scientific data

DATA DESCRIPTOR

Check for updates

OPEN Comprehensive visceral transcriptome profiling of three pig breeds along altitudinal gradients in Yunnan

Jinlong Huo^{1,11}, Zhipeng Liu^{1,11}, Xia Zhang^{2,11}, Changyao Li^{1,11}, Decai Xiang³, Guowen Fu⁴, Wan Lin¹, Lingxiang Wu¹, Shaorong Gong⁵, Jiading Zhao⁵, Zhen Wang⁶, Xiaohong Wang⁷, Zhiping Xiao⁸, Fanfan Hao⁹, Yue Ren⁹, Yu H. Sun 10¹⁰ & Guiying Zhao 10¹

High-altitude hypoxia acclimatization requires comprehensive physiological regulation in highland immigrants, yet the genetic mechanisms remain unclear. Yunnan's vertical zoning with pig breeds distributed across varying elevations provides an excellent model for investigating hypoxic adaptation. Here, we examined three indigenous Yunnan pig breeds: Diannan small-ear pigs (DSE, 500 m), Baoshan pigs (BS, 1500 m), and Diging Tibetan pigs (DT, 3200 m). Using PacBio Iso-Seq, we obtained comprehensive full-length transcriptomes from five tissues (heart, kidney, liver, lung, and spleen), identifying 51,774 transcripts, including 34,813 novel ones, 74,843 Alternative Splicing (AS) events across 10,686 AS genes and pinpointed five actin-binding genes through weighted gene coexpression network analysis (WGCNA). Our research significantly improved porcine genome annotation and provided a high-guality transcriptome resource for investigating the genetic mechanisms of high-altitude hypoxia adaptation. This work lays a solid foundation for future studies in genetics, evolutionary biology, and environmental adaptation.

Background & Summary

The established demarcation for the onset of high-altitude illnesses has been conventionally set at 2500 meters, with sporadic instances of mild conditions manifesting at elevations between 2000 meters and the threshold¹⁻³. Genomic investigations into the adaptive mechanisms for enduring hypoxia have unveiled a plethora of positively selected genetic and non-genetic elements that correlate with cardiorespiratory and metabolic adaptations in both humans inhabiting highland regions and other high-altitude vertebrates⁴⁻⁷. The Yunnan province in southwest China is characterized by its complex topography, vast altitudinal span (ranging from 76 meters to 6740 meters), variable climate, and rich biodiversity. Based on these criteria, our research focused on three indigenous pig breeds derived from regions with distinct altitudes and climates. These breeds include the Diannan small-ear pig (DSE) from the Xishuangbanna Autonomous Prefecture (~500 m in altitude), which exhibits thermotolerance; the Baoshan pig (BS) from Baoshan city (~1500 m in altitude), noted for its high environmental adaptability; and the Diqing Tibetan pig (DT) from Diqing Prefecture (~3200 m in altitude), which demonstrates adaptations to hypoxia and low temperatures.

¹College of Animal Science and Technology, Yunnan Agricultural University, Kunming, 650201, Yunnan, China. ²Department of Biological and Food Engineering, Lyuliang University, Lyliang, 033001, Shanxi, China. ³Yunnan Academy of Animal Husbandry and Veterinary Sciences, Kunming, 650224, Yunnan, China. ⁴College of Veterinary Medicine, Yunnan Agricultural University, Kunming, 650201, Yunnan, China. ⁵Baoshan Pig Research Institute, Baoshan, 678200, Yunnan, China. ⁶Institute of Animal Husbandry and Veterinary Science of Diging Tibetan Autonomous Prefecture, Diqing, 674499, Yunnan, China. ⁷Animal Health Supervision Institute, Bureau of Agriculture and Rural Affairs of Shangri-la, Shangri-la, 674499, Yunnan, China. ⁸Pure Land Agricultural Development Co., LTD, Shangri-Ia, 674401, Yunnan, China. 9School of Medicine and Dentistry, University of Rochester Medical center, Rochester, New York, 14642, USA. ¹⁰Department of Biology, University of Rochester, Rochester, New York, 14627, USA. ¹¹These authors contributed equally: Jinlong Huo, Zhipeng Liu, Xia Zhang, Changyao Li. ¹²e-mail: jinlonghuo973@163.com; yu.sun.rochester@gmail.com; zhaoguiying2006@163.com

Current pig gene annotation databases, such as Ensembl⁸ and NCBI RefSeq⁹ are primarily based on assembly methods for short-read RNA-Seq data. Due to the restricted read length, this technology often produces isoform annotations that are fragmented, incomplete, or improperly compressed^{10,11}. The advent of PacBio Iso-Seq method rectifies these shortcomings by enabling the sequencing of full-length mRNAs with a span that can extend up to 25–60 kb^{12–14}. The applications of Iso-Seq have identified a wide range of hitherto uncharacterized isoforms across several taxa, including but not limited to humans, mice, and plants^{12,15–17}. Many of these studies have synergistically combined Iso-Seq with short-read RNA-Seq datasets to enhance the robustness of isoform expression quantification^{18–20}.

