DNA Research, 2021, 28(3), 1–11
doi: 10.1093/dnares/dsab005
Advance Access Publication Date: 14 May 2021
Resource Article: Genomes Explored

OXFORD

Resource Article: Genomes Explored

Characterization and complexity of transcriptome in *Gymnocypris przewalskii* using single-molecule long-read sequencing and RNA-seq

Xindan Li[1,2†], Jinming Wu[1†], Xinping Xiao[1], Yifeng Rong[1,2], Haile Yang[1], Junyi Li[1], Qiong Zhou[1], Weiguo Zhou[3], Jianquan Shi[3], Hongfang Qi[3], and Hao Du[1,2,3*]

[1]Key Laboratory of freshwater biodiversity conservation, Ministry of Agriculture, Chinese Academy of Fishery Sciences, Wuhan, Hubei 430223, China, [2]College of Marine Science, Shanghai Ocean University, Shanghai 201306, China, and [3]The Rescue and Rehabilitation Center of Naked Carps in Lake Qinghai, Xining, Qinghai 810016, China

*To whom correspondence should be addressed: Hao Du, Key Laboratory of freshwater biodiversity conservation, ministry of agriculture, Chinese Academy of Fishery Sciences. Tel: +86-27-81780118, Fax: +86-27-81780215, Email: duhao@yfi.ac.cn

[†]Co-first authors.

Received 7 December 2020; Editorial decision 11 May 2021; Accepted 11 May 2021

Abstract

The Tibetan Schizothoracinae fish *Gymnocypris przewalskii* has the ability to adapt to the extreme plateau environment, making it an ideal biological material for evolutionary biology research. However, the lack of well-annotated reference genomes has limited the study of the molecular genetics of *G. przewalskii*. To characterize its transcriptome features, we first used long-read sequencing technology in combination with RNA-seq for transcriptomic analysis. A total of 159,053 full-length (FL) transcripts were captured by Iso-Seq, having a mean length of 3,445 bp with N50 value of 4,348. Of all FL transcripts, 145,169 were well-annotated in the public database and 134,537 contained complete open reading frames. There were 4,149 pairs of alternative splicing events, of which three randomly selected were defined by RT–PCR and sequencing, and 13,293 long non-coding RNAs detected, based on all-vs.-all BLAST. A total of 118,185 perfect simple sequence repeats were identified from FL transcripts. The FL transcriptome might provide basis for further research of *G. przewalskii*.

Key words: *Gymnocypris przewalskii*, single-molecule sequencing, alternative splicing, gene expression, full-length transcriptome

1. Introduction

Qinghai Lake, located in the northeast of Qinghai–Tibet Plateau (QTP) with an average elevation of ~4,000 m above sea level,[1,2] is the largest inland saline and alkaline lake in China. As a typical salty lake with unusual high sodium, potassium and magnesium concentrations, the pH value of Qinghai Lake is 9.4, and the salinity is 13%.[3] Tibetan naked carp (*Gymnocypris przewalskii*), belonging to the family *Cyprinidae*, subfamily *Schizothoracinae*, is an endangered

and endemic fish species living in Qinghai Lake with important fishery and ecological significance in the fish–bird–grassland system.[4] *Gymnocypris przewalskii* gradually evolved from the freshwater fish to tolerate high salinity and alkalinity because of the long-term geographical isolation in the Qinghai Lake.[5] As an anadromous spawning species, Tibetan naked carp inhabits saline and alkaline lake (Qinghai Lake) for most of its life history, then returns to freshwater rivers (such as the Buha River, Shaliu River, Quanji River, Heima River and Haergai River) during the reproductive season (from April to August), which indicates its adaptation to both saline and freshwater environments.[6,7] In addition, both the growth rate and gonadal development of this species are slow, gaining approximately 500 g every 10 years,[8] and males reach sexual maturity at the age of 3–4 followed by females after 4 years old.[7] Due to its unique environmental adaptability and migration reproductive characteristics, *G. przewalskii* could be used as an important model species for studies on adaptive evolution and population genetics of wild animals living in the plateau environment. Previous studies on Tibetan naked carp have mainly focussed on life history,[9] adaptive evolution,[10] physiology,[11] and immunology,[12] reproduction[8] and growth.[13,14] However, the genomics and genetic resources of this species are still insufficient, limiting further research on the molecular mechanism of its physiological response, behavioural patterns and adaptive evolution.

The gene expression, alternative splicing (AS) events and evolution selection are believed to be closely correlated with physiological responses, behavioural patterns and local adaptation.[15–17] During the last decade, transcriptomics technologies have been widely applied to obtain an insight of the organism's transcriptome, providing valuable information for studying gene expression and evolution selection, detecting non-coding RNAs (ncRNAs) and AS events, and developing molecular markers.[18] Although most of the transcriptomes of both model and non-model species are generated using RNA sequencing with short-read sequencing (RNA-seq), the sensitivity and precision of RNA-seq have been questioned, especially for transcriptome reconstruction and alternative spliced isoform detection.[19] A newly developed long-read sequencing technology, named isoform sequencing (Iso-seq), could contribute to overcoming these limitations by obtaining sequence information of full-length (FL) cDNA without further assembly.[19] Despite the relatively high cost, Iso-Seq technology is becoming increasingly popular for in-deep research with transcriptome data.[20]

To reveal the complexity of the FL reference transcriptome of *G. przewalskii* more comprehensively, PacBio Iso-Seq and Illumina RNA-seq technologies were performed in the present study. The main research purposes are as follows: (i) FL reference transcriptome sequencing and functional annotation; (ii) detection of alternatively spliced transcript isoforms; (iii) comparison of gene expression patterns among different tissues; and (iv) development of gene-associated microsatellite tag data. This research contributes to unravelling the characteristics of *G. przewalskii* transcriptome and provides a valuable genetic resource for further studies on behaviour patterns, phylogeny, adaptive evolution, population genetics, conservation and rejuvenation in this species and other Gymnocypris fishes.

2. Materials and methods

2.1 Sample collection and RNA preparation

Six wild females (body weight = 188.4 ± 79.3 g) and three males (body weight = 92.9 ± 37.9 g) of *G. przewalskii* were collected in June 2018 (spawning season) from Qinghai Lake, Quanji River and Erhai (Fig. 1). After euthanization with 200 mg/L MS222 (Sigma, USA), five tissues (liver, muscle, brain, gill and kidney) of each fish were sampled and transferred immediately to liquid nitrogen for RNA extraction.

In this study, efforts were made to minimize suffering as much as possible, and all animal experiments were conducted in accordance with the procedures and guidelines of the Animal Ethics Committee of Yangtze River Fisheries Research Institute of Chinese Academy of Fishery Sciences. The present study was approved by the Animal Care and Use Committee of Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences.

The total RNA of each individual tissue was isolated using TRIzol Reagent (Invitrogen, USA) in accordance with the manufacturer's instructions. Gel electrophoresis, Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and Agilent Bioanalyzer 2100 system (Agilent Technologies, USA) were used to assess RNA quality and quantity. All the sequencing operations were conducted at Frasergen Information Co., Ltd. (Wuhan, China).

2.2 PacBio Iso-Seq library construction and sequencing

The total RNA of each sample was pooled together in equimolar ratios, and approximately 2 μg of the total RNA was subsequently used for the FL cDNA synthesis by a SMARTer^{TM} PCR cDNA Synthesis Kit (Takara Clontech Biotech, Dalian, China) according to the standard protocol. Size fractionation and selection (1–2 kb, 2–3 kb, 3–6 kb and 5–10 kb) of the FL cDNA was carried out using the BluePippin^{TM} Size Selection System (Sage Science, Beverly, MA), and a re-amplification was than conducted for the selected FL cDNA fragments with PCR. The three SMRTbell Template libraries (1–2 kb, 2–3 kb and >3 kb) were constructed with SMRTBell Template Prep Kit 2.0 (Pacific Biosciences, USA) following the manufacturer's recommendations. The library quantification was performed using Qubit System,[21] and the size range of the library was detected with Agilent Bioanalyzer 2100 system (Agilent Technologies, USA). The single-molecule real-time (SMRT) cells for the three libraries were then sequenced on the Pacific Bioscience RS II platform using P6-C4 reagent with 4 h sequencing movies (Pacific Biosciences, Menlo Park, CA, USA).

