scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Transcriptome analysis of testes gene expression to explore genetic diversity of Mangalica and Camborough boars

Keigo Yamane[1], Sangwoo Kim[1], Akari Koide[1], Erina Yoneda[1], Mitsunori Kayano[2] & Yuki Muranishi[1,3] ✉

The testis, an important reproductive organ, is involved in spermatogenesis and steroid hormone secretion and has been the subject of a wide variety of studies. Pigs are often used as model animals for studies on human physiology and disease, and studies on the testicular development of pigs could shed light on human reproduction. Mangalica, an indigenous Hungarian pig breed, has reproductive traits that are different from those of commercial pig breeds. This specificity could reveal important differences in the cascades and reproductive genes between humans and other animals. In this study, we conducted RNA-sequencing analysis of the testes of 14 days old Mangalica and Camborough boars. We also performed clustering and pathway analysis of differentially expressed genes. These datasets and analyses are expected to provide important gene sets for pig testis development that can be applied in future studies on human reproductive mechanisms.

Background & Summary

Testicular research is essential in a wide range of fields, from medicine to life science. The testis is an important organ involved in spermatogenesis[1] and steroid hormone secretion[2], and has been the focus of many studies on testicular diseases[3,4], stem cells[5], and spermatogenesis[6]. In recent years, with the development of next-generation sequencing, transcriptome analyses of human and mammalian testes have been used to evaluate gene expression involved in testicular development and spermatogenesis[7,8].

Pigs are considered appropriate animal models for studies on human health and disease since they have similar anatomies, physiologies, and genetics to humans[9–11]. There is ongoing research on the xenotransplantation of pig organs in humans[12] and generation of pig models that replicate human diseases[13,14]; furthermore, boar testicular stem cells are being isolated and cultured for potential use in treatments on infertility in humans[15,16].

Mangalica, an indigenous Hungarian pig breed, is an interesting animal model with many characteristics; it is a fatty-type breed with curly hair and three different lines (blonde, red, and swallow-belly)[17]. In particular, Mangalica has been reported to have lower fertility than commercial pigs and relatively different steroid hormone levels[18]. Furthermore, the populations of many indigenous pig breeds are decreasing due to the rise of commercial breeds[19]; therefore, the elucidation of breeding traits unique to Mangalica can also contribute to their conservation.

In contrast, Camborough pigs are a widely used commercial hybrid breed developed for rapid growth, lean meat production, and high reproductive efficiency[20]. Due to intensive artificial selection in livestock breeding, they exhibit physiological and reproductive traits that differ markedly from those of indigenous breeds[21]. Including Camborough pigs in this study enables us to compare a traditional, low-intervention breed with a modern, high-performance commercial line. This contrast provides insights into how selective breeding has influenced the acquisition of testicular function and gene expression programs during development.

[1]Graduate school of Animal and Veterinary Sciences and Agriculture, Obihiro University of Agriculture and Veterinary Medicine, Inadacho, Obihiro, 080-8555, Hokkaido, Japan. [2]Research Center for Global Agromedicine, Obihiro University of Agriculture and Veterinary Medicine, Inadacho, Obihiro, 080-8555, Hokkaido, Japan. [3]Laboratory for Molecular and Developmental Biology, Institute for Protein Research, The University of Osaka, Suita-shi, 565-0871, Osaka, Japan. ✉e-mail: muranishi@obihiro.ac.jp

