

Article

# Transcriptome Analysis of the Tadpole Shrimp (*Triops longicaudatus*) by Illumina Paired-End Sequencing: Assembly, Annotation, and Marker Discovery

Jiyeon Seong<sup>1,†</sup>, Se Won Kang<sup>2,†</sup>, Bharat Bhusan Patnaik<sup>2,3</sup>, So Young Park<sup>4</sup>, Hee Ju Hwang<sup>2</sup>, Jong Min Chung<sup>2</sup>, Dae Kwon Song<sup>2</sup>, Mi Young Noh<sup>5</sup>, Seung-Hwan Park<sup>6</sup>, Gwang Joo Jeon<sup>1</sup>, Hong Sik Kong<sup>1</sup>, Soonok Kim<sup>7</sup>, Ui Wook Hwang<sup>8</sup>, Hong Seog Park<sup>9</sup>, Yeon Soo Han<sup>10</sup> and Yong Seok Lee<sup>2,\*</sup>

<sup>1</sup> Genomic Informatics Center, Hankyong National University, 327 Chungang-no, Anseong-si, Gyeonggi-do 17579, Korea; s-jiyeon@hanmail.net (J.S.); jeon5894@gmail.com (G.J.J.); kebinkhs@empal.com (H.S.K.)

<sup>2</sup> Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, 22 Soonchunhyangro, Shinchang-myeon, Asan, Chungchungnam-do 31538, Korea; bioksw@naver.com (S.W.K.); drbharatbhusan4@gmail.com (B.B.P.); hwamux@naver.com (H.J.H.); jong6922@daum.net (J.M.C.); elegangce@naver.com (D.K.S.)

<sup>3</sup> Trident School of Biotech Sciences, Trident Academy of Creative Technology (TACT), Chandaka Industrial Estate, Chandrasekharapur, Bhubaneswar, Odisha 751024, India

<sup>4</sup> Biodiversity Conservation & Change Research Division, Nakdonggang National Institute of Biological Resources, 137 Donam 2-gil, Sangju, Gyeongsangbuk-do 37242, Korea; cindysory@naver.com

<sup>5</sup> Department of Applied Biology, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Korea; annemi@chonnam.ac.kr

<sup>6</sup> Biological Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), 181 Ipsin-gil, Jeongeup-si, Jeollabuk-do 56212, Korea; biopark@kribb.re.kr

<sup>7</sup> National Institute of Biological Resources, 42, Hwangyeong-ro, Seo-gu, Incheon 22689, Korea; sokim90@korea.kr

<sup>8</sup> Department of Biology Education, Kyungpook National University, 80 Daehakro, Bukgu, Daegu 41566, Korea; uwhwang@knu.ac.kr

<sup>9</sup> Research Institute, GnC BIO Co., LTD., 36 Banseokro, Yuseong-gu, Daejeon 34069, Korea; 5022daniel@gmail.com

<sup>10</sup> College of Agriculture and Life Science, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Korea; hanys@chonnam.ac.kr

\* Correspondence: yslee@sch.ac.kr; Tel.: +82-41-530-3040; Fax: +82-41-530-1256

† These authors contributed equally to this work.

Academic Editor: Paolo Cinelli

Received: 12 September 2016; Accepted: 24 November 2016; Published: 2 December 2016

**Abstract:** The tadpole shrimp (*Triops longicaudatus*) is an aquatic crustacean that helps control pest populations. It inhabits freshwater ponds and pools and has been described as a living fossil. *T. longicaudatus* was officially declared an endangered species South Korea in 2005; however, through subsequent protection and conservation management, it was removed from the endangered species list in 2012. The limited number of available genetic resources on *T. longicaudatus* makes it difficult to obtain valuable genetic information for marker-aided selection programs. In this study, whole-transcriptome sequencing of *T. longicaudatus* generated 39.74 GB of clean data and a total of 269,822 contigs using the Illumina HiSeq 2500 platform. After clustering, a total of 208,813 unigenes with an N<sub>50</sub> length of 1089 bp were generated. A total of 95,105 unigenes were successfully annotated against Protostome (PANM), Unigene, Eukaryotic Orthologous Groups (KOG), Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using BLASTX with a cut-off of 1E−5. A total of 57,731 unigenes were assigned to GO terms, and 7247 unigenes were mapped to 129 KEGG pathways. Furthermore, 1595 simple sequence repeats (SSRs) were detected

from the unigenes with 1387 potential SSR markers. This is the first report of high-throughput transcriptome analysis of *T. longicaudatus*, and it provides valuable insights for genetic research and molecular-assisted breeding of this important species.

**Keywords:** *Triops longicaudatus*; tadpole shrimp; transcriptome; Illumina sequencing; SSRs (simple sequence repeats)

---

## 1. Introduction

The tadpole shrimp, *Triops* spp. (order: Notostraca; class: Branchiopoda) is a crustacean that inhabits freshwater, ephemeral ponds in arid regions worldwide [1], and it has been described as a living fossil from the late Cretaceous period similar to other members of this ancient crustacean order. This is allegedly due to their virtually unchanged morphology during an evolutionary time scale spanning more than 70 million years [2,3]. This includes the ability to control the size of mosquito populations by consuming *Culex* larvae [4,5], and its utilization as a biological agent to control weeds in paddy fields [6]. The diversification of cryptic species within the genus occurred more recently than this, based on the subtle differences in genetic composition and morphology [7–9]. *Triops longicaudatus* is the most widespread notostracan crustacean, being found in North America, South America, the Caribbean, Saudi Arabia, Japan, and the Pacific Islands [10–16]. There are a number of reports on its distribution morphology, and reproduction [2,10,11,15,17]. In South Korea, the species has been reported since 1986, where it was collected from paddy fields in the cities of Changnyeong and Samcheonpo (Gyeongsangnam-Do Province) [13]. It was registered as an endangered species in South Korea by the Ministry of Environment in 2004. Since then, populations of *T. longicaudatus* have increased through regional conservation measures and it was removed from the endangered species list in 2012. *T. longicaudatus* is economically important species to be used for environmental friendly agriculture. It is proposed that genetic studies involving genome, transcriptome, and gene function analysis will be necessary to preserve the genotypes of this species by assisting in determining their developmental and regulatory functions. Furthermore, the elucidation of cDNA simple sequence repeat (SSR) markers in the putative coding transcripts will be necessary to assess population genetic structure and diversity.

