RESEARCH ARTICLE
Open Access

Journal of Veterinary and Animal Sciences
ISSN (Print): 0971-0701, (Online): 2582-0605
https://doi.org/10.51966/jvas.2023.54.2.524-531

Identification of genetic variants by whole genome sequencing in Ankamali pigs of Kerala#

Michelle Elizabeth Roy[1], M. Manoj[2]*, P.M. Rojan[3], Tina Sadan[4]
T.V. Aravindakshan[5], A.P. Usha[6] and M.P. Unnikrishnan[2]
Department of Animal Breeding Genetics and Biostatistics
College of Veterinary and Animal Sciences
Mannuthy, Thrissur- 680 651
Kerala Veterinary and Animal Sciences University
Kerala, India

Citation: Michelle, E.R., Manoj, M., Rojan, P.M., Tina, S., Aravindakshan, T.V., Usha, A.P. and Unnikrishnan, M.P. 2023. Identification of genetic variants by whole genome sequencing in Ankamali pigs of Kerala. *J. Vet. Anim. Sci.* **54**(2):524-531
DOI: https://doi.org/10.51966/jvas.2023.54.2.524-531

Received: 15.11.2022 *Accepted: 31.01.2023* *Published: 30.06.2023*

Abstract

Ankamali pig is a domesticated native variety of Kerala which is well known for its disease resistance, lean meat and adaptability to hot tropical environments. Recent breakthrough in genome sequencing technologies have created unparalleled prospects to characterize individual genomic landscapes and identifying mutations between and within populations. The current study aims to determine the genetic variations in Ankamali pigs using whole genome sequencing. The GATK HaplotypeCaller was used to identify the variants. There were over 26 million (26,604,589) single nucleotide variants (SNVs), including more than 21 million SNPs and over 5 million indels. In Ankamali pigs, the total genome length obtained was more than 2.5 billion with an average variant rate of one variant in every 94 bases. The significance of different variant rate on 18 chromosomes were analysed using the chi-square statistics. The variant rates in Sus scrofa chromosomes10 and 13 were significantly different (p<0.01%) in Ankamali pigs. The significantly higher variable rate on chromosome 10 was observed with one variant per 64 bases. Whereas, significantly lower variable rate was observed on chromosome 13, with one variant in every 122 bases. The variant rate reported on Sus scrofa chromosome 12 (SSC12) was also significantly higher (p<0.05%), having one variant per 72 bases. The variability of many genes and QTLs associated with several haematological traits and meat quality traits located on these chromosomes may contribute the phenotypic and genetic uniqueness of Ankamali animals.

# Part of M.V.Sc thesis submitted to Kerala Veterinary and Animal Sciences University, Pookode, Wayanad, Kerala

1. *M.V.Sc Scholar*
2. *Assistant Professor, CPPR, Mannuthy, KVASU*
3. *Assistant Professor*
4. *Ph. D Scholar*
5. *Senior Professor and Head*
6. *Senior Professor and Head, CPPR, Mannuthy, KVASU*

** Corresponding author : manojm@kvasu.ac.in, Ph. 8547638755*

Keywords: *GATK Haplotype caller, single nucleotide variants, indels, haematological traits, missense, quantitative trait loci (QTL)*

Domestication of pigs is thought to have begun in the Near East around 9000 years ago and may have occurred repeatedly from local populations of wild boars. The domestic pig (*Sus scrofa*) belongs to the Suidae family and order Artiodactyla. Pigs are one of the most common livestock species reared for meat in Kerala. Ankamali pig is a domesticated native variety of Kerala notable for its lean meat, disease resistance and adaptability to humid tropical environments (Behl *et al*., 2006). However, exotic breeds are preferred over desi pig breeds due to their excellent production performance. Due to this there is a drastic decline in the Ankamali pig population. Thus, conservation of Ankamali pigs is essential for future animal production and conservation of the genetic variety.

