



OPEN Full-Length transcriptome assembly and SSR marker development for *Spinibarbus hollandi* using PacBio SMRT sequencing

Sixun Li, Jie Lai, Mengmeng Wu, Zihang Xie, Min Zhang, Guojun Cai, Ziyang Deng, Binhua Deng, Chong Han[✉] & Qiang Li[✉]

Spinibarbus hollandi is an economically significant subtropical fish in China, valued for both ornamental and aquaculture purposes. However, the full-length transcriptomic resources for this species remain scarce. To address this gap, we constructed the first multi-tissue, full-length transcriptome of *S. hollandi* using PacBio single-molecule real-time sequencing. We assembled 23,403 non-redundant transcripts, including 15,197 unigenes, with a mean length of 2,147 bp and an N50 of 2,868 bp. Among these, 14,567 unigenes (95.85%) were annotated in public databases. Our analysis identified 373 alternative splicing events, with retained intron being the most common type. We also predicted 2,397 long non-coding RNAs. Furthermore, a comprehensive screening revealed 7,449 simple sequence repeat (SSR) loci, comprising 1,198 compound and 6,251 perfect SSRs, with an occurrence frequency of 30.96%. Di- and tri-nucleotide repeats were the predominant types. Thirteen highly polymorphic SSR loci showed robust polymorphism across four geographical populations. This study provides a crucial data foundation and genetic resource for functional gene research, molecular marker-assisted breeding, and germplasm conservation in *S. hollandi*.

Keywords *Spinibarbus hollandi*, PacBio sequencing, Full-length transcriptome, Gene annotation, Alternative splicing, Simple sequence repeat

Spinibarbus hollandi, a cyprinid fish belonging to the subfamily Barbinae, is an economically important freshwater species in China, primarily distributed in the southeastern provinces, including Zhejiang, Guangdong, Guangxi, and Fujian. As an omnivore, it feeds on small fish, shrimp, aquatic insects, and organic detritus¹. *S. hollandi* is also a popular food fish, recognized for its desirable taste and nutritional quality^{2,3}. In recent years, the aquaculture scale of *S. hollandi* has expanded annually, particularly in southern China. However, its commercial development is constrained by several biological limitations, including an exceptionally slow growth rate (requiring 1.5–2.5 years to reach market size)⁴, delayed sexual maturation (3–4 years), and low fecundity⁵.

To elucidate the molecular mechanisms underlying these traits, several transcriptomic studies have been conducted. For instance, brain and muscle transcriptomes have been analyzed to investigate starvation response and compensatory growth^{6,7}. Han et al. not only used genomic data to identify sex-specific markers⁸ but also analyzed gonadal transcriptome data, revealing many sex-related genes⁹. Although a complete mitochondrial genome was published several years ago¹⁰, the currently available genomic and high-quality transcriptomic data remain limited. This paucity of genetic resources restricts further research on functional genes related to key traits such as growth, stress tolerance and gonad development, and hinders the development of aquaculture for this species.

SMRT sequencing technology (Pacific Biosciences), also known as third-generation sequencing, provides a powerful strategy for obtaining comprehensive gene sequence and structural information by constructing a full-length transcriptome, especially in the absence of a high-quality reference genome. Full-length transcriptome sequencing has been widely applied to obtain complete, non-fragmented transcripts (including 5'UTR, 3'UTR, and polyA tails)¹¹, discover novel functional genes¹², and distinguish alternative splicing isoforms¹³. This approach

South China Biodiversity Research Center, School of Life Sciences, Guangzhou University, Guangzhou, China. ✉email: hanch@gzhu.edu.cn; lq512328@163.com

has demonstrated considerable advantages and has been successfully employed in genetic resource development for various aquatic economic species, such as *Lateolabrax maculatus*¹⁴ and *Danio rerio*¹⁵. Furthermore, high-quality full-length transcriptome data enable efficient and cost-effective development of a large number of SSR molecular markers, which can be used for genetic analysis of the sequenced species and related taxa, as well as for research in conservation biology and molecular-assisted breeding.

Microsatellites, also known as simple sequence repeats (SSR), are DNA sequences consisting of tandem repeats of 1–6 nucleotide motifs^{7,8}. Variations in motif type and repeat number at each SSR locus generate polymorphism. SSR markers can be classified into genomic SSR and expressed sequence tag (EST)-derived SSR based on their source. Although EST-SSRs generally exhibit lower polymorphism than genomic SSR, they are more accessible in non-model organisms lacking reference genomes and often show higher conservation and transferability across related species¹⁶.

In this study, we employed PacBio Iso-Seq sequencing to generate the first multi-tissue, full-length transcriptome of *S. hollandi*. This resource enabled comprehensive gene functional annotation, alternative splicing analysis, and SSR marker screening. This resource will provide a foundation for transcriptomic studies, functional gene mining, and molecular marker development in this species, and will serve as a key reference for future genome assembly and annotation efforts.

Materials and methods

Sample collection and RNA extraction

In this study, six 4-year-old female ($n=3$) and male ($n=3$) *S. hollandi* were obtained from Jinyue Aquatic Technology Co., Ltd. in Shaoguan, Guangdong Province, China. Prior to sampling, all fish were acclimatized in culture ponds for at least one week and showed no signs of disease, each with an average body weight of 1.5 kg. The fish were anesthetized with MS-222 until opercular movements ceased, indicating deep anesthesia, and then euthanized by overdose. From each fish, six tissues (heart, gill, brain, fin, liver, and gonad) were collected. All tissues were immediately snap-frozen in liquid nitrogen and stored at -80°C for long-term preservation. The entire procedure—from anesthesia and sacrifice of the fish to tissue dissection and freezing in liquid nitrogen—was completed within 10–15 min to maximize RNA integrity. This study was reported in accordance with ARRIVE guidelines. All experimental protocols involving animals were performed in accordance with the guidelines and regulations approved by the Experimental Animal Ethics Committee of Guangzhou University (No. GURBBB241205).

Total RNA was extracted from each tissue using RNA isolator Total RNA Extraction Reagent (Vazyme, China), following the manufacturer's instructions. After extraction, the concentration and purity of all RNAs were detected by Nanodrop2000 (Thermo Scientific, USA). RNA integrity and quantity were assessed with an Agilent 4200 Bioanalyzer (Agilent Technologies, USA).

Library construction and Pacbio SMRT sequencing

The RNA used for library construction was a pooled sample prepared by mixing equal quantities of RNA extracted from all six tissues collected from each of the six individuals. Polyadenylated mRNA was enriched using Oligo(dT) magnetic beads. First-strand cDNA was synthesized from the mRNA template using the SMARTer™ PCR cDNA Synthesis Kit, followed by PCR amplification to enrich full-length cDNA. The resulting full-length cDNA fragments were size-selected using the BluePippin system and subjected to an additional round of PCR amplification. The Iso-seq library was then prepared through end repair and ligation of SMRTbell adapters.

Following library construction, concentration was quantified using a Qubit 2.0 fluorometer, and insert size distribution was assessed with an Agilent 2100 Bioanalyzer. Sequencing primers and DNA polymerase were bound to the SMRTbell templates using the PacBio Calculator, and the library was sequenced on a PacBio Sequel system (Pacific Biosciences, USA). The Raw data of the original sequence in Pacbio are called Polymerase Read sequences. Sequencing was performed using DNA polymerase with SMRTbell™ circular template chain. The raw data generated by the PacBio platform, termed “Polymerase Reads,” were processed using SMRT Link v8.0 software. Circular Consensus Sequences (CCSs) were generated from subreads with the following criteria: minimum number of full passes ≥ 0 , sequence length > 50 bp, and predicted accuracy > 0.75 . Sequences meeting these criteria were classified as high-quality CCSs for downstream analysis. Adapter sequences were identified and removed, and the resulting subsequences (subreads) were filtered to exclude those shorter than 50 bp, yielding the final subread dataset.