Among the limited number of genomic resources on *T. longicaudatus*, only the mitochondrial DNA sequence is known [18]. The variation in mitochondrial genes has been successfully utilized to identify cryptic lineages of the genus *Triops* [9]. Despite these studies, genetic and genomic information on the species is limited due to the lack of whole genome sequencing, RNA sequencing, expression profiles of transcripts, and microsatellite markers. The traditional method of expressed sequence tag (EST) construction using Sanger sequencing is time consuming and inefficient, producing at best 10,000 sequences. This is likely an insufficient representation of the size of the genome and thus a major limitation of functional research applications [19]. High-throughput next-generation sequencing (NGS) technologies, such as 454 (Roche), Solexa/Illumina (Illumina), and SOLiD (ABI), collect massive amounts of sequencing data in a single run with increased efficiency at an affordable level [20,21]. This technology has enabled genome and transcriptome-level computational analyses [22], leading to the discovery of molecular markers such as SSRs, single nucleotide polymorphisms (SNPs), and quantitative trait loci (QTL) [23]. Because genome sequences are currently unavailable or unreliable in many non-model species, transcriptome sequencing provides direct relevance to the genetic level by measuring the expression of relevant traits [24–26]. Among the NGS technologies, Illumina sequencing is a preferred choice due to the generation of short-read sequences with greater coverage [27–30].

Over the last four to five years, significant progress has been made in characterizing the transcriptome of economically important crustacean species such as *Litopenaeus vannamei*, *Fenneropenaeus chinensis*, *Eriocheir sinensis*, *Macrobrachium nipponense*, *Portunus trituberculatus*, and

*Carcinus maenas*. These analyses have provided insights into species biology, the functional regulation of defense signaling pathways, growth and reproduction, and strategies to improve culture productivity. In this study, we present the first massive sequencing data for the tadpole shrimp, *T. longicaudatus*, using the Illumina HiSeq 2500 NGS platform. The assembled and annotated sequencing data were utilized for the large-scale identification of putative functional transcripts. Furthermore, the identification and analysis of SSR loci and SSR markers in the transcriptome will be useful for population genomics and variability studies, further assisting in the marker assisted selection breeding of *T. longicaudatus*.

## 2. Materials and Methods

### 2.1. Ethics Statement

The experiments in this study were performed in accordance with relevant national and international guidelines. Because *T. longicaudatus* is not an endangered or protected species, in Korea, sample collection did not require special permits. Our project was approved by the National Institute of Biological Resources (NIBR), Korea.

### 2.2. Sample Collection and RNA Isolation

Ten individual *T. longicaudatus* were collected from the Metropolitan City (Palgongsan), Gooam, Dong-gu, Daego-si, South Korea, on 10 June 2014.

The adult whole-body tissues of hermaphrodite *T. longicaudatus* ( $n = 10$ ) were pooled and total RNAs were extracted using Trizol reagent (Invitrogen, Carlsbad, CA, USA) and extracted in accordance with the manufacturer's protocol. The extracted RNA was treated with RNase-free DNase I (Qiagen) to remove the genomic DNA. RNA purity and concentration were measured using a Nanodrop-2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA) measures RNA quantity and agarose gel electrophoresis. Total RNAs were pooled, purified to obtain mRNA using oligo (dT) magnetic beads, and subsequently fragmented using an RNA Fragmentation Kit (Ambion, Austin, TX, USA).

### 2.3. cDNA Synthesis and HiSeq 2500 Sequencing

First-strand cDNA synthesis was performed using reverse-transcriptase (Invitrogen) and random hexamer-primers. Second-strand cDNA was synthesized using RNase H (Invitrogen) and DNA polymerase I (New England BioLabs, Ipswich, MA, USA). The double-stranded cDNA was end-repaired using T4 DNA polymerase, the Klenow fragment (New England BioLabs), and T4 polynucleotide kinase (New England BioLabs). The end-repaired cDNA fragments were ligated to the PE (paired-end) Adapter Oligo Mix with T4 DNA ligase (New England BioLabs) at room temperature for 15 min. The ligated products were purified and separated by size on a 2% agarose gel. DNA fragments of the desired size ( $200 \pm 25$  bp) were excised and sequenced on the Illumina HiSeq 2500 sequencing platform with  $2 \times 126$  bp after validation.

### 2.4. De Novo Assembly and Assessment of De Novo Assemblies

Before de novo transcriptome assembly, the raw reads were cleaned by removing adaptor-only reads (nt length of the recognized adaptor  $\leq 13$  and the remaining adaptor-excluded nt length of  $\leq 35$ ), repeated reads, and low-quality reads (Phred quality score  $\leq 20$ ) using Sickle (<http://github.com/najoshi/sickle>) [31] and Cutadapt (<http://cutadapt.readthedocs.io/en/stable/index.html>) [32]. High-quality reads were assembled using Trinity software (software version 2013-02-25) with default parameters (100 GB of memory, path reinforcement distance of 50, and minimum allowed length of 200 bp) [33]. The Trinity program assembles reads of a certain length that overlap to form longer fragments without gaps; these are called contigs. The total number of contigs, as well as the mean length, the  $N_{50}$  length, and GC% were recorded. The contigs were further assembled into sequences that could not be extended at either end; these are called unigenes (having 94% identity, 30 bp

overlap) [34,35]. Such unigenes were subjected to annotation analysis against public protein and nucleotide databases. The assessment of the assembly and annotation completeness we applied the software tool BUSCO (software version 1.1b) [36].

### 2.5. Transcriptome Annotation

For functional annotation, all transcripts were handled as queries and annotated against subject sequences in the Protostome database (PANM-DB) [37], as well as the Unigene and Eukaryotic Orthologous Groups (KOG) databases using the BLASTX and BLASTN programs, with a significant threshold *E*-value of <0.00001 [38]. Gene Ontology (GO) (<http://www.geneontology.org>) terms were analyzed using BLAST2GO (<http://www.blast2go.org/>) software, professional version [39]. Subsequently, the GO functional classification of the annotated transcripts was performed using WEGO software (<http://wego.genomics.org.cn/cgi-bin/wego/index.pl>) [40]. Kyoto Encyclopedia of Genes and Genomes (KEGG) (<http://www.genome.jp/kegg/>) metabolic pathway analysis was determined by searching against the KEGG database with an *E*-value cutoff of  $1E-5$  [41].

### 2.6. Identification of SSRs

SSRs in the unigenes (length > 1000 bp) of *T. longicaudatus* were identified using the Perl script program MicroSATellite (MISA) (<http://pgrc.ipk-gatersleben.de/misa/>). Search criteria included the number of repetitions for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides. The minimum repeats were as follows: six for dinucleotides, five for trinucleotides, and four for tetra-, penta-, and hexa-nucleotides. Primers for each SSR were designed using the BatchPrimer3 (<http://wheat.pw.usda.gov/demos/BatchPrimer3/>) program.

