Received: 28 January 2022 | Accepted: 1 November 2022

DOI: 10.1111/age.13274

RESEARCH ARTICLE

ANIMAL GENETICS WILEY

Transcriptome annotation of 17 porcine tissues using nanopore sequencing technology

Jinghui Li[1] | Dailu Guan[1] | Michelle M. Halstead[1] | Alma D. Islas-Trejo[1] | Daniel E. Goszczynski[1] | Catherine W. Ernst[2] | Hao Cheng[1] | Pablo Ross[1] | Huaijun Zhou[1]

[1]Department of Animal Science, University of California Davis, Davis, California, USA

[2]Department of Animal Science, Michigan State University, East Lansing, Michigan, USA

Correspondence
Pablo Ross and Huaijun Zhou, Department of Animal Science, University of California Davis, Davis, CA 95616, USA.
Email: pross@ucdavis.edu and hzhou@ucdavis.edu

Funding information
United States Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative, Grant/Award Number: 2015-43567015-22940, 2017-67015-26297, 2018-67015-27957 and 2021-67015-33412

Abstract
The annotation of animal genomes plays an important role in elucidating molecular mechanisms behind the genetic control of economically important traits. Here, we employed long-read sequencing technology, Oxford Nanopore Technology, to annotate the pig transcriptome across 17 tissues from two Yorkshire littermate pigs. More than 9.8 million reads were obtained from a single flow cell, and 69 781 unique transcripts at 50 108 loci were identified. Of these transcripts, 16 255 were found to be novel isoforms, and 22 344 were found at loci that were novel and unannotated in the Ensembl (release 102) and NCBI (release 106) annotations. Novel transcripts were mostly expressed in cerebellum, followed by lung, liver, spleen, and hypothalamus. By comparing the unannotated transcripts to existing databases, there were 21 285 (95.3%) transcripts matched to the NT database (v5) and 13 676 (61.2%) matched to the NR database (v5). Moreover, there were 4324 (19.4%) transcripts matched to the SwissProt database (v5), corresponding to 11 356 proteins. Tissue-specific gene expression analyses showed that 9749 transcripts were highly tissue-specific, and cerebellum contained the most tissue-specific transcripts. As the same samples were used for the annotation of *cis*-regulatory elements in the pig genome, the transcriptome annotation generated by this study provides an additional and complementary annotation resource for the Functional Annotation of Animal Genomes effort to comprehensively annotate the pig genome.

KEYWORDS
long-read sequencing, nanopore sequencing, pig tissues, transcriptome annotation

INTRODUCTION

Genome annotation aims to identify and understand the functional elements of a genome (Stein, 2001), which not only reveals important biological processes, such as gene expression variability and evolutionary conservation among species (Fair et al., 2020; Kaplow et al., 2021), but also informs genomic predictions and genome-wide association studies (Nani et al., 2019; Weissbrod et al., 2020). Several international consortia, such as the Encyclopedia of DNA Elements (ENCODE Project Consortium, 2004) and Functional Annotation of the Mammalian Genome (FANTOM; Kawaji et al., 2009), focus on identifying and annotating functional elements in the human genome and have expanded their work to several model species. Motivated by these projects, the Functional Annotation of Animal Genomes Consortium (FAANG) has been focusing on the functional annotation of genomes in domesticated animal species (Andersson et al., 2015; Clark et al., 2020; Giuffra et al., 2019).

Jinghui Li and Dailu Guan contributed equally to the paper.

The domestic pig (*Sus scrofa*) is an economically significant livestock species and a relevant biomedical model (Lassaletta et al., 2019; Lunney, 2007; Lunney et al., 2021). Consequently, there are both agricultural and medical interests in improving the functional annotation of the porcine genome. Most annotations of the pig transcriptome were mainly based on short-read sequencing technologies (Jin et al., 2021; Summers et al., 2020), which require the reconstruction of transcripts, and thereby are subject to the difficulties with assembly, phasing and identification of transcript isoforms (Hu et al., 2021). Recently, long-read sequencing data have greatly improved the pig genome annotation (Beiki et al., 2019; Li et al., 2018). There are currently two long-read sequencing platforms, developed by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), respectively. Li et al. (2018) found more than 26 000 novel genes and 92 000 novel alternative splicing events in the pig genome based on PacBio Iso-Seq data collected from Large White pigs. Beiki et al. (2019) applied the same sequencing technology for transcriptome annotation of nine tissues from a single White cross-bred pig and characterized the transcript variability among them. Despite the novel findings by these studies, there is still great potential to further improve pig genome annotations, especially of full-length and isoform transcripts. For example, after adding the novel genes (10 465) identified by Beiki et al. (2019) to the reference genome, the number of annotated genes (~36 000) is still much lower than that in human (~57 000; Zerbino et al., 2018). Even in human, thousands of novel transcripts were found using long-read sequencing technologies recently (Leung et al., 2021; Veiga et al., 2022), which indicates the complexity of transcriptome and a huge potential to further explore it.

ONT sequencing can generate more and longer reads per flow cell than PacBio Sequel II (Garalde et al., 2018), and has been widely used in a variety of species, including human (Jain et al., 2018), mouse (Sessegolo et al., 2019), cattle (Halstead et al., 2021), maize (Peng et al., 2021), and yeast (Istace et al., 2017). The sequence error of ONT is higher than short-read methods (Halstead et al., 2021), but ONT can sequence the entire transcript at once and capture the complex transcript structure (Glinos et al., 2022). By comparing to reference genomes (e.g., Ensembl and NCBI), the individual sequence error can be largely minimized through appropriate filtering. In this study, we aim to show the potential of ONT sequencing in porcine transcriptome studies by applying it to a comprehensive collection of pig tissues, and to expand the porcine transcriptome annotation by comparing our results to existing Ensembl and NCBI annotations.