2.3 PacBio sequencing data processing and error correction

PacBio raw data were preprocessed and filtered using SMRT Link v7.0 pipelines (https://www.pacb.com/ support/software- downloads/ (21 July 2021, date last accessed)). Briefly, the raw polymerase reads were filtered and trimmed to generate the Subreads using the RS_Subreads protocol (parameters: 50 bp ≤ subread length ≤ 15,000 bp, minimum number of passes = 3, minimum predicted accuracy = 0.8, minimal read score = 0.65, minimum accuracy of polished isoforms = 0.99). Subsequently, circular consensus sequence (CCS) reads were generated from the subreads using the P_CCS model with default parameters. The full-length non-chimeric (FLNC) reads were identified from CCS reads by searching for the 5′/3′ cDNA primers and the poly (A) tail in the read of inserts (ROI). An isoform-level clustering algorithm Iterative Clustering for Error Correction (ICE) and the Quiver software module were performed to generate the polished FL consensus isoforms from the FLNC reads. Furthermore, the high-quality polished isoform sequences (post-correction accuracy ≥ 99%) were used for the subsequent analysis, while the low-quality polished isoform sequences (post-correction

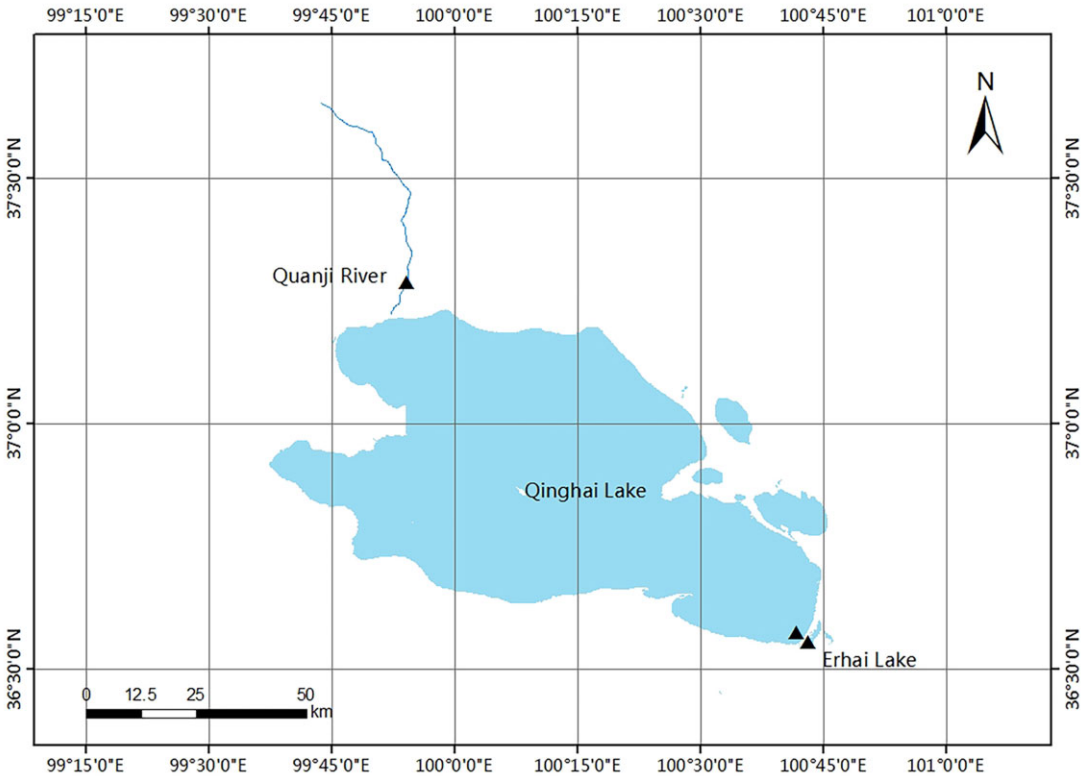

Figure 1. Sampling points' diagram of *G. przewalskii* in Qinghai province. The sampling map was created by the ArcGIS v10.1 (ESRI, CA, USA) and the sample site represented by the black triangle.

accuracy < 99%) were further corrected by Illumina reads as the reference using the Proovread v2.13.13 with the default parameters.[22] After the combination of the high-quality and low-quality transcripts, CD-HIT 4.6.1 software was applied to remove the redundant isoforms (parameters: -c 0.99 -T 24 -G 0 -aL 0.90 -AL 100 -aS 0.99 -AS 30 -M 0).[23] Finally, the first high-quality FL transcript data set for *G. przewalskii* was obtained in the present study.

2.4 FL transcript annotation

Functional annotation of the FL transcripts of *G. przewalskii* was carried out by Diamond 0.8.33 software[24] with a series of public databases including NR (NCBI non-redundant protein sequences), GO (Gene Ontology), KOG (euKaryotic Ortholog Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes), KO (KEGG Orthology Database) and Swiss-Prot (Swiss-Prot Protein Sequence Database).

2.5 Identification of ORF and LncRNA

Open reading frame (ORF) prediction of the transcripts was performed using TransDecoder software to identify the potential coding sequences from transcripts. In addition, the protein sequences of the predicted ORFs were further mapped to Swiss-Prot and Pfam databases using BlastP and Hmmscan, which could contribute to improving the sensitivity of ORF prediction. Those transcripts containing complete ORFs, 5′-UTR (untranslated regions) and 3′-UTR were considered as FL transcripts.

Based on the transcript annotation results, Coding-Potential Assessment Tool (CPAT) software was used to evaluate the coding potential of transcripts without annotation information in the protein databases.[25] The transcripts with coding potential greater than the cut-off or length less than 200-bp were filtered out, and the rest could be regarded as the long non-coding RNA (LncRNA) in this study. The cut-off of coding potential was determined by nonparametric two-graph ROC curves from the known data set of lncRNA and coding RNA in *Danio rerio*.

2.6 Detection and validation of AS events

AS event is one of the primary sources of transcript and proteome diversity. The annotated reference genome of Tibetan naked carp has not yet been published; therefore, the de novo AS detection pipeline described by Liu et al.[26] was used to characterize and identify the alternatively spliced transcripts in *G. przewalskii* without the reference genome sequences. The validation of AS events was performed with RT–PCR using 1 μg of total mixed RNA extracted from five different tissues in 20 μl reactions, which was the same as that used for the PacBio Iso-Seq library construction. In an exon skipping event, there should be two high-scoring segment pairs (HSPs) in the alignment. In the shorter transcript, the base pair coordinates representing the end of HSP1 and the start of HSP2 should be sequentially continuous. And, in the another, the base pair coordinates between the end of HSP1 and the start of HSP2 should be the skipped exon (recorded here as 'AS gap'). The specific primers were designed in the flanking region of 'AS gap' with Primer Premier 6. PCR products were detected using 2% agarose gel electrophoresis, then isolated and

purified from gels using a gel extraction kit for subsequent cloning and sequencing. PCR products were subsequently isolated from gels and purified using a gel extraction kit, cloned and sequenced. The obtained sequences were aligned with related isoforms to verify the predicted AS isoforms.

2.7 Illumina short-read sequencing and *de novo* assembly of the transcriptome

For Illumina RNA-Seq, an equal amount of total RNA from nine fish was pooled for each tissue, and the Illumina cDNA libraries were then prepared for each tissue sample using a NEBNext® Ultra™ RNA Library Prep Kit (NEB, Beverly, MA, USA) in accordance with the manufacturer's instructions. The qualified libraries were subsequently sequenced with the paired-end sequencing method (each end 150 bp) on the Illumina HiSeq X Ten System (Illumina, Inc., San Diego, CA, USA) following the manufacturer's recommendations.