In the present study, we generated high-quality PacBio Iso-Seq long reads data and paired-end RNA-Seq short reads data for five tissues (heart, kidney, liver, lung, and spleen) from three breeds (Fig. 1a). Finally, we constructed a high-quality transcriptome of Yunnan indigenous pigs associated with altitude. In total, we identified 51,774 transcripts, including 34,813 novel transcripts. Weighted gene co-expression network analysis (WGCNA) revealed five actin-binding genes including FHOD3, TNNC1, ACTN2, PDLIM5, and TNNI3. The constructed transcriptome will be a valuable dataset for further explorations into the genetic underpinnings of hypoxic adaptation and facilitate broader comparative transcriptomic assessments within porcine species.

Methods

Sampling and sequencing. This study received ethical approval from the Life Science Ethics Committee of Yunnan Agricultural University and adhered to the standards set forth by the People's Republic of China for the care and use of laboratory animals (Approval No. 2006-398). Sample collection was conducted at three different breeding locations: the Diannan Small Ear-Pig Breeding Farm in Xishuangbanna for DSE samples, the Baoshan Pig Research Institute for BS samples, and the Diqing Tibetan Pig Breeding Farm for DT samples. Heart, kidney, liver, lung, and spleen tissues were collected from each of the three pig breeds at 12 months of age. The feeding conditions for the three pig breeds were standardized during the rearing process. All pigs were provided with the same diet, feeding schedule, housing conditions, and veterinary care to ensure uniformity in environmental factors. The collected biological specimens were immediately stabilized by snap-freezing in liquid nitrogen and subsequently maintained in a -80 °C storage facility. Extraction of total RNA from a compendium of sixty tissue samples was performed using TRNzol reagent (TIANGEN, China). Post-extraction, the RNA samples underwent purity assessment with the NanoDrop (Thermo Fisher Scientific, USA) and precise quantification using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA). The integrity of the RNA was systematically evaluated by ascertaining the RNA Integrity Number (RIN) and the 28S/18S ribosomal RNA ratio with the aid of an Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

PacBio Iso-Seq sequencing began with the reverse transcription of RNA into complementary DNA (cDNA) using the NEBNext[®] Single Cell/Low Input cDNA Synthesis& Amplification Module. Polymerase chain reaction (PCR) amplification was then performed utilizing the Iso-Seq Express Oligo Kit (Pacific Biosciences, USA). Following PCR amplification, PacBio SMRTbell libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA). Finally, the generated libraries underwent Isoform Sequencing (Iso-Seq) on the Sequel II system using Sequel II Binding Kit 2.0 (Pacific Biosciences, USA).

For short-read RNA sequencing, poly-A RNAs were initially enriched using Dynabeads mRNA Purification Kit (Invitrogen, USA). Subsequently, the libraries were constructed using the MGIEasy RNA Library Prep Kit V3.1 (MGI, China), encompassing key steps such as mRNA fragmentation, first- and second-strand cDNA synthesis, end repair, A-tailing addition, and adapter ligation. The cDNA libraries were then amplified by PCR and purified using MGIEasy DNA Clean beads (MGI, China). Following heat denatured and circularized using MGIEasy Circularization Module (MGI, China), the final libraries underwent paired-end sequencing (150 base pairs) on the DNBSEQ-T7 platform.

Long-read data processing. SMRTLink (v25.1)²¹ tool was used to process PacBio subreads to generate high-quality full-length transcript sequences. Initially, circular consensus sequences (CCS) were generated using ccs command with the specified parameter "-top-passes 60-min-rq 0.9", primer removal and demultiplexing using lima (v2.9.0) with parameters "-isoseq-peek-guess". Subsequent step aimed at curating full-length non-chimeric (FLNC) reads, we eliminated chimeric reads and trimmed poly-A tails using the "isoseq refine" with parameter "-require-polya". To eliminate redundancy caused by the generation of identical transcripts in each sample, we combined 15 FLNC bam files with "pbmerge" and performed the clustering phase using "isoseq cluster2" with the parameter "-use-qvs". The clustered HiFi (High Fidelity) isoforms achieved a prediction accuracy of \geq 0.99. Further analysis required aligning these HiFi isoforms, which effectively represented a draft of the full-length transcriptome, against the *Sus scrofa* reference genome assembly (Sscrofa11.1). This alignment was accomplished using pbmm2 (v1.13.1), a minimap2²² wrapper, with the alignment settings set to "-preset ISOSEQ-sort". The final stage of this process involved the use of cDNA_Cupcake (version 29.0.0)²³ to combine the full-length isoforms, which improves the accuracy of assembled transcriptome.