MATERIALS AND METHODS

Sample collection

Tissues used in this study were from the US FAANG pilot study as described in Tixier-Boichard et al. (2021), which generated a collection of tissues from different livestock animals for the investigation of functional elements of animal genomes. A total of 17 tissues, including brain cortex, cerebellum, hypothalamus, thyroid, heart, thymus, kidney, liver, lung, lymph, muscle, spleen, adipose, duodenum, jejunum, ileum, and colon, were collected from two littermate castrated male Yorkshire pigs at the age of 6 months. The animals were raised at the Michigan State University Swine Teaching and Research Center in East Lansing, MI, and tissues were collected following a humane slaughter under the USDA inspection. Six tissues (brain cortex, thyroid, heart, thymus, duodenum, and colon) were only available from one animal, while the rest were from both animals, resulting in 28 samples in total. All the tissues were stored at −80°C until RNA extraction. The sample collection was conducted according to the Protocol for Animal Care and Use no. 18464 (approved by Institutional Animal Care and Use Committee at the University of California, Davis).

RNA sequencing

Total RNA for all samples was extracted using a protocol reported in (Halstead et al., 2021; Kern et al., 2018). Briefly, we homogenized frozen tissues with Trizol reagent (Invitrogen) and extracted RNA using the Direct-zol RNA Mini Prep Plus kit (Zymo Research). After checking the integrity of extracted RNA on the Experion electrophoresis system (Bio-Rad), we transferred 50 ng of RNA to a PCR tube (0.2 ml) and mixed with nuclease-free water to 9 μl. The RNA sample was incubated with 1 μl 10 μM VNP primer and 1 μl 10 mM dNTPs for 5 min at 65°C, and then cooled on a freezer block. Strand-switching buffer (4 μl of 5× RT buffer, 1 μl of RNaseOUT, 1 μl of nuclease-free water, and 2 μl of 10 μM strand-switching primer) was then added to the annealed RNA, and incubated at 42°C for 2 min. After that, 1 μl of Maxima H Minus Reverse Transcriptase was added, and the reaction was incubated at 42°C for 90 min, 85°C for 5 min, then held at 4°C to obtain the cDNA sample. A round of PCR was used to introduce barcodes to the cDNA using the Oxford Nanopore PCR barcoding expansion 1–96 kit (Cat. No. EXP-PBC096). Barcoding PCR reactions were set up for each cDNA sample with 1 μl of PCR barcode, 19 μl of first-strand cDNA and 20 μl of LongAmp Taq 2× master mix, and incubated for 3 min at 95°C (×1 cycle), 15 s at 95°C, 15 s at 62°C and 7 min at 65°C (×13 cycles), and 15 min at 65°C (×1 cycle), then held at 4°C. After purifying (1× Ampure XP Beads), eluting (20 µl of nuclease free water) and quantifying (Qubit), all the barcoded cDNA samples were pooled to a final volume of 47 µl. The DNA Technologies Core and Expression Analysis Laboratory at the University of California, Davis, performed adapter ligation on the cDNA pool with the SQK-DCS109 kit. Finally, 50 fmol of adapter ligated library was loaded onto a PromethION flow cell (vR9.4.1) to obtain the raw sequencing data.

Preliminary analysis of sequencing data

The raw sequencing data were first processed using PYCHOPPER (v2.5) to identify, orient, and trim the full-length reads, the qualities of which, including number of reads, read length, read quality, and read length N50, were then summarized using NANOPLOT (v1.32.1; De Coster et al., 2018). The full-length reads were mapped to the Sscrofa 11.1 genome assembly using MINIMAP2 (v2.17; Li, 2018), and then the chimeric and multi-mapped reads, as well as the reads with a minimum quality score smaller than 10 were discarded. The mapped reads after filtering were counted using HTSEQ (v0.13.15; Anders et al., 2015) based on the Ensembl annotation (release 102) to obtain the raw counts for gene features. To examine if the sampling procedures were correct and to investigate any potential outliers, a principal components analysis was carried out based on raw gene counts transformed with the variance stabilizing transformation algorithm implemented in DESEQ2 (v1.26.0; Love et al., 2014).

Predicting transcript isoforms

We first merged the reads from all samples, and then used the StringTie PIPELINE (v2.1.5; Kovaka et al., 2019, https://github.com/nanoporetech/pipeline-nanopore-ref-isoforms) to assemble all the mapped reads into potential transcripts. Only transcripts with exon depth ≥2 and coverage = 100%, as well as those on placed scaffolds were retained for subsequent analyses. After filtering, the remaining predicted transcripts were compared to two reference genome annotations, Ensembl (release 102) and NCBI (release 106), using GFFCOMPARE (v0.11; Pertea & Pertea, 2020). Based on the classification codes given by GFFCOMPARE (v0.11), predicted transcripts were categorized into four groups: exact match with the reference annotation (class code '='), novel isoforms (class codes 'o', 'n', 'm', 'k', 'j', and 'c'), transcripts found at novel loci that were not annotated in the reference annotation (class codes 'y', 'x', 'u', and 'i'), and potential artifacts (class codes 'e', 's', and 'p') (Pertea & Pertea, 2020).

Characterization of predicted transcripts

We used SUPPA (v2.3; Trincado et al., 2018) to identify alternative splicing events of the transcript isoforms, including alternative 3′ splice-site (A3), alternative 5′ splice-site (A5), alternative first exon (AF), alternative last exon (AL), mutually exclusive exons (MX), retained intron (RI), and skipping exon (SE). Besides, CPPred (Tong & Liu, 2019) was used to predict the coding potential of each transcript. To interpret the function of predicted transcripts at novel loci, their sequences were compared against three databases: NT (NCBI non-redundant nucleotide, v5), NR (NCBI non-redundant protein, v5), and SwissProt (protein sequence database, v5). NT is a collection of sequences from multiple species, and NR and SwissProt include protein coding sequences from multiple species. For transcripts matching with the SwissProt database, the functions of corresponding proteins were identified using the DAVID database (v6.8; Huang et al., 2009a, 2009b) by including the Kyoto Encyclopedia of Genes and Genomes term (KEGG), and gene ontology terms (biological process, cellular component, and molecular function).

