elements for mice was 47.67% (sequence similarity >0.5) (Figure 1A), consistent with earlier observations on promoter sequence conservation between humans and rodents (Dermitzakis & Clark, 2002). The highest precision was 90.85% for H3K4me3 in the lung and the lowest was 1.12% for H3K9me3 in the lung (sequence similarity >0.5). The 64 matched datasets included five H3K9me3 datasets, which were all found in the bottom seven predictions with lowest precision (Figure 1A). H3K9me3 is enriched in repeat-rich regions of constitutive heterochromatin (Nakayama et al., 2001), and the low precision in our prediction may have resulted from less functional conservation of such sequences between humans and mice. In addition, reproductive- (testis and placenta) and immune-related (thymus and spleen) tissues showed relatively low precision of predictions compared to the lung, liver, heart, and kidney (Supplementary Figure S1A, C). If the H3K9me3 and reproductive/immune-related tissues were excluded, the average precision (for 38 datasets) increased to 60.28%. Discovery ratio analysis also revealed a weak prediction robustness in the H3K9me3 and reproductive/immune-related tissues (Figure 1B and Supplementary Figure S1B, D). Our results indicated that OrthReg could provide considerably reliable predictions, except for the reproductive and immune-related tissues and H3K9me3 signals. In addition, we found that a higher level of sequence identity resulted in higher precision and discovery ratio (Figure 1A, B), consistent with higher functional constraints on ultra-conserved non-coding sequences.

To demonstrate the efficiency of OrthReg, regulatory annotation of the pig genome was conducted using the human ENCODE data as a reference, as pigs show high levels of sequence identity with humans. A total of 28 190 886 putative regulatory annotations in the pig genome were obtained. Classification of the putative regulatory elements and their chromosomal distribution statistics are shown in Supplementary Table S2. A density variation was observed among chromosomes when the predicted regulatory sequences were counted in 10 kb non-overlapping sliding windows (Supplementary Figure S2). The chromosomes 8, 11, and X contained a lower density of regulatory elements, whereas chromosome 12 had a higher density than the genome-wide average. Further analysis showed that clustering of putative regulatory elements was observed on many chromosomes. An evident clustering of transcription factor binding site ("TFBS"), transcription factor recognizing motif ("Motif"), and histone chemical modification site ("Histone") sequences were observed on pig chromosome 4, spanning the region harboring the *ZBTB7B* gene. The *ZBTB7B* gene encodes an important transcription factor, and its genomic region in the human genome showed an abundance of ENCODE regulatory sequences (Supplementary Figure S3). Thus, data consistency indicated a high level of accuracy in the mapping of the regulatory sequences from the human to pig genomes.

Transcriptional activity verification of the predicted regulatory elements was also performed to confirm the reliability of OrthReg. Four predicted non-coding regulatory elements in the pig genome were selected to test their transcriptional activities with luciferase assays (Supplementary Table S3). All four motifs contained a non-coding SNP in the pig population. The luciferase assays demonstrated that all predicted wild-type motifs (pGL3-Basic-Pro-Motif-WT) showed statistically significant changes (P<0.05) in luciferase expression when compared to the promoter vector (pGL3-Basic-Pro) (Figure 1C–F). Three motifs showed transcriptional activation and one showed transcriptional repression. Furthermore, all mutant-type motifs (pGL3-Basic-Pro-Motif-MT) showed statistically significant changes (P<0.05) in luciferase expression when compared to the wild-type vector (pGL3-Basic-Pro-Motif-WT) (Figure 1C–F). Thus, these results imply that the four selected non-coding sequences were *cis*-regulatory elements.