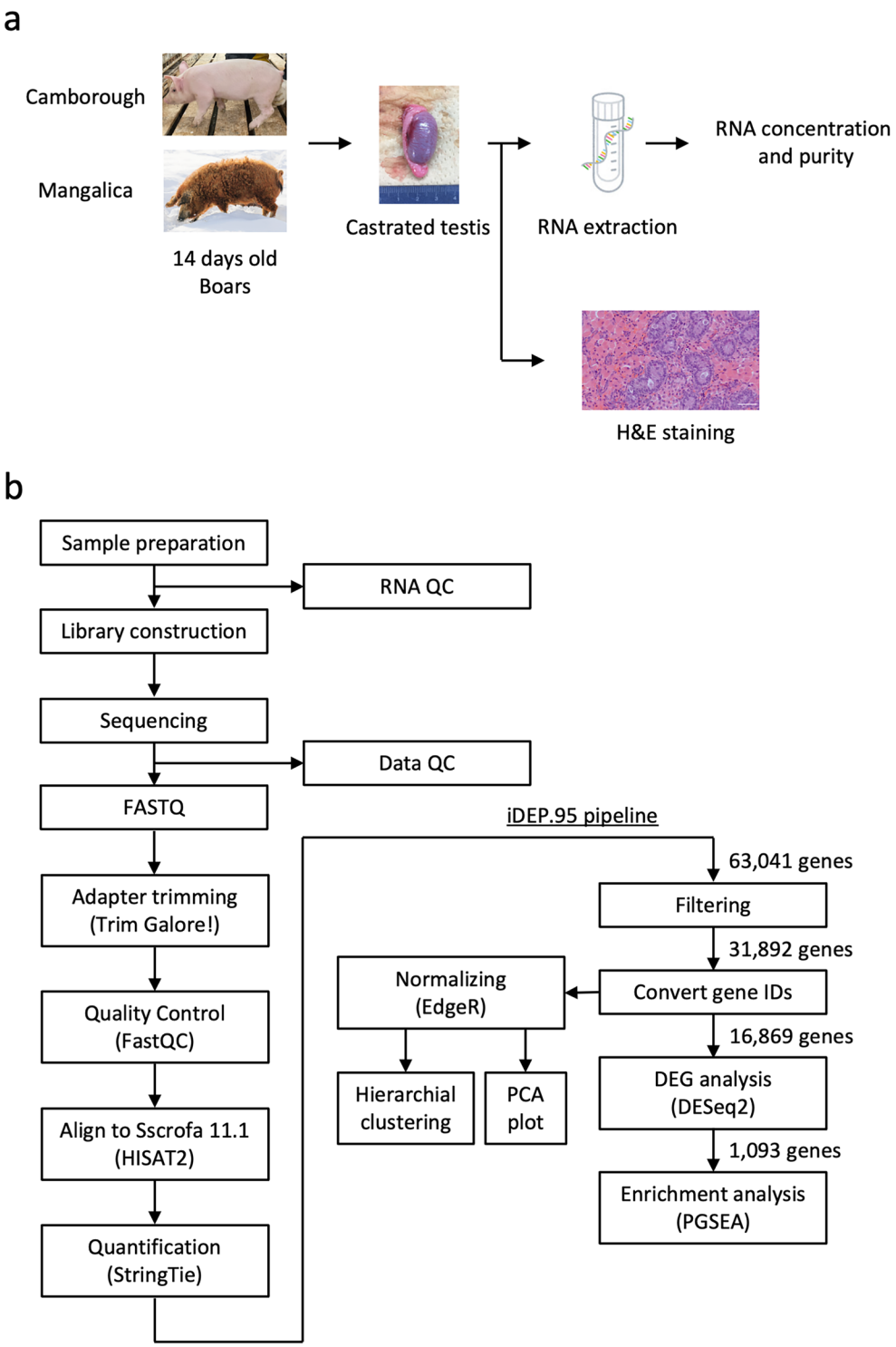

Fig. 1 Overview of the study design and data analyses workflow. (**a**) The boars and testis sample collection for sequencing. (**b**) The data analysis workflow of the sequencing data.

Spermatogenesis in boars starts early after birth, with key processes like gonocyte transformation into spermatogonial stem cells (SSCs) occurring within the first few weeks[16,22]. At 14 days, rapid testicular changes take place, including SSC formation and activation of signaling pathways for maturation[23]. Analyzing testicular transcriptomes at this stage helps reveal genetic programs essential for testis development and reproductive physiology. Early testicular development also marks the emergence of breed-specific reproductive traits[24]. Comparing gene expression at this stage clarifies how genetic background influences testicular maturation.

Sample Name	Sampling date	Line	Postnatal day	Body weight (kg)	Testis weight (g)	Epididymis weight (g)
C1	10/13/2021	Camborough	14	4.40	5.10	1.50
C2	10/13/2021	Camborough	14	4.65	3.70	1.90
C3	10/13/2021	Camborough	14	4.55	4.20	1.10
M1	10/20/2020	Mangalica	14	3.78	2.40	0.80
M2	10/23/2020	Mangalica	14	2.48	2.30	1.00
M3	6/30/2020	Mangalica	14	3.76	3.80	0.60

Table 1. Summary of the sample data. Sample Name: Identifiers for each sample. "C" indicates Camborough boars, while "M" represents Mangalica boars. Sampling date: The date when the tissue samples were collected. Line: The breed of the boars used in the study (Camborough or Mangalica). Postnatal day: The age of the boars in days after birth when the samples were taken. Body weight: The body weight of each boar at the time of sampling. Testis weight: The weight of the testes. Epididymis weight: The weight of the epididymis.