## 3. Results

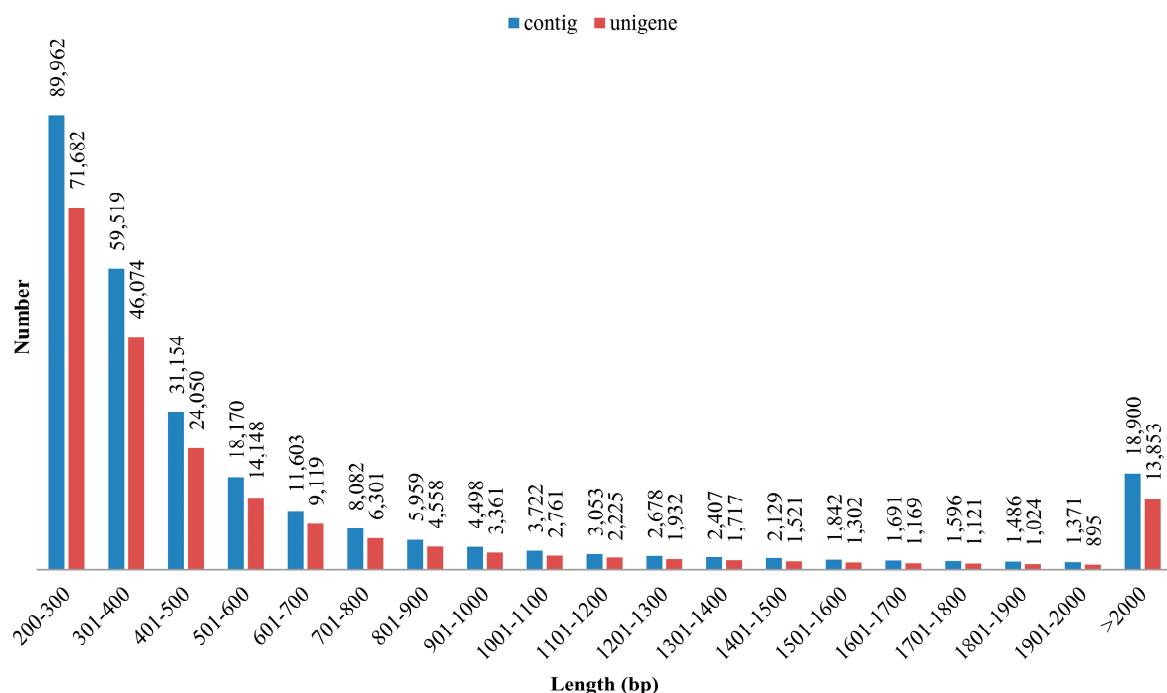
### 3.1. Illumina Reads and Sequence Assembly

Transcriptome information for the *T. longicaudatus* was characterized from adult whole-body. The Illumina HiSeq 2500 platform generated a total of 323,319,608 paired-end reads (40,738,270,608 bases) were generated with a read length of 100 bp. All raw sequencing data were deposited into the NCBI Sequence Read Archive (SRA) under accession number SRR3961747. After adaptor trimming, a total of 318,610,596 clean sequencing reads (98.54%) were filtered, which were used for further analysis. The mean length, the  $N_{50}$  length, and GC% of the obtained clean reads were 124.7 bp, 126 bp, and 48.39%, respectively.

Because the reference genome sequence is unavailable, de novo assembly of the transcriptome was performed. Trinity assembly with default parameters was used to resolve the clean transcripts to overlapping contiguous sequences. De novo assembly of the high-quality sequences generated a total of 269,822 contigs (192,327,026) with a mean length of 712.8 bp and an  $N_{50}$  length of 1148 bp. Of the total assembled contigs, 89,407 were  $\geq 500$  bp, with the longest contig size of 40,450 bp. The clustering of the contigs generated 208,813 unigenes with a mean length of 700 bp and an  $N_{50}$  length of 1089 bp. The lengths of the unigenes varied from 224 bp to 40,450 bp. Table 1 summarizes the transcriptome sequencing, de novo assembly, and clustering of contigs. Among the unigenes, 85.86%, 7.50%, and 6.63% showed lengths of 200–1000 bp, 1001–2000 bp, and >2000 bp, respectively. The size distribution of the contigs and unigenes are shown in Figure 1. The unigenes represent a comprehensive resource of functional information on the *T. longicaudatus* genome and may facilitate the discovery of relevant phenotypes in this species.

**Table 1.** Summary statistics from Illumina sequencing of the *T. longicaudatus*.

Total Number of Raw Reads	
Number of sequences	323,319,608
Number of bases	40,738,270,608
Total number of clean reads	
Number of sequences	318,610,596
Number of bases	39,745,513,470
Mean length of contig (bp)	124.7
N <sub>50</sub> length of contig (bp)	126
GC % of contig	48.39
High-quality reads (%)	98.54 (sequences), 97.56 (bases)
Contig information	
Total number of contig	269,822
Number of bases	192,327,026
Mean length of contig (bp)	712.8
N <sub>50</sub> length of contig (bp)	1148
GC % of contig	46.82
Largest contig (bp)	40,450
No. of large contigs (≥500 bp)	89,407
Unigene information	
Total number of unigenes	208,813
Number of bases	146,173,633
Mean length of unigene (bp)	700.0
N <sub>50</sub> length of unigene (bp)	1089
GC % of unigene	46.97
Length ranges (bp)	224–40,450

**Figure 1.** Size distribution of contigs (blue) and unigenes (red) after assembly and clustering of the quality reads from the transcriptome of *T. longicaudatus*.

### 3.2. Sequence Annotation of Unigenes

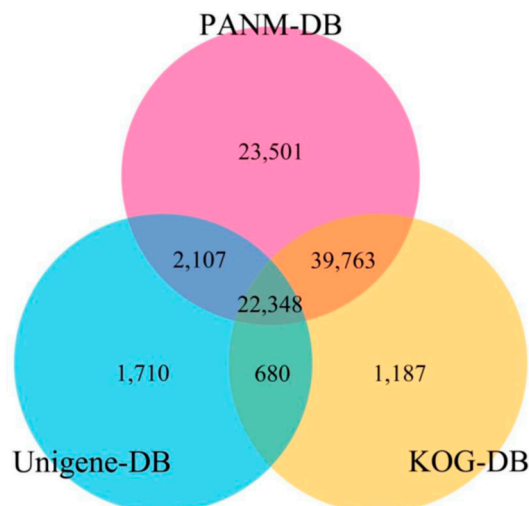
Several public databases comprised of known protein and nucleotide sequences were used as subject databases for the sequence annotation of *T. longicaudatus* unigenes. The unigene sequences (as

queries) were searched to identify homologous sequences using BLASTX and BLASTN ( $E$ -value cut-off of  $1E-5$ ) for protein and nucleotide databases, respectively. The PANM-DB, KOG, GO, and KEGG databases were used as protein databases, while the Unigene database was used as the nucleotide database. Of the total of 208,813 unigenes, 95,105 (45.55%) were annotated to any one of the databases with a great number of unigenes having lengths of 300–1000 bp. The number of matches to PANM-DB was the greatest (87,719 unigenes), followed by the KOG (63,978 unigenes). The annotation results of unigenes to the public databases are shown in Table 2. The results also show that 23,732 (27.1%), 7729 (28.8%), 20,131 (31.55%), 16,663 (28.8%), and 2112 (29.1%) of the unigenes that were over 1000 bp in length had BLAST matches in the PANM, Unigene, KOG, GO, and KEGG databases, respectively. Next, to understand the overlap of the unigene sequence annotations between PANM-DB and Unigene and KOG databases, we constructed a three-way Venn diagram (Figure 2). We found that a maximum number of 39,763 unigenes matched in both PANM-DB and KOG database, and 22,348 unigenes matched in all three databases. The number of unigenes annotated exclusively to PANM-DB, and the Unigene and KOG databases without any overlap were 23,501, 1710, and 1187, respectively.