The quality filtering of raw paired-end reads was performed using NGS QC Toolkit.[27] Briefly, the first five bases from the 5' end of the read were trimmed and the reads consisting of the low-quality bases (quality score ≤ 30) >20% or ambiguous bases >1% were removed. Ends of read were trimmed and the reads consisting of low-quality bases (quality score ≤ 30) >20% or ambiguous bases >1% were removed. The clean paired-end short reads from each library were mapped to the PacBio Iso-Seq sequences of *G. przewalskii* using bowtie2,[28] and then used for *de novo* assembly of transcripts by using Cufflinks v2.2.1[29] and Trinity v2.8.4 software[30] following the default parameters.

2.8 Quantification of gene expression levels

The expression level of each transcript for each tissue was quantified and estimated with Fragments Per Kilobase per Million (FPKM) bases values, which were calculated and normalized using RNA-Seq by Expectation-Maximization (RSEM) software.[31] The DESeq2 R package was applied to determine the differential expression,[32] and the false discovery rate (FDR) was controlled by adjusted *P*-values according to the method described by Benjamini and Hochberg.[33] Transcripts with *P*-values <0.001 ($P < 0.001$) and fold change greater than 2 (log2 FC > 2) were defined as differentially expressed transcripts (DETs). MA plot, volcano plot and heat maps were generated using R package to visualize the differential expression of transcripts among each tissue. Additionally, gene ontology (GO) and KEGG pathway enrichment analyses for all DETs were both performed with Phyper of R package. The transcription factors were detected using Diamond 0.8.33 software[24] based on the Animal TFDB Database.[34] To evaluate the accuracy and reliability of the quantification result, the q-PCR was performed for 10 randomly selected transcripts.

3. Results

3.1 Transcriptome from PacBio isoform sequencing

The mean lengths of subreads from different libraries were 1,722, 3,574 and 2,613 bp, respectively. After the polymerase reads were obtained from three libraries (Supplementary Fig. S1), we integrated the data of the three different libraries. Three Iso-Seq libraries were constructed for the total RNA generating 12,898,077 subreads (Table 1). A total of 568,004 CCS reads were generated from PacBio Iso-Seq with a mean length of 3,064 bp (Supplementary Table S1).

Table 1. Description of the transcriptome of *G. przewalskii* by PacBio Iso-Seq and Illumina RNA-seq

Parameter	PacBio Iso-Seq	Illumina RNA-seq
Number of subreads or raw reads	12,898,077	285,490,628
Reads of CCS or clean reads	568,004	281,915,120
Number of FLNC	508,704	–
Full-length transcriptome	–	–
Number of transcripts	159,053	164,142
Mean length (bp)	3,445	1,426
Smallest length (bp)	175	188
largest length (bp)	14,106	67,560
N50 length (bp)	4,348	2,940

CCS reads contained 508,704 FLNC reads (Fig. 2) with a mean length of 2,864 bp. After clustering (ICE algorithm) and polishing (Arrow algorithm), a total of 214,911 FL polished consensus isoforms were generated from FLNC sequence. The FL polished consensus isoforms had the mean length of 3,208 bp, ranging from 175 to 14,075 bp.

3.2 Error correction with Illumina RNA-seq

The error of single molecule sequencing technology is mainly due to extra-base insertion and deletion of single bases, which can be effectively corrected by multiple sequencing.[35] In addition, polished isoforms can be further corrected by proovread error correction software with the Illumina reads. In this study, a total of 214,911 corrected isoforms with the mean length of 3,029 bp (ranging from 175 to 14,106 bp) were generated (Table 2).

After clustering and removing redundancy from the corrected isoforms, a total of 159,053 FL transcripts were generated (Table 1). The FL transcripts were applied for subsequent analysis. The length of FL transcripts ranged from 175 to 14,106 bp. The mean length and N50 value of FL transcripts were 3,445 bp and 4,348, which were all longer than those of the transcripts captured by Illumina RNA-seq (Table 1).

3.3 Functional annotation

To predict and analyse the function of the 159,053 FL transcripts, the FL transcripts were searched against public databases using Diamond.[24] A total of 145,095 (91.22%), 86,225 (54.21%), 100,175 (62.98%), 82,451 (51.84%) and 134,509(84.57%) FL transcripts were assigned to NR, GO, KO, KOG and Swiss-Prot databases, respectively (Fig. 3), while 13,884 (8.73%) transcripts were not assigned to public databases. Through Nr annotation, 91.22% homologous hits were assigned to five fish species, including *Sinocyclocheilus rhinocerous*, *Sinocyclocheilus angustiporus*, *Sinocyclocheilus grahami*, *Cyprinus carpio* and *D. rerio* (Fig. 4) (Supplementary Table S2).

Functional classification of the transcripts was carried out through GO database. The GO-annotated transcripts were mainly assigned to 54 Level 2 GO terms (Supplementary Fig. S2). In the biological process category, the most abundant term was 'cellular process' (53,967), followed by 'single-organism process' (46,614) and 'metabolic process' (37,155). For the cellular component category, 'cell part' (48,469) and 'cell' (48,463) were the most abundant terms. Within the molecular function category, 'binding' (47,034) and 'catalytic activity' (31,054)

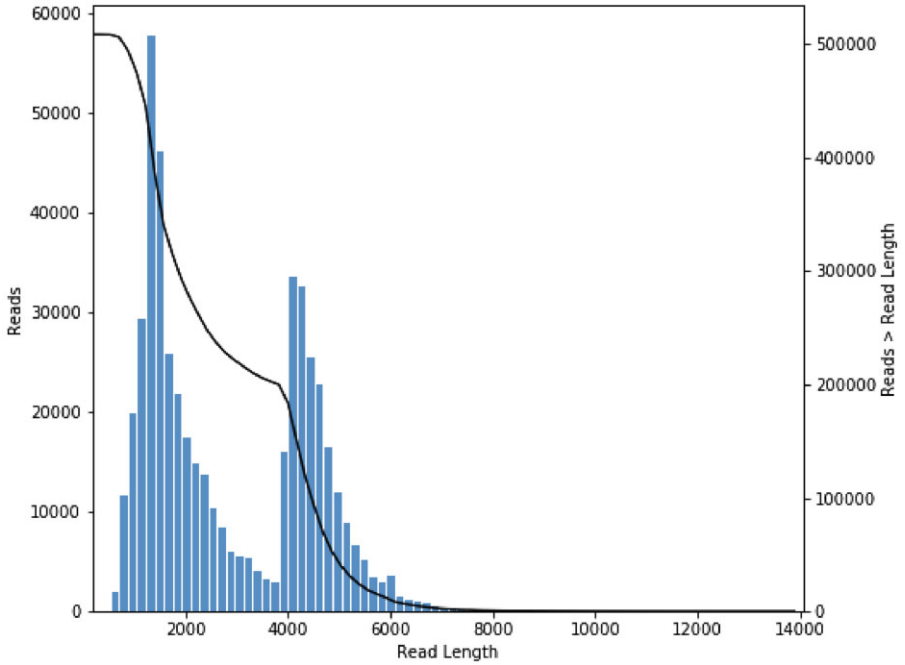

Figure 2. The length distribution of FLNC obtained by Iso-Seq. The *x*-axis represents the FLNC length, and the *y*-axis represents the number of the FLNC.

Table 2. Comparison of isoforms before and after RNA-seq data correction

Parameter	Before correction by RNA-seq data	After correction by RNA-seq data
Isoforms number	214,911	214,911
Average length	3,208	3,209
Maxnum length	14,075	14,106
Minnum length	175	175
N50	4,281	4,281

had the largest number of transcripts. KEGG-annotated transcripts were classified into 34 Level 2 KEGG groups. Among them, the greatest number of transcripts was in the signal transduction pathway (19,896), followed by immune system (11,480), transport and catabolism (10,247), and endocrine system (8,259) (Supplementary Fig. S3). For KOG annotation, of the 26 categories, the most annotated were signal transduction mechanisms (18,563), followed by general function prediction only (14,113) and posttranslational modification protein turnover (8,605), and intracellular trafficking, secretion and vesicular transport (6,832) (Supplementary Fig. S4).