Transcript expression and tissue specificity analysis

To determine the expression of predicted transcripts, we extracted reference sequences of the predicted transcripts using GFFREAD (v0.12; Pertea & Pertea, 2020). The full-length reads generated by PyChopper were mapped to the predicted transcriptome (in FASTA format) using MINIMAP2 (v2.17; Li, 2018). Then SAMTOOLS (v1.10; Danecek et al., 2021) was used to extract the alignments, and the expression of predicted transcripts in transcripts per million (TPM) was determined using NANOCOUNT (v0.2.4; Leger, 2021). Based on the average TPM of a transcript over all tissues, transcripts were classified as highly (average TPM ≥10), moderately (1 ≤ average TPM <10), and lowly expressed (average TPM <1) (Halstead et al., 2021).

Two measurements of tissue specificity were calculated based on TPM. The global tissue specificity index (gTSI; Halstead et al., 2021; Julien et al., 2012) was calculated for each transcript as follows:

$$\mathrm{gTSI} = \frac{\max\left(\mathrm{TPM}_i\right)}{\sum_i \mathrm{TPM}_i}, i = 1, \ldots, n$$

where TPM_i is the average TPM of a certain transcript for tissue i, and n is total number of tissues (17 in this study). Since there were two replicates for some tissues while only one for the other, TPM_i is either the mean TPM of two replicates or the TPM value itself of one replicate. The transcripts with max $(\mathrm{TPM}_i) < 0.1$ were considered as sequencing error and excluded from the calculation. The

TABLE 1 Sensitivity and precision estimates of predicted transcripts compared to Ensembl (release 102) and NCBI (release 106).

	Predicted vs. Ensembl		Predicted vs. NCBI	
Transcript annotations	**Sensitivity**	**Precision**	**Sensitivity**	**Precision**
Exon base	46.5	28.7	47.1	28.9
Exon interval	45.3	43.6	47.2	43.4
Intron interval	47.7	57.1	50.8	58.1
Intron chain	13.2	11.3	16.4	19.8
Transcript	13.8	8.0	16.4	13.4
Locus	34.6	9.7	50.6	14.8
Matching intron chains	5284		9270	
Matching transcripts	5610		9351	
Matching loci	4795		7395	
Missed exons	81 173/222 894 (36.4%)		77 490/210 821 (36.8%)	
Novel exons	97 533/214 967 (45.4%)		96 488/214 967 (44.9%)	
Missed introns	73 428/190 892 (38.5%)		67 104/182 458 (36.8%)	
Novel introns	59 574/159 394 (37.4%)		59 411/159 394 (37.3%)	

value of gTSI is between 0 and 1, and a gTSI close to 1 indicates a high tissue specificity due to the dominance of a certain tissue in the expression of the transcript. Based on gTSI, transcripts were categorized as tissue-specific (gTSI ≥0.8), broadly expressed (gTSI <0.5), and biased toward a group of tissues ($0.5 \leq \text{gTSI} < 0.8$). Another measurement of tissue specificity, individualized tissue specificity index (iTSI, Julien et al., 2012), was calculated as follows:

$$\text{iTSI} = \frac{\text{TPM}_i}{\sum_i \text{TPM}_i}.$$

The difference between gTSI and iTSI is that each transcript only corresponds to one gTSI value, but to a number (equal to the number of tissues) of iTSI values. In terms of interpretation, gTSI evaluates the tissue specificity of a transcript across all tissues, whereas iTSI evaluates the tissue specificity of a transcript for each tissue. For transcripts with iTSI ≥0.8, the corresponding biological functions, including gene ontology (molecular function, cellular component, and biological process) and KEGG pathway terms, were identified using g:Profiler (version e104_eg51_p15_3922dba; Raudvere et al., 2019) for each tissue.

RESULTS

Data summary

The ONT sequencing data generated in this study can be found in the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information with the identifier PRJNA671673 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA671673). The raw sequencing data contained 9.8 million reads in total, with an average GC content of 45%. After orienting and trimming, on average each sample had 303 295 full-length reads, a read length of 730.6 bp, a read length N50 of 873.2 bp, and a read quality of 11.2. Detailed summary information of each sample is shown in Table S1. The principal components analysis based on the raw gene counts showed that biological replicates for each tissue clustered together (Figure S1), which indicates proper sampling and sequencing procedures.

Based on the StringTie pipeline and filtering criteria, 69 781 unique transcripts at 50 108 loci were predicted. Of these loci, 10 543 contained more than one transcript, resulting in ~1.4 transcripts per locus. There were 46 737 multi-exon and 23 044 single-exon transcripts, resulting in 4.6 exons per transcript on average. The average lengths of transcripts, exons, and introns were 2033.0, 443.6 and 4779.0 bp, respectively (Figure S2). Our transcriptome annotation is publicly available in the GitHub repository https://github.com/liangend/Porcine_Nanopore.

Comparison to reference genome annotations

The comparison of the predicted transcripts to Ensembl (release 102) and NCBI (release 106) annotations is shown in Table 1. The results based on the two annotations were similar, except more matching intron chains, transcripts and exons were found using NCBI database. The characterization of each predicted transcript is shown in Figure 1a. More exact matches (class code '=') were found when comparing to NCBI (Ensembl: 5610 transcripts, 8.0%; NCBI: 9351 transcripts, 13.4%). There were 24 708 and 23 925 transcripts (22 344 in common) categorized as novel loci