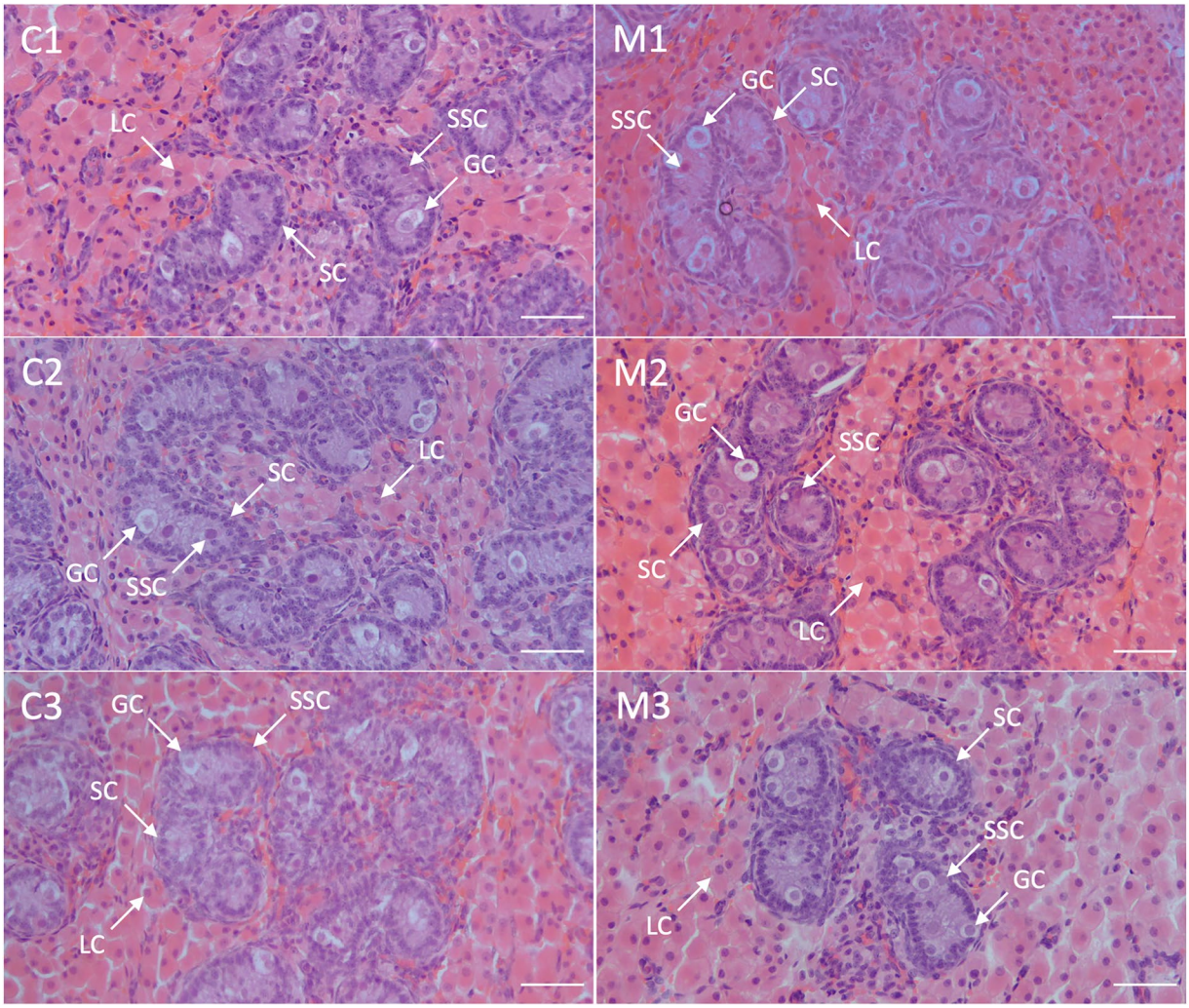

Fig. 2 The histological observation of the testes from 14 days old Red Mangalica and Camborough boars. Gonocytes, spermatogonial stem cells and Sertoli cells were identified in seminiferous tubule. Leydig cells were identified in interstitium. Bars = 100 μm. GC: gonocyte, SSC: spermatogonial stem cell, SC: Sertoli cell, LC: Leydig cell.

RNA-seq analysis in the testes of Chinese Meishan boars has been conducted and novel genes involved in their early sexual maturity[24]. In contrast, transcriptomic data for testicular tissue remain limited for indigenous pig breeds, including Mangalica.

In this study, we performed an RNA-seq analysis to clarify the role of reproductive differential gene expression between Mangalica and Camborough pig breeds using 14 days old testes (Fig. 1) in order to provide gene sets that are important for understanding pig testis development and human testicular mechanisms. This study

Sample Name	Concentration (ng/μl)	A260	260/280	260/230
C1	49.50	1.24	1.69	2.04
C2	41.21	1.03	1.72	1.84
C3	53.56	1.34	1.73	1.98
M1	35.43	0.89	1.69	1.85
M2	67.71	1.69	1.80	1.51
M3	55.41	1.39	1.72	1.63

Table 2. The data of RNA quality and quantity using spectrophotometer. Sample Name: Same as in Table 1. Concentration (ng/μl): The RNA concentration of each sample. A260: The absorbance at 260 nm, which indicates the nucleic acid concentration. 260/280: The ratio of absorbance at 260 nm and 280 nm, used to assess RNA purity. 260/230: The ratio of absorbance at 260 nm and 230 nm, used to evaluate contamination by organic compounds or salts.

Sample Name	Concentrarion (ng/μl)	Volume(μl)	Total amount(μg)	RIN	28s/18s
C1	53	24	1.272	8.8	1.4
C2	57	24	1.368	8.8	1.7
C3	79	24	1.896	8.5	1.5
M1	47	24	1.128	8.7	1.5
M2	88	24	2.112	8.9	1.6
M3	73	24	1.752	8.7	1.6

Table 3. The data of Integrity of RNA samples using bioanalyser. Sample Name: Same as in Table 1. Concentration (ng/μl): Same as in Table 2. Volume: The total volume of the RNA sample. Total amount: The total RNA amount in each sample. RIN: RNA Integrity Number, which indicates RNA quality based on the 28S/18S ribosomal RNA ratio.

Sample Name	Raw reads	Effective (%)	Error (%)	Q20 (%)	Q30 (%)	GC (%)
C1	31267882	95.0	0.03	97.2	92.7	50.6
C2	30833302	97.1	0.03	97.2	92.6	50.5
C3	27116642	95.0	0.03	97.5	93.1	50.4
M1	25895402	97.1	0.03	97.2	92.5	51.1
M2	25745920	97.1	0.03	97.1	92.4	50.3
M3	24909718	97.0	0.03	97.4	93.0	51.0

Table 4. The Summary of sequence data. Sample Name: Same as in Table 1. Raw reads: The total number of raw sequencing reads generated from each sample. Effective (%): The percentage of valid reads that passed quality control filtering. Error (%): The percentage of sequencing errors detected. Q20 (%): The percentage of bases with a Phred quality score $\geq$ 20. Q30 (%): The percentage of bases with a Phred quality score $\geq$ 30. GC (%): The GC content percentage, representing the proportion of guanine (G) and cytosine (C) bases in the total sequence.

provides valuable insights into testicular development in pigs and may contribute to future research on reproductive biology in humans.

Methods

Animals and testis sample collection. Testis samples were taken from Mangalica and Camborough boars on hog farms in Hokkaido, Japan. The boars were castrated at 14 days old, and testis samples were collected from three different littermates each of Mangalica and Camborough boars (Table 1). The testis samples were stored in TRIzol reagent (Invitrogen, Carlsbad, CA, USA) at −80 °C until RNA extraction. This study was performed under Regulations Regarding Animal Experiments by the Obihiro University of Agriculture and Veterinary Medicine (accepted No. 19–26, 20–29, 21-19).