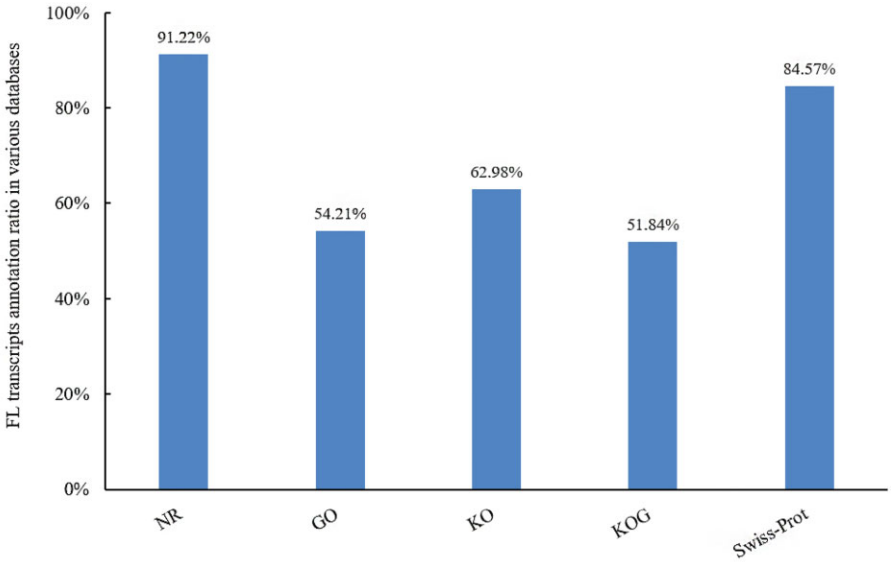

Figure 3. FL transcripts annotation percentage in NR, GO, KOG and Swiss-Prot databases.

3.4 ORF and LncRNA prediction

Transcripts with coding potentials greater than truncation or with length <200 bp were filtered out and verified with the non-parametric double-graph ROC curve obtained from known *D. rerio* lncRNA and coding RNA data sets (Fig. 5). A total of 13,293 potential lncRNAs were detected based on the transcript annotation results (Supplementary Table S3). The lncRNAs had lengths ranging from 316 to 10,490 bp, and the mean length of lncRNAs was 2,486.6 bp. After removing lncRNAs, 134,537 ORFs (ORFs ≥ 100 aa) were detected in the PacBio transcripts by TransDecoder software accounting for 92.30% of the total isoforms (145,760).

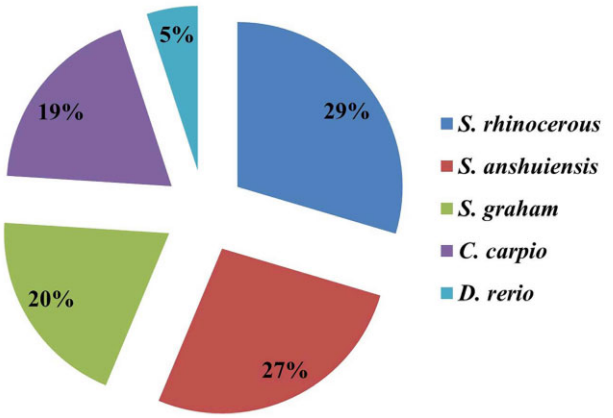

Figure 4.. Homologous species annotation. The species identified by homology search against the NCBI NR databases. Note that only the top five for transcripts are covered in the analysis.

3.5 Detection of microsatellite markers

There were 118,185 perfect and 19,221 complicated SSRs identified from 159,053 FL transcripts in all. Those composed of single-nucleotide repeats, called perfect SSRs, such as $(CA)_{20}$, and SSRs that are not pure repeats, called complicated SSRs, such as $(CA)_4$-$(T)_7$-$(CTT)_3$.[36] The perfect SSRs consisted of 60,051 mononucleotide SSRs, 42,027 dinucleotide SSRs, 14,213 trinucleotide SSRs, 1,605 tetranucleotide SSRs, 221 pentanucleotide SSRs and 68 hexanucleotide SSRs (Fig. 6A). There is a positive relationship between the degree of polymorphism and repeat unit length in SSRs.[37] The most abundant motif in dinucleotide SSRs was AC/GT (22,818, 54.29%), followed by AT (9,926, 23.62%) and AG/TC (9,204, 21.90%); the most abundant motifs in trinucleotide, tetranucleotide and pentanucleotide SSRs were AAT/ATT (4018, 28.27%), ATCT/AGAT (327, 20.37%) and TTCTC/GAGAA (48, 21.72%), respectively. Among the hexanucleotide SSRs, the motifs of AGCCAC, CCCAAC and CCCAAC were the same (6, 8.82%) (Fig. 6B).

3.6 Detection and validation of AS events

A total of 4,149 pairs of potential AS events were detected from the FL transcripts by all-vs.-all BLAST with high identity settings (e-value of 1e-20, pairwise identity of 95%) (Supplementary Table S4). The mean length of the 'AS gap' in AS events was 655.3 bp.

In order to verify the accuracy of the identified splicing isomers experimentally, three AS events were randomly selected for RT–PCR and sequencing analysis. Primers were designed and synthesized (Table 3), and RT–PCR was performed using mixed RNA from five different tissues. The results showed that the fragment size and the bands on the agarose gel were consistent with the AS isomers (Supplementary Fig. S5). In addition, DNA fragments corresponding to the predicted size were cloned, and their subtypes were verified by sequencing.

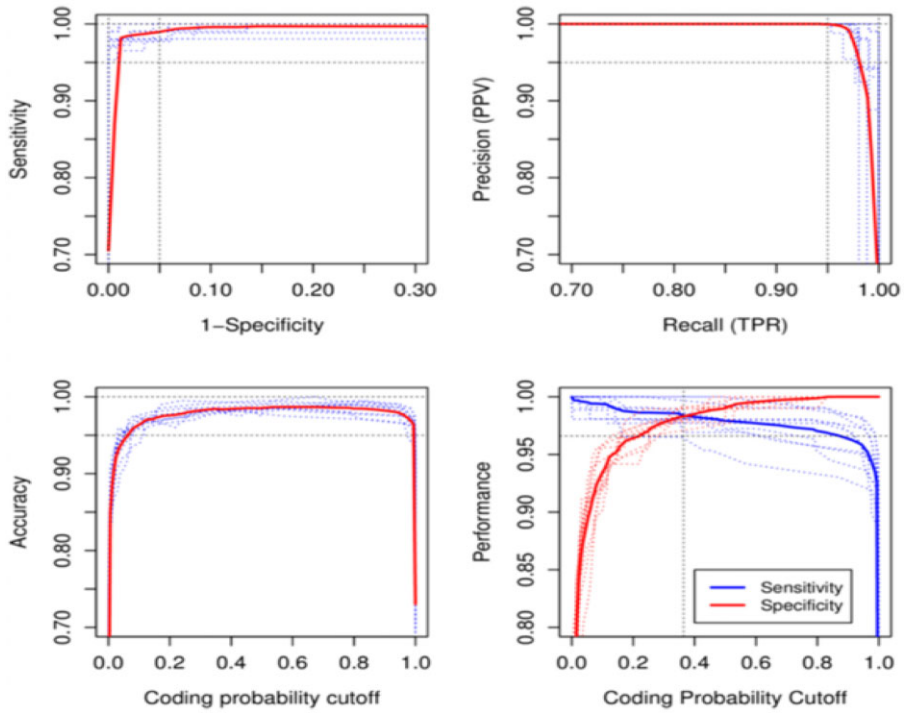

Figure 5.. Determination of cut-off of encoding potential. Performance evaluation using 10-fold.