Full-Length transcriptome assembly

Circular Consensus Sequences (CCS), also referred to as Reads of Insert (ROI), were generated from the subreads using the CCS tool (v6.0.0). The Lima version 2.0.0 and Isoseq3 Refine version 3.4.0 software program to obtain the Length to classify the chimeric sequences (Full-Length chimeric Read, FLC) and full-length non-chimeric Read (FLNC). To reduce redundancy, FLNC reads were clustered using the cluster program within IsoSeq3, with similar sequences grouped into clusters, each represented by a single transcript. Unclustered FLNC data were designated as Singletons. Finally, unigenes were obtained by clustering isoform sequences using CD-HIT (v4.6.7).

Functional annotation and classification

Unigene annotation was carried using homology searches against the major public databases including Nr database, the Clusters of euKaryotic Orthologous Groups database, the Kyoto Encyclopedia of Genes and Genomes¹⁷ and the Uniprot protein database by BLASTx. Blast2GO was applied to annotate the NR annotation results, and the number of unigenes related to each gene ontology was calculated based on biological process, cell composition and molecular function classification.

Gene structure prediction

To obtain high quality non-redundant transcripts, Cogent (Coding GENome reconstruction Tool) v6.0.0 software was adopted. Then, CPC2.py v0.1 software¹⁸ was used to determine the coding potential of the non-redundant transcript sequence, and the encoded transcript sequence was further validated by BLASTx search against the NR database. Transcripts that lacked significant homology to known mRNAs in the NR database and were longer than 200 nt were classified as long non-coding RNAs (lncRNAs). Alternative splicing (AS) isoforms were further analyzed using the SUPPA program (2.2)¹⁹.

SSR locus screening

SSR loci were identified from unigene sequences longer than 500 bp using MISA v2.1 software. Screening parameters were set as follows: minimum of 6 repeats for dinucleotides, and 5 repeats for tri-, tetra-, penta-, and hexa-nucleotides SSRs separated by less than 100 bp were classified as compound SSRs. Primers for the identified SSR loci were designed using Primer5 with the following parameters: primer length 18–22 bp, annealing temperature 55–65 °C, GC content 40%–60%, and product size 100–300 bp. Primer pairs were screened against the transcriptome database using BLAST to exclude those likely to cause non-specific amplification.

Screening of SSR primers and PCR amplification

A preliminary screening of all 50 designed primer pairs was conducted using DNA templates from eight randomly selected *S. hollandi* individuals representing four distinct populations. PCR reactions were carried out in a 20 µL volume containing 10 µL of 2×PCR master mix (Vazyme, Nanjing, China), 0.5 µL each of forward and reverse primers, 2 µL of template DNA, and ddH₂O to the final volume. The amplification protocol was as follows: initial denaturation at 95 °C for 5 min; The thermal cycling protocol consisted of an initial denaturation at 95 °C for 5 min; 35 cycles of 95 °C for 30 s, 56 °C for 30 s, and 72 °C for 15 s; and a final extension at 72 °C for 5 min. Amplification products were separated by 1% agarose gel electrophoresis. Primer pairs that yielded a single, clear band of the expected size were selected for subsequent polymorphism analysis.

Thirteen representative SSR primer pairs, exhibiting high polymorphism and consistent amplification, were fluorescently labeled. These primers were used to genotype 51 individuals from four river systems: GL (Li River, Pearl River system; *N* = 12; 110°25'49.191"E, 25°6'50.221"N), YS (Lian River, Pearl River system; *N* = 12; 112°43'52.681"E, 24°25'56.787"N), HC (Long River, Pearl River system; *N* = 15; 108°57'37.749"E, 24°28'24.123"N) and GZ (Gan River, Yangtze River system; *N* = 12; 114°41'12.859"E, 26°32'39.500"N). Post-amplification, products were first verified on 1% agarose gels. PCR products showing clear target bands were then sent to Sangon Biotech (Shanghai) Co., Ltd. for high-throughput capillary electrophoresis genotyping on an ABI 3730xl DNA Analyzer.

Population genetics analysis

The population genetic analysis comprised the following steps. First, raw capillary electrophoresis data were genotyped using GeneMarker[®] software to determine allele sizes for each individual at each locus. Low-signal peaks, non-specific amplifications, and monomorphic loci were filtered out. A preliminary SSR genotype matrix was constructed and manually curated against molecular weight standards to finalize allele counts and genotypes. Data quality was assessed using MICRO-CHECKER 2.2.3 to detect potential genotyping errors such as null alleles, stuttering, or large allele dropout. Genetic diversity parameters-including the number of alleles, observed heterozygosity, expected heterozygosity, and polymorphism information content-were calculated using Cervus 3.0. Deviations from Hardy-Weinberg equilibrium per locus and pairwise linkage disequilibrium among the retained loci were assessed with Arlequin 3.0 using 10,000 permutations. Finally, to elucidate genetic structure, a UPGMA dendrogram based on Nei's genetic distance was constructed in MEGA 5.0, and genetic relationships among populations and individuals were visualized via principal coordinate analysis implemented in GENALEX 6.5.

Results

Quality control of the Full-Length transcriptomes

A total of 41,844,240 reads (89.86 Gb nucleotides) were obtained, with 2,147 bp of mean length and 2,868 bp of N50 value (Supplementary Tables 1 and 2). The raw sequencing data have been deposited in the NCBI SRA database under the accession number PRJNA915485. Cumulative distribution of Bases numbers (Fig. 1A) and reads numbers (Fig. 1B) by length was mainly in the range of 0–4000 bp (Supplementary Table 3). Data processing yielded 638,500 Circular Consensus Sequences (CCSs) with a mean length of 2,852 bp. Subsequent classification of these CCSs produced 441,782 full-length non-chimeric (FLNC) reads, which were further clustered into 39,944 high-quality isoforms for downstream analysis. (Figure 1C and D; Supplementary Table 4). Finally, we used Cogent v6.0.0 to process high quality reads into clusters to remove redundant sequences. A total of 23,403 Non redundant isoforms, 15,188 UniTransModel transcripts and 15,197 unigenes were assembled (Supplementary Tables 5 and Table 6).

Functional annotation

Of the 15,197 unigenes, 14,567 (95.85%) were successfully annotated in at least one public database (Fig. 2A). Specifically, 14,486 (95.32%), 14,540 (95.68%), 11,194 (73.66%), and 10,843 (71.35%) unigenes were annotated in the NR, Uniprot, KOG, and KEGG databases, respectively. BLASTx analysis revealed that the vast majority of annotated unigenes showed the highest similarity to species within the Cyprinidae family. The most frequent top-hit species was *Sinocyclocheilus rhinocerosus* (27.02%), followed by *Sinocyclocheilus anshuiensis* (25.14%), *Cyprinus carpio* (17.25%), and *Sinocyclocheilus grahami* (16.71%) (Fig. 2B).

A total of 12,967 unigenes (85.33%) were assigned to 49 GO subcategories. The most abundant ten subcategories were: binding (7,069), cellular process (6,991), cell part (6,540), single-organism process (4,738),