Furthermore, SQANTI3 (v5.2)²⁴ was employed to evaluate the transcriptome assemblies derived from the IsoSeq and cDNA_Cupcake processing pipelines. This evaluation aimed to benchmark our assembled transcriptome against the reference isoform annotations detailed in the *Sus scrofa* annotation release-109. Through the utilization of SQANTI3, a comprehensive identification and characterization of novel isoforms were undertaken. These novel isoforms encompassed sequences emerging from previously uncharacterized genomic loci, antisense isoforms generated from established genes, and isoforms resulting from the fusion of distinct gene sequences.



Fig. 1 Identification and characterization of isoforms in pig breeds by Iso-Seq. (**a**) Schematic representation of isoform profiling in five tissues across three pig breeds, utilizing both Iso-Seq and short-read RNA-Seq. (**b**) Classification of Iso-Seq isoforms from all samples. (**c**) The left tree represents hierarchical clustering of Iso-Seq data using the Jaccard pairwise similarity coefficient, while the right bar depicts discovered isoforms in each sample, color-coded by category. (**d**) Distribution of read lengths for all discovered isoforms across all samples. (**e**,**f**) Examination of the correlation between the exon count and the isoform number. (**e**) isoforms in Ensembl 109; (**f**) all Iso-Seq isoforms detected in this study. (**g**) Bar graph illustrating isoform distribution across different classifications according to their predicted protein-coding potential and the identification of known Pfam domains within the encoded peptides. (**h**) Violin plot showing the ORF length distribution of isoforms per category. (**i**) Percentage of Iso-Seq isoforms categorized as NMD sensitive. (**j**) Proportion of amino acid sequence identity between Iso-Seq derived ORFs and pig protein isoforms recorded in the UniProt database.

Isoform quality control. To ensure the reliability of our transcriptome, a rigorous artifact filtration strategy was implemented following the comparative analysis with known isoforms using the SQANTI3. This approach entailed two critical filtering procedures: initially, the exclusion of isoforms indicative of reverse transcription (RT) template switching, as well as those exhibiting noncanonical splicing junctions. Subsequently, the genomic 3' ends of transcripts that deviated by over 100 base pairs from an annotated transcription termination site (TTS) were deemed unreliable and, consequently, removed from further analysis.

Rarefaction curve analysis. The assessment of isoform diversity through rarefaction curve analysis is a critical process to ascertain whether the genes or isoforms have been adequately detected in this study. This



Fig. 2 Examination of coding or non-coding RNA isoforms. (a) Comparative expression analysis between coding and non-coding isoforms. The significance was calculated by ANOVA. (b) The number of coding and non-coding isoforms identified by Iso-Seq (also divided by known or novel). (c) Classification of novel non-coding isoforms (left) and comparing their position with protein-coding genes (right). (d) The length of isoforms and the number of exons in coding or non-coding and known or novel isoforms. (e) The chromosomal landscape of isoforms in all Iso-Seq data at the genome level. I: Pig chromosomes, length in Mb; II: density of known coding isoforms, in 1 Mb intervals; III: density of novel coding isoforms; IV: length of coding isoforms; V: density of known non-coding isoforms, in 1 Mb intervals; VI: density of novel non-coding isoforms, in 1 Mb intervals; VII: length of non-coding isoforms.

analysis was carried out using the subsample.py and subsample_with_category.py scripts from the cDNA_Cupcake GitHub repository.

Isoform clustering. To elucidate the relationships among samples characterized by Isoform Sequencing (Iso-seq), a hierarchical clustering analysis was conducted within the R environment. First, we calculated the Jaccard pairwise similarity coefficient using "vegdis" function (parameter: method = "jaccard") of the vegan $(v2.6-4)^{25}$ package; and then clustered the resulting coefficient matrix using "hclust" function (parameter: method = "single") from the stats package.



Fig. 3 Examination of short-read RNA-Seq data. (a) The number of detected genes (left) and the percentage of reads mapped to genes (right). (b) The quantification of reads alignments corresponding to distinct gene regions. (c) Principal Component Analysis (PCA) of gene expression, colored by tissue, shaped by breed. (d) Hierarchical clustering of gene expression based on the Jaccard pairwise similarity coefficient.

Protein domain scan. GeneMarkS-T (GMST)²⁶, a tool integrated into SQANTI3, can identify the most probable protein-coding regions from RNA transcripts. For each potential coding isoform, protein domains were scanned using pfam_scan.pl (v1.6) based on Pfam database (v36.0)²⁷.