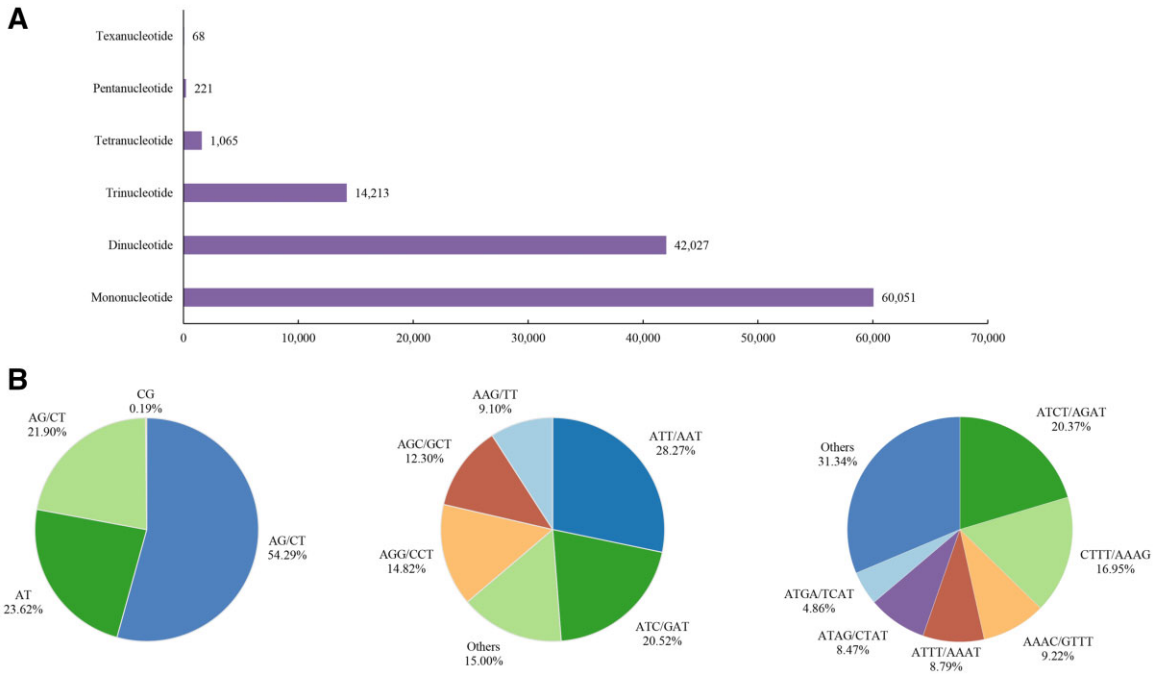

Figure 6.. Overview of SSRs isolated from FL transcripts of *G. przewalskii*. (A) The number of SSRs with different repeats and motifs. (B) The dominant motifs of dinucleotide, trinucleotide and tetranucleotide SSRs.

Table 3.. Primer sequences used in validation of AS events

Primer	Sequences 5'-3'
Gym.prz_8755 F	AGGATGATGATGGCGAGGAT
Gym.prz_8755 R	CGGATTGCCGTTAGCACTAG
Gym.prz_151000F	CAAGTTGAAGGAGCAAGAGTGC
Gym.prz_151000 R	CTTCATTAGGAATGGGCTGTGA
Gym.prz_131234 F	GGCTGCTCTGTTCGTTAGCC
Gym.prz_131234 R	CCTCCTCCTTTCTTTGCGTTAA

3.7 Transcriptional expression level analysis

Transcriptional expression levels in each tissue were analysed based on Illumina short reads mapping to long reads captured by PacBio. Most of the detected transcripts showed very low expression (0-1 FPKM) and low expression (1-3 FPKM) in all tissues, while only approximately 12.25% and 0.75% of the transcripts showed high (3-60 FPKM) and very high (>60 FPKM) expression, respectively (Supplementary Table S5). Of the 159,053 isoforms, the number of transcripts with a cut-off > 0 FPKM detected in each tissue ranged from 96,395 (60.61%) in the liver to 119,786 (75.31%) in the brain. We identified a total of 19,099 (12.0%) housekeeping transcripts from 159,053 transcripts that were expressed in all five tissues with no less than 1 FPKM in each tissue. The level of transcript expression of each of these tissue-specific transcripts was at least 10-fold higher in one tissue relative to the others. Different amounts of expressed transcripts were detected in various tissues, with the highest expression in the brain (17,059), followed by the muscles (5,252), liver (3,696) and gill (3,100), and the lowest expression in the kidney (1,157) (Fig. 7). We randomly selected 10 transcripts from tissue-specific transcripts to verify their expression levels in various tissues. The experimental results confirmed by q-PCR were consistent with the transcriptome sequencing results (Fig. 8) (Supplementary Table S6).

4. Discussion

In recent years, with the development of sequencing technology and bioinformatics, it has become possible to study the transcriptome of non-model organisms in the absence of a reference genome.[38] To date, most *G. przewalskii* transcriptome studies have been based on next-generation sequencing;[5,13] however, the short reads resulting from this approach have prevented the accurate assembly of FL transcripts in the absence of genomic sequence information.[39] By contrast, SMRT does not require further assembly, which is particularly suitable for the transcriptional analysis of non-model organisms lacking genome sequences. SMRT has the advantage of producing continuous long-read fragments without PCR amplification. However, single-molecule sequencing comes with a high error rate. In this study, after further error correction of polished isoforms sequence with Illumina RNA-seq data, the differences between polished isoforms sequence and before further error correction were slight, which indicates that after re-sequencing and polishing, effective error correction has been obtained.

The FL transcripts of *G. przewalskii*, with mean length and N50 length of 3,445 and 4,348 bp, were obtained by PacBio Iso-seq firstly, which were longer than *G. przewalskii* (average length of 875 and 1988 bp, N50 length of 1593 and 3076 bp) captured by the de novo assembled.[7,13] Regardless of N50 length or average length, FL transcripts were superior to RNA-seq data, although the de novo

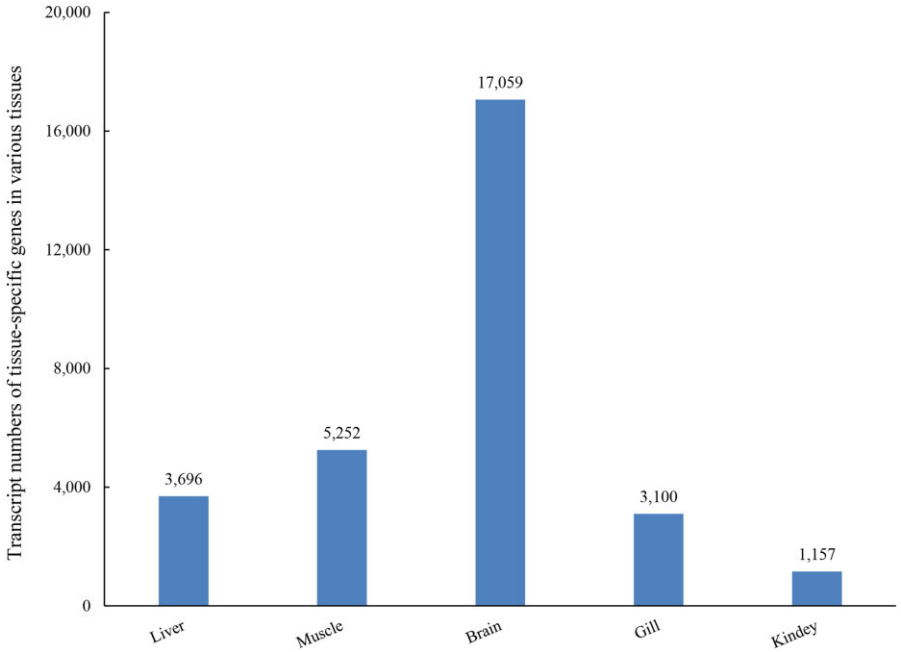

Figure 7.. Transcript numbers of tissue-specific genes in various tissues. Brain showed the greatest number of tissue-specific transcripts, and kidney exhibited the least.