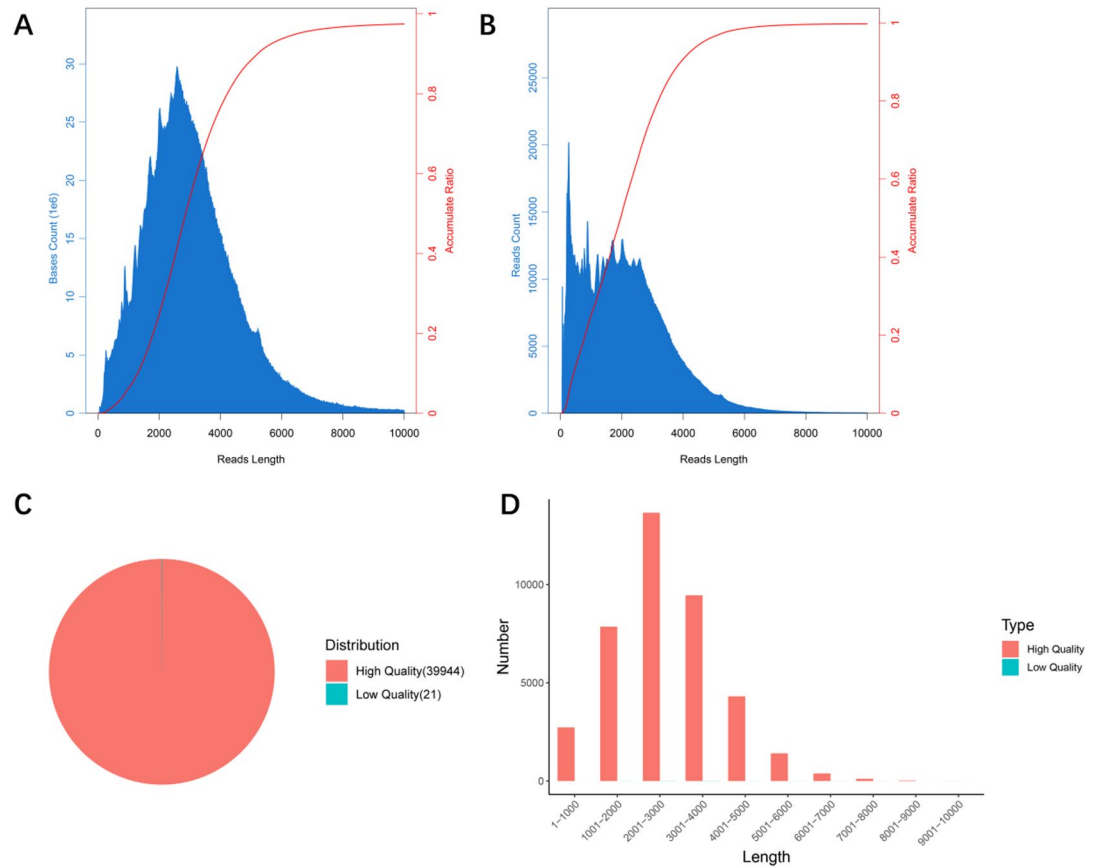


Fig. 1. Cumulative distribution of base number by length (A) and read number by length (B), and the ratio (C) and length distribution (D) of High quality and low quality subreads.

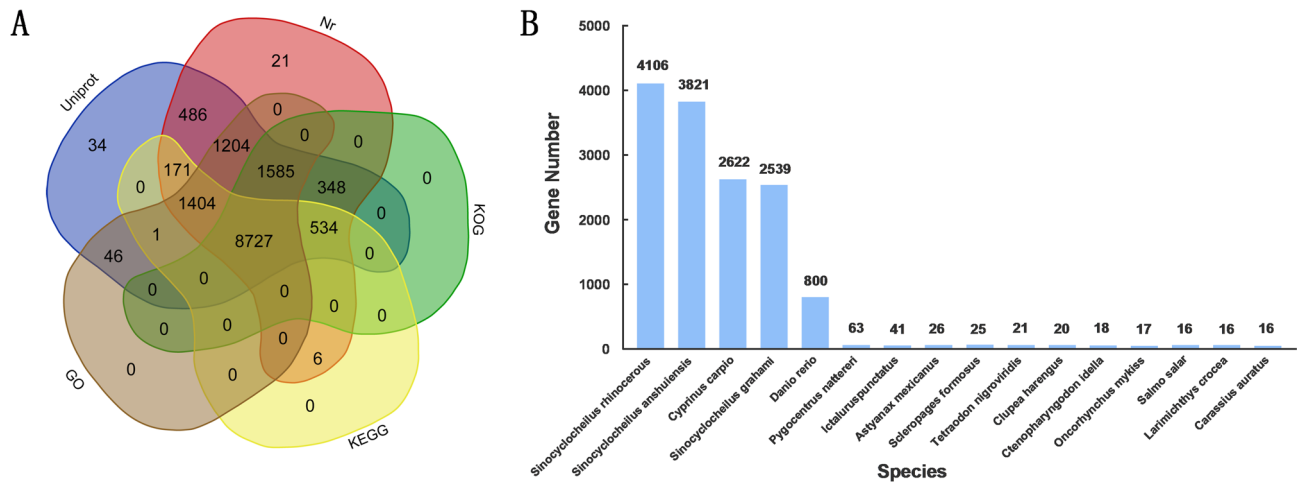


Fig. 2. Functional annotation and species distribution. (A) Venn diagram of function annotated transcripts. (B) Species distribution of the top BLASTx hits for unigenes against the NR database.

metabolic process (4,681), biological regulation (4,476), catalytic activity (4,113), membrane part (3,469), organelle (3,355), and developmental process (2,199). “Within the biological process domain, single-organism process (4,738) and metabolic process (4,681) were most represented. For cellular process component, cell part (6,540) was predominant, and in molecular function, binding (7,069) was the most abundant term (Fig. 3A).

A total of 10,843 (68.12%) unigenes annotated in KEGG were primarily classified into five functional categories: Cellular Processes (2,872), Environmental Information Processing (2,407), Genetic Information Processing (1,462), Metabolism (3,273), and Organismal Systems (4,853). The most enriched pathway was

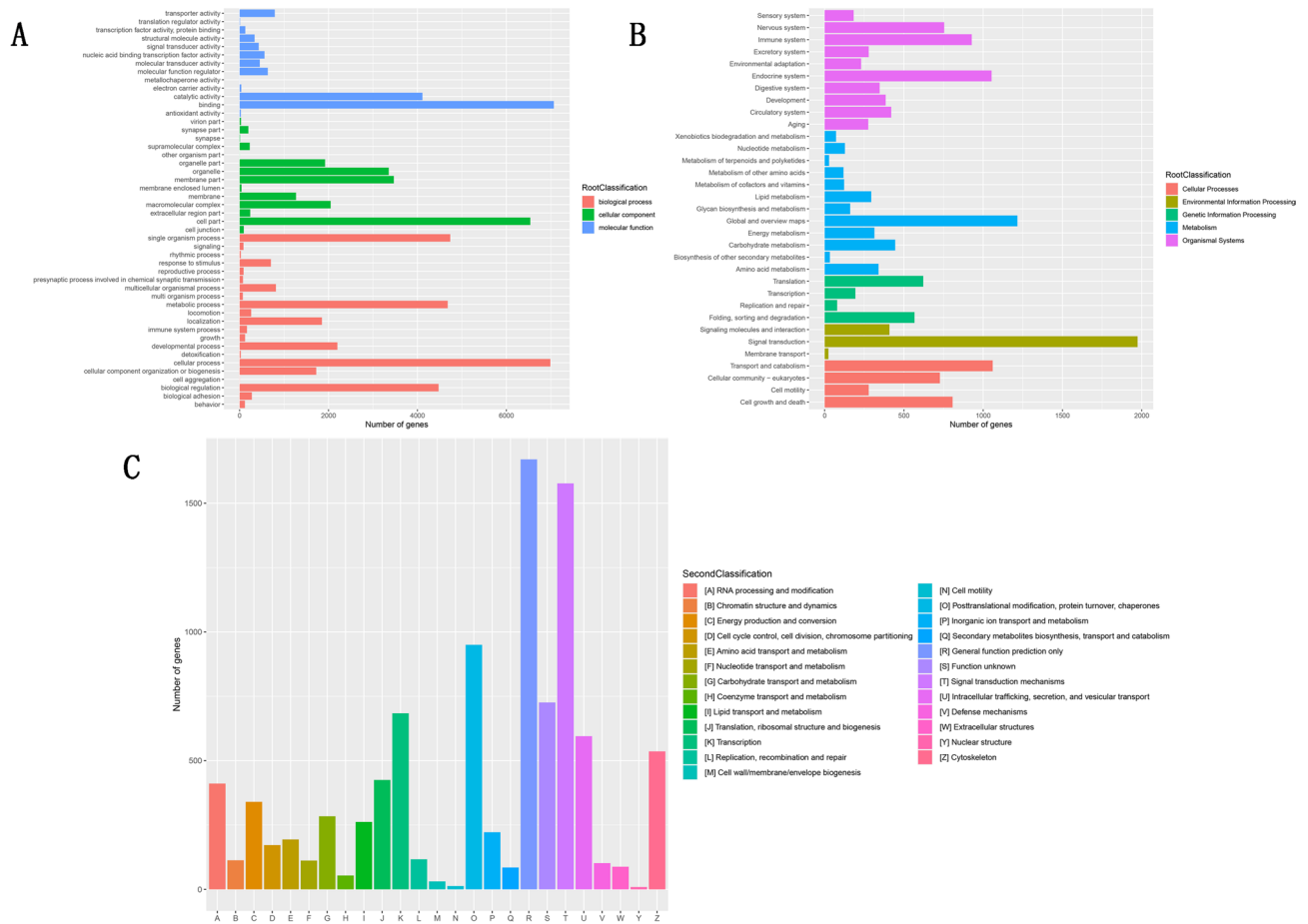


Fig. 3. Functional classification of annotated unigenes. **(A)** Gene Ontology (GO) classification across biological process, cellular component, and molecular function categories. **(B)** KEGG pathway classification. **(C)** Eukaryotic Orthologous Groups (KOG) functional classification.