Recent breakthrough in genome sequencing technologies have created unparalleled prospects for characterizing individual genomic landscapes and identifying mutations. Whole exome sequencing (WES) is a cost-effective approach as it sequence only the coding regions of the genome (Reshma *et al*., 2020). However, Whole-genome sequencing (WGS) is becoming increasingly attractive as an alternative, due to its broader coverage and decreasing cost. Whole genome sequencing (WGS) approach allows the detection of nearly an organism's entire genome sequence (Park and Kim, 2016). One of the primary goals of sequencing-based studies is to find genetic variants that differ between individuals. Several studies have been conducted in many native breeds of different species including Vechur cattle, Malabari and Attapady black goats to gain a better understanding of the genetic variances responsible for the unique features such as short stature of Vechur cattle, high prolificacy and low-fat meat of Malabari goat and high disease resistance of Attapady black goats (Reshma *et al*., 2020; Marykutty *et al*., 2021). Although a wide range of major and small-scale genomic sequence changes can affect gene function, variants that change amino acid sequences via missense, nonsense, frameshift and splice site variants are among those most likely to affect function (Bickhart and Liu, 2014). Animal genomic sequences can be easily aligned to a high-quality, annotated reference genome assembly *Sus scrofa* (Sscrofa 11.1), to identify DNA polymorphisms encoding these protein variations. Several variant calling tools have been created over the years, including SAMtools/BCFtools, CLC Genomics Workbench (Qiagen), FreeBayes, GATK, LoFreq, SNVer, VarDict and VarScan. GATK HaplotypeCaller (HC) is a common haplotype based variant caller that finds variants between a DNA sequence and a reference sequence (Lefouili and Nam, 2022). GATK HaplotypeCaller (GATK HC) was created by the Broad Institute of MIT and Harvard. It has high accuracy in variant calling compared to other variant calling tools, but its feasibility is limited by the long execution time needed for the analysis. The haplotype-based variant detection method can be used with single or many samples (Ren *et al*., 2019). The aim of the present study was to identify the genetic variants in Ankamali pigs using whole genome sequencing and analyzing the rate of variants among different chromosomes.

Materials and methods

DNA extraction and whole – genome sequencing

The blood samples were collected from 12 Ankamali pigs reared at Centre for pig production and research, Mannuthy. Approximately, 5 ml of blood was collected using a sterile 20-gauge needle and syringe from the ear vein of each animal into vacutainer tubes containing EDTA as an anticoagulant (1mg/ml of blood). The samples were preserved in ice and brought to the laboratory and were stored at -40 °C. The genomic DNA were extracted through Qiagen DNeasy blood and tissue kit (Qiagen, Germany). Quality of the isolated genomic DNA was checked by agarose gel electrophoresis and the concentration and purity was estimated by spectrophotometer (NanoDrop^{TM} 2000C) and pooled in equimolar quantities (0.25 µl/ per sample). A single library was generated for the pooled sample using QIAseq FX DNA Library Kit for Illumina (Average fragment size a 342bp) and sequenced on NovaSeq 6000 instrument

(Paired – end 2× 150bp) (Clevergene, Bengaluru)

Sequence quality checking and filtering

Trimmomatic v0.39 was used to eliminate adaptor sequences and low-quality bases before downstream analysis with following parameters, SLIDINGWINDOW 4:20, MINLEN:36, LEADING:3, TRAILING:3 (Bolger *et al*., 2014).

Using Burrow Wheeler Aligner (BWA v0.717) software high quality paired-end reads were mapped to the pig reference genome Sscrofa11.1 (Li and Durbin, 2009). Following alignment, SAMtools were used to convert from SAM format to BAM format (Li *et al*., 2009). The resultant BAM files were sorted using Picard 'SortSam', and duplicate reads were removed by Picard 'MarkDuplicates'.

Variant calling and annotation

The GATK Haplotype caller with default parameters was used to call variants for single nucleotide polymorphisms (SNPs) (McKenna *et al*., 2010). SnpEff was used to annotate the identified SNPs (Cingolani *et al*., 2012). The variant rate in Ankamali genome was estimated from the sequence length and total number of variants. The significance of different variant rate on 18 chromosomes were analysed using the chi-square statistics (Snedecor and Cochran, 1989). The value of the Chi-square (χ^2) statistic was calculated using the formula:

$$\chi^2 = \sum (Oi - Ei)^2/Ei$$

χ2 = Chi-square deviation
Oi = Observed value
Ei = Expected value

Transition and transversion ratio

Transition and transversion are two-point mutations that occur in DNA due to substitution errors. Transition mutation occurs due to an interchange of two-ring purines (A ↔ G) or one-ring pyrimidines (C ↔ T). Transversion mutation occurs due to interchanges of pyrimidine for purines or purines for pyrimidines. Transition mutations are more frequent than transversions and two out of three SNPs are caused by transitional mutations. However, transition mutations are less likely to cause amino acid sequence changes. Transversion occurs in two possible ways since two pyrimidines and two purines are present. This is more likely to cause amino acid sequence changes (Baes *et al*., 2014)