Hematoxylin and Eosin staining. Testis samples were fixed in 10% formalin solution, dehydrated using an ethyl alcohol series, cleared in xylene and embedded in paraffin wax. The paraffin-embedded samples were cut into 4-μm thick sections using a SM2000R microtome (Leica Instrument Company, Germany). To confirm whether the testicular tissue had normal morphology, the sections were stained with hematoxylin and eosin (H&E) and observed under a microscope (Fig. 2).

RNA preparation and sequencing. The total RNA was extracted from testis tissues using TRIzol reagent following the manufacturer's instructions. The RNA quality and quantity were measured using the DS-11 spectrophotometer (DeNovix, Wilmington, USA) (Table 2). The total RNA samples were sent to Novogene

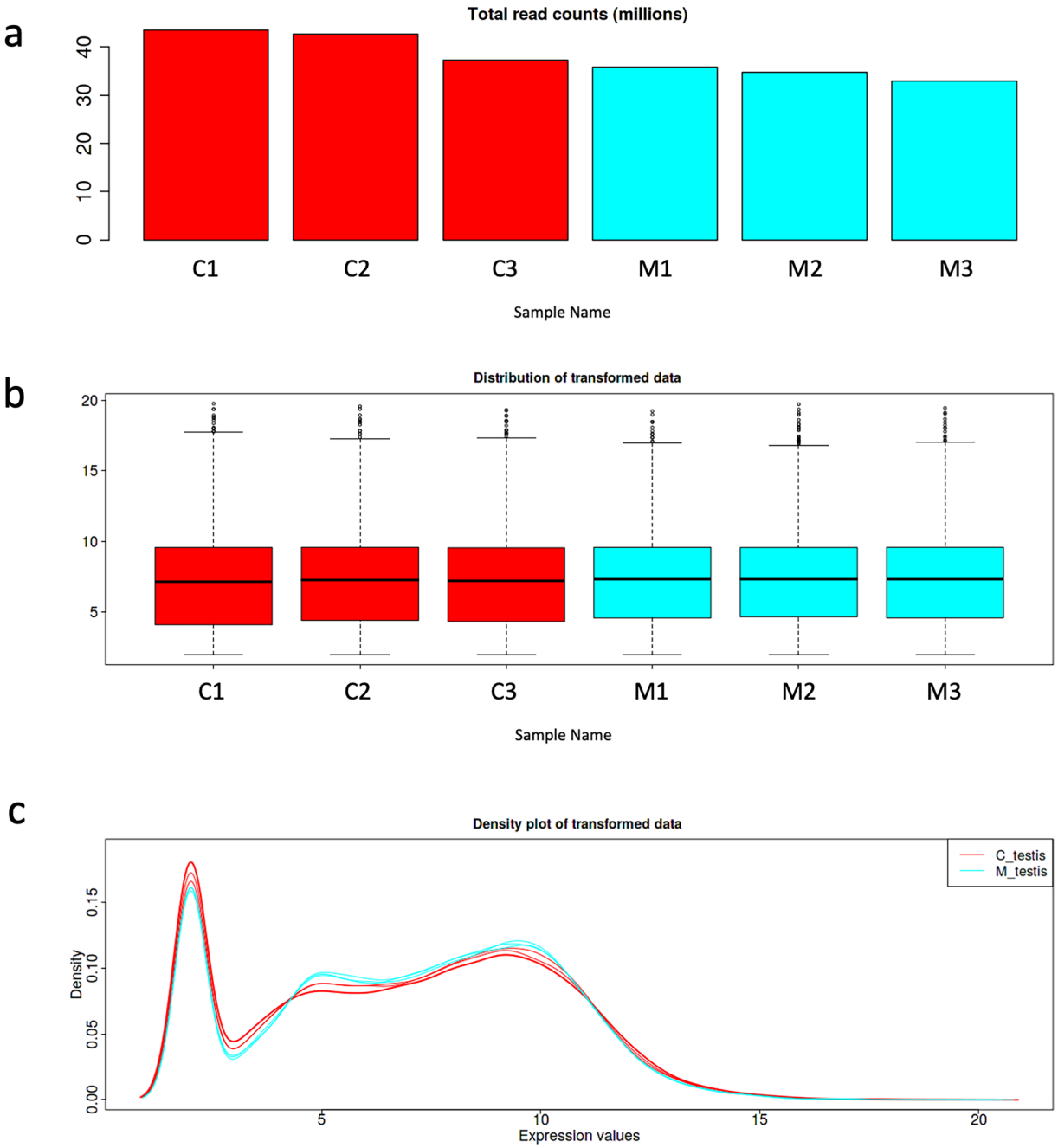

Fig. 3 Normalization of the transcriptome data for the clustering and principal component analysis (PCA). (**a**) The box plot of data before normalization. (**b**) Box plot of the normalized data. (**c**) Distribution of normalized data.

(Novogene, Beijing, China) for RNA-seq analysis. RNA integrity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) (Table 3). Note that differences in RNA concentration values between Tables 2 and 3 are due to the use of different measurement platforms: a DS-11 spectrophotometer for Table 2 and an Agilent 2100 Bioanalyzer for Table 3. Sequencing was performed on the Illumina NovaSeq 6000 system (Illumina, San Diego, CA, USA) with the S4 flow cell, resulting in 150 bp paired-end reads. Approximately 20 million raw reads were generated per library[25] (Table 4).