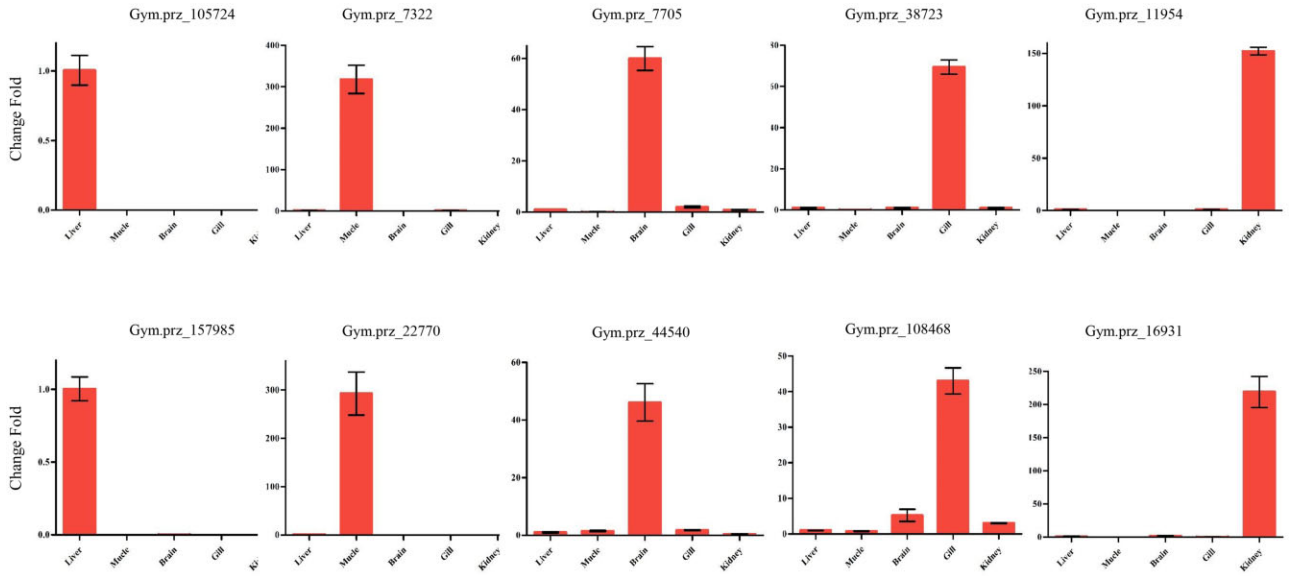

Figure 8.. Quantitative real-time PCR confirmation of the transcript expression obtained by high-throughput sequencing. According to the housekeeping gene, the expression amount of the gene in the tissues was normalized, and the liver was homogenized into 1 serve as reference.

assembled transcripts had a larger number than the FL transcripts. This phenomenon has also been demonstrated in other fish, such as *Gymnocypris namensis* and *Gymnocypris selincuoensis*.[40,41]

In this study, with the help of RNA-seq data, the advantages of single-molecule sequencing, longer length and higher quality, have been fully demonstrated. To some extent, long reads produced by PacBio not only improve the transcript accuracy and quality of *G. przewalskii*, but also facilitate the identification of gene isoforms and gene annotation. After modified, long reads captured by PacBio produced longer ORFs, better predictive results, and better performance in integrity assessment.[42] The results of this study showed that the percentage of complete ORFs contained in the FL transcript of *G.*

przewalskii was 92.3%, which was more abundant than that in other fish with RNA-seq, such as *Oncorhynchus mykiss* (57.1%)[43] and *Oreochromis mossambicus* (13.6%).[44] The ORF is of central importance to gene identification, and the ORF obtained in this study would contribute to the discovery of new genes.[45]

Illumina RNA-seq is now widely used for transcriptome analysis due to its high reading accuracy and low cost;[46] however, short sequence splicing cannot provide a large number of long fragments of transcripts, and some important information may be lost, such as AS. In the field of variable splicing recognition, the acquisition of short-read sequences requires additional assembly from scratch, and the accuracy of the gene model is also questioned.[38] In this study, an all-vs.-all-BLAST pipeline[26] was used to identify AS events from long-read sequences, and 4,149 pairs of AS events were detected from 159,053 FL transcripts. The AS events we randomly selected were validated by RT–PCR and Sanger sequencing. In the absence of a PCR amplification template, a large number of thousand-base reads were obtained from PacBio Iso-seq. Single-molecule sequencing demonstrated great potential for AS event recognition.

The flexibility of AS contributes to environmental adaption and phenotypic plasticity in organisms.[47] It has previously been reported that tropical and polar octopuses differ in RNA editing of voltage-gated potassium channels, resulting in channels adapted to their functional characteristics in different habitats.[48] We identified a large number of protein ubiquitination and DNA repair function genes involved in AS events. Ubiquitin protein genes and DNA repair genes are closely associated with high stress tolerance in grey whales.[49] Generally, ubiquitin protein genes and DNA repair genes are upregulated in response to low temperature.[48] AS facilitates the adaptation of *G. przewalskii* to the environment before genes qualitative change, but its specific mechanism needs to be further studied.

Long-reads captured by single molecule sequencing not only have advantages in AS event recognition, but also contribute to SSR identification. In this study, each FL transcript of *G. przewalskii* contained 0.74 perfect SSR on average, and the frequency of SSR detection was much higher than that detected from the transcripts with RNA-seq in *G. przewalskii* (0.15),[5] *Ctenopharyngodon idella* (0.05)[50] and *Hypophthalmichthys molitrix* (0.16).[51] Among the perfect SSRs, mononucleotide SSRs were the most abundant, and the motifs of mononucleotide, dinucleotide and trinucleotide SSRs were A/T, AC/GT and AAT/ATT, respectively. Similar results have been reported in other fish species.[48,51] It was found that SSRs were mostly located in the non-translated region of mRNA and closely related to gene expression regulation.[52] At present, SSR has been widely used in the fields of genetic diversity detection, genetic relationship identification and population genetics, etc.[51] The SSRs obtained in this study contribute to the accumulation of rich biological data for the genetic research of *G. przewalskii*.

In this study, 145,169 (91.27%) out of 159,053 isoforms were annotated in NR, GO, KO, KOG and Swiss-Prot public databases. The NR database has the largest annotation ratio (91.22%), probably because it represents the largest protein database in the world.[48] Given the absence of reference genomic information, the remaining 13,884 isoforms, including lncRNAs, may suggest putative novel genes in *G. przewalskii*. In the present study, a total of 13,293 lncRNAs were detected. With the development of chip and transcriptome sequencing technology, more and more lncRNAs have been found in mammals and fish.[53,54] For example, in previous neural studies of the model organism *D. rerio*, it has been found that two classes of lncRNA, sox2-ot and cyrano, play active regulatory roles in the development and growth of nerve cells.[54,55] Moreover, lncRNAs have been applied to the study of immune regulation in *D. rerio*, and transcriptomic data analysis has proven that lncRNA-regulated plk3 and syt10 are related to the immune response.[56] The study of lncRNAs in *D. rerio* has a certain reference function for aquatic animals. The lncRNAs obtained in this study are involved in life activities in *G. przewalskii,* and their role needs to be further studied.