Transition-transversion ratio is the ratio of the number of transitions to the number of transversions for a pair of sequences. The ratio becomes 0.5 when there is no bias towards either transitional or transversional substitution. Across the entire genome the ratio of transitions to transversions is typically around 2 and in protein coding regions, this ratio is typically higher, often a little above 3 (Bainbridge *et al.*, 2011).

From the base change matrix of SNPs obtained by the WGS data, transition-transversion ratio was calculated as:

Ts/Tv = sum of 4 transitions/sum of 8 transversions

Results and discussion

The whole genome sequencing of Ankamali pig generated 205.66 Gbs raw paired-end reads. During sequencing, 95.47 per cent of R1 reads and 94.27 per cent of R2 reads were obtained through paired end sequencing having the base quality score (Q) of >30. In DNA sequencing, the Phred quality score was used to determine the measure of base quality. Phred score of 30 (Q30) indicate, the risk of an inaccurate base call is one in 1000 with base call accuracy of 99.9 per cent (Cock *et al*., 2009). The GC content of both R1 and R2 reads were 43 per cent. Total GC content, which ranged between 39 and 59 percent, might be utilized as a measure for assessing the quality of sequenced reads (Guo *et al*., 2013). Any variation from this range suggested that contamination from other sources was present. The GC content of raw sequence data obtained in this study was 43 per cent.

After variant calling, more than 26 million SNVs were identified in the Ankamali pig's genome. Of the total number of variants discovered, 21,303,641 (80.08 %) were single nucleotide polymorphisms (SNPs), 3,056,981 (11.4%) were insertions and 2,243,967 were deletions (8.43%) (Table 1). In Wannan black pig more than 21 million SNV's were identified of which 16,249,548 (76.23%) were single nucleotide polymorphisms (SNPs), 2,898,582 (13.59%) of insertions and 2,168,624 (10.17%) of deletions (Zhang *et al*., 2020). The number of variants in Ankamali pigs and Wannan black revealed that the number of variants were almost similar in both breeds. Because the Ankamali and Wannan black pigs (Chinese local breed) were compared to a European breed-based reference (Duroc), the practically identical number of variations reported in these breeds could be explained.

The number of different transitions and transversions in Ankamali pig genome were presented in Table. 2 and Fig. 1. The total number of transitions and transversions obtained were more than 14 million (14,824,878) and 6 million (6,478,763), respectively. The number of A to G and T to C transitions were 3,906,144 (18.34%) and 3,511,198 (18.33%), respectively. The number of A to T and A to C transversions were 729,849 (3.43%) and 901,472 (4.23%), respectively. Similarly, the number of G to C and G to T transversions were 734,780 (3.45%) and 877,366 (4.12%), respectively. Ts/Tv ratio for SNPs found using whole-genome sequencing should be at least 2 (DePristo *et al*., 2011). In general, a greater Ts/Tv ratio means more accuracy. The average Ts/Tv ratio observed in Ankamali pigs was 2.29, which was consistent with the previous report in Anquing-six-end pigs (Zhang *et al*., 2020). Given that only 2% of the genome codes for proteins, the majority of the discovered SNVs were found in non-coding areas (46.6 per cent in intron and 39.2 per cent in intergenic regions). Intergenic and intron regions are anticipated to play essential roles in a range of cell activities, notably in gene expression, transcriptional control and gene splicing (Bartonicek *et al*., 2017). This is consistent with data from several whole genome sequencing studies, which showed more than 50 per cent of single nucleotide variants in cattle and pig were found in non-coding regions (Mei *et al*., 2018; Yu *et al*., 2020).