“Signal transduction” (1,974), followed by “Global and overview maps” (1,216) and “Transport and catabolism” (1,060) (Fig. 3B). Additionally, 11,194 (70.33%) KOG-annotated genes were categorized into 25 families. The most abundant category was observed in “General function prediction only” (1,669), followed by “Signal transduction mechanisms” (1,577) and “Posttranslational modification, protein turnover, chaperones” (950). The smallest family was “Nuclear structure” (9) (Fig. 3C).

Gene structure predictions

Alternative splicing generates different mRNA isoforms from a single gene. AS is one mode of gene expression that is common in most eukaryotic biological cells, which leads to differences in the number of genes and proteins. There are seven common types of AS, including Alternative 3’ splice Site (A3), Alternative 5’ splice Site (A5), Mutually exclusive exons (MX), Skipping Exon (SE), Retained Intron (RI), Alternative First exon (AF) and Alternative Last exon (AL). We identified 373 AS events, which were categorized into six types. Retained intron (RI) was the most prevalent (249 events, 66.8%) (Fig. 4A).

lncRNAs that contained polyA tails were also analyzed. A total of 2,397 lncRNAs containing polyA tails were identified, with an average length of 1,942 bp (Supplementary Table 5). Since the reference genome of *S. hollandi* has not been published, the exons of lncRNA were not predicted. The length distribution showed that most lncRNAs was in the length of 0–4000 bp (Fig. 4B).

SSR loci and repeat types

Screening of the 15,197 unigene sequences identified 7,449 SSR loci from 4,705 unigenes. The occurrence frequency (proportion of sequences containing SSRs) was 30.96%, and the distribution density (ratio of total SSR loci to total unigenes) was 49.01%, with an average density of one SSR per 6.144 kb. A total of 1,505 unigenes contained two or more SSR loci, and 1,198 compound SSRs were detected (Table 1).

Dinucleotide repeats were the most abundant (4,888 loci; 65.62%). Among the four dinucleotide types, AC/GT was dominant (2,773 loci; 37.23%). Trinucleotide repeats comprised 2,069 loci (27.78%) across 10 motif types, with AAT/ATT being most frequent (550 loci; 7.38%). Tetra-nucleotide (393 loci; 5.28%) and penta-nucleotide (68 loci; 0.91%) repeats were less common, with 23 and 25 motif types, respectively. Hexa-nucleotide repeats had the lowest proportion (0.42%), with 21 motif types detected (Fig. 5 and Fig. 6).

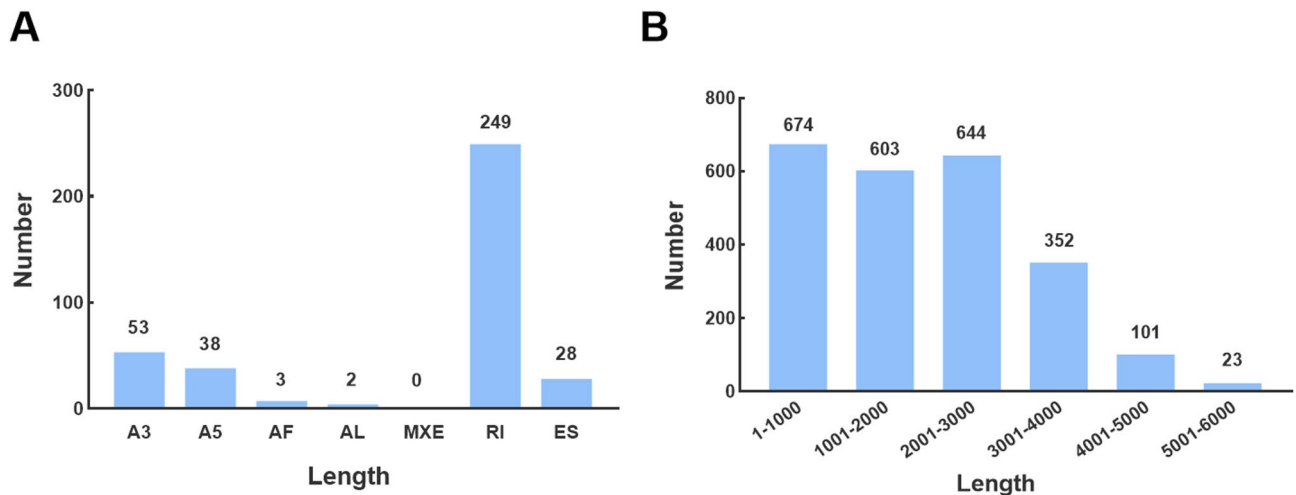


Fig. 4. Analysis of alternative splicing and long non-coding RNAs. **(A)** Types and counts of alternative splicing (AS) events identified. **(B)** Length distribution of predicted long non-coding RNAs (lncRNAs).

Type	Number
Total number of sequences examined:	15,197
Total size of examined sequences (bp):	45,767,809
Total number of identified SSRs:	7,449
Number of SSR containing sequences:	4,705
Number of sequences containing more than 1 SSR:	1,505
Number of SSRs present in compound formation:	1,198

Table 1. Summary statistics of SSR loci identified in the *S. hollandi* transcriptome.

SSR length and marker development

The length distribution of all identified SSRs showed significant variation, with the shortest being 12 bp and the longest exceeding 100 bp. Overall, the majority of SSRs (65.54%) ranged from 12 to 20 bp in length. Within this range, the most abundant repeat motifs were dinucleotides (3,304 loci), followed by trinucleotides (1,415 loci) and tetranucleotides (163 loci). Among the identified SSRs, 2,567 loci (34.46%) exceeded 20 bp in length. Notably, the majority of these longer SSRs (2,238 loci, 87.18%) consisted of di- and tri-nucleotide repeats (Table 2).

Based on the results of SSR detection, 50 primer pairs were successfully designed and 36 primer pairs resulted in PCR products of the expected size. Thirteen candidate SSRs were selected based on the PCR band intensity and size polymorphism resolved on agarose gel electrophoresis. The genotyping quality of the 13 SSR markers was rigorously validated prior to diversity analysis. First, all loci conformed to HWE across the four populations (all $P > 0.05$). Additionally, MICRO-CHECKER analysis detected no significant evidence for null alleles or other genotyping artifacts (e.g., large allele dropout, stuttering), confirming the reliability of the dataset. Among the 13 primer pairs, observed heterozygosity (H_o) ranged from 0.389 to 0.783 (mean = 0.603), expected heterozygosity (H_e) varied from 0.513 to 0.843 (mean = 0.696), and polymorphism information content (PIC) values spanned 0.491–0.832 (mean = 0.651). Eleven loci showed high informativeness ($PIC > 0.500$), while one displayed moderate polymorphism ($0.500 > PIC > 0.250$), indicating good overall polymorphism. A total of 152 alleles were amplified by the 13 primer pairs, with an average of 11.69 alleles per locus (N_a). The mean Shannon's information index (I) was 1.077 (range: 0.632–1.341) (Table 3).

The UPGMA dendrogram based on Nei's genetic distance (Fig. 7A) revealed that populations YS and GL clustered first, showing clear genetic differentiation from populations HC and GZ. Principal coordinate analysis (PCoA) performed on the 51 *S. hollandi* individuals from the four river systems (Fig. 7B) indicated that the first and second principal coordinates explained 16.9% and 13.6% of the total genetic variation, respectively. In the PCoA plot, closer spatial proximity corresponds to closer genetic relatedness. While individuals from all four populations showed partial overlap, indicating some genetic similarity, populations HC and GZ formed relatively distinct clusters. In contrast, populations YS and GL exhibited extensive overlap, with GL showing the least within-population variation.