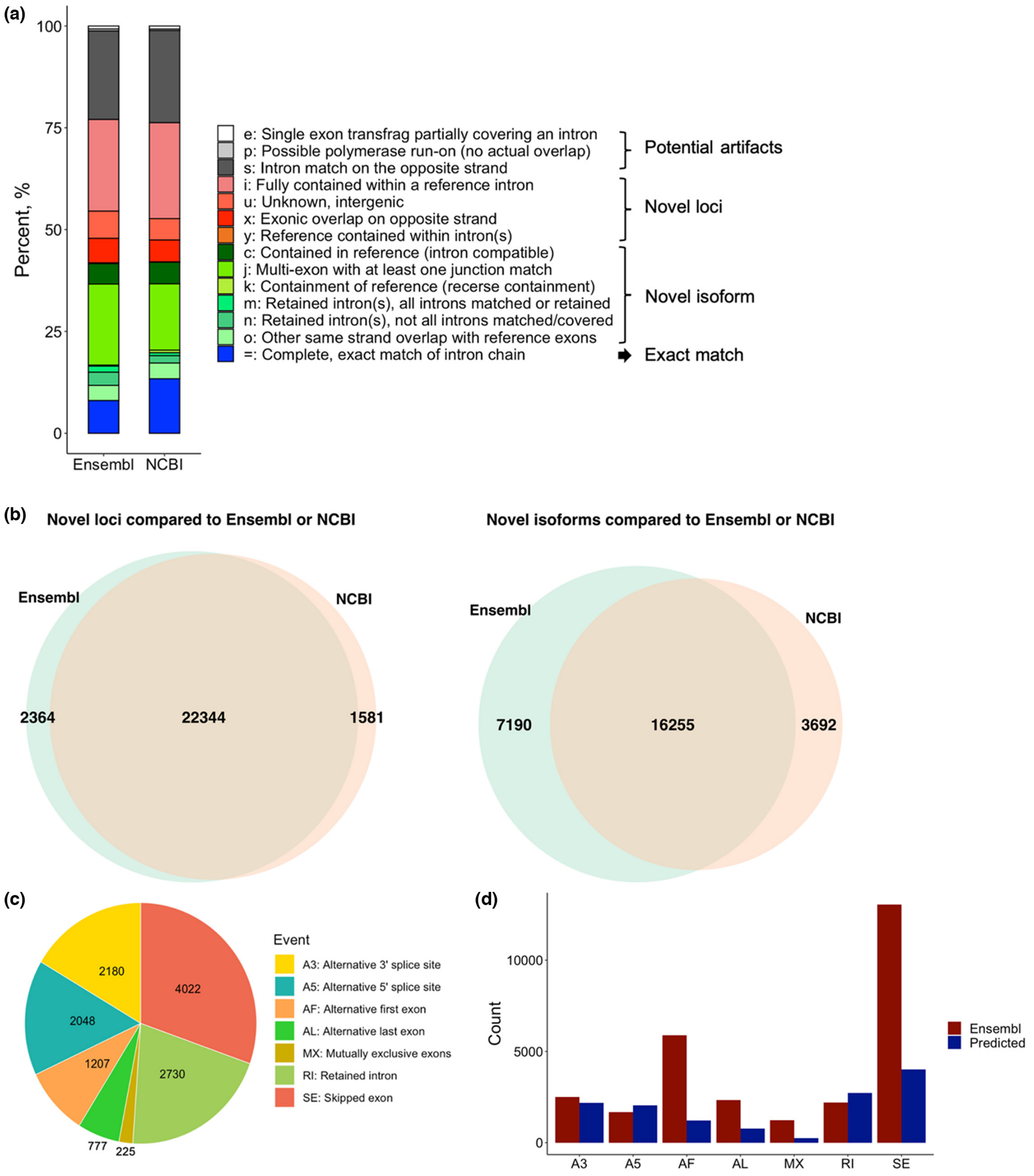

FIGURE 1 Characterization of predicted transcripts. (a) Classification of predicted transcripts compared to Ensembl (release 102) and NCBI (release 106). (b) Transcripts classified as novel loci or novel isoforms compared to Ensembl (release 102) and NCBI (release 106). (c) Frequency of alternative splicing events in predicted transcripts. (d) Alternative splicing events in predicted transcripts and Ensembl (release 102).

(see Methods), and 23 445 and 19 947 transcripts (16 255 in common) were categorized as novel isoforms compared to Ensembl (release 102) and NCBI (release 106), respectively (Figure 1b). Most of the transcripts found at novel loci (~23%) were classified as fully contained within a reference intron. Of the transcripts that were novel to both Ensembl and NCBI (22 344 found at novel loci and 16 255 novel isoforms), most were expressed in cerebellum (16 402), followed by lung (13 545), liver (11 954), spleen (11 941), and hypothalamus (11 918). In addition, cerebellum also expressed the most unique novel transcripts (Figure 2).