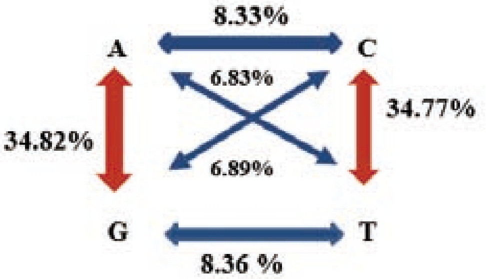

Fig. 1. Number of Transitions and transversions in Ankamali pig genome

On whole genome sequencing of Ankamali pigs, the total genome length obtained was more than 2.5 billion with an average variant rate of one variant in every 94 bases. Chromosome wise distribution of variants is depicted in the Fig 2. The overall

Table 1. Classification of number of variants by type in Ankamali pig genome

Sl. No.	Type	Total
1	Single nucleotide polymorphisms (SNPs)	21,303,641
2	Insertion (INS)	3,056,981
3	Deletion (DEL)	2,243,967
4	TOTAL	26,604,589

Table 2. Number of transitions and transversions in Ankamali pig genome

Nucleotides	A	C	G	T
A	-	8,72,068	35,11,198	7,25,487
C	9,01,472	-	7,34,780	39,04,409
G	39,06,144	7,33,166	-	9,04,575
T	7,29,849	35,03,127	8,77,366	-

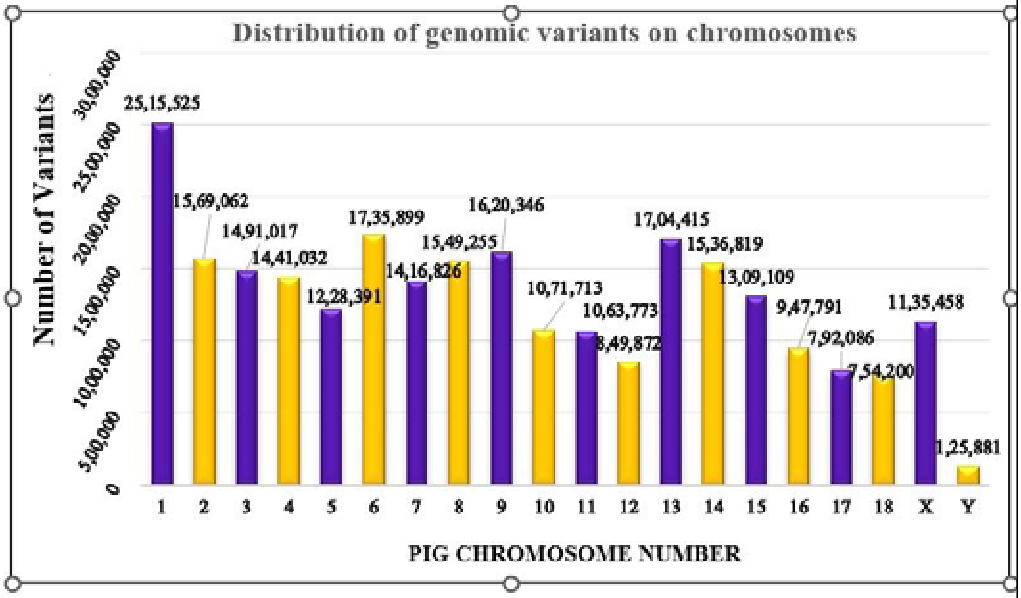

Fig. 2. Chromosome-wise distribution of variants in Ankamali pig genome.

number of variants in the X and Y chromosome was 1,135,458 and 125,881, with a variant rate of one variant in every 110 bases and one variant in every 345 bases, respectively. The total number of variants in mitochondrial DNA was 250 with one variant in every 66 bases. A total of 587 unplaced scaffold were obtained, with an average length of 66,528,145, total variants of 8,745,869 and a variant rate of one variant in every 89 bases. The total number of variants across 18 autosomes were 24.59 million (24,597,381) with one variant in every 92 bases. The chi-square test was used to determine the significance of the variations across 18 autosomes. The variant rates in Sus scrofa chromosomes10 (SSC10) and Sus scrofa chromosome 13 (SSC13) were significantly different (p<0.01%) in Ankamali pigs. The significantly higher variant rate on chromosome 10 was observed with one variant per 64 bases. Whereas, significantly lower variable rate was observed on chromosome 13, with one variant in every 122 bases. The variant rate reported on Sus scrofa chromosome12 (SSC12) was also significantly higher (p<0.05%), having one variant per 72 bases. The overall variant rate observed in Nero siciliano pig was one variant in every 276 bases (D'Alessandro *et al*., 2019).