Discussion

Quality of the full-length transcriptome

The third-generation PacBio SMRT Sequencing eliminates PCR amplification, avoiding allelic bias²⁰, and its long-read capability enables complete capture of transcript structures²¹—essential for transcriptome assembly in

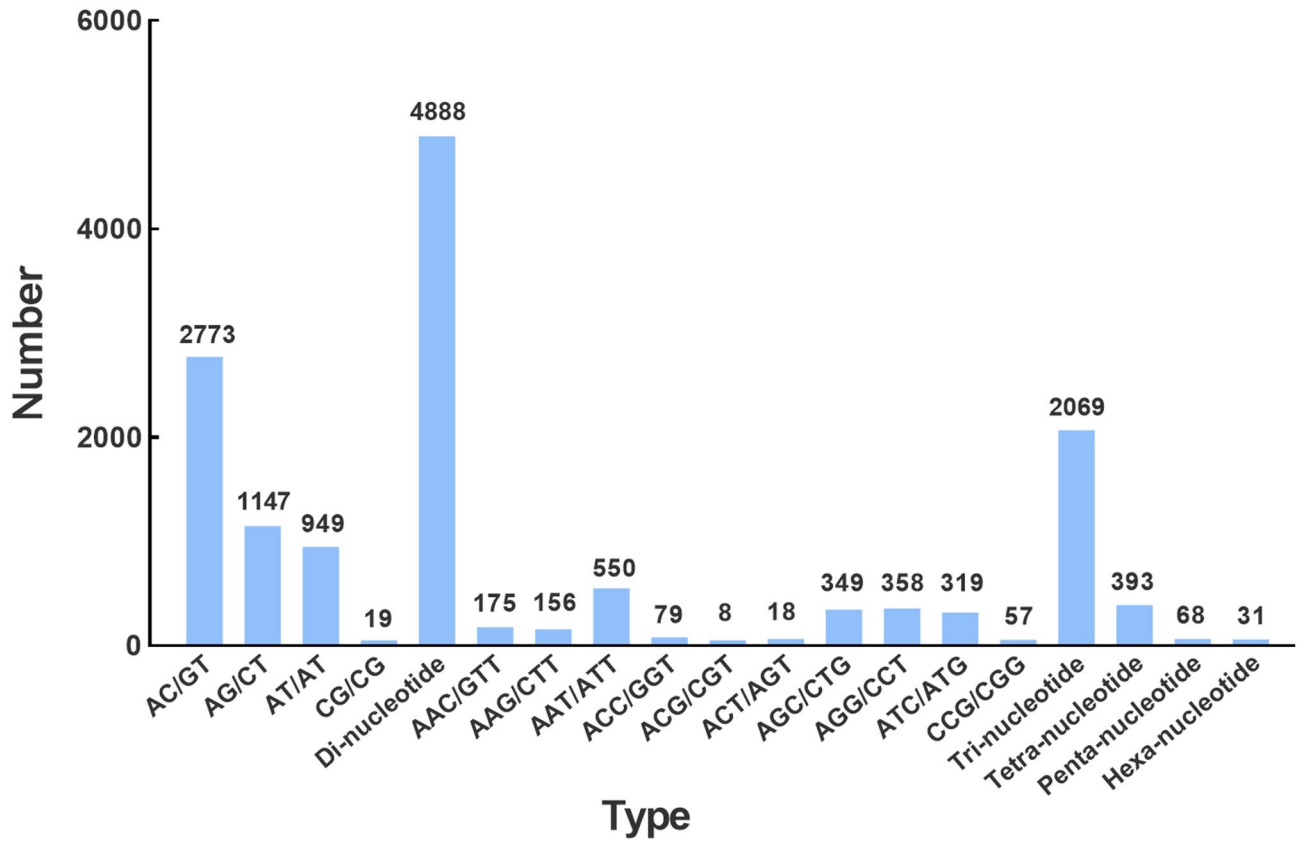


Fig. 5. Frequency and distribution of SSR motif types.

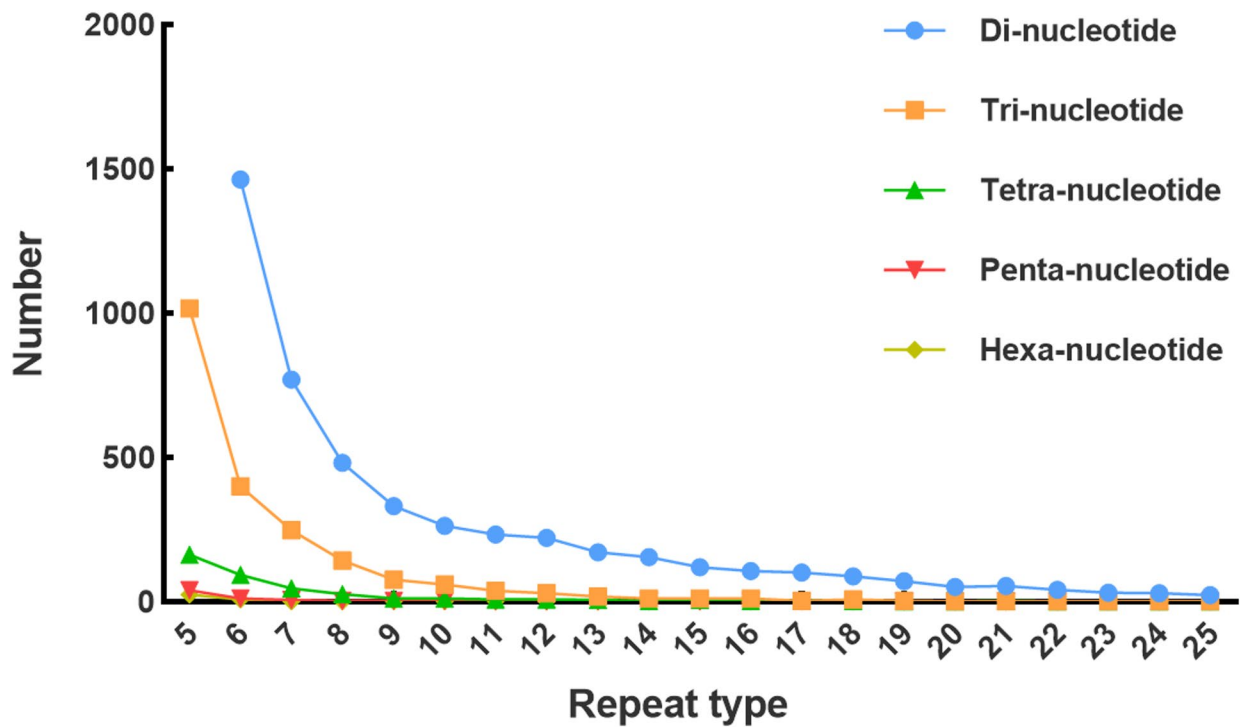


Fig. 6. Detailed distribution of different specific SSR motifs identified.

Repeat Type	12–20 bp	21–30 bp	31–40 bp	41–50 bp	51–100 bp	> 100 bp	Total
Di-nucleotide	3304	896	415	177	93	0	4888
Tri-nucleotide	1415	467	142	31	12	2	2069
Tetra-nucleotide	163	138	48	16	28	0	393
Penta-nucleotide	0	50	7	4	7	0	68
Hexa-nucleotide	0	24	7	0	0	0	31
Total	4882	1575	619	228	140	2	7449

Table 2. Distribution of SSR motifs by repeat length in *S. hollandi* transcriptome. Note: The table presents the number of SSR loci categorized by their motif type and total repeat length (in base pairs). Compound SSRs are not included in this count.