Pre-processing of sequencing data. To ensure high-quality reads, adapter sequences were removed and quality trimming was performed using Trim Galore![26] with its default settings: a Phred quality score threshold of 20 and a minimum read length of 20 bp. The trimmed reads were mapped on the reference genome (Sscrofa 11.1) using HISAT2[27], and the mapped reads were sorted and converted from SAM files to BAM files using SAMtools[28]. Transcript quantification was conducted with converted BAM files and the GTF file (Sscrofa 11.1) using StringTie[29]. The quantified BAM files were converted to a CSV file using prepDE.py

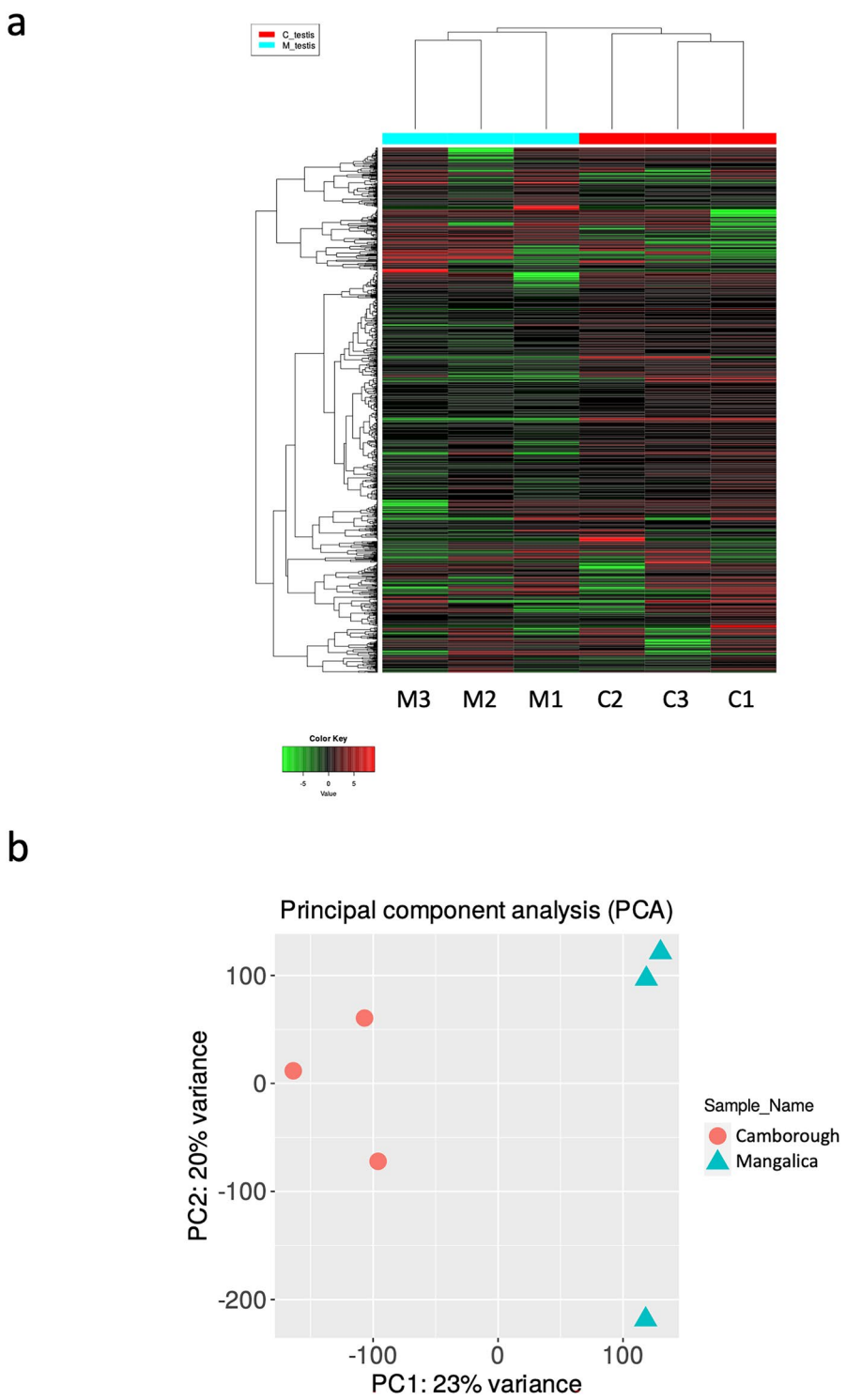

Fig. 4 The hierarchical clustering, correlation matrix and PCA. (**a**) Hierarchical clustering of the top 1000 genes by standard deviation. There are two different clusters among the breeds. (**b**) PCA plot. The variance of PC1 and PC2 show that 23% and 20% respectively. Variance represents the proportion of total variation in the dataset explained by each principal component.

(https://ccb.jhu.edu/software/stringtie/dl/prepDE.py) for gene expression analysis (DEseq2[30]). The converted CSV files were uploaded to integrated Differential Expression and Pathway analysis (iDEP.95)[31] and downstream analyses were performed using iDEP (The data of pre-processed sequencing in Figshare[25]).