Several quantitative trait loci (QTL) for haematological traits such as haematocrit (HCT), haemoglobin (HCB), mean corpuscular volume (MCV) and blood platelet counts (PLT) were discovered on SSC10, all of which are important components of the animal immune system. A QTL related to teat number was also identified on SSC10. A sow must provide a fair opportunity for all of its piglets to access nipple throughout the suckling phase, so teat number has been recognised as one of the most important factors in measuring swine mothering skill (Rohrer, 2000). On SSC13, some QTLs associated with back fat thickness, growth and meat quality was discovered (Yu *et al*., 1999). Multiple meat quality traits (SHEAR, MOIST, DRIP, MARB, and CFAT) were impacted by SSC12 QTL This co-localization of numerous QTL for meat quality-related characteristics shows a genetic correlation between these traits. The variability of many genes and QTLs located on chromosomes 10 and 12 may contribute the phenotypic and genetic uniqueness of Ankamali animals.

Table 3. Chromosome-wise distribution of variants in Ankamali pig genome

Sl. No.	Chromosomes	Length	Variants	Variant rate
1	1	274,330,532	2,515,525	109
2	2	151,935,994	1,569,062	96
3	3	132,848,913	1,491,017	89
4	4	130,910,915	1,441,032	90
5	5	104,526,007	1,228,391	85
6	6	170,843,587	1,735,899	98
7	7	121,844,099	1,416,826	85
8	8	138,966,237	1,549,255	89
9	9	139,512,083	1,620,346	86
10	10	69,359,453	1,071,713	64**
11	11	79,169,978	1,063,773	74
12	12	61,602,749	849,872	72*
13	13	208,334,590	1,704,415	122**
14	14	141,755,446	1,536,819	92
15	15	140,412,725	1,309,109	107
16	16	79,944,280	947,791	84
17	17	63,494,081	792,086	80
18	18	55,982,971	754,20	74
19	X	125,939,595	1,135,458	110
20	Y	43,547,828	125,881	354
21	MT	16,613	250	66
22	587 Unplaced genomic scaffolds	66,528,145	745,869	89.20
23	Total	2,501,806,821	26,604,589	94.04

** significant at ≤ 1% level and * significant at ≤ 5% level

Conclusion

Whole genome sequencing of Ankamali pigs yielded more than 1.37 billion paired end reads and 99.77% of QC passed reads were successfully aligned to the *Sus scrofa* (Sscrofa11.1) reference genome. The sequencing revealed more than 21 million SNPs and five million indels. Majority of the identified SNVs (85.8 per cent) were present in the non- coding regions The Ts/Tv ratio obtained was 2.29 and overall variable rate was one in every 94 bases. Significantly lower variable rate was observed on chromosome 13 whereas; significantly higher variable rate was observed on chromosome 10 and 12. The variability of many genes and QTLs located on these chromosomes might contribute to the unique phenotypic and genetic characteristics of Ankamali pigs. The study reveals the importance of conservation of this native variety.

Acknowledgements

We are thankful to the Kerala Veterinary and Animal Sciences University for providing the facilities for the conduct of research. We thank Clever gene Lab Pvt Ltd, Bangaluru, Karnataka for the sequencing and bioinformatics services.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Baes, C.F., Dolezal, M.A., Koltes, J.E., Bapst, B., Fritz-Waters, E., Jansen, S., Flury, C., Signer-Hasler, H., Stricker, C., Fernando, R. and Fries, R. 2014. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC genomics*. **15:** 1-18.

Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L. and Gibbs, R.A. 2011. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Gen. biol*. **12**: 1-12.