Isoform functional characterization. To identify functional coding sequences within long-read isoforms, a comprehensive reference proteome of pig was assembled, encompassing both canonical entries (SwissProt and TrEmbl) sourced from UniProt (release 2024_01)²⁸, containing a total of 4,285,756 protein sequences. Subsequently, RNA sequences procured from Iso-Seq were aligned to this reference proteome utilizing the diamond tool (v2.1.8)²⁹ through the "blastx" module, applying parameters "-outfmt 6-long-reads-max-target-seqs. 1-id 50-more-sensitive".

Coding or non-coding isoforms analysis. GMST divides isoform sequences into coding or non-coding, and we evaluated the expression difference for coding and non-coding isoforms in each sample using ANOVA model. Moreover, we located the position between our novel non-coding isoforms and known protein-coding genes using bedtools (v2.30.0)³⁰. The chromosome level density and distribution of coding or non-coding isoforms were visualized using the R package circlize (v0.4.15)³¹.

Short reads paired-end sequencing data processing. The raw FASTQ files derived from RNA-Seq were subjected to preprocessing using fastp (v0.23.4)³², which eliminated adapter reads, low-quality reads, and poly-N-containing reads. The resulting high-quality reads were then meticulously aligned to the pig reference genome (Sscrofa11.1) using STAR (v2.7.10a)³³ software. The key alignment options were listed as following: "-alignEndsType EndToEnd-outFilterIntronMotifs RemoveNoncanonicalUnannotated-outSAMattributes All-outSAMtype BAM". Notably, the alignment process incorporated annotations from both the Ensembl reference database and our de-novo assembled transcriptome. The produced alignment bam files were then sorted and indexed utilizing samtools (v1.6)³⁴. The quantification of read distributions across genes was executed through featureCounts (v2.0.1)³⁵ with the specified parameters "-p -t exon -g gene_id". The RseQC (v5.0.1)³⁶ package's read_distribution.py and tin.py were used to calculate the transcript integrity number (TIN) and read distribution



Fig. 4 Weighted correlation network analysis. (a) Dendrogram depicting the outcomes of gene co-expression network analysis through hierarchical clustering. (b) Number of genes in each module. (c) Gene type in each module. (d) Illustrative network diagram delineating the interconnections among genes from distinct modules within the gene co-expression network, including the top 17 hub genes (adjacency threshold > 0.5).

over genomic features, respectively. Gene and isoform expression quantification was achieved using the "quant" module (salmon v1.10.1)³⁷. Principal component analysis (PCA) was performed using "prcomp" function from the stats package of R (v4.3.2). The sample hierarchical clustering procedure was identical to that described in the **Isoform clustering** section.

Weighted gene co-expression network analysis. Normalized expression data from short-read RNA-Seq samples were utilized to construct gene co-expression networks employing the WGCNA (v1.72-1)³⁸ package, R environment (v4.3.2). Genes were subjected to a filtering criterion based on a threshold of median absolute deviation (MAD) greater than 0.5 across all samples prior to network construction. Co-expression modules were subsequently delineated, characterized by genes that exhibit positive correlation and significant topological overlap. This identification process leveraged the signed weighted correlation matrix, utilizing the blockwiseModules function within the WGCNA package (parameters: power = 6, minModuleSize = 30, mergeCutHeight = 0.3). The edge and node files of co-expression network were constructed using the export-NetworkToCytoscape function of WGCNA based on TOM matrix. The co-expression network was visualized using Cytoscape³⁹.

Identification of alternative splicing (AS) events. Alternative splicing events across all samples were profiled using SUPPA2 (v2.3)⁴⁰. First, we identified seven types of alternative splicing (AS) events using the SUPPA2 generateEvent function based on the reference GTF and our de novo assembled transcriptome GTF: skipping exon (SE), alternative 5' splice sites (A5), alternative 3' splice sites (A3), retained intron (RI), mutually exclusive exon (MX), alternative first exon (AF), and alternative last exon (AL). Further, transcript-level transcripts per million (TPM) values, derived from salmon (v1.10.1), were employed to quantify the percent spliced



Fig. 5 Identification of Alternative Splicing (AS) events. (a) Schematic representation of seven distinct categories of alternative splicing events. (b) The number of AS events generated by Ensembl (left) and Iso-Seq novel isoforms (right). (c) The number of AS events detected in each sample (events PSI > 0.1). (d) Profile of AS event expression across individual samples.

in (PSI) values across seven distinct types of alternative splicing events for each sample. This quantification was facilitated using the psiPerEvent function of the SUPPA2 software.

Data Records

The long-read Iso-Seq raw data and short-read RNA-Seq raw data from the five visceral organs (heart, liver, spleen, lung, and kidney) of Diannan small-ear pigs, Baoshan pigs, and Diqing Tibetan pigs have been archived in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) with the accession number GSE269429⁴¹ and SRP512775⁴², respectively. The annotation of the assembled full-length transcriptome, the sequences of full-length transcripts, isoform classification, functional annotations including Pfam scan results, splicing data, and the quantification of genes and transcripts have now been archived in the Figshare database⁴³.