SSR	Primer sequence	Na	P(HWE)	Ho	He	I	PIC
(TCA)11	F: TTGTGGGTGCTTGTTTTCA R: TCTGGCCTTTTATGATGG	16	0.377	0.783	0.843	1.281	0.832
(ATC)11	F: ACTGCAGTAGGTGCATGTGG R: TCTGGCCTTTTATGATGG	12	0.316	0.597	0.631	1.139	0.614
(GAG)10	F: TCTCTGTCCCGTTCGAGT R: GGATAGCTTCATCAAAGCGG	10	0.249	0.686	0.781	0.981	0.638
(ATC)9	F: GATGACGCTGCATGTGTT R: TTCCTCCTTCATGTCGCT	7	0.421	0.389	0.513	0.847	0.495
(GT)19	F: TCATTGTAACGCACGACCTC R: ACACCTGGCTCTCTTCCAA	14	0.264	0.695	0.833	1.288	0.804
(AG)19	F: TTTGCCATCACACAACAGGT R: TATCTGCCTCCAGATCCCAC	12	0.364	0.617	0.754	1.267	0.737
(AC)17	F: CAAAATACCCTGAAGCCCAG R: CAAACAATCCAGTCGTCCAA	13	0.403	0.632	0.778	1.341	0.711
(AC)15	F: CTGTCAGGCCTATGAGCGAT R: TCAAGCCTTTCAGCACATTG	12	0.252	0.476	0.530	1.019	0.513
(TTCA)11	F: TGAGATCATGGTTTCAAATGGA R: TTTGTGCTCACAATTACTGCAT	8	0.347	0.468	0.566	0.779	0.491
(GAAT)10	F: TGCATGAGAACCAATGAATGA R: CGTCATCCCAGCGACTACT	16	0.135	0.618	0.704	1.096	0.688
(TGTT)9	F: CCTGAAAATAAAAGTGTCTGGA R: GGCGGTGATGAAAGAACAAG	11	0.454	0.580	0.768	1.258	0.703
(TTCT)8	F: TTGCTCTGGTGTAGAGCGTG R: TGCCTCAAGTCTGAGAGGGT	10	0.587	0.651	0.687	1.067	0.603
(TCCA)8	F: TGCATCCATCTGTCCATTGT R: TGGATGCATAACATGAATGG	11	0.428	0.647	0.662	0.632	0.633

Table 3. Characteristics of thirteen polymorphic SSR primers developed for *S. hollandi*. Note: Na: observed number of alleles; Ho: observed heterozygosity; He: expected heterozygosity; I: Shannon's Information Index. PIC: polymorphism information content; P(HWE): *p* value of Hardy-Weinberg equilibrium test.

species lacking a reference genome, such as *S. hollandi*²². Previous genomic resources for *S. hollandi* were limited by the fragmented nature of short-read assemblies. For instance, Han et al.⁹ generated a gonadal transcriptome using Illumina sequencing, yielding 167,152 unigenes with an average length of 871 bp and an N50 of 1,198 bp. In contrast, our study produced 23,403 high-quality non-redundant transcripts and 15,197 unigenes, with a mean length of 2,147 bp and an N50 of 2,868 bp—markedly superior to the Illumina-based assembly. Similarly, for the ancient clam (*Coelomactra antiquata*), Illumina sequencing produced 214,732 unigenes with an average length of 616.2 bp²³, while PacBio SMRT sequencing yielded 39,209 unigenes with a mean length of 2,732 bp²⁴. These results demonstrate that third-generation sequencing outperforms second-generation sequencing in data quality, transcript length, and annotated genetic information.

Functional annotation insights

The annotation success rate of transcripts in public databases reached 95.85%—a key metric for evaluating transcriptome quality. This high rate is consistent with PacBio full-length transcriptome studies of other fish species, such as Japanese flounder (*Paralichthys olivaceus*) (90.08%)²⁵, and white cloud mountain minnow (*Tanichthys albonubes*) (97.72%)²⁶. Furthermore, it is significantly higher than the 53.64% annotation rate of *S. hollandi* transcriptomes assembled using second-generation sequencing²⁷. This discrepancy likely stems from the severe sequence fragmentation inherent to short-read technologies, which impedes the effective annotation of numerous transcripts. BLAST alignment showed the highest sequence similarity to cyprinid fishes (e.g., *Sinocyclocheilus rhinoceros* and *Cyprinus carpio*), confirming the correct assembly and annotation of our transcript library.

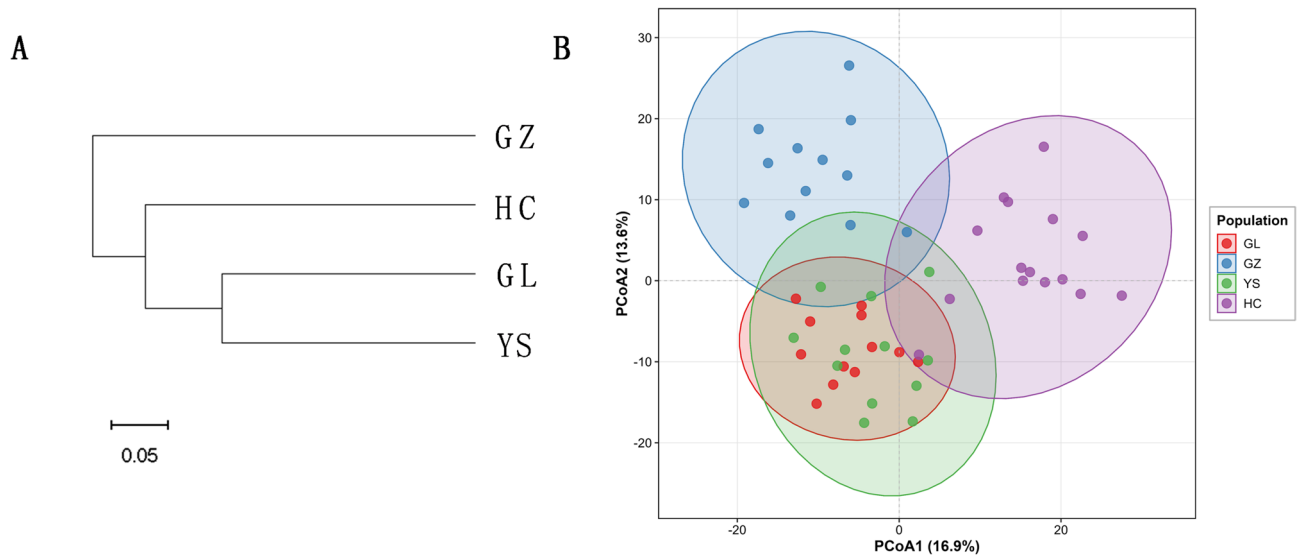


Fig. 7. Genetic structure of four *S. hollandi* populations. **(A)** UPGMA dendrogram based on Nei's genetic distance. **(B)** Principal coordinate analysis (PCoA) plot based on SSR genotype data.

Gene structure features

Regarding gene structure, We identified 373 AS events, with RI being the most common type and MX the rarest—consistent with findings in *Megalobrama terminalis*²⁸. This pattern differs from that of other fish species; for instance, exon skipping (SE) is the most prevalent AS type (34.2%) in *Danio rerio*²⁹. RI can generate distinct mRNA isoforms by retaining introns without altering the coding region, thereby increasing proteome diversity³⁰. *S. hollandi* may extensively employ this mechanism for rapid and precise post-transcriptional regulation of gene expression. Future functional experiments should validate the roles of key RI events to gain a more comprehensive understanding of the biological characteristics of this species.

lncRNAs, which are longer than 200 nt and do not encode proteins, play critical roles in transcriptional control, post-transcriptional processing, and epigenetic regulation³¹. In fish, lncRNAs mediate immune responses in *Oncorhynchus mykiss* during adaptation to plant-based diets³², regulate skeletal muscle development in *Sparus aurata*³³, and the lncRNA DMRT2-AS modulates dmrt2 expression to influence sex differentiation in half-smooth tongue sole (*Cynoglossus semilaevis*)³⁴. The proportion of transcripts classified as lncRNAs in this study was 10.24%, which differs from values reported in PacBio-based studies of other fish species: 3.16% in *Ictiobus cyprinellus*³⁵, and 1.36% in *Clarias gariepinus*³⁶. Although inherent differences in genome size and non-coding region proportion influence lncRNA transcript ratios, the relatively high proportion of transcripts classified as lncRNAs (10.24%) in *S. hollandi* highlights a substantial non-coding component of its transcriptome. The high proportion of lncRNAs (10.24%) suggests a substantial non-coding regulatory component in the *S. hollandi* transcriptome, the specific roles of which in processes like growth and sexual maturation remain to be elucidated.