Clustering. The updated data in iDEP were normalized using EdgeR[32] (counts per million (CPM) < 0.5, pseudo count = 4) for clustering (Fig. 3 and the normalized data using EdgeR for clustering in Figshare[25]). The

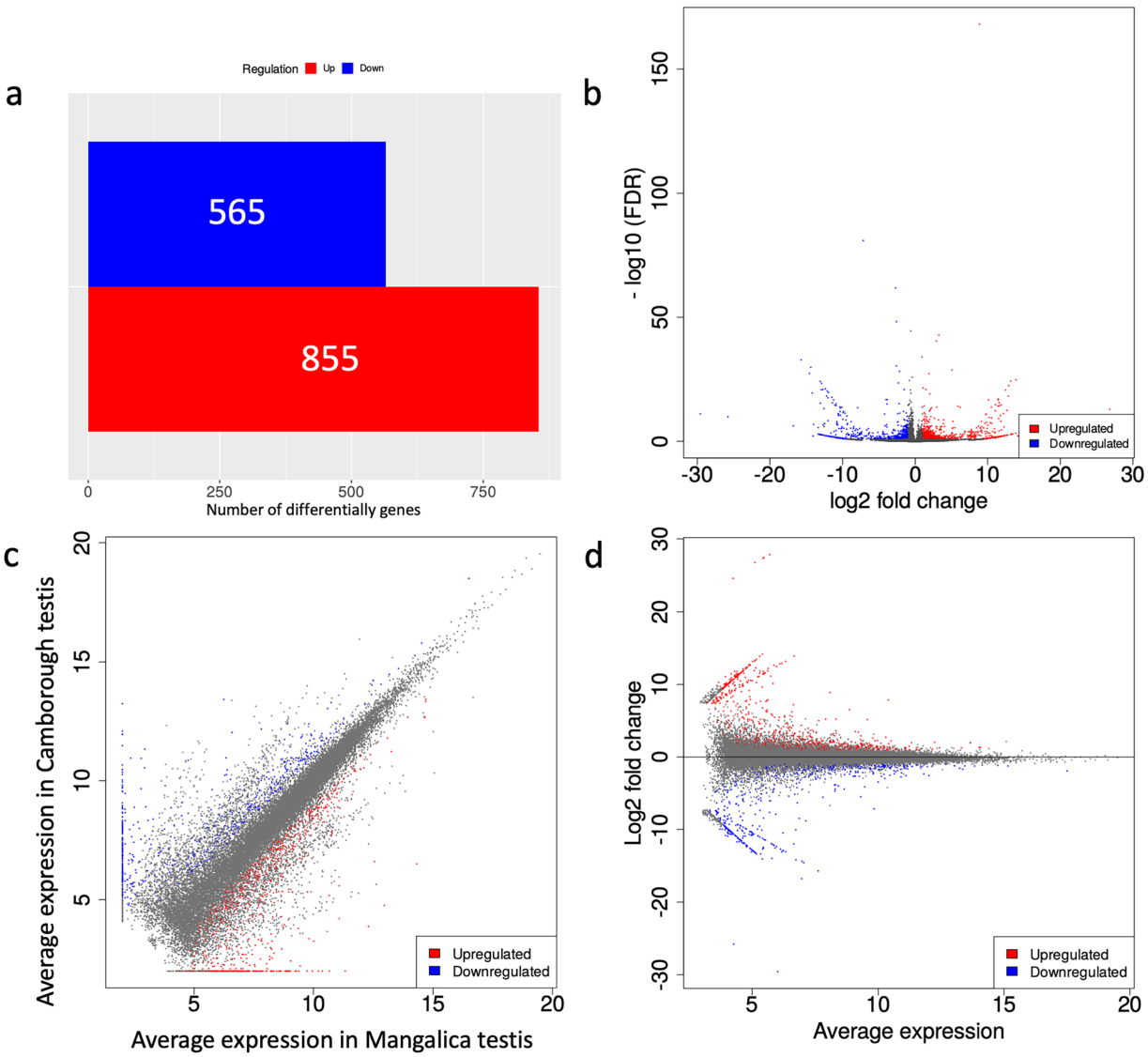

Fig. 5 Differentially expressed genes (DEGs) analysis. The red and blue plots show up and down regulated genes respectably. (**a**) The number up and down regulated DEGs is shown in a bar graph (FDR ≦ 0.1 and fold-change >2). (**b**) Volcano plot of DEGs show expression ratios (log2 fold change) and statistically significant differences (FDR). (**c**) Scatter plot of DEGs show average expression between Mangalica and Camborough testis. (**d**) MA plot (log ratios (M) vs. mean average (A) plot) of DEGs show expression ratios (log2 fold change) and average expression.

normalized data were ranked according to their standard deviation, as a higher variance indicates greater differences between samples, which can help identify biologically meaningful patterns. To balance biological informativeness with visualization clarity, the top 1000 genes were selected. This approach ensures that key expression patterns are preserved while minimizing noise (Fig. 4a and "The top 1000 genes ranked by standard deviation in hierarchical clustering"[25]). Figure 4b shows the correlation of the matrix between samples (The correlation of the matrix between samples[25]). The Principal component analysis (PCA) plots were visualized using the first and second principal components to separate the gene expression of the testis in both boar breeds (Fig. 4b and The data of principal components analyses[25]).

Differential expressed genes (DEGs). We filtered low-expression genes by retaining only those with a minimum of 0.5 CPM in at least one library. The filtered genes were mapped to pig gene IDs and converted to Ensembl gene IDs (Sscrofa11.1) using the iDEP database. DEGs were identified using DESeq2, applying a false discovery rate (FDR) < 0.1 and a fold-change > 2. The DEGs were visualized using volcano plots, scatter plots, and Mean-Average (MA) plots[33] (Fig. 5a–d). As shown in Fig. 5c, upregulated genes are primarily specific to the Mangalica breed, while downregulated genes are associated with the Camborough breed. For GO enrichment analysis, we used all expressed genes as the background set to avoid biased interpretations, ensuring that the enrichment results accurately reflect the biological context.