Technical Validation

High-quality data from long-read sequencing. To delineate the isoform landscape within the viscera of distinct pig breeds, we extracted RNA from the heart, kidney, liver, lung, and spleen and performed Iso-Seq (Fig. 1a). We generated a total of approximately 936 Gb of PacBio subreads. As a quality control step after IsoSeq pipeline, a total of 7,800,974 full-length non-concatemer long reads were obtained. Moreover, the mapping rate of clustered HiFi isoforms is more than 99%. The length distribution of CCS, full-length isoforms, full-length non-chimeric isoforms, and clustered HiFi isoforms ranged from 1–10 Kb across all five tissues, including heart, kidney, liver, lung, and spleen. Upon eliminating redundant sequences and 5' degraded products via

cDNA_Cupcake and excluding unreliable 3' ends, RT-switching, and noncanonical junctions based on SQANTI3 output, we obtained 51,774 isoforms, of which the most was novel not in catalog (NNC, constituting novel isoforms with at least one new splice site, 19,478, 37.6%), succeeded by novel in catalog (NIC, referring to novel isoforms constituted by a new combination of known splice sites, 14,641, 28.3%), full-splice matches (FSM, isoforms that correspond precisely with reference splice junctions, 13,067, 25.2%), incomplete-splice matches (ISM, isoforms that partially correspond with reference splice junctions, 3,894, 7.5%), and Other (encompassing fusion and genic isoforms, 694, 1.3%) (Fig. 1b). Sample clustering based on the Jaccard pairwise similarity coefficient reveals that tissues were grouped independently (Fig. 1c). Moreover, novel isoforms constituted approximately 60% of the sequenced transcripts across the individual samples (Fig. 1c). We observed that the peak of isoform read lengths occurred at approximately 3 Kb (Fig. 1d). A positive correlation was observed between the number of exons and number of ENSEMBL known isoforms (Pearson r = 0.6882) and Iso-Seq isoforms (Pearson r = 0.6876) (Fig. 1e,f), signifying that genes exhibiting greater exon complexity are associated with a wider variety of isoforms. Leveraging the protein-coding predictions of GMST, isoforms across all categories exhibited exceptionally high coding potential, with values spanning from 94.9% to 97.1% (Fig. 1g). Next, we used PfamScan²⁷ to identify protein domains contained in the potential protein-coding sequences. Most of the isoforms transcribed from recognized genes were anticipated to produce proteins that contain at least one known domain (Fig. 1g), with FSM being the most frequently observed category (91.4%), followed by NIC (87.8%), NNC (87.7%), ISM (85.7%) and Other (85.7%). The median lengths of ORFs (open reading frames) spanned from 380 to 448 across each category (Fig. 1h). Intriguingly, isoforms resulting from NIC and NNC splicing appear more susceptible to mRNA degradation via NMD sensitivity, because observed 20.87% of ORFs translated from NIC and 18.05% from NNC contain premature termination codons, in stark contrast to only 1.09% of ORFs derived from FSM (Fig. 1i). To ascertain the novelty of protein sequences encoded by Iso-Seq isoforms, we conducted comparisons of their ORFs against the pig UniProt database utilizing the diamond tool. A significant proportion of FSM and ISM isoforms, constituting 96.1% and 92.04% respectively, encode ORFs that exhibit over 99% similarity to entries within the UniProt database (Fig. 1j). We also calculated the number of isoforms per gene, identifying 13,088 genes with multiple isoforms and 2,964 genes with more than ten isoforms. For each sample, > = 10 isoforms of genes range 70 to 294.

Rarefaction curve. To evaluate whether isoforms and genes were sufficiently identified in the sequencing, we conducted rarefaction analysis for multiple categories. The conducted analysis for each gene and isoform revealed that the gene rarefaction curves reached saturation across all samples, while rarefaction curves for isoforms approached saturation. Specifically, rarefaction curves for known isoforms achieved saturation, and those delineating known or novel genes showed a tendency towards saturation. Furthermore, rarefaction curves corresponding to isoforms within each categorized group likewise exhibited trends towards saturation.

Analysis of coding or non-coding isoforms. To assess the divergence between coding and non-coding isoforms within our Iso-Seq transcriptomic dataset, we commenced by analyzing their expression levels. It was observed that coding isoforms exhibited a significantly lower abundance compared to non-coding isoforms across all samples (ANOVA, P < 0.05) (Fig. 2a). Subsequently, these isoforms were categorized into "known" or "novel," revealing a predominance of novel isoforms within both coding and non-coding categories (Fig. 2b). Further analysis was conducted on the classification and locational distribution of novel non-coding isoforms, which showed a majority classified as NNC and predominantly sense isoforms (Fig. 2c). In an additional comparison between the number of exons and the length of isoforms, it was found that the distribution of exon number and isoform length exhibited a parallel pattern among novel coding and known coding isoforms, as well as between novel non-coding and known non-coding isoforms (Fig. 2d). Furthermore, the distribution of these isoforms was noted to be uniform across all chromosomes (Fig. 2e).