**Table 2.** Annotation of *T. longicaudatus* assembled unigene sequences against public databases.

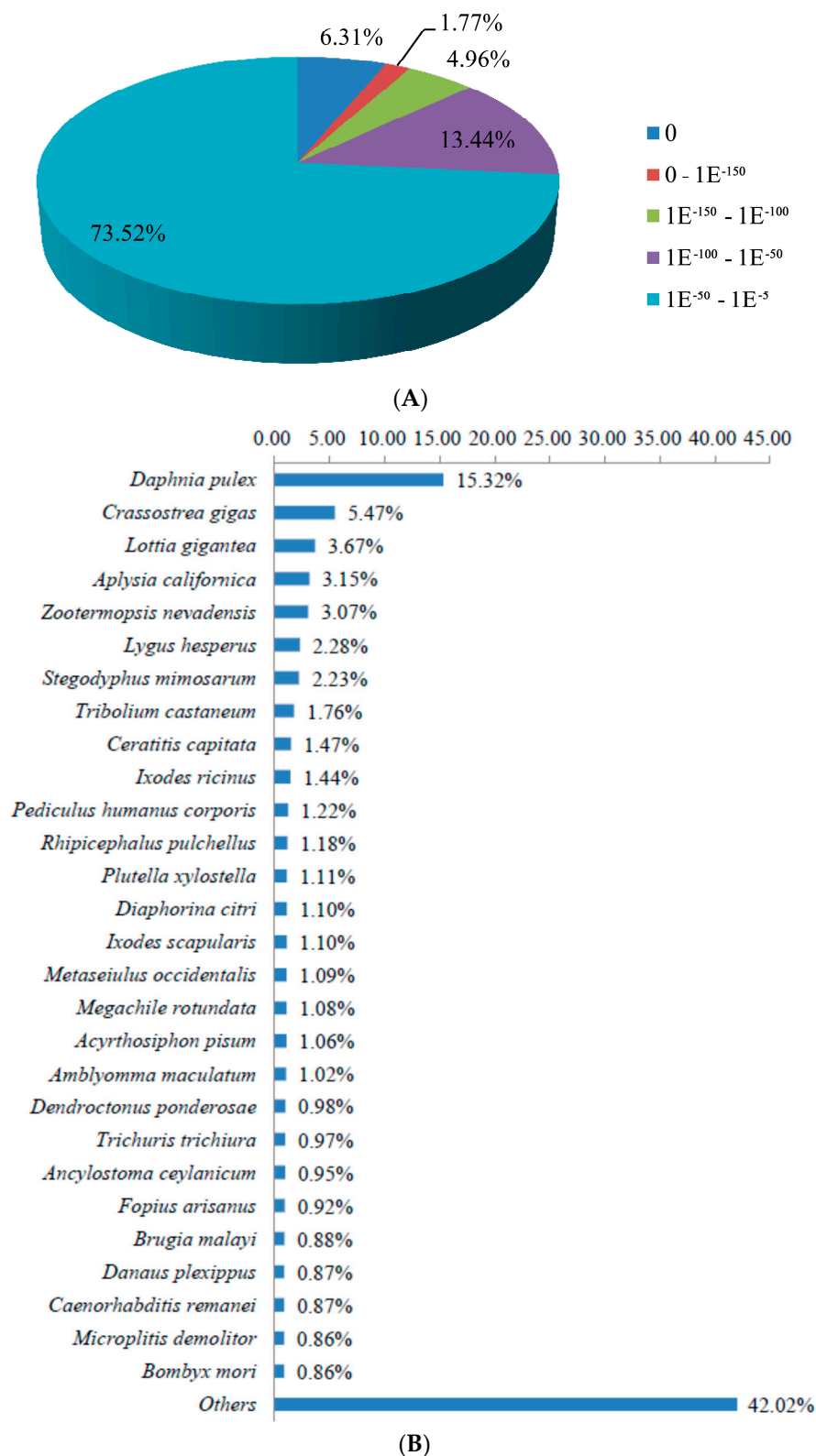
Databases	All	≤300 bp	300–1000 bp	≥1000 bp
PANM-DB	87,719	20,029	43,958	23,732
UNIGENE	26,845	6231	12,885	7729
KOG	63,978	12,955	30,892	20,131
GO	57,731	12,915	28,153	16,663
KEGG	7247	1735	3400	2112
ALL	95,105	22,935	48,081	24,089

The number of unigenes hits using BLASTX search ( $E$ -value <  $1E-5$ ).



**Figure 2.** The sequence annotation profile of *T. longicaudatus* unigenes against PANM-DB, Unigene DB and KOG DB.

The homology search of the unigene sequences of *T. longicaudatus* against PANM-DB using BLASTX was represented of top-hit  $E$ -values and top-hit species distribution. The  $E$ -value distribution revealed that 64,493 (73.52%) unigenes showed significant homology to the deposited sequences, with an  $E$ -values ranging from  $1E-50$  to  $1E-5$  (Figure 3A). For top species distribution, 13,440 (15.32%) unigenes showed similarities with *Daphnia pulex* followed by *Crassostrea gigas* (4794 unigenes; 5.47%), *Lottia gigantea* (3218 unigenes; 3.67%), *Aplysia californica* (2759 unigenes; 3.15%), and others (Figure 3B).



**Figure 3.** Homology searches of *T. longicaudatus* unigenes against the PANM-DB. (A) E-value distribution; (B) Top-hit species distribution.

We also examined homology search characteristics such as score, identity and similarity distribution. The score distribution, which represents the quality of the BLAST alignment, showed that

45,158 (51.48%) unigenes had a score <100 (Figure S1A). The identity distribution revealed that 36,329 (41.42%) unigenes showed an identity of 40%–60%, followed by identities of 33.17% and 19.08% for 15%–40% and 60%–80% unigenes, respectively (Figure S1B). According to the similarity distribution analysis, 36,411 (41.51%) unigenes showed a similarity of 60%–80% with homologous sequences in the PANM-DB. Only 18.03% of unigene sequences showed similarity of 80%–100% to sequences in PANM-DB (Figure S1C). The BLASTX annotation hits to homologous protein sequences in PANM-DB increased with increasing unigene length. More than 90% of unigenes with a sequence length >2000 bp showed annotation hits against PANM-DB (Figure S1D).