By using a combination of Illumina short reads and PacBio capture long reads, gene expression levels in each tissue were compared in this study. The percentage of housekeeping transcripts (larger than 1 FPKM in each tissue) (12.0%) was lower than *O. mykiss* (17.0%),[43] and slightly lower than *G. selincuoensis* (13.9%).[48] This difference may be due to various factors such as sequencing technology, the number of research tissues and different species. For example, *G. przewalskii* and *G. selincuoensis* belong to the Schizothorax subfamily, and they were studied through the same method, leading to little difference in the proportion of housekeeper transcripts. The housekeeping gene was expressed in all tissues and was affected little by the environment, so it has often been used as an internal reference gene for gene expression research. The housekeeper genes obtained in this study provided a certain reference value for the screening of the internal reference gene of *G. przewalskii*. Similar to *O. mykiss*[43] and *G. selincuoensis,*[48] the specific expression of transcription tissue in *G. przewalskii* (18,576) was higher than that in gill, kidney, liver and muscle. On one hand, this phenomenon may occur because of the special function of the brain, as the nerve centre of the brain performs more complex homeostatic activities. In the brain, genes for neurokinin, inositol compounds and other signalling substances are expressed, such as inositol-3-phosphate synthase, which are more abundant in brain tissue than in other tissues. Inositol 3 phosphate, the second messenger of lipids, is involved in a series of physiological activities regulated by ion channels in animal cells. *G. przewalskii* lives in Qinghai Lake with high salinity, and good ion regulation function is needed to maintain homeostasis, which is a significant strategy for its survival. On the other hand, due to screening criteria, tissue size and other reasons, we screened out different amounts of specific genes expressed in each tissue. The expression of antioxidation-related genes (glutathione peroxidase) and immune-related genes (macroglobulin, complement C3) is specific to liver tissues. Creatine kinase and troponin C have been detected as specific transcripts with high muscle expression, and they are associated with intracellular energy movement and muscle contraction.[56] In addition, electrogenic sodium bicarbonate cotransporters and structural proteins (keratin) are specifically expressed in the kidney and gill, respectively. Specifically expressed genes are often predictive of tissue-related functions. The results of tissue-specific expression require further work to reveal their expression and regulation patterns in different tissues.

In addition, we identified a large number of ion channel genes (e.g. potassium channel subfamily), transmembrane protein (TM) family members (TM9, TM6, TM4, TM7) and solute carrier (SLC) family members among genes from the *G. przewalskii* transcripts. Ion channels and transporter genes could help *G. przewalskii* better maintain the body homeostasis. In the transcriptome of *G. przewalskii*, we identified an expanded SLC12 family. Previous scholars have also found the SLC12 family to be expanded in *G. przewalskii*.[7] However, the SLC gene family 12 is critical for cation-coupled

chloride transport and chloride concentration modulation. This suggests that SLC12 might play a positive role in *G. przewalskii* to cope with severe environment stress.

5. Conclusion

In summary, we identified the first FL transcriptome specific to *G. przewalskii* in combination with PacBio Iso-seq and Illumina RNA-seq. A total of 159,053 FL transcripts presented in our study were identified with average length of 3,445 bp and N50 value of 4,348. In total, 91.27% were annotated to Nr, GO, KO, KOG and Swiss-Prot databases. Gene expression profiles were obtained by mapping RNA-seq short reads to FL transcripts. In addition, a total of 4,149 pairs of AS events, 118,185 perfect SSRs and 13,293 lncRNAs were identified. Compared with previous reports using RNA-seq, the length and accuracy of the transcriptome obtained in this study have been improved. The results of this work provide an important and valuable basis for further research on the potential genetic mechanism of *G. przewalskii* to adapt to plateau environment and its population genetics, population protection and phylogeny.

Acknowledgements

We thank the rescue and rehabilitation centre of naked carps in Lake Qinghai for their assistance in the sampling process and Frasergen Information Co., Ltd. (Wuhan, China) for their help in the sequencing work. This work was supported by Project of Natural Science Foundation of Qinghai Province (2018-ZJ-908) and Basic Scientific Research Funds for The Central Public Welfare (2020TD08; 2019 HY-JC01).

Accession numbers

The Iso-seq and short-read RNA-seq data sets that support the findings of this study have been deposited in GenBank with the following accession numbers: PRJNA649368 for Iso-seq data; SRR11429411 to SRR11429420 and SRR11429422 to SRR11429426 for RNA-Seq data set for kidney, gill, brain, muscle and liver transcriptome.

Conflict of interest

The authors declare that they have no competing interests.

References

1. Qu, Y., Zhao, H., Han, N., et al. 2013, Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau, *Nat. Commun*., **4**, 1–9.
2. Ge, R.L., Cai, Q., Shen, Y.Y., et al. 2013, Draft genome sequence of the Tibetan antelope, *Nat. Commun*., **4**, 1–7.
3. Jiang, H.C., Dong, H., Yu, B., et al. 2008, Dominance of putative marine benthic Archaea in Qinghai Lake, north-western China, *Environ. Microbiol*., **10**, 2355–67.
4. Wei, F., Wang, C., Wang, Z., et al. 2014, Cloning and characterization of two subunits of calcineurin cDNA in naked carp (Gymnocypris przewalskii) from Lake Qinghai, China, *Folia Histochem. Cytobiol*., **52**, 232–43.
5. Tong, C., Lin, Y., Zhang, C., Shi, J., Qi, H. and Zhao, K. 2015, Transcriptome-wide identification, molecular evolution and expression analysis of Toll-like receptor family in a Tibet fish, *Gymnocypris przewalskii*, *Fish Shellfish Immunol*., **46**, 334–45.
6. Walker, K.F., Dunn, I.G., Edwards, D., Petr, T. and Yang, H.Z. 1995, A fishery in a changing lake environment: the naked carp Gymnocypris przewalskii (Kessler) (Cyprinidae: schizothoracinae) in Qinghai Hu, China, *Int. J. Salt Lake Res*., **4**, 169–222.
7. Tian, T., Liu, S., Shi, J., Qi, H., Zhao, K. and Xie, B. 2019, Transcriptomic profiling reveals molecular regulation of seasonal reproduction in Tibetan highland fish, Gymnocypris przewalskii, *BMC Genomics*., **20**, 2.
8. Chen, D.Q., Wang, X., Xiong, F. and Tang, H.Y. 2006, Studies on growth characteristics of Gymnocypris przewalskii, *Acta Hydrobiologica Sinica*., **30**, 173–9.
9. Chen, D., Zhang, X., Tan, X., Wang, K., Qiao, Y. and Chang, Y. 2009, Hydroacoustic study of spatial and temporal distribution of Gymnocypris przewalskii (Kessler, 1876) in Qinghai Lake, *Environ. Biol. Fish*., **84**, 231–9.
10. Zhang, R., Ludwig, A., Zhang, C., et al. 2015, Local adaptation of Gymnocypris przewalskii, (*Cyprinidae) Tibetan Plateau. Sci. Rep*., **5**, 1–10.
11. Wood, C.M., Du, J., Rogers, J., et al. 2007, Przewalski's naked carp (Gymnocypris przewalskii): an endangered species taking a metabolic holiday in Lake Qinghai, China. *Physiol. Biochem. Zool*., **80**, 59–77.
12. Tian, F., Tong, C., Feng, C., Wanghe, K. and Zhao, K. 2017, Transcriptomic profiling of Tibetan highland fish (Gymnocypris przewalskii) in response to the infection of parasite ciliate Ichthyophthirius multifiliis, *Fish Shellfish Immunol*., **70**, 524–35.
13. Tong, C., Fei, T., Zhang, C. and Zhao, K. 2017, Comprehensive transcriptomic analysis of Tibetan Schizothoracinae fish Gymnocypris przewalskii reveals how it adapts to a high-altitude aquatic life, *BMC Evol. Biol*., **17**, 1–11.
14. Jun, L. 2005, Life history pattern of Gymnocypris przewalskii przewalskii (Kessler) by fuzzy pattern recognition, *Sichuan J. Zool*., **24**, 455–8.
15. Singh, P., Börger, C., More, H. and Sturmbauer, C. 2017, The role of alternative splicing and differential gene expression in cichlid adaptive radiation, *Genome Biol. Evol*., **9**, 2764–81.
16. Garrett, S. and Rosenthal, J.J. 2012, RNA editing underlies temperature adaptation in K+ channels from polar octopuses, *Science*, **335**, 848–51.
17. Rosenthal, J.J. 2015, The emerging role of RNA editing in plasticity, *J. Exp. Biol*., **218**, 1812–21.
18. Rohan, L., Neil, S., Bleackley, M., Stephen, D. and Thomas, S. 2017, Transcriptomics technologies, *PLoS Comput. Biol*., **13**, e1005457.
19. Manuel, L.G.G. Transcriptomics and gene regulation. In: Wu, J. (ed.), *Translational Bioinformatics*, 2016, *vol.* **9**. Springer, Dordrecht, pp. 141–60.
20. Gao, Y., Xi, F., Zhang, H., et al. 2019, Single-molecule real-time (SMRT) isoform sequencing (Iso-Seq) in plants: the status of the bioinformatics tools to unravel the transcriptome complexity, *Curr. Bioinf*., **14**, 566–73.
21. Elaine, M. and McCombie, W.R. 2017, Library quantification: fluorometric quantitation of double-stranded or single-stranded DNA samples using the qubit system. In: *Cold Spring Harbor Laboratory*, Cold Spring Harbor, New York.
22. Hackl, T., Hedrich, R., Schultz, J. and Förster, F.J.B. 2014, proovread: large-scale high-accuracy PacBio correction through iterative short read consensus, *Bioinformatics*, **30**, 3004–11.
23. Li, W. and Adam, G. 2006, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658–9.
24. Buchfink, B., Xie, C. and Huson, D.H. 2015, Fast and sensitive protein alignment using DIAMOND, *Nat. Methods*, **12**, 59–60.
25. Wang, L., Park, H.J., Dasari, S., Wang, S., Wang, S. and Kocher, J.P. 2013, CPAT: coding-potential assessment tool using an alignment-free logistic regression model, *Nucleic Acids Res*., **41**, e74.
26. Liu, X., Mei, W., Soltis, P.S., Soltis, D.E. and Barbazuk, W.B. 2017, Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome, *Mol. Ecol. Resour*., **17**, 1243–56.
27. Patel, R.K. and Mukesh, J. 2012, NGS QC Toolkit: a toolkit for quality control of next generation sequencing data, *PLoS One*, **7**, e30619.
28. Langmead, B. 2010, Aligning short sequencing reads with Bowtie, *Curr. Protoc. Bioinf*., **32**, 11–7.

29. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat. Protoc.*, **7**, 562–78.
30. Haas, B.J., Papanicolaou, A., Yassour, M., et al. 2013, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nat. Protoc.*, **8**, 1494–512.
31. Li, B. and Dewey, C.N. 2011, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinformatics*., **12**, 323.
32. Love, M.I., Huber, W. and Anders, S. 2014, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, *Genome Biol.*, **15**, 550.
33. Benjamini, Y. and Hochberg, Y. 1995, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc.*, **57**, 289–300.
34. Zhang, H.M., Liu, T., Liu, C.J., et al. 2015, AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors, *Nucleic Acids Res.*, **43**, D76–D81.
35. Xu, Q., Zhu, J., Zhao, S., et al. 2017, Transcriptome profiling using single-molecule direct RNA sequencing approach for in-depth understanding of genes in secondary metabolism pathways of Camellia sinensis, *Front. Plant Sci.*, **8**, 1205.
36. Buschiazzo, E. and Gemmell, N.J. 2006, The rise, fall and renaissance of microsatellites in eukaryotic genomes, *Bioessays*, **28**, 1040–50.
37. Cuc, L.M., Mace, E.S., Crouch, J.H., Quang, V.D., Long, T.D. and Varshney, R.K. 2008, Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (Arachis hypogaea), *BMC Plant Biol.*, **8**, 55–11.
38. Au, K.F., Underwood, J.G., Lee, L. and Wong, W.H. 2012, Improving PacBio long read accuracy by short read alignment, *PLoS One*, **7**, e46679.
39. Zhang, X., Zhou, J., Li, L., et al. 2020, Full-length transcriptome sequencing and comparative transcriptomic analysis to uncover genes involved in early gametogenesis in the gonads of Amur sturgeon (Acipenser schrenckii), *Front. Zool.*, **17**, 11.
40. Luo, H., Liu, H., Zhang, J., et al. 2020, Full-length transcript sequencing accelerates the transcriptome research of Gymnocypris namensis, an iconic fish of the Tibetan Plateau, *Sci. Rep.*, **10**, 1–11.
41. Xiu, F., Yintao, J., Ren, Z., Kang, C. and Yifeng, C. 2019, Characterization and analysis of the transcriptome in Gymnocypris selincuoensis on the Qinghai-Tibetan Plateau using single-molecule long-read sequencing and RNA-seq, *DNA Res.*, **26**, 353–63.
42. Hoang, N.V., Furtado, A., Mason, P.J., et al. 2017, A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing, *BMC Genomics.*, **18**, 1–22.
43. Salem, M., Paneru, B., Al-Tobasei, R., et al. 2015, Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout, *PloS One*, **10**, e0121778.
44. Zhu, W., Wang, L., Dong, Z., et al. 2016, Comparative transcriptome analysis identifies candidate genes related to skin color differentiation in red tilapia, *Sci. Rep.*, **6**, 1–12.
45. Sieber, P., Platzer, M. and Schuster, S. 2018, The definition of open reading frame revisited, *Trends Genet.*, **34**, 167–70.
46. Qian, X., Ba, Y., Zhuang, Q. and Zhong, G. 2014, RNA-Seq technology and its application in fish transcriptomics, *Omics*, **18**, 98–110.
47. Mastrangelo, A.M., Marone, D., Laidò, G., De Leonardis, A.M. and De Vita, P. 2012, Alternative splicing: enhancing ability to cope with stress via transcriptome plasticity, *Plant Sci.*, **185**, 40–9.
48. Healy, T.M. and Schulte, P.M. 2019, Patterns of alternative splicing in response to cold acclimation in fish, *J. Exp. Biol.*, **222**, jeb. 193516.
49. Toren, D., Kulaga, A., Jethva, M., et al. 2020, Gray whale transcriptome reveals longevity adaptations associated with DNA repair and ubiquitination, *Aging Cell*, **19**, e13158.
50. Wan, Q. and Su, J. 2015, Transcriptome analysis provides insights into the regulatory function of alternative splicing in antiviral immunity in grass carp (Ctenopharyngodon idella), *Sci. Rep.*, **5**, 12946.
51. Dewoody, J.A. and Avise, J.C. 2000, Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals, *J. Fish Biol.*, **56**, 461–73.
52. Tranbarger, T.J., Kluabmongkol, W., Sangsrakru, D., et al. 2012, SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of Elaeis guineensis, *BMC Plant Biol.*, **12**, 1.
53. Angrand, P.O., Vennin, C., Le Bourhis, X. and Adriaenssens, E. 2015, The role of long non-coding RNAs in genome formatting and expression, *Front. Genet.*, **6**, 165.
54. Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H. and Bartel, D.P. 2011, Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution, *Cell*, **147**, 1537–50.
55. Amaral, P.P., Neyt, C., Wilkins, S.J., et al. 2009, Complex architecture and regulated expression of the Sox2ot locus during vertebrate development, *RNA*, **15**, 2013–27.
56. Wallimann, T., Tokarska-Schlattner, M. and Schlattner, U. 2011, The creatine kinase system and pleiotropic effects of creatine, *Amino Acids.*, **40**, 1271–96.
57. Fu, B. and He, S. 2012, Transcriptome analysis of silver carp (Hypophthalmichthys molitrix) by paired-end RNA sequencing, *DNA Res.*, **19**, 131–42.
58. Wang, X., Lin, J., Li, F., et al. 2017, Screening and functional identification of lncRNAs under β-diketone antibiotic exposure to zebrafish (Danio rerio) using high-throughput, *Aquat. Toxicol.*, **182**, 214–25.