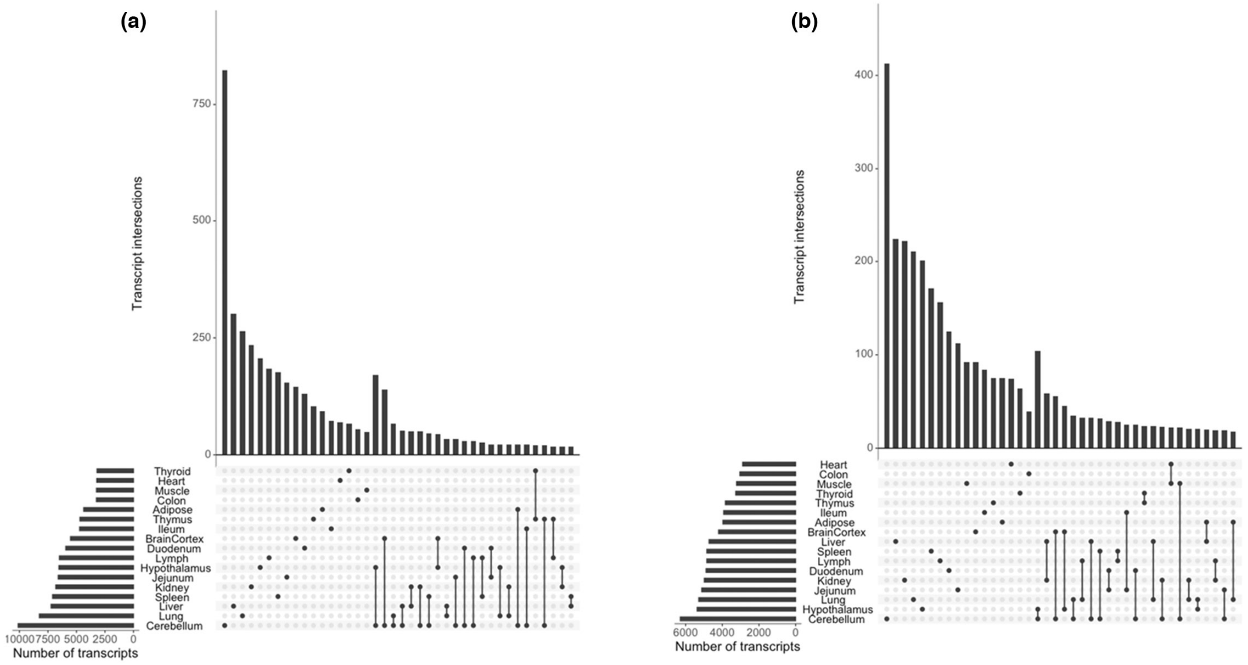

FIGURE 2 The sharing of novel transcripts across tissues. (a) Transcripts found at novel loci. (b) Transcripts categorized as novel isoforms. Each row of the matrix on bottom corresponds to a tissue, and the bar chart on left shows the number of novel transcripts found in the tissue. Each column corresponds to a possible intersection of tissues showed by the filled-in cells, and the bar chart on top shows the number of transcripts in each intersection.

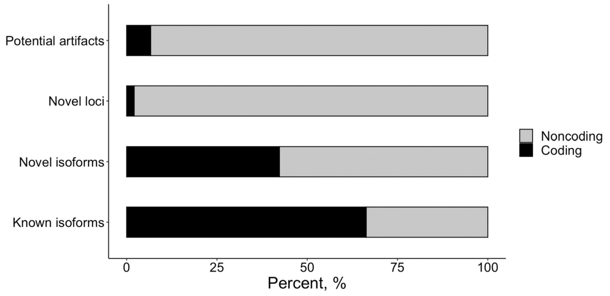

FIGURE 3 Functional analysis of predicted transcripts. Coding potential of predicted transcripts categorized as known isoforms ($n = 5610$), novel isoforms ($n = 24\,708$), found at novel loci ($n = 23\,445$), and potential artifacts ($n = 16\,018$) based on the comparison to Ensembl (release 102).

Characterization of predicted transcripts

A total of 13 219 alternative splicing events were identified for predicted transcript isoforms, most of the which were generated from skipped exon, followed by retained intron, alternative 3′-splice site, and alternative 5′-splice site (Figure 1c). For the Ensembl database, many more alternative splicing events (28 887), especially skipped exon and alternative first exon, were identified compared to the predicted transcript set (Figure 1d).

Coding potential of each predicted transcript, categorized based on the comparison to Ensembl, is shown in Figure 3. Based on the CPPred model (Tong & Liu, 2019), 66.3% and 42.3% of known and novel isoforms were predicted to be coding, respectively. However, very few transcripts found at novel loci were predicted to be coding (2.1%). We then compared the transcripts, which were found at loci novel to both Ensembl and NCBI annotations, to existing databases. There were 21 285 (95.3%) transcripts matched to NT and 13 676 (61.2%) matched to NR. Furthermore, there were 4324 (19.4%) transcripts matched to SwissProt, resulting in 11 356 corresponding proteins, which were predicted to have 722 significant biological functions (details in Table S2), including 370 related to biological process, 99 related to cellular component, 150 related to molecular function, and 103 related to KEGG pathway, based on the DAVID database (v6.8; Huang et al., 2009a, 2009b).

Tissue-specificity of predicted transcripts

The results of tissue specificity are shown in Figure 4. After removing the transcripts with a maximum TPM_i <0.1, there were 8025 (18.6%) transcripts highly expressed (average TPM ≥ 10), 22 149 (51.5%) moderately expressed (1 ≤ average TPM < 10), and 12 875 (29.9%) lowly expressed (average TPM < 1). For gTSI distribution, one peak around 0.25 and another peak at 1 were observed (Figure 4a). For highly expressed transcripts, there was a peak around 0.16 of gTSI, whereas a peak at 1 was observed for lowly expressed transcripts (Figure 4b). Based on gTSI, 9749 (22.6%) transcripts were highly tissue-specific (gTSI ≥ 0.8), and 25 034 (58.2%) transcripts