SSR marker development and utility

SSRs can regulate gene expression and play crucial roles in environmental adaptation and genetic variation underlying adaptive evolution³⁷. The frequency of SSR loci observed in this study was 30.96%, significantly higher than the SSR development efficiency of Illumina sequencing—for example, 6.16% in *Harpodon nehereus*³⁸ and 1.49% in *Pampus argenteus*³⁹. This indicates that long-read sequencing captures SSRs more effectively in genomic regions difficult to assemble with short-read technologies, further confirming the reliability of the data in this study. We detected 7,449 SSR loci, with AC/GT being the most abundant dinucleotide motif, followed by AG/CT and AT/AT; CG/CG was the least abundant. This pattern is consistent with the SSR in the transcriptomes of other aquatic organisms, such as *Siniperca chuatsi*⁴⁰ and *Cyprinus carpio*⁴¹. Previous studies have shown that GC-rich repeat types are rare in vertebrate genomes; the low number of CG SSR may be due to cytosine methylation and deamination converting C to T⁴².

SSRs with lengths greater than or equal to 20 bp generally exhibit high polymorphism, those between 12 and 20 bp show moderate polymorphism, while those shorter than 12 bp possess low polymorphism^{43,44}. In our study, SSRs longer than 20 bp accounted for 33.37% of the total. Among these, 2,238 loci consisted of low-order repeat (di- and tri-nucleotides). These SSR loci, characterized by high repeat numbers and longer sequences, demonstrated considerable polymorphism and may play a significant role in generating genetic variation for adaptive evolution. For instance, SSRs have been shown to regulate gene expression in the *Litopenaeus vannamei* contributing to its environmental adaptation⁴⁵.

Among the 50 pairs of primers designed, 13 pairs of SSR markers were successfully and stably amplified across all four geographic populations, yielding a success rate of 21.7%. The failure of amplification for some primers may be attributed to their target regions containing relatively large introns. According to the classification standard established by Botstein et al.⁴⁶, loci with a PIC > 0.5 are considered highly polymorphic,

those between 0.25 and 0.5 as moderately polymorphic, and those below 0.25 as lowly polymorphic. The average PIC value of the 13 polymorphic loci screened in this study reached 0.651, all of which belong to the highly polymorphic category. This indicates that the developed markers possess rich polymorphism within the tested germplasm and also suggests that *S. hollandi* may inherently harbor high genetic diversity and strong potential for environmental adaptation.

Based on these 13 highly polymorphic SSR markers, we conducted an analysis of the genetic structure of the four geographic populations. Both UPGMA dendrogram and PCoA analyses consistently revealed that the YS and GL populations, both belonging to the Pearl River system, are genetically the closest and cluster together first. In contrast, the GZ population from the Yangtze River system and the HC population from the Pearl River system exhibit significant genetic differentiation. This genetic pattern closely corresponds to the natural geographical isolation between the two major river systems in southern China—the Yangtze River and the Pearl River. The Nanling-Wuyi Mountain Range, situated between these two basins, has long served as a natural barrier, effectively blocking fish migration and gene flow between the watersheds and thereby driving genetic differentiation among populations from different river systems^{47,48}. Within the Pearl River system, the high genetic similarity between the YS and GL populations may be attributed to historical or existing waterway connectivity that maintains gene flow. Although the HC population also belongs to the Pearl River basin, it displays a certain degree of genetic uniqueness, implying that it may have undergone local adaptation or been subject to strong genetic drift.

The strategy employed in this study for efficiently developing SSR markers using PacBio SMRT long-read sequencing has also been successfully validated in other aquatic species. For instance, in *Acanthopagrus latus*, this technology enabled the one-time development of a large number of SSR markers, with an average PIC as high as 0.851, successfully elucidating the genetic structure of different geographic groups⁴⁹. In *Scomberomorus commerson*, not only were 32 highly polymorphic loci (PIC > 0.5) efficiently obtained, but a highly efficient multiplex PCR system was also established, revealing subtle yet significant genetic differentiation among populations⁵⁰. These cases collectively demonstrate the reliability and advantages of PacBio SMRT sequencing in developing high-quality SSR markers.

In conclusion, this study established the first multi-tissue full-length transcriptome for *S. hollandi* using PacBio SMRT sequencing. This comprehensive dataset enabled detailed functional annotation, alternative splicing analysis, and the development of SSR markers, with validation of 13 highly polymorphic loci providing a valuable resource for population genetic studies. Importantly, these high-quality full-length transcripts, when integrated with short-read sequencing data, offer a robust foundation for future genome annotation in *S. hollandi*. This work thereby establishes essential molecular resources for elucidating the genetic mechanisms underlying key biological traits in this species.

Limitations of the study

Although the multi-tissue pooled sample enabled the construction of a comprehensive transcript catalog, it does not allow for the resolution of gene expression with tissue specificity. The developed SSR markers, while validated as effective across four geographical populations, require further testing for their utility in broader populations or closely related species. PacBio long-read sequencing offers advantages in obtaining full-length transcripts, but the capture of genes with extremely low expression levels may still be incomplete. Regarding sample size and tissue representation: this study utilized only six adult individuals and constructed a pooled library from six tissues (heart, gill, brain, fin, liver, and gonad). The limited sample number and the absence of coverage across all developmental stages or physiological conditions may have resulted in incomplete capture of the full transcriptional diversity across all biological processes in this species.

Data availability

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at the Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra>) of NCBI [accession: PRJNA915485].

Received: 18 October 2025; Accepted: 13 January 2026

Published online: 17 January 2026

References

1. Yang, S. D., Lin, T. S., Liou, C. H. & Peng, H. K. Influence of dietary protein levels on growth performance, carcass composition and liver lipid classes of juvenile *Spinibarbus hollandi* (Oshima): dietary protein and *Spinibarbus hollandi*. *Aquac Res.* **34** (8), 661–666 (2003).
2. Chen, Y. M. et al. Analysis of the flesh content and muscle nutritional components of *Spinibarbus hollandi*. *J. Hydrol.* **21** (2), 22–24 (2001). (In Chinese with English Abstract).
3. Cai, Z. D. et al. Study on the reproductive biology of *Spinibarbus hollandi*. *J. South. Agric.* **38** (2), 200–204 (2007). (In Chinese with English Abstract).
4. Yang, Y. et al. Association between expression levels and growth trait-related SNPs located in promoters of the MC4R and MSTN genes in *Spinibarbus hollandi*. *Genes Genom.* **40** (11), 1119–1125 (2018).
5. Luo, K. Biological Characteristics and Germplasm Resource Evaluation of *Spinibarbus hollandi* (Doctoral dissertation, Master's Thesis Guiyang: Guizhou University) (2008).
6. Yang, Y. et al. Effect of fasting and subsequent refeeding on the transcriptional profiles of brain in juvenile *Spinibarbus hollandi*. *PLoS One.* **14** (3), e0214589 (2019).
7. Yang, Y., Zhou, H., Hou, L., Xing, K. & Shu, H. Transcriptional profiling of skeletal muscle reveals starvation response and compensatory growth in *Spinibarbus hollandi*. *BMC Genom.* **20** (1), 938 (2019).
8. Han, C. et al. Screening and characterization of sex-specific markers by NGS sequencing in *Spinibarbus hollandi* with implication of XY sex determination system. *Aquaculture* **565**, 739147 (2023).