The predicted regulatory elements were used to identify regulatory SNPs involved in the domestication of European and East Asian pigs. Earlier research has indicated that European and East Asian domestic pigs were domesticated independently from local wild boars (Larson et al., 2005; Wu et al., 2007). In total, 179 pigs (Supplementary Table S5) from previous studies were downloaded for analysis. After variant calling, genetic differentiation (F_{ST}) was calculated for each site between wild and domestic populations in Europe and Asia, respectively. Two body-length-related sweeps harboring the *LCORL* and *PLAG1* genes in European domestic pigs established in previous research (Rubin et al., 2012) were analyzed in this study. Screening of the *LCORL* sweep (chromosome 8: 12.61–12.76 Mb, Supplementary Figure S4A) identified a total of 657 SNPs. Among them, 24 SNPs with high level differentiation (top 1%, F_{ST}≥0.72) were only located in either intergenic or intronic sequences (Supplementary Table S6). We found that the SNP with the highest F_{ST} (chr8: 12 755 647, A>G; F_{ST}=0.97) in *LCORL* intron 1 was covered by many predictive regulatory sequences, including "TFBS" of GATA3, DNase I hypersensitive sites, and diverse histone chemical modifications (Supplementary Table S6). The histone H3K4me1 and H3K79me2 modifications in this site were observed in human osteoblasts and skeletal muscle myoblasts, respectively (Davis et al., 2018). This site is likely found within an enhancer sequence because H3K4me1 is indicative of the presence of an enhancer (Heintzman & Ren, 2009). Furthermore, the nucleotide in this site was highly conserved from hedgehogs to humans (Supplementary Figure

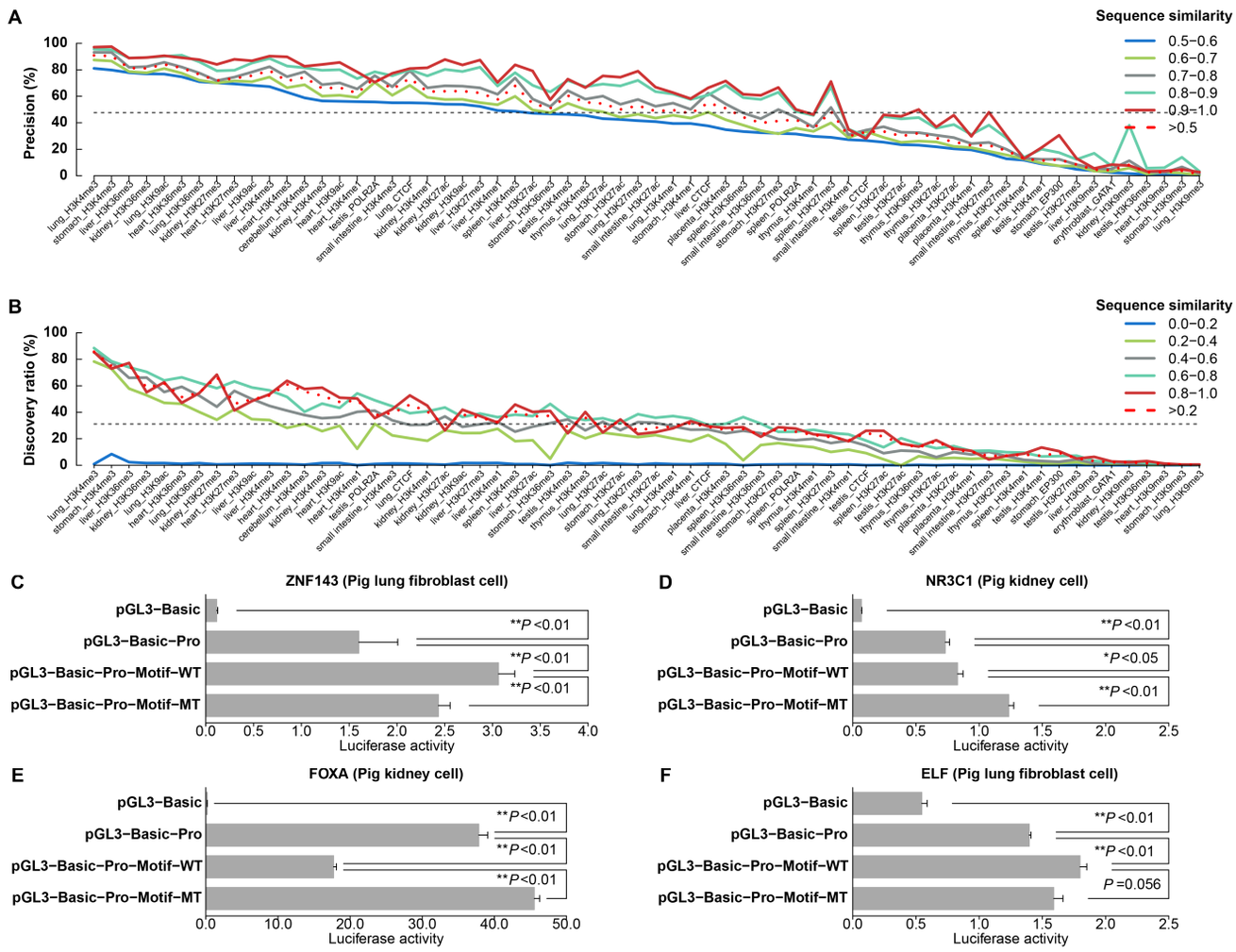

Figure 1 Reliability of OrthReg and luciferase assays for four predicted transcription factor recognizing motifs in pig genome