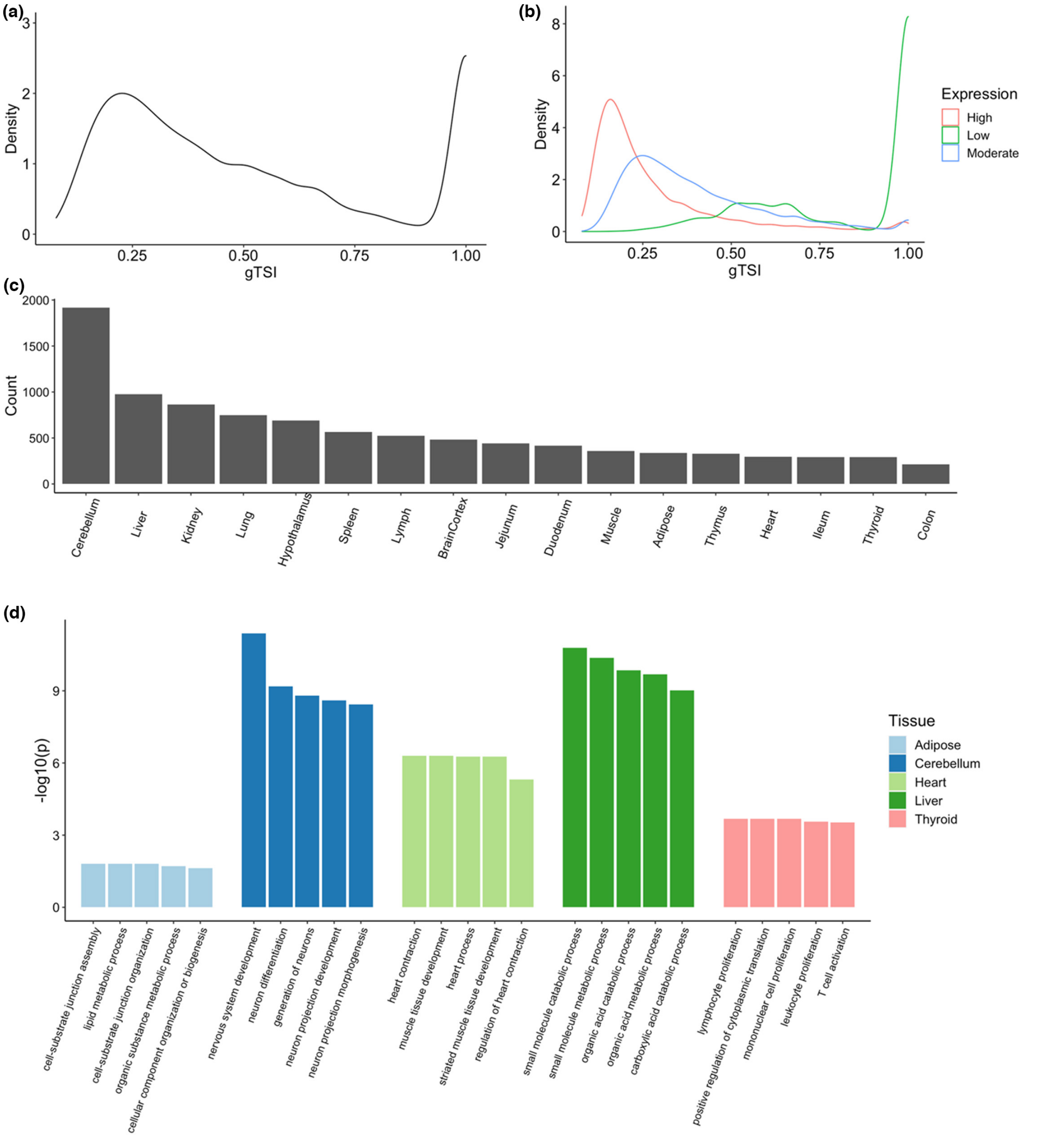

FIGURE 4 Tissue specific analysis of predicted transcripts. (a) Distribution of global tissue specificity index (gTSI) for all transcripts. (b) Distribution gTSI for transcripts categorized as highly (average TPM ≥ 10), lowly (average TPM < 1), and moderately (1 ≤ average TPM < 10) expressed. (c) Number of tissue-specific transcripts (local tissue specificity index ≥0.8) for each tissue. (d) Top five most significant biological process terms of tissue-specific transcripts in adipose, cerebellum, heart, liver, and thyroid based on the g:Profiler (version e104_eg51_p15_3922dba) reference.

were broadly expressed (gTSI < 0.5). The remaining 8266 (19.2%) transcripts were biased toward a group of tissues (0.5 ≤ gTSI < 0.8). The proportions of transcripts with coding potential based on the CPPred model were 21% for highly, 25% for moderately and 33% for lowly tissue-specific, which shows a tendency that more broadly expressed transcripts had more coding potential. The transcripts with iTSI ≥ 0.8 were extracted for each tissue (Figure 4c), and cerebellum expressed the most tissue-specific transcripts. These tissue-specific transcripts also exhibited tissue-specific functions. For example, transcripts with functions related to lipid metabolic processes were expressed in adipose, whereas transcripts with functions related to neuron development and differentiation were expressed in cerebellum based on the g:Profiler (version e104_eg51_p15_3922dba; Raudvere et al., 2019) reference (Figure 4d).

DISCUSSION

In the recent years, much progress has been made towards annotating the functional elements of the pig genome (Pan et al., 2021; Zhao et al., 2021). We used ONT sequencing to investigate the transcriptome annotation of 17 porcine tissues, which are shared within the FAANG community for other studies (Kern et al., 2018, 2021; Pan et al., 2021). Therefore, the transcripts identified in this study can be directly related to other functional elements, such as the porcine epigenome generated by the FAANG community (Pan et al., 2021).

Müller et al. (2021) recently characterized the pig cardiac transcriptome and found 4790 (~30% of the total transcript assembly) novel transcripts using ONT sequencing. By sequencing tissues from additional organ systems, we found in total 38 599 novel transcripts (16 255 novel isoforms and 22 344 found at novel loci) that were not annotated by either Ensembl (release 102) or NCBI (release 106). We further compared these novel transcripts to the porcine transcriptome annotated by Beiki et al. (2019), who generated their annotation using the PacBio Iso-Seq technology. Most of the novel transcripts (20 427/22 344 of transcripts at novel loci, and 11 072/16 255 novel isoforms) in this study were also novel in the analysis performed by Beiki et al. (2019) (Figure S3). However, the average transcript isoforms per locus in this study was only 1.4, which is fewer than those from previous studies (~1.92 in Beiki et al., 2019; ~1.93 in Li et al., 2018), and much fewer than the Ensembl (2.9) and NCBI (3.9) annotations, because although more transcript isoforms (69 781) were identified in this study (Ensembl: 40 568 and NCBI: 56 972), more loci (50 108) were also identified (Ensembl: 13 846 and NCBI: 14 605). In addition, fewer alternative splicing events were observed in the current study than the Ensembl annotation. These results indicate that the transcriptome annotation obtained in this study is not complete. Future studies could include more tissues from different development stages and different physiological status to further improve the pig transcriptome annotation, since many transcripts are spatially and temporally specific (Lukk et al., 2010).

Most of the transcripts found at novel loci (97.9%) were predicted to be noncoding based on the CPPred model (Tong & Liu, 2019), indicating that most of the coding loci were captured by the Ensembl annotation. Similar results were reported by Beiki et al. (2019). However, most of the transcripts found at novel loci (95.3%) can be found in the NT database, and ~20% of them were predicted to be protein coding and have important biological functions based on the SwissProt and DAVID databases. Such discrepancy between the CPPred model and existing databases could be because the CPPred model has not been fully validated on pig yet, despite a decent prediction accuracy in several other species (Tong & Liu, 2019), while NT and SwissProt match the sequence based on the data including multiple species.