### 3.3. KOG, GO and KEGG Classifications

For a functional classification of the *T. longicaudatus* unigenes, we conducted a BLAST search against the KOG, GO, and KEGG databases. Under the KOG classification, a total of 63,978 unigenes were predicted under 25 functional categories excluding the “multi” category. Within the 25 categories, the unigenes were predominantly distributed to “translation, ribosomal structure and biogenesis (7210 unigenes)”, followed by “general function prediction only” (6591 unigenes), “post-translational modification, protein turnover and chaperones” (6005 unigenes), and “signal transduction mechanisms” (5017 unigenes). The least represented groups included “cell motility” (90 unigenes) and “nuclear structure” (103 unigenes) (Figure 4).

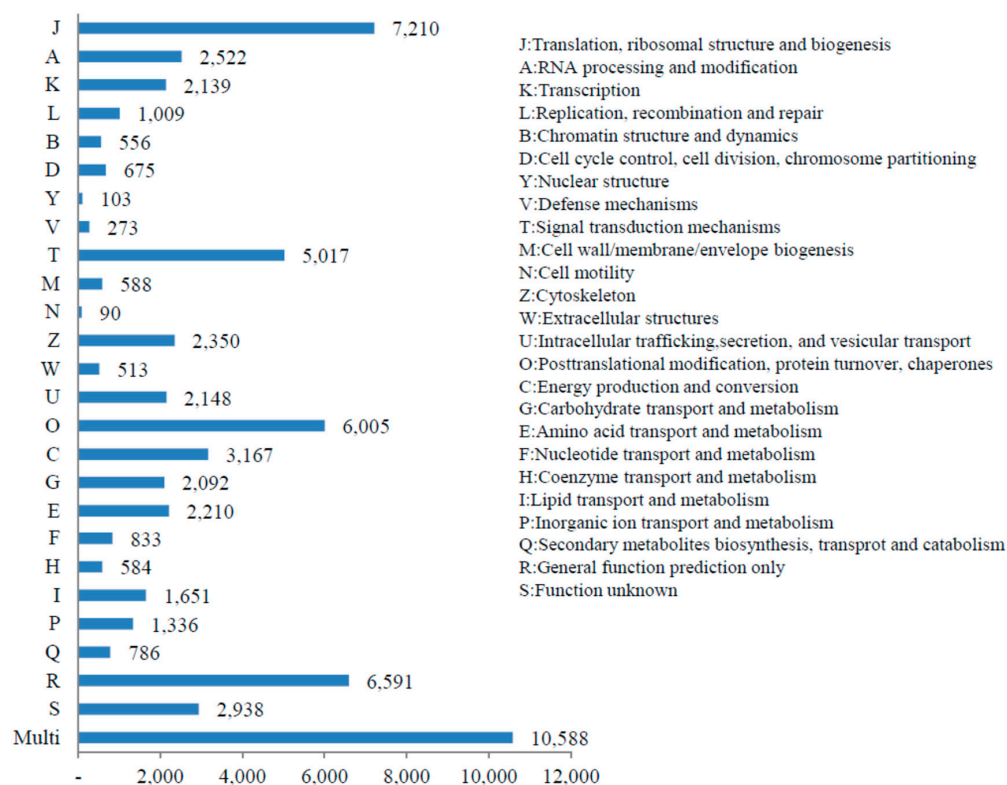
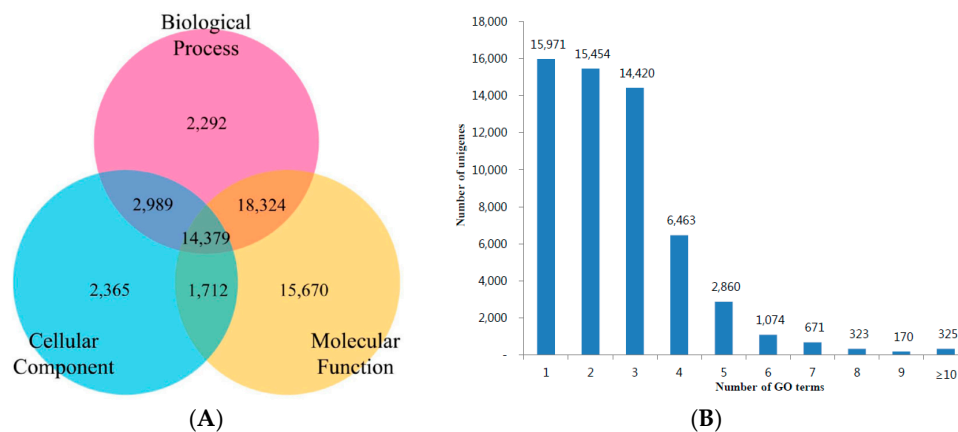


Figure 4. KOG DB based functional analysis of *T. longicaudatus* unigenes.

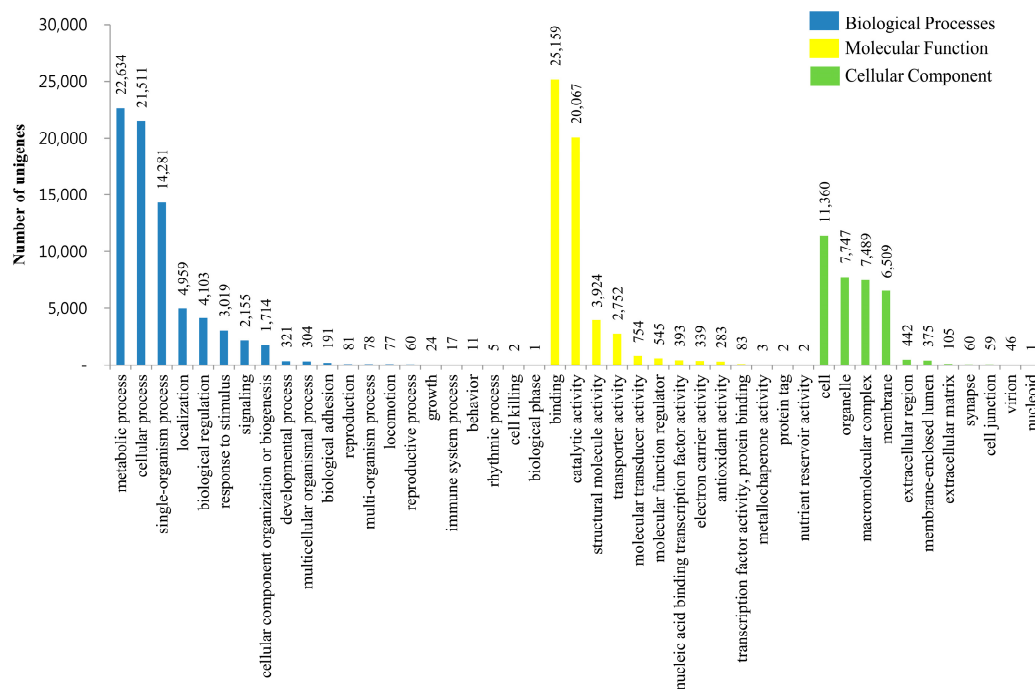
GO is an international standardization of the gene functional classification system. The GO classification system comprises three large categories: molecular function, biological process and cellular components. Among all the unigenes with GO annotations, we found that 57,731 (27.65% of all unigenes) unigenes matched to GO terms and 14,379 unigenes showed functional attributes shared within the three main categories. The unigenes predominantly shared the biological process and molecular function categories (Figure 5A). Approximately 15,971 (27.7%) unigenes were represented by one GO term; 15,454 (26.8%) unigenes were represented by two GO terms; and 14,420 (25.0%) unigenes