A: Precision of predicted regulatory elements in mice. Black dotted line represents average precision (sequence similarity >0.5). B: Discovery ratio of true regulatory elements in mice. Black dotted line represents average discovery ratio (sequence similarity >0.2). C–F: Transcription activity assay of different alleles within predicted recognizing motifs for ZNF143 (C), NR3C1 (D), FOXA (E), and ELF (F). Motif-WT: Predicted motifs with wild-type allele; Motif-MT: Predicted motifs with mutant allele. Two-tailed *t*-test was used for statistical assessment of transcription activity change.

S5A). In cross-population comparison, we found that European domestic pigs showed homozygosity for the derived allele (guanine), and that European wild boars and East Asian wild and domestic pigs showed high frequencies of the ancestral allele (Supplementary Figure S5B). For the *PLAG1* sweep (chromosome 4: 82.56–82.71 Mb), a total of 226 SNPs were identified (Supplementary Figure S4B). Among these SNPs, 147 showed a high level of differentiation (top 1%, F_{ST}≥ 0.72) between European domestic pigs and wild boars. Of these 147 SNPs, we identified several important regulatory candidates (Supplementary Table S7). A candidate SNP (chr4: 82,601,069, A>C, F_{ST}=0.97) located upstream of the *PLAG1* gene in a 54 bp intergenic DNA sequence with predicted regulatory signals of histone H3K4me1, H3K4me2, H3K4me3, H3k27ac, and H3k9ac chemical modifications was evident in 10 cell lines, including osteoblasts and skeletal muscle myoblasts (Supplementary Table S7). This site was also highly conserved among species (Supplementary Figure S5C). Population analysis showed that European domestic pigs were homozygous for cytosine, whereas other populations had high frequencies of adenine in this site (Supplementary Figure S5D). In East Asian domestic pigs, we focused on a sweep region (chromosome 1: 148.27–149.59 Mb) containing a total of 7 563 SNPs (Supplementary Figure S4C). More than 1 762 SNPs showed a moderately high level of differentiation (top 1%, F_{ST}≥0.38) in East Asian domestic pigs from wild boars (Supplementary Table S8). One SNP (chr1: 148 429 411, C>A, F_{ST}=0.65), located 13 kb downstream of the *SPRED1* gene, was located inside a predictive DNase I hypersensitive site and CEBPB and EZH2 binding sites (Supplementary Table S8). This SNP nucleotide was conserved across different species (Supplementary

Figure S6).

OrthReg offers a software package containing several tools that can predict regulatory sequences and annotate SNPs against the predicted regulatory dataset. With the increasing availability of next-generation sequencing data, OrthReg could help in the generation of genomic regulatory element annotations and facilitate the screening of functional non-coding variants. This could offer an opportunity to explore non-coding SNPs in comparative genomic studies and advance our understanding of regulatory SNPs in phenotypic variations. OrthReg was developed to incorporate variant data from whole genome resequencing projects and the source code is freely available at https://github.com/haibing-evo/OrthReg.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Y.P.Z., Y.F.M., H.B.X., and C.P.H. designed the research. H.B.X. and Y.F.M. implemented the code. Y.F.M., C.P.H., and F.R.L. analyzed the data. J.X.L. and X.M.H. performed the experiments. Y.G. and J.K.D. collected samples. Y.F.M., H.B.X., C.P.H., and Y.P.Z. wrote the manuscript. A.C.A. revised the manuscript. All authors read and approved the final version of the manuscript.