Bartonicek, N., Clark, M.B., Quek, X.C., Torpy, J.R., Pritchard, A.L., Maag, J.L.V., Gloss, B.S., Crawford, J., Taft, R.J., Hayward, N.K. and Montgomery, G.W. 2017. Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biol*. **18**: 1-16.

Behl, R., Sheoran, N., Behl, J. and Vijh, R.K. 2006. Genetic analysis of Ankamali pigs of India using microsatellite markers and their comparison with other domesticated Indian pig types. *J. Anim. Breed. Genet*. **123**: 131-135.

Bickhart, D.M. and Liu, G.E. 2014. The challenges and importance of structural variation detection in livestock. *Front. Genet.* **5**: 37.

Bolger, A.M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *J. Bioinform*. **30**: 2114-2120.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff SNPs in the genome of Drosophila melanogaster strain. *Fly*. **6**: 80-92.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*. **38**: 1767-1771.

D'Alessandro, E., Sapienza, I., Giosa, D., Giuffrè, L. and Zumbo, A. 2019. *In silico* analysis of meat quality candidate genes among Nero Siciliano, and Italian heavy pigs genomes. *Large Anim. Rev*. **25**: 137-140.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M. and McKenna, A. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet*. **43**: 491-498.

Guo, Y., Ye, F., Sheng, Q., Clark, T. and Samuels, D.C. 2013. Three-stage quality control strategies for DNA re-sequencing data. *Brief. Bioinform*. **15**(6): 879-89.

Lefouili, M. and Nam, K., 2022. The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. *Sci. Rep. **12***:1-8.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *J. Bioinform*. **25**: 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. 2009. The sequence alignment/map format and SAMtools. *J. Bioinform*. **25**: 2078-2079.

Marykutty, T., Radhika, G., Aravindakshan, T.V., Thirupathy, R., Raji, K. and Shynu, M. 2021. Linkage disequilibrium over short physical genomic distances measured using medium density SNP beadchip in native goat breeds of Kerala. *J. Vet. Anim. Sci.* **52**: 14-18.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M.A. 2010. The

Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20**: 1297-1303.

Mei, C., Wang, H., Liao, Q., Wang, L., Cheng, G., Wang, H., Zhao, C., Zhao, S., Song, J., Guang, X. and Liu, G.E. 2018. Genetic architecture and selection of Chinese cattle revealed by whole genome resequencing. *Mol. Biol. Evol*. **35**: 688-699.

Park, S.T. and Kim, J. 2016. Trends in next-generation sequencing and a new era for whole genome sequencing. *Int. Neurol. J.* **20**: S76.

Ren, S., Ahmed, N., Bertels, K. and Al-Ars, Z. 2019. GPU accelerated sequence alignment with traceback for GATK HaplotypeCaller. *BMC Genom*. **20**: 103-116.

Reshma, R.S., Aravindakshan, T.V., Radhika, G., Naicy, T. and Raji, K. 2020. Exome wide variant discovery by next generation DNA sequencing in Vechur cattle of Kerala. *J. Vet. Anim. Sci*. **51**: 201-206.

Rohrer, G.A. 2000. Identification of quantitative trait loci affecting birth characters and accumulation of backfat and weight in a Meishan-White Composite resource population. *J. Anim. Sci.* **78**: 2547-2553.

Snedecor, G.W. and Cochran, W.G. 1989. *Statistical Methods*. (8th Ed.). Iowa State University Press, Ames, Iowa.

Yu, B.T., Wang, L., Tuggle, C.K. and Rothschild, M.F. 1999. Mapping genes for fatness and growth on pig chromosome 13: a search in the region close to the pig *PIT1* gene. *J. Anim. Breed. Genet*. **116**:269-280.

Yu, J., Zhao, P., Zheng, X., Zhou, L., Wang, C. and Liu, J.F. 2020. Genome-wide detection of selection signatures in Duroc revealed candidate genes relating to growth and meat quality. *G3: Genes Genom. Genet*. **10**: 3765-3773

Zhang, W., Yang, M., Zhou, M., Wang, Y., Wu, X., Zhang, X., Ding, Y., Zhao, G., Yin, Z. and Wang, C. 2020. Identification of signatures of selection by whole-genome resequencing of a Chinese native pig. *Front. Genet*. **11**: 566255. ■