Analysis of paired-end sequencing data. A total of 533.61 Gb of clean reads were generated using the DNBSEQ-T7 platform. The mean TIN (Transcript Integrity Number) score was 84.85, with a median score ranging from 74.2 to 86.98. In excess of 73.8% of sequencing reads were mapped to genes (Fig. 3a), with a predominant localization within exonic regions compared to a lesser extent in intronic regions (Fig. 3b). This distribution pattern signifies the procurement of high-quality RNA-sequencing data from the five porcine tissue samples analyzed.

To facilitate a nuanced examination of the similarities and variances among samples, we employed the principal component analysis (PCA) based on normalized gene expression. The analysis revealed that the primary and secondary principal components (PCs), accounting for 52.7% and 26.9% of the explained variance, respectively, markedly differentiated between heart, kidney, liver, and other tissue types. Further, the tertiary PC, elucidating an additional 4.8% of the variance, distinctly segregated lung and spleen tissues (Fig. 3c). The utilization of the Jaccard pairwise similarity coefficient for sample clustering demonstrated that five tissues could be clustered separately, with the Jaccard pairwise similarity coefficient between the same tissue was greater than 0.9 (Fig. 3d). Moreover, an investigation into the correlation between Iso-Seq and RNA-Seq expressions across samples unveiled a robust association, with Pearson correlation coefficients ranging between 0.45 to 0.72. This strong correlation further substantiates the reliability and accuracy of the sequencing data and analytic methodologies employed in this study.

Co-expression network analysis of known and novel genes. Weighted gene co-expression network analysis (WGCNA) was performed, resulting in the identification of eight distinct modules of co-expressed genes (Fig. 4a). We further investigated these modules, which displayed the number of genes in each module, as well as diverse categories of genes (Fig. 4b,c). We visualized the network of these modules and identified the hub

genes with high connectivity degree (degree > 100) and significant Pearson correlation with the corresponding module eigengene ($k_{\rm ME}$, P < 0.05). Out of the 17 hub genes discovered across all modules, 5 were identified as actin-binding genes, including FHOD3, TNNC1, ACTN2, PDLIM5, and TNNI3, highlighting their potential pivotal roles within the co-expression network (Fig. 4d).

Characterization of AS. To elucidate the alternative splicing (AS) patterns of full-length isoforms across five organs in three pig breeds, we quantified the occurrence of SE, A5, A3, RI, MX, AF, and AL events (Fig. 5a). Significantly, SE was identified as the dominant form of alternative splicing in the reference dataset, AF events as the most frequent AS occurrences in our dataset (Fig. 5b). Furthermore, SE represented the predominant AS event in each sample, with a Percent Spliced In (PSI) value greater than 0.1 (Fig. 5c). In addition, the transcript per million (TPM) values for mutually exclusive exons (MX) events were observed to be lower across five tissues (Fig. 5d).

Usage Notes

The Iso-seq data enables comprehensive characterization of full-length transcript and accurate detection of alternative splicing events of five visceral organs in three pig breeds across altitudes. The paired-end RNA-seq data accurately quantifies the gene expression and can be used for differential analysis and differential alternative splicing analysis.

The present study provides a valuable transcriptome resource of five visceral organs in three pig breeds across altitudes. The high-quality long-read and short-read dataset is significant for quantifying gene or transcript expression, detecting differential expression genes or differential alternative splicing genes across three pig breeds, identifying novel transcripts, and enhancing porcine genome annotation.

Code availability

All software used in the present study is open access, and detailed parameters of certain software were described in the Method section. Subsequently, the utilization of the research software is delineated.

SMRTLink (v25.1) was used to process raw Iso-seq data: https://www.pacb.com/support/software-downloads/ Isoseq (v4.3.0) was used for the quality control, clustering, and deduplication of HiFi reads: https://github.com/ PacificBiosciences/IsoSeq

pbmm2 (v1.13.1) was used to align the full-length isoforms of Iso-seq sequences to the *Sus scrofa* reference genome: https://github.com/PacificBiosciences/pbmm2

cDNA_Cupcake (v29.0.0) was used to collapse redundant isoforms to unique isoform: https://github.com/ Magdoll/cDNA_Cupcake