Yun-Fei Ma[1,2,3,#], Cui-Ping Huang[1,2,3,#], Fang-Ru Lu[1,2,3,#], Jin-Xiu Li[1,5,6,#], Xu-Man Han[1], Adeniyi C. Adeola[1], Yun Gao[1], Jia-Kun Deng[1], Hai-Bing Xie[1,*], Ya-Ping Zhang[1,4,*]

[1] *State Key Laboratory of Genetic Resources and Evolution*, *Kunming Institute of Zoology*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650223, *China*

[2] *Kunming College of Life Science*, *University of Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650204, *China*

[3] *University of Chinese Academy of Sciences*, *Beijing*, 100049, *China*

[4] *Center for Excellence in Animal Evolution and Genetics*, *Chinese Academy of Sciences*, *Kunming*, *Yunnan* 650223, *China*

[5] *State Key Laboratory for Conservation and Utilization of Bio-Resource in Yunnan*, *Yunnan University*, *Kunming*, *Yunnan* 650091, *China*

[6] *Key Laboratory for Animal Genetic Diversity and Evolution of High Education in Yunnan Province*, *School of Life Sciences*, *Yunnan University*, *Kunming*, *Yunnan* 650091, *China*

[#]Authors contributed equally to this work

*Corresponding authors, E-mail: xiehb@mail.kiz.ac.cn; zhangyp@mail.kiz.ac.cn

REFERENCES

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar Å, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, **495**(7441): 360−364.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science*, **304**(5675): 1321−1325.

Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, **15**(7): 901−913.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, **46**(D1): D794−D801.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Molecular Biology and Evolution*, **19**(7): 1114−1121.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics*, **38**(2): 223−227.

Gumucio DL, Heilstedt-Williamson H, Gray TA, Tarlé SA, Shelton DA, Tagle DA, Slightom JL, Goodman M, Collins FS. 1992. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Molecular and Cellular Biology*, **12**(11): 4919−4929.

Heintzman ND, Ren B. 2009. Finding distal regulatory elements in the human genome. *Current Opinion in Genetics & Development*, **19**(6): 541−549.

Larson G, Dobney K, Albarella U, Fang MY, Matisoo-Smith E, Robins J, Lowden S, Finlayson H, Brand T, Willerslev E, Rowley-Conwy P, Andersson L, Cooper A. 2005. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science*, **307**(5715): 1618−1621.

Ma YF, Adeola AC, Sun YB, Xie HB, Zhang YP. 2020. CaptureProbe: a java tool for designing probes for capture Hi-C applications. *Zoological Research*, **41**(1): 94−96.

Ma YF, Han XM, Huang CP, Zhong L, Adeola AC, Irwin DM, Xie HB, Zhang YP. 2019. Population genomics analysis revealed origin and high-altitude adaptation of Tibetan pigs. *Scientific Reports*, **9**(1): 11463.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu HZ, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**(6099): 1190−1195.

Nakayama JI, Rice JC, Strahl BD, Allis CD, Grewal SIS. 2001. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, **292**(5514): 110−113.

Rubin CJ, Megens HJ, Martinez Barrio A, Maqbool K, Sayyab S,

Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB, Archibald AL, Fredholm M, Groenen MAM, Andersson L. 2012. Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(48): 19529−19536.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou MM, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**(8): 1034−1050.

The FAANG Consortium, Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, Couldrey C, Dalrymple BP, Elsik CG, Foissac S, Giuffra E, Groenen MA, Hayes BJ, Huang LS, Khatib H, Kijas JW, Kim H, Lunney JK, McCarthy FM, McEwan JC, Moore S, Nanduri B, Notredame C, Palti Y, Plastow GS, Reecy JM, Rohrer GA, Sarropoulou E, Schmidt CJ, Silverstein J, Tellam RL, Tixier-Boichard M, Tosser-Klopp G, Tuggle CK, Vilkki J, White SN, Zhao SH, Zhou HJ. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, **16**(1): 57.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**(1): 5−22.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, Walter K, Abnizova I, Gilks W, Edwards YJK, Cooke JE, Elgar G. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, **3**(1): e7.

Wu GS, Yao YG, Qu KX, Ding ZL, Li H, Palanichamy MG, Duan ZY, Li N, Chen YS, Zhang YP. 2007. Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome Biology*, **8**(11): R245.

Yang Y, Adeola AC, Xie HB, Zhang YP. 2018. Genomic and transcriptomic analyses reveal selection of genes for puberty in Bama Xiang pigs. *Zoological Research*, **39**(6): 424−430.

Yang Y, Liu CR, Adeola AC, Sulaiman X, Xie HB, Zhang YP. 2019. Artificial selection drives differential gene expression during pig domestication. *Journal of Genetics and Genomics*, **46**(2): 97−100.