were represented by three GO terms of predicted functions (Figure 5B). Additionally, biological processes, molecular functions and cellular components were associated with 75,548, 54,306, and 34,193 unigenes, respectively. In the biological process category, metabolic process (22,634 unigenes), cellular process (21,511 unigenes), and single-organism process (14,281 unigenes) were the most abundant groups, whereas cell killing (2 unigenes) and biological phase (1 unigene) were the least abundant groups. Under the molecular function category, binding (25,159 unigenes) and catalytic activity (20,067 unigenes) were the most abundant groups, while antioxidant activity (283 unigenes) and metallochaperone activity (3 unigenes) were also observed. In cellular component terms, cell (11,360 unigenes), organelle (7,747 unigenes), macromolecular complex (7,489 unigenes), and membrane (6,509 unigenes) were the dominant groups. An account of the suggested function of *T. longicaudatus* unigenes under the GO term categories is shown in Figure 6.

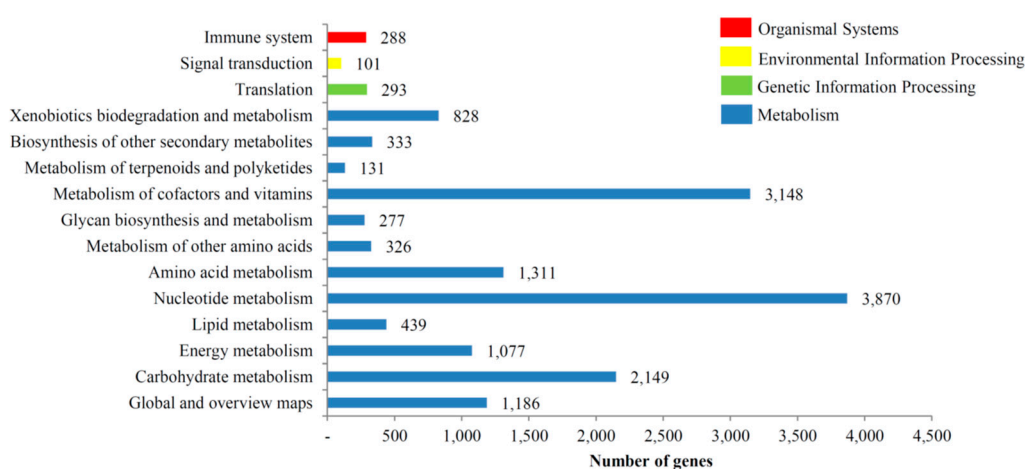


**Figure 5.** GO term classification for *T. longicaudatus*. (A) Predicted functional interpretation of unigenes into represented biological process, cellular component, and molecular function; (B) Number of unigene sequences annotated with numbers of GO terms per sequence.



**Figure 6.** GO annotation of unigenes from *T. longicaudatus* based on biological processes, molecular functions and cellular components.

We classified unigenes into biological pathways by annotating the unigene sequences against the KEGG database. A total of 7247 unigenes were predicted to function in a total of 129 pathways. Predominantly, the unigene sequences were classified into the metabolism pathway group, wherein “nucleotide metabolism”, “metabolism of cofactors and vitamins”, and “carbohydrate metabolism” constituted the major groups (Table S1). A total of 293 unigene sequences were predicted to be classified under translation group, followed by 288 under the immune system and 101 under the signal transduction group. The identified KEGG pathways for *T. longicaudatus* unigenes are presented in Figure 7. Using the InterPro Scan analysis feature in BLAST2GO, we identified the most prominent protein domains predicted for *T. longicaudatus* unigenes. A total of 1252 unigenes showed top-hits to the P-loop-containing nucleoside triphosphate hydrolase (P-loop NTPase) domain. Other top domains identified based on unigene homology included the insulin-like growth factor binding protein, N-terminal domain, zinc finger, C2H2-like domain, heat shock protein 70 family, EGF-like domain, and helicase C-terminal domain (Table S2).



**Figure 7.** Identified KEGG pathways of assembled unigenes from *T. longicaudatus*.

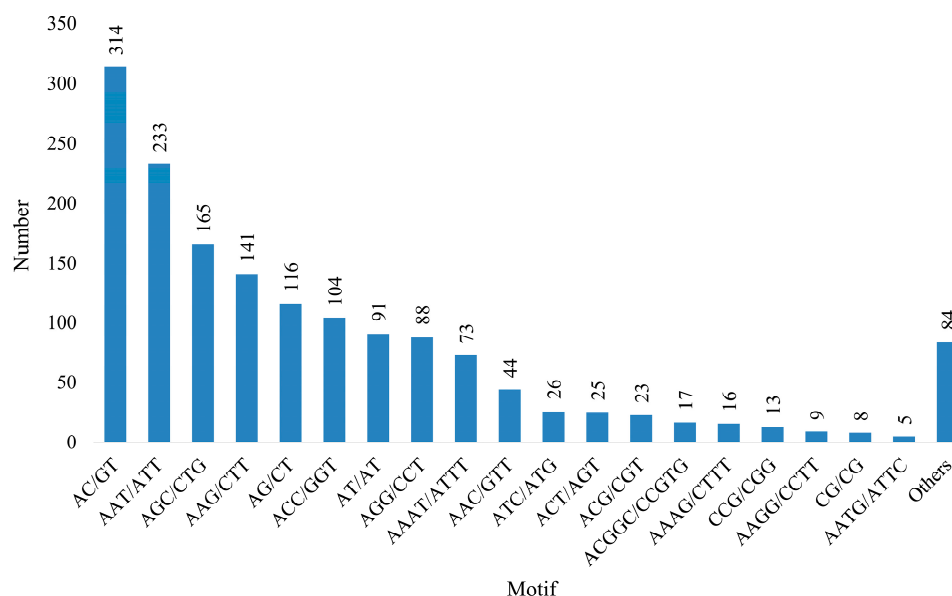
### 3.4. Development and SSR Locus Analysis

To identify SSRs, we scanned 29,547 unigene sequences (75,658,821 bp) of *T. longicaudatus* with lengths >1000 bp. A total of 1595 potential SSR loci were detected including 529 (33.2%), 862 (54%), 144 (9%), 33 (2.1%) and 27 (1.7%), di-, tri-, tetra-, penta-, and hexa-nucleotide repeats, respectively (Table 3). The SSR repeats identified were present predominantly in six, five, four, four, and four iterations, respectively, for di-, tri-, tetra-, penta- and hexa-nucleotide repeats (Table S3).