SQANTI3 (v5.2) was used for the quality control and classification of Iso-seq full-length transcripts: https://github.com/ConesaLab/SQANTI3

vegan (v2.6-4) was used for hierarchical clustering of gene expression: https://github.com/vegandevs/vegan Diamond tool (v2.1.8) was used to align the novel isoform sequence to NR and UniProt: https://github.com/ bbuchfink/diamond

bedtools (v2.30.0) was used to locate the position between novel non-coding isoforms and known proteincoding genes: https://github.com/arq5x/bedtools2

circlize (v0.4.15) is an R package used for visualizing the density and distribution of coding or non-coding isoforms at the chromosomal level in a circular layout: https://jokergoo.github.io/circlize/

fastp (v0.23.4) was used for quality control of RNA-seq data: https://github.com/OpenGene/fastp STAR (v2.7.10a) was used to map paired-end RNA-seq data to the *Sus scrofa* reference genome: https://github.com/ alexdobin/STAR

samtools (v1.6) was used to sort and build an index for short-reads aligned BAM file: https://github.com/samtools/samtools

featureCounts (v2.0.1) was used to calculate the gene counts: https://github.com/ShiLab-Bioinformatics/ subread

RseQC (v5.0.1) was used to evaluate the short-reads sequence quality: https://github.com/MonashBioinfor maticsPlatform/RSeQC

salmon v1.10.1) was used to calculate the expression of genes and transcripts: https://github.com/COMBINE-lab/salmon

WGCNA (v1.72-1) was used to construct gene co-expression networks: https://github.com/topics/wgcna SUPPA2 (v2.3) was used for alternative splicing events detection and alternative splicing analysis: https://github.com/comprna/SUPPA

The pipeline code for Iso-seq: https://github.com/zhipengliux/PipIsoseq

The pipeline code for RNA-seq: https://github.com/sunyumail93/PipeRNAseq

Received: 4 March 2025; Accepted: 24 April 2025; Published online: 03 May 2025

References

- Bärtsch, P. et al. Acute High-Altitude Illnesses. New England Journal of Medicine 368, 2294–2302, https://doi.org/10.1056/ NEJMcp1214870 (2013).
- 2. Yang, J. et al. Proteomic and clinical biomarkers for acute mountain sickness in a longitudinal cohort. Commun Biol 5, 548, https://doi.org/10.1038/s42003-022-03514-6 (2022).
- 3. Luks, A. M. et al. Acute high-altitude sickness. Eur Respir Rev 26, https://doi.org/10.1183/16000617.0096-2016 (2017).
- 4. Azad, P. *et al.* High-altitude adaptation in humans: from genomics to integrative physiology. *Journal of Molecular Medicine* **95**, 1269–1282, https://doi.org/10.1007/s00109-017-1584-7 (2017).