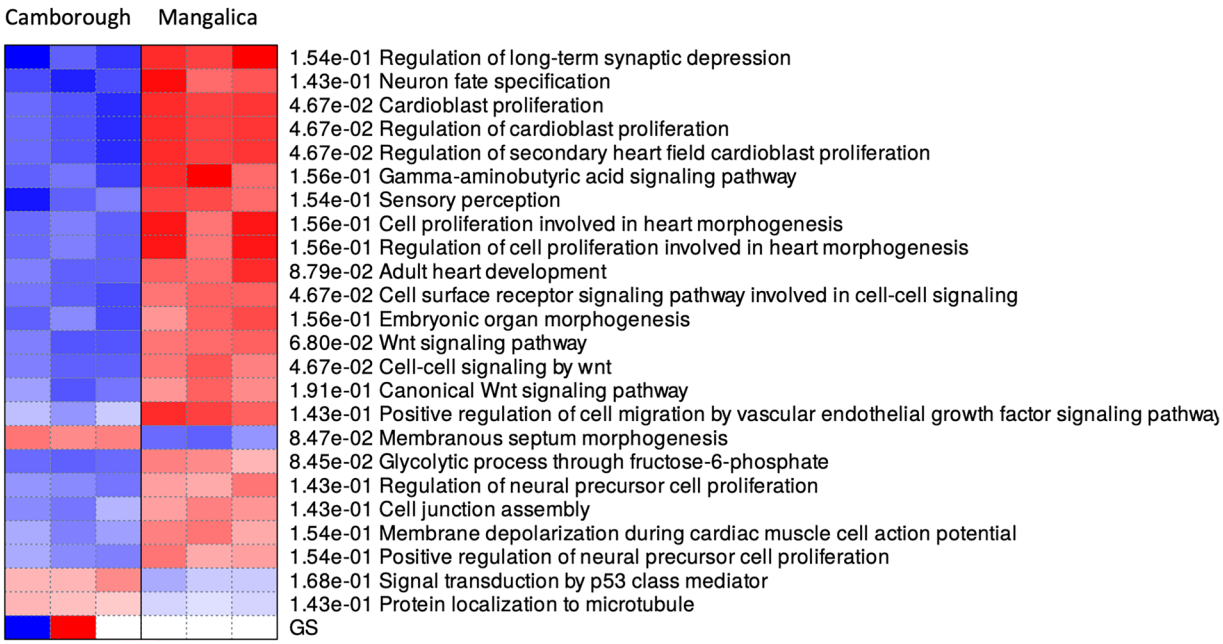

Fig. 6 Pathway analysis of the DEGs. Red and blue indicate activated and suppressed pathways, respectively. There are different activated and suppressed pathways among the breeds. Additionally, several pathways involved in cell proliferation, differentiation, and metabolic processes were identified, which are critical for testicular development and spermatogenesis[36–38].

Enrichment and pathway analysis. The downstream analysis of DEGs was conducted using the iDEP95 platform. Enrichment analyses of the DEGs were performed by selecting the GO biological process (The list of DEGs's enrichment analyses[25]). For pathway analyses of the DEGs, we utilized the PGSEA (Parametric Gene Set Enrichment Analysis)[34] method and selected the GO biological process as the gene sets, with the specified gene-set size of >15 and an FDR threshold of <0.2[25] (Fig. 6 and The list of DEG's pathway analyses[25]).

Data Records

All sequence reads were submitted to registration in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (BioProject accession: PRJDB12748)[35]. The files are available in Figshare at https://figshare.com/s/4408ed09a959bcc4f80c25.

Technical Validation

Testis morphology. The stained sections by H&E confirmed normal morphology of the testicular tissues. All of the testicular samples had normal morphology compared with testicular tissues from similarly-aged boars reported in the past[22] (Fig. 2).

RNA integrity. The integrity of the total RNA sample quality was measured using Agilent 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA). The samples with RNA Integrity Number (RIN) $\geqq 8$ and rRNA ratio (28 s/18 s) $\geqq 1.4$ were subjected to sequencing library construction (Table 3).

Quality validation and analyses. We measured the sequence quality per base, GC content per sequence, sequence duplication levels, and quality score distribution over all sequences of the fastq files (Fig. S1; Figshare[25]).

Code availability

No custom code was used in this study.
Trimgalore! (Version 0.6.5): https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
FastQC (Version 0.11.9): https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
Sscrofa 11.1: http://asia.ensembl.org/Sus_scrofa/Info/Index.
Hisat2 (Version 2.2.1): http://daehwankimlab.github.io/hisat2/.
SAMtools: http://www.htslib.org/.
StringTie (Version 2.2.0): https://ccb.jhu.edu/software/stringtie/.
EdgeR (3.36.0): https://bioconductor.org/packages/release/bioc/html/edgeR.html.
prepDE.py: https://github.com/gpertea/stringtie/blob/master/prepDE.py.
DEseq2 (1.34.0): https://bioconductor.org/packages/release/bioc/html/DESeq2.html.
iDEP.95: http://bioinformatics.sdstate.edu/idep95/.