As expected, the distribution of gTSI across different tissues in pig is similar to the previous study in cattle (Halstead et al., 2021). Transcripts that were highly expressed tended to have lower tissue specificity, indicating that they were expressed across different tissues in general. By contrast, transcripts that were lowly expressed tended to have a higher tissue specificity, because they were only expressed in certain tissues. This reflects that sampling a small set of tissues limits the ability to annotate the full spectrum of transcript isoforms. In other words, a diverse set of tissues is a critical factor affecting the discovery of transcript diversity. We observed that cerebellum expressed the most tissue-specific transcripts, which was consistent with the previous finding that many regionally enriched genes were expressed in pig cerebellum (Sjöstedt et al., 2020). In addition, cerebellum also contained the most novel transcripts. Given that pig is a preferred animal model for some human brain diseases, such as Alzheimer's disease (Richter et al., 2021; Sauleau et al., 2009), our findings could be a helpful resource for future studies using pig as a model for human diseases.

However, there were several limitations of this study. The ratio of average transcript isoforms per locus (1.4) in the current study was much lower than in other mammalian species, e.g. 3.8 in human (Ensembl v101) and 3.6 in cattle genome annotation (Halstead et al., 2021). In addition, more than 30% of exons and introns in the references were not captured in our samples (Table 1). One reason could be that both Ensembl and (release 102) NCBI annotations (release 106) were based on a Duroc sow, and our study used two Yorkshire pigs that are genetically closer to Landrace but different from Duroc (Tang et al., 2020). In addition, the average read quality (11.2) was not high, so the inadequacy of sequencing may hamper the detection of rare transcripts (Tarazona et al., 2011). Another limitation is that our ONT sequencing approach was based on cDNA, which is limited by the capacity of reverse transcriptase to amplify long transcripts. Although ONT generates longer reads, full-length transcripts for some of the longest genes could be missed. For example, the maximum read length of our sequencing result was 17 106 bp, whereas the longest transcripts in the Ensembl and NCBI annotations are 18 565 and 52 937 bp, respectively.

In conclusion, despite some limitations, our ONT sequencing results revealed a great number of novel transcripts and loci in the pig genome, which enhances the existing pig transcriptome annotation and complements efforts to annotate regulatory elements in pig. The pig genome at the transcript level has a high diversity that remains undiscovered, and additional studies are required to further characterize the transcription in additional cell types, breeds, developmental stages, and physiological states.

ACKNOWLEDGMENTS

We thank Dr. Ying Wang, Dr. Perot Saelao, and Ganrea Chanthavixay for their assistance in sample collection, organization, and storage. We also thank the staff at the University of California, Davis, DNA Technologies Core for their guidance and assistance with Nanopore library construction and sequencing. Funding for sample collection was provided by the United States Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative grant no. 2015-43567015-22940 awarded to H.Z. and P.R. Funding for library generation and sequencing was provided by the United States Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative grant no. 2017-67015-26297 awarded to P.R. and H.Z. Funding for bioinformatics analysis was provided by Agriculture and Food Research Initiative grant no. 2018-67015-27957 and 2021-67015-33412 awarded to H.C.

CONFLICT OF INTEREST

The authors declare they do not have any conflict of interest.

DATA AVAILABILITY STATEMENT

The data presented in this study are available in the SRA database of NCBI with identifier PRJNA671673 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA671673).

ORCID

Jinghui Li https://orcid.org/0000-0002-4854-2613
Catherine W. Ernst https://orcid.org/0000-0003-2833-0995
Huaijun Zhou https://orcid.org/0000-0001-6023-9521

REFERENCES

Anders, S., Pyl, P.T. & Huber, W. (2015) HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31, 166–169.

Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W. et al. (2015) Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biology*, 16, 57.

Beiki, H., Liu, H., Huang, J., Manchanda, N., Nonneman, D., Smith, T.P.L. et al. (2019) Improved annotation of the domestic pig genome through integration of iso-seq and RNA-seq data. *BMC Genomics*, 20, 344.

Clark, E.L., Archibald, A.L., Daetwyler, H.D., Groenen, M.A.M., Harrison, P.W., Houston, R.D. et al. (2020) From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biology*, 21, 285.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O. et al. (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, 10, giab008.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. & Van Broeckhoven, C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34, 2666–2669.

ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306, 636–640.

Fair, B.J., Blake, L.E., Sarkar, A., Pavlovic, B.J., Cuevas, C. & Gilad, Y. (2020) Gene expression variability in human and chimpanzee populations share common determinants. *eLife*, 9, e59929.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M. et al. (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15, 201–206.

Giuffra, E., Tuggle, C.K. & FAANG Consortium. (2019) Functional annotation of animal genomes (FAANG): current achievements and roadmap. *Annual Review of Animal Biosciences*, 7, 65–88.

Glinos, D.A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A. et al. (2022) Transcriptome variation in human tissues revealed by long-read sequencing. *Nature*, 608, 353–359.

Halstead, M.M., Islas-Trejo, A., Goszczynski, D.E., Medrano, J.F., Zhou, H. & Ross, P.J. (2021) Large-scale multiplexing permits full-length transcriptome annotation of 32 bovine tissues from a single nanopore flow cell. *Frontiers in Genetics*, 12, 621.

Hu, T., Chitnis, N., Monos, D. & Dinh, A. (2021) Next-generation sequencing technologies: an overview. *Human Immunology*, 82, 801–811.

Huang, D.W., Sherman, B.T. & Lempicki, R.A. (2009a) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4, 44–57.

Huang, D.W., Sherman, B.T. & Lempicki, R.A. (2009b) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37, 1–13.

Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O. et al. (2017) de novo assembly and population genomic survey of natural yeast isolates with the Oxford nanopore MinION sequencer. *Gigascience*, 6, giw018.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A. et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36, 338–345.

Jin, L., Tang, Q., Hu, S., Chen, Z., Zhou, X., Zeng, B. et al. (2021) A pig BodyMap transcriptome reveals diverse tissue physiologies and evolutionary dynamics of transcription. *Nature Communications*, 12, 3715.

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F. et al. (2012) Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biology*, 10, e1001328.

Kaplow, I.M., Schäffer, D.E., Wirthlin, M.E., Lawler, A.J., Brown, A.R., Kleyman, M. et al. (2021) Inferring mammalian tissue-specific regulatory conservation by predicting tissue-specific differences in open chromatin. *bioRxiv*, 2020-12.

Kawaji, H., Severin, J., Lizio, M., Waterhouse, A., Katayama, S., Irvine, K.M. et al. (2009) The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biology*, 10, R40.

Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H. et al. (2018) Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics*, 19, 684.

Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthavixay, G. et al. (2021) Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nature Communications*, 12, 1821.

Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L. & Pertea, M. (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20, 278.

Lassaletta, L., Estellés, F., Beusen, A.H.W., Bouwman, L., Calvet, S., van Grinsven, H.J.M. et al. (2019) Future global pig production systems according to the shared socioeconomic pathways. *Science of the Total Environment*, 665, 739–751.

Leger, A. (2021) *a-slide/NanoCount 0.2.4.post1*. Zenodo.

Leung, S.K., Jeffries, A.R., Castanho, I., Jordan, B.T., Moore, K., Davies, J.P. et al. (2021) Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports*, 37, 110022.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34, 3094–3100.

Li, Y., Fang, C., Fu, Y., Hu, A., Li, C., Zou, C. et al. (2018) A survey of transcriptome complexity in sus scrofa using single-molecule long-read sequencing. *DNA Research*, 25, 421–437.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W. et al. (2010) A global map of human gene expression. *Nature Biotechnology*, 28, 322–324.

Lunney, J.K. (2007) Advances in swine biomedical model genomics. *International Journal of Biological Sciences*, 3, 179–184.

Lunney, J.K., Van Goor, A., Walker, K.E., Hailstock, T., Franklin, J. & Dai, C. (2021) Importance of the pig as a human biomedical model. *Science Translational Medicine*, 13, eabd5758.

Müller, T., Boileau, E., Talyan, S., Kehr, D., Varadi, K., Busch, M. et al. (2021) Updated and enhanced pig cardiac transcriptome based on long-read RNA sequencing and proteomics. *Journal of Molecular and Cellular Cardiology*, 150, 23–31.

Nani, J.P., Rezende, F.M. & Peñagaricano, F. (2019) Predicting male fertility in dairy cattle using markers with large effect and functional annotation data. *BMC Genomics*, 20, 258.

Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L. et al. (2021) Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nature Communications*, 12, 5848.

Peng, C., Mei, Y., Ding, L., Wang, X., Chen, X., Wang, J. et al. (2021) Using combined methods of genetic mapping and nanopore-based sequencing technology to analyze the insertion positions of G10evo-EPSPS and Cry1Ab/Cry2Aj transgenes in maize. *Frontiers in Plant Science*, 12, 1519.

Pertea, G. & Pertea, M. (2020) GFF utilities: GffRead and GffCompare. *F1000Research*, 9. https://doi.org/10.12688/f1000research.23297.2

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H. et al. (2019) G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47, W191–W198.

Richter, C.-P., La Faire, P., Tan, X., Fiebig, P., Landsberger, D.M. & Micco, A.G. (2021) Listening to speech with a Guinea pig-to-human brain-to-brain interface. *Scientific Reports*, 11, 12231.

Sauleau, P., Lapouble, E., Val-Laillet, D. & Malbert, C.H. (2009) The pig model in brain imaging and neurosurgery. *Animal*, 3, 1138–1151.

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T. et al. (2019) Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Scientific Reports*, 9, 14908.

Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C. et al. (2020) An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science*, 367, eaay5947.

Stein, L. (2001) Genome annotation: from sequence to biology. *Nature Reviews Genetics*, 2, 493–503.

Summers, K.M., Bush, S.J., Wu, C., Su, A.I., Muriuki, C., Clark, E.L. et al. (2020) Functional annotation of the transcriptome of the pig, sus scrofa, based upon network analysis of an RNAseq transcriptional atlas. *Frontiers in Genetics*, 10, 1335.

Tang, Z., Fu, Y., Xu, J., Zhu, M., Li, X., Yu, M. et al. (2020) Discovery of selection-driven genetic differences of Duroc, landrace, and Yorkshire pig breeds by EigenGWAS and Fst analyses. *Animal Genetics*, 51, 531–540.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Research*, 21, 2213–2223.

Tixier-Boichard, M., Fabre, S., Dhorne-Pollet, S., Goubil, A., Acloque, H., Vincent-Naulleau, S. et al. (2021) Tissue resources for the functional annotation of animal genomes. *Frontiers in Genetics*, 12, 847.

Tong, X. & Liu, S. (2019) CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Research*, 47, e43.

Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J. et al. (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19, 40.

Veiga, D.F.T., Nesta, A., Zhao, Y., Mays, A.D., Huynh, R., Rossi, R. et al. (2022) A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Science Advances*, 8, eabg6711.

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S. et al. (2020) Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics*, 52, 1355–1363.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J. et al. (2018) Ensembl 2018. *Nucleic Acids Research*, 46, D754–D761.

Zhao, Y., Hou, Y., Xu, Y., Luan, Y., Zhou, H., Qi, X. et al. (2021) A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. *Nature Communications*, 12, 2217.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Li, J., Guan, D., Halstead, M.M., Islas-Trejo, A.D., Goszczynski, D.E., Ernst, C.W. et al. (2023) Transcriptome annotation of 17 porcine tissues using nanopore sequencing technology. *Animal Genetics*, 54, 35–44. Available from: https://doi.org/10.1111/age.13274