- Storz, J. F. et al. High-Altitude Adaptation: Mechanistic Insights from Integrated Genomics and Physiology. Molecular Biology and Evolution 38, 2677–2691, https://doi.org/10.1093/molbev/msab064 (2021).
- Yan, Z. et al. A time-resolved multi-omics atlas of transcriptional regulation in response to high-altitude hypoxia across whole-body tissues. Nature Communications 15, 3970, https://doi.org/10.1038/s41467-024-48261-w (2024).
- Samanta, D. et al. Systems biology of oxygen homeostasis. Wiley Interdiscip Rev Syst Biol Med 9, https://doi.org/10.1002/wsbm.1382 (2017).
- 8. Martin, F. J. et al. Ensembl 2023. Nucleic Acids Research 51, D933-D941, https://doi.org/10.1093/nar/gkac958 (2023).
- O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44, D733–D745, https://doi.org/10.1093/nar/gkv1189 (2016).
- Bao, E. et al. BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences. Bioinformatics 29, 1250–1259, https://doi.org/10.1093/bioinformatics/btt127 (2013).
- Torre, D. et al. Isoform-resolved transcriptome of the human preimplantation embryo. Nature Communications 14, 6902, https://doi.org/10.1038/s41467-023-42558-y (2023).
- Sun, Y. H. et al. Single-molecule long-read sequencing reveals a conserved intact long RNA profile in sperm. Nature Communications 12, 1361, https://doi.org/10.1038/s41467-021-21524-6 (2021).
- Rhoads, A. et al. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13, 278–289, https://doi.org/10.1016/j. gpb.2015.08.002 (2015).
- Cho, H. et al. High-resolution transcriptome analysis with long-read RNA sequencing. PLoS One 9, e108095, https://doi. org/10.1371/journal.pone.0108095 (2014).
- 15. Veiga, D. F. T. *et al.* A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Science Advances* 8, eabg6711, https://doi.org/10.1126/sciadv.abg6711 (2024).
- Shimada, M. et al. Identification of region-specific gene isoforms in the human brain using long-read transcriptome sequencing. Science Advances 10, eadj5279, https://doi.org/10.1126/sciadv.adj5279 (2024).
- Yao, S. *et al.* A global survey of the transcriptome of allopolyploid Brassica napus based on single-molecule long-read isoform sequencing and Illumina-based RNA sequencing data. *The Plant Journal* 103, 843–857, https://doi.org/10.1111/tpj.14754 (2020).
- Huang, K. K. et al. Long-read transcriptome sequencing reveals abundant promoter diversity in distinct molecular subtypes of gastric cancer. Genome Biology 22, 44, https://doi.org/10.1186/s13059-021-02261-x (2021).
- Liu, Z. et al. Long- and short-read RNA sequencing from five reproductive organs of boar. Scientific Data 10, 678, https://doi. org/10.1038/s41597-023-02595-0 (2023).
- Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 21, 30, https://doi. org/10.1186/s13059-020-1935-5 (2020).
- 21. SMRT Link. https://www.pacb.com/smrt-link/
- 22. Li, H. et al. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
- 23. Tseng, E. et al. cDNA cupcake. 697, 3094-3100 https://github.com/Magdoll/cDNA_Cupcake/wiki (2020).
- Pardo-Palacios, F. J. et al. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. Nature Methods https://doi.org/10.1038/s41592-024-02229-2 (2024).
- Dixon, P. et al. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science 14, 927–930, https://doi. org/10.1111/j.1654-1103.2003.tb02228.x (2003).
- Tang, S. et al. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res 43, e78, https://doi.org/10.1093/nar/ gkv227 (2015).
- Mistry, J. et al. Pfam: The protein families database in 2021. Nucleic Acids Research 49, D412–D419, https://doi.org/10.1093/nar/gkaa913 (2021).
- Consortium, T. U. et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51, D523–D531, https://doi.org/10.1093/nar/gkac1052 (2022).
- Buchfink, B. et al. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods 18, 366–368, https://doi. org/10.1038/s41592-021-01101-x (2021).
- Quinlan, A. R. et al. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842, https://doi.org/10.1093/bioinformatics/btq033 (2010).
- Gu, Z. et al. circlize Implements and enhances circular visualization in R. Bioinformatics 30, 2811–2812, https://doi.org/10.1093/ bioinformatics/btu393 (2014).
- Chen, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890, https://doi.org/10.1093/ bioinformatics/bty560 (2018).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21, https://doi.org/10.1093/bioinformatics/bts635 (2013).
- 34. Danecek, P. et al. Twelve years of SAMtools and BCFtools. Gigascience 10 https://doi.org/10.1093/gigascience/giab008 (2021).
- Liao, Y. et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930, https://doi.org/10.1093/bioinformatics/btt656 (2014).
- Wang, L. et al. RSeQC: quality control of RNA-seq experiments. Bioinformatics 28, 2184-2185, https://doi.org/10.1093/ bioinformatics/bts356 (2012).
- Patro, R. et al. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14, 417–419, https://doi. org/10.1038/nmeth.4197 (2017).
- Langfelder, P. et al. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559, https://doi. org/10.1186/1471-2105-9-559 (2008).
- Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498–2504, https://doi.org/10.1101/gr.1239303 (2003).
- Trincado, J. L. *et al.* SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* 19, 40, https://doi.org/10.1186/s13059-018-1417-1 (2018).
- 41. Liu, Z. Raw sequencing data. Gene Expression Omnibus https://identifiers.org/geo:GSE269429 (2024).
- 42. NCBI, S. R. A. https://identifiers.org/ncbi/insdc.sra:SRP512775 (2024).
- Liu, Z. Comprehensive visceral transcriptome profiling of three pig breeds along altitudinal gradients in Yunnan, figshare., https:// doi.org/10.6084/m9.figshare.28517582 (2025).

Acknowledgements

This study was supported by the Science and Technology Major Project of the Yunnan Province Science and Technology Department (no. 202302AE090016 to G.Y.Z.); Science, Education and Culture Project of Yunnan Provincial Department of Finance (no. [2022]281-A3032022211 to G.Y.Z.); Key Research and Development Project of Yunnan Province (no. 2018BB003 to G.Y.Z.); Agricultural Joint Project of the Yunnan Province Science and Technology Department (no. 202401BD070001-004 to J.L.H.); Basic Research

Project of Shanxi Province, China (no. 202403021222371 to X.Z.); Basic Research Key Project of Yunnan Province, China (no. 202501AS070041 to J.L.H.); Yunnan Provincial academician and expert workstation (no.202405AF140009, no. 202305AF150168).

Author contributions

G.Y.Z., J.L.H. and Y.H.S. designed and conceived the study. Z.P.L., C.Y.L., W.L., F.F.H. and Y.R. performed data analysis. Z.P.L. and X.Z generated all the figures. D.C.X., G.W.F., L.X.W., S.R.G., J.D.Z., Z.W., X.H.W. and Z.P.X. collected the samples. J.L.H., Z.P.L. and Y.H.S. wrote and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.H., Y.H.S. or G.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025