9. Han, C. et al. De Novo Assembly, characterization and comparative transcriptome analysis of the mature gonads in *Spinibarbus hollandi*. *Animals* **13** (1), 166 (2023).
10. Chen, I. S., Bai, J. C. & Fang, H. C. C. The complete mitochondrial genome of holland's spinibarbel *Spinibarbus hollandi* Oshima (Cypriniformes, Cyprinidae). *MITOCHONDRIAL DNA A*. **27** (3), 1721–1723 (2016).
11. Minoche, A. E. et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol.* **16**, 184 (2015).
12. Wang, B., Kumar, V., Olson, A. & Ware, D. Reviving the transcriptome studies: an insight into the emergence of Single-Molecule transcriptome sequencing. *Front. Genet.* **10**, 384 (2019).
13. Arzalluz-Luque, A. & Conesa, A. Single-cell RNAseq for the study of isoforms-how is that possible? *Genome Biol.* **19**, 110 (2018).
14. Tian, Y. et al. Characterization of Full-Length transcriptome sequences and splice variants of lateolabrax maculatus by Single-Molecule Long-Read sequencing and their involvement in salinity regulation. *Front. Genet.* **10**, 1126 (2019).
15. Mehjabin, R. et al. Full-Length transcriptome sequencing and the discovery of new transcripts in the unfertilized eggs of zebrafish (*Danio rerio*). *G3-GENES GENOM GENET.* **9** (6), 1831–1838 (2019).
16. Liu, S. et al. Transcriptome sequencing and SSR prediction of clematis calyx based on SMRT sequencing platform. *Sci. Rep.* **14** (1), 28949 (2024).
17. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *NUCLEIC ACIDS RES.* **53** (1), 672–677 (2025).
18. Kang, Y. J. et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45** (W1), W12–W16 (2017).
19. Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19** (1), 40 (2018).
20. Eid, J. et al. Real-Time DNA sequencing from single polymerase molecules. *Sci* **323** (5910), 133–138 (2009).
21. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *GPB* **13** (5), 278–289 (2015).
22. Wang, B. et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7** (1), 11708 (2016).
23. Yi, L. et al. Insights into cryptic diversity and adaptive evolution of the clam *Coelomastra antiquata* (Spengler, 1802) from comparative transcriptomics. *Mar. Biodiv.* **49** (5), 2311–2322 (2019).
24. Deng, A. et al. SMRT sequencing of the Full-Length transcriptome of the *Coelomastra antiquata*. *Front. Genet.* **12**, 741243 (2020).
25. Xiu, Y., Li, Y., Liu, X. & Li, C. Full-length transcriptome sequencing from multiple immune-related tissues of *Paralichthys Olivaceus*. *FISH. SHELLFISH IMMUN.* **106**, 930–937 (2020).
26. Li, C. et al. Full-length transcriptome data for the white cloud mountain minnow (*Tanichthys albonubes*) from a wild population based on isoform sequencing. *Front. Mar. Sci.* **9**, 831148; (2022). <https://doi.org/10.3389/fmars.831148> (2022).
27. Li, S. X. et al. Transcriptome sequencing and characteristic analysis of dorsal muscle tissue in spinibarbus hollandi. *Anhui Agric. Sci. Bull.* **31**, 63–68 (2025).
28. Lu, J., Peatman, E., Wang, W. & Yang, Q. Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons. *Mol. Genet. Genomics.* **283**, 531–539 (2010).
29. Liu, K. & Xie, N. Full-length transcriptome assembly of black Amur Bream (*Megalobrama terminalis*) as a reference resource. *Mol. Biol. Rep.* **51**, 0301–4851 (2024).
30. Ner-Gaon, H. et al. Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant. J.* **39**, 877–885 (2004).
31. Novikova, I. V., Hennelly, S. P., Tung, C. S. & Sanbonmatsu, K. Y. Rise of the RNA machines: exploring the structure of long non-coding RNAs. *J. Mol. Biol.* **425**, 3731–3746 (2013).
32. Abernathy, J. & Overturf, K. Expression of antisense long noncoding RNAs as potential regulators in rainbow trout with different tolerance to plant-based diets. *Anim. Biotechnol.* **30**, 87–94 (2019).
33. García-Pérez, I. et al. The emerging role of long non-coding RNAs in development and function of Gilthead sea Bream (*Sparus aurata*) fast skeletal muscle. *Cells* **11**, 428–441 (2022).
34. Feng, B. et al. LncRNA DMRT2-AS acts as a transcriptional regulator of dmrt2 involving in sex differentiation in the Chinese tongue sole (*Cynoglossus semilaevis*). *Comp. Biochem. Physiol. B-Biochem Mol. Biol.* **253**, 1096–4959 (2021).
35. Ge, H. et al. Full-length transcriptome sequencing from the longest-lived freshwater bony fish of the world: Bigmouth Buffalo (*Ictiobus cyprinellus*). *Front. Mar. Sci.* **8**, 736188. <https://doi.org/10.3389/fmars.2021.736188> (2021).
36. Zhang, F. et al. Construction of a full-length transcriptome resource for the African Sharptooth catfish (*Clarias gariepinus*), a prototypical air-breathing fish, based on isoform sequencing (Iso-Seq). *Gene* **930**, 0378–1119 (2024).
37. Dang, Z. et al. Identification of genic SSRs provide a perspective for studying environmental adaptation in the endemic shrub *Tetraena Mongolica*. *Genes* **11**, 322. <https://doi.org/10.3390/genes11030322> (2020).
38. Huang, X. X. et al. Analysis of microsatellite markers in Harpadon Nehereus based on transcriptome sequencing illumina HiseqTM 2500. *J Zhejiang Univ. -Sc A.* **40**, 189–195 (2021). (In Chinese with English Abstract).
39. Liu, L. et al. Bioinformatic analysis of SSR markers based on RNA-seq of *Pampus Argenteus*. *J. Anhui Agric. Sci.* **44**, 102–105 (2016).
40. Sun, H. Y. et al. Transcriptome sequencing and development and application of novel SSR markers for *Siniperca Chuatsi*. *GAB* **38**, 4413–4421 (2019).
41. Wang, D., Liao, X., Cheng, L., Yu, X. & Tong, J. Development of novel EST-SSR markers in common carp by data mining from public EST sequences. *Aquaculture* **271**, 558–574 (2007).
42. Schorderet, D. F. & Gartler, S. M. Analysis of CpG suppression in methylated and nonmethylated species. *Proc. Natl. Acad. Sci. U S A.* **89**, 957–961 (1992).
43. Wang, H. et al. Genome-wide characterization and comparative analyses of simple sequence repeats among four miniature pig breeds. *Animals* **10**, 1792. <https://doi.org/10.3390/ani10101792> (2020).
44. Xiao, T. et al. Screening of SSR markers associated with scale cover pattern and mapped to a genetic linkage map of common carp (*Cyprinus Carpio L.*). *J. Appl. Genet.* **56**, 261–269 (2015).
45. Yuan, J. et al. Simple sequence repeats drive genome plasticity and promote adaptive evolution in Penaeid shrimp. *Commun. Biol.* **4** <https://doi.org/10.1038/s42003-021-01716-y> (2021).
46. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
47. Yang, J. Q. & Liu, H. Z. Analysis of Cytb gene sequence variability in two species of Bagridae from the Yangtze river and Pearl river basins. *Acta Hydrobiol. Sin.* **27** (3), 253–257. <https://doi.org/10.12131/20220007> (2003). (In Chinese with English Abstract).
48. Chen, W. T. et al. Multigene analysis of genetic structure of ochetobius elongatus in the Yangtze river and Pearl river. *South. China Fisheries Sci.* **18** (6), 19–25 (2022). (In Chinese with English Abstract).
49. Zhang, J. et al. Full-length transcriptome sequence and SSR marker development for genetic diversity research in Yellowfin seabream acanthopagrus latus. *J. Ocean. Limnol.* **41**, 1073–1083 (2023).
50. Linu, J. et al. Microsatellite marker development in Spanish mackerel *Scomberomorus commerson* using third generation sequencing technology. *Mol. Biol. Rep.* **47**, 10005–10014 (2020).

Author contributions

Sixun Li: Formal analysis, Methodology, Data curation, Project administration, Writing-original draft. Jie Lai: Formal analysis, Methodology, Data curation, Validation. Mengmeng Wu: Formal analysis, Software. Zihang

Xie: Methodology, Software. Min Zhang: Validation. Guojun Cai: Validation, Investigation. Ziyang Deng: Data curation, Software. Binhua Deng: Visualization and Validation. Qiang Li, Chong Han: Conceptualization, Supervision, Writing-review and editing.

Funding

Characteristics Innovation program of Guangdong Provincial Department of Education (2025KTSCX092) and college student Innovation and Entrepreneurship Training program of Guangzhou University (S202411078013; 202511078046).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-36468-4>.

Correspondence and requests for materials should be addressed to C.H. or Q.L.

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

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

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

© The Author(s) 2026