**Table 3.** SSRs identified from the unigene sequences of *T. longicaudatus*.

SSR parameters	Number
Total number of sequences examined	29,547
Total size of examined sequences (bp)	75,658,821
Total number of identified SSRs	1595
Di-nucleotide	529
Tri-nucleotide	862
Tetra-nucleotide	144
Penta-nucleotide	33
Hexa-nucleotide	27
Number of SSR containing sequences	1432
Number of sequences containing more than 1 SSR	140
Number of SSRs present in compound formation	74

Among the di-nucleotide repeats, AC/GT (314 unigenes), AG/CT (116 unigenes), and AT/AT (91 unigenes) were the dominant motifs. Within the tri-nucleotide repeats, AAT/ATT (233 unigenes), followed by AGC/CTG (165 unigenes) and AAG/CTT (141 unigenes), were the most repeated motifs (Figure 8). All 1162 SSR-containing unigenes were functionally annotated. In addition, a total of 1387 SSR sites were randomly selected from the SSR-containing sequences to design SSR primers for genotyping. Among the 1387 SSR sites, 1123 were included known functional regions. A list of PCR primers and conditions is shown in Table S4.



**Figure 8.** The number of SSRs discovered in the unigenes from *T. longicaudatus* based on motif sequence types.

#### 4. Discussion

In this study, we used high-throughput mRNA-Seq technology to analyze expressed transcripts of the longtail tadpole shrimp *T. longicaudatus*. RNA-Seq platform technology has been used for the rapid characterization of genomic and genetic resources in related non-model species including the Pacific white shrimp (*Litopenaeus vannamei*) [28,42], the Banana shrimp (*Fenneropenaeus merguensis*) [43], the Brine shrimp (*Artemia franciscana*) [44], and the *Triops newberryi* [1]. Transcriptome studies have also provided advances in establishing putative genes involved in the growth, reproduction and innate immune system pathways in the European shore crab (*Carcinus maenas*) [45], the Mud crab (*Scylla paramamosain*) [46], and the swimming crab (*Portunus trituberculatus*) [47]. These studies have researched the need for genetic data on these species through the screening and exploitation of microsatellites in a cost-efficient and timely manner. In this study, using the Illumina HiSeq 2500 sequencing method and Trinity de novo assembly, 269,822 contigs and 208,813 unigenes were generated. The N<sub>50</sub> length (1148 bp) and the average length (712.8 bp) of the contigs and unigenes (N<sub>50</sub> length of 1089 bp and an average length of 700 bp) are greater than in the transcriptomic analysis of other crustacean species such as *L. vannamei* (42,336 unigenes with an N<sub>50</sub> of 736 bp and an average length of 561 bp) [48], brine shrimp, *A. franciscana* (36,896 contigs with an average length of 746 bp) [44], crayfish, *Cherax quadricarinatus* (36,128 contigs with an N<sub>50</sub> of 936 bp and an average length of 800 bp) [49], and pandalid shrimp, *Pandalus latirostris* (45,467 contigs with an N<sub>50</sub> of 493 bp) [50], and are lower than in the transcriptome of *Parhyale hawaiiensis* (35,301 contigs with an N<sub>50</sub> of 1510 bp) [51]. For further we applied the BUSCO, which is reference based software for assessing quality of de novo assemblies. Out of 2675 single copy orthologs for arthropods our assembly is 88.56% complete (1708 complete

single copy BUSCOs and 661 complete duplicated BUSCOs), while 5.35% of contigs are fragmented (143 fragmented BUSCOs) and 6.09% are missing (163 missing BUSCOs).

We annotated the *T. longicaudatus* unigene sequences against the PANM, Unigene, KOG, GO, and KEGG databases by BLASTX with a cut-off value of  $1E-5$ . Approximately 45.55% of unigenes matched to homologous sequences in the databases, which is less than half of the unigenes present in the *T. longicaudatus* transcriptome could be annotated. Lineage-specific genes are often difficult to annotate because their function is specific to the species [1,52]. We also characterized the homology search using PANM-DB due to the greater degree of annotation of unigene sequences obtained with this database. PANM-DB is preferred over the NCBI nr database due to faster processing of NGS datasets (15 times faster than that of the NCBI nr database) and a higher number of annotation hits [37]. The locally curated PANM-DB was an addition to the Molluscs database, and covers the available sequences of the Protostomia group in a multi-FASTA format [53]. Furthermore, our results showed that more than 90% of unigenes with a sequence length  $>2000$  bp matched with a homologous protein in the databases, which is possible because the protein-coding genes generally give rise to longer full-length transcripts [54]. The BLASTX top-hit species distribution showed putative homology of the annotated unigene sequences across species in the PANM-DB. Most sequences matched the crustacean, *Daphnia pulex* (15.32%), followed by *Crassostrea gigas* (5.47%) and *Lottia gigantea* (3.67%).

Functional annotations of the assembled unigenes using KOG, GO terms, KEGG pathway analysis, as well as an InterPro conserved domain scan, were conducted to obtain a comprehensive description of the properties of these genes and their products in the species. GO classification only suggests that a unigene is related to a predicted function, as all GO terms are not of equal validity [55]. Most of the evidence codes are based on electronic annotations and are not manually created. The computational source of evidence constitutes more than 95% of the total GO annotation results in non-model species [56,57]. KEGG pathway analysis suggests the classification of unigenes into regulatory biological pathways that include metabolism, genetic information processing, environmental information processing, and organismal systems. The *T. longicaudatus* unigenes were mapped to 129 reference canonical pathways, among which distribution to the metabolism pathways was predominant. In the transcriptome analysis of *Litopenaeus vannamei*, a total of 9621 unigenes were mapped to 317 pathways, wherein the most enriched sequences were assigned to metabolic pathways, followed by the biosynthesis of secondary metabolites and spliceosome and RNA transport [58]. In the mud crab (*Scylla paramamosain*) transcriptome using 454 sequencing, 4878 unigenes were classified into 281 KEGG pathways, and the identified genes were found to be involved in growth, development, and disease resistance pathways [46]. Among the top-hit InterPro domain obtained in the present analysis, P-loop NTPases were predominant. These represent a large protein family that is involved in a variety of cellular functions, such as signal transduction, translation, protein transport and localization, signal-sequence recognition, chromosome partitioning, and membrane transport [59]. The  $C_2H_2$  type zinc finger domains are widely found in DNA binding motifs in eukaryotic transcription factors [60].