Received: 9 December 2024; Accepted: 14 May 2025;
Published online: 28 May 2025

References

1. Griswold, M. D. Spermatogenesis: The Commitment to Meiosis. *Physiol. Rev.* **96**, 1–17 (2016).
2. Eacker, S. M. *et al*. Hormonal regulation of testicular steroid and cholesterol homeostasis. *Mol. Endocrinol.* **22**, 623–635 (2008).
3. M Nistal, R. P. Non-neoplastic diseases of the testis. *Urologic Surgical Pathology* 614, https://doi.org/10.1016/B978-0-323-01970-5.50014-2 (2008).
4. McLachlan, R. I., Rajpert-De Meyts, E., Hoei-Hansen, C. E., de Kretser, D. M. & Skakkebaek, N. E. Histological evaluation of the human testis–approaches to optimizing the clinical value of the assessment: mini review. *Hum. Reprod.* **22**, 2–16 (2007).
5. Bhang, D. H. *et al*. Testicular endothelial cells are a critical population in the germline stem cell niche. *Nat. Commun.* **9**, 4379 (2018).
6. Neto, F. T. L., Bach, P. V., Najari, B. B., Li, P. S. & Goldstein, M. Spermatogenesis in humans and its affecting factors. *Semin. Cell Dev. Biol.* **59**, 10–26 (2016).
7. Hermann, B. P. *et al*. The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. *Cell Rep.* **25**, 1650–1667.e8 (2018).
8. Guo, J. *et al*. The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. *Cell Stem Cell* **26**, 262–276.e4 (2020).
9. Lunney, J. K. Advances in swine biomedical model genomics. *Int. J. Biol. Sci.* **3**, 179–184 (2007).
10. Walters, E. M. & Prather, R. S. Advancing swine models for human health and diseases. *Mo. Med.* **110**, 212–215 (2013).
11. Swindle, M. M., Makin, A., Herron, A. J., Clubb, F. J. Jr & Frazier, K. S. Swine as models in biomedical research and toxicology testing. *Vet. Pathol.* **49**, 344–356 (2012).
12. Ibrahim, Z. *et al*. Selected physiologic compatibilities and incompatibilities between human and porcine organ systems. *Xenotransplantation* **13**, 488–499 (2006).
13. Walters, E. M., Wells, K. D., Bryda, E. C., Schommer, S. & Prather, R. S. Swine models, genomic tools and services to enhance our understanding of human health and diseases. *Lab Anim.* **46**, 167–172 (2017).
14. Meurens, F., Summerfield, A., Nauwynck, H., Saif, L. & Gerdts, V. The pig: a model for human infectious diseases. *Trends Microbiol.* **20**, 50–57 (2012).
15. Zheng, Y. *et al*. Establishment of cell lines with porcine spermatogonial stem cell properties. *J. Anim. Sci. Biotechnol.* **11**, 33 (2020).
16. Goel, S. *et al*. Identification, isolation, and *in vitro* culture of porcine gonocytes. *Biol. Reprod.* **77**, 127–137 (2007).
17. Egerszegi, I., Rátky, J., Solti, L. & Brüssow, K.-P. Mangalica–an indigenous swine breed from Hungary. *Archives Animal Breeding* **46**, 245–256 (2003).
18. Egerszegi, I., Brüssow, K. P., Sarlós, P., Manabe, N. & Rátky, J. Reproductive Capacity of Mangalica pigs – What is the Reality? *Arch. Zootec*. 81–86 (2018).
19. Chen, K., Baxter, T., Muir, W. M., Groenen, M. A. & Schook, L. B. Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). *Int. J. Biol. Sci.* **3**, 153–165 (2007).
20. Knauer, M. & Hostetler, C. US swine industry productivity analysis, 2005 to 2010. *J. Swine Health Prod.* **21**, 248–252 (2013).
21. PIC. Camborough: The world's most prolific sow. PIC official website. https://www.pic.com (2024).
22. Culty, M. Gonocytes, the forgotten cells of the germ cell lineage. *Birth Defects Res. C Embryo Today* **87**, 1–26 (2009).
23. Mäkelä, J.-A., Koskenniemi, J. J., Virtanen, H. E. & Toppari, J. Testis Development. *Endocr. Rev.* **40**, 857–905 (2019).
24. Ding, H., Luo, Y., Liu, M., Huang, J. & Xu, D. Histological and transcriptome analyses of testes from Duroc and Meishan boars. *Sci. Rep.* **6**, 20758 (2016).
25. Keigo, Y. *et al*. Transcriptome analysis of testes gene expression to explore genetic diversity of Mangalica and Camborough boars. *figshare* https://doi.org/10.6084/m9.figshare.20712592.v1 (2023).
26. Krueger, F. Trim galore. *A wrapper tool around Cutadapt and FastQC to* (2015).
27. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
28. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
29. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
31. Ge, S. X., Son, E. W. & Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics* **19**, 534 (2018).
32. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
33. Yang, Y. H. *et al*. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic acids research* **30**(4), e15–e15 (2002).
34. Kim, S.-Y. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**, 144 (2005).
35. Keigo, Y. *et al*. Transcriptome data of testes from Red Mangalica and Camborough boars., *DDBJ Sequence Read Archive, BioProject accession: PRJDB12748.*, https://identifiers.org/ncbi/insdc.sra:DRP010019 (2023).
36. Matsui, Y. & Hayashi, Y. Metabolic pathways regulating the development and non-genomic heritable traits of germ cells. *Journal of Reproduction and Development* **68**(2), 96–103 (2022).
37. Hayashi, Y. & Matsui, Y. Metabolic control of germline formation and differentiation in mammals. *Cells* **11**(2), 1–15 (2022).
38. Helsel, A. R. *et al*. ID4 levels dictate the stem cell state in mouse spermatogonia. *Development* **144**(4), 624–634 (2017).

Acknowledgements

We would like to thank to the staff of the Royal Mangalica Farm for providing everyday maintenance and care for the animals. We also thank the commercial pig company, Dream Pork. This work was supported by the Northern Advancement Center for Science & Technology (NOASTEC) and Obihiro City - Obihiro University of Agriculture and Veterinary Medicine Collaboration Project.

Author contributions

K.Y. and Y.M. contributed to the conception and experimental design of the study. Y.K. and Y.M. performed the experiments. M.K. advised technically bioinformatic analysis and data base information. Data were analysed by K.Y., Y.M., S.K., A.K., E.Y., M.K. The article was written by K.Y. and Y.M. and critically revised by S.K., A.K., E.Y., M.K.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.M.