Polymorphic microsatellite markers such as SSRs have been utilized for a variety of genetic and breeding studies [61]. NGS technologies can be used to develop abundant SSR or SNP markers with high efficiency and accuracy [62]. In this study, we screened 1595 SSRs of 2–6 bp in length from unigene sequences  $>1000$  bp in length. The tri-nucleotide repeats were predominant, followed by di- and tetra-nucleotide repeats. The tri-nucleotide SSR motifs have been consistently found as the predominant markers in the transcriptome sequences of many monocotyledonous plants [63,64]; however, in animals, the di-nucleotide repeats are predominant [65]. One nucleotide repeat motifs were detected but were not considered as these may be the result of single nucleotide stretch errors generated by sequencing [66,67]. These SSR loci provide an abundant marker resource for studying the genetic variation, population, and conservation genomics of species. In a previous study that constructed a genetic linkage map of *L. vannamei* using AFLP and SSR markers, 25 SSR markers were found to be informative in mapping a population of *L. vannamei* and are available for map construction [68].

The abundance of AC/GT motifs found in the present study is consistent with the SSR motif study in the mud crab, *Scylla paramamosain* [46]. The tri-nucleotide motifs AGC/CTG and ACC/GGT found in this study were also the preferred motifs in the SSRs isolated from the transcriptome of the Red Swamp Crayfish *Procambarus clarkii* [69]. A total of 1387 potential SSR markers identified in this study will provide important research advances for genetic studies including the assessment of genetic diversity, the development of genetic maps, comparative genomics, and marker-assisted selection breeding. The primer pairs designed for polymorphism identification would add towards genotyping of the species diversity and exploitation of the economic potential of the species.

## 5. Conclusions

This is the first report of high-throughput transcriptome analysis of *T. longicaudatus*. In total, 95,105 unigenes were annotated for putative functions using BLASTX with a cut-off of  $1E-5$ . A total of 57,731 unigenes were assigned to GO terms, and 7247 unigenes were mapped to 129 KEGG pathways. Furthermore, 1595 SSRs were detected from the unigenes with 1387 potential SSR markers. A total of 1387 potential SSR markers identified in this study will provide important research advances for genetic studies including the assessment of genetic diversity, the development of genetic maps, comparative genomics, and marker assisted selection breeding.

**Supplementary Materials:** The following are available online at [www.mdpi.com/2073-4425/7/12/114/s1](http://www.mdpi.com/2073-4425/7/12/114/s1). Figure S1: Summary of homology search of assembled unigenes of *T. longicaudatus* against PANM-DB. (A) score distribution, (B) identity distribution, (C) similarity distribution, (D) distribution of hit and non-hit sequences as compared with the length of unigenes; Table S1: KEGG mappings for *T. longicaudatus* unigenes; Table S2: List of top-hit InterPro domains in *T. longicaudatus*, Table S3: Summary of SSR types in the *T. longicaudatus* transcriptome, Table S4: Sequences of 1,387 primer pairs for SSR markers.

**Acknowledgments:** This work was supported by the grant entitled “The Genetic and Genomic Evaluation of Indigenous Biological Resources” funded by the National Institute of Biological Resources (NIBR201503202) and Soonchunhyang University Research Fund.

**Author Contributions:** Jiyeon Seong, Bharat Bhusan Patnaik, Hee Ju Hwang, Soonok Kim, Ui Wook Hwang and Yong Seok Lee designed the experiments; So Young Park, Jong Min Chung, Dae Kwon Song, and Mi Young Noh performed the experiments; Jiyeon Seong, Se Won Kang, Hee Ju Hwang, Seung-Hwan Park, Gwang Joo Jeon and Hong Sik Kong analyzed the data; Jiyeon Seong, Se Won Kang, Bharat Bhusan Patnaik and Hee Ju Hwang wrote the paper; Soonok Kim, Ui Wook Hwang, Hong Seog Park and Yeon Soo Han contributed reagents/materials/analysis tools; Yong Seok Lee supervised the entire study.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Horn, R.L.; Ramaraj, T.; Devitt, N.P.; Schilkey, F.D.; Cowley, D.E. De novo assembly of a tadpole shrimp (*Triops newberryi*) transcriptome and preliminary differential gene expression analysis. *Mol. Ecol. Resour.* **2016**. [[CrossRef](#)] [[PubMed](#)]
2. Suno-Uchi, N.; Sasaki, F.; Chiba, S.; Kawata, M. Morphological stasis and phylogenetic relationships in Tadpole shrimps, *Triops* (Crustacea: Notostraca). *Biol. J. Linn. Soc.* **1997**, *61*, 439–457. [[CrossRef](#)]
3. Wooten, D. *Triops longicaudatus*. In *Zooplankton of the Great Lakes*; Central Michigan University: Mount Pleasant, MI, USA, 2010.
4. Fry, L.L.; Mulla, M.S.; Adams, C.W. Field introductions and establishment of the Tadpole shrimp, *Triops longicaudatus* (Notostraca: Triopsidae), a biological control agent of mosquitoes. *Biol. Control* **1994**, *4*, 113–124. [[CrossRef](#)]
5. Becker, N.; Petric, D.; Zgomba, M.; Boase, C.; Madon, M.; Dahl, C.; Kaiser, A. *Mosquitoes and Their Control*; Springer-Verlag: Heidelberg, Germany, 2010.
6. Yonekura, M. Weeding efficacy of tadpole shrimp (*Triops* spp.) in transplanted rice fields. In Proceedings of the 7th Asian Pacific Weed Science Society Conference, Sydney, Australia, 26–30 November 1979; p. 240.
7. Korn, M.; Marrone, F.; Perez-Bote, J.L.; Machado, M.; Cristo, M.; Da Fonseca, L.C.; Hundsdorfer, A.K. Sister species within the *Triops cancriformis* lineage (Crustacea, Notostraca). *Zool. Scr.* **2006**, *35*, 301–322. [[CrossRef](#)]

8. Macdonald, K.S.; Sallenave, R.; Cowley, D.E. Morphologic and Genetic variation in *Triops* (Branchiopoda: Notostraca) from ephemeral waters of the Northern Chihuahuan Desert of North America. *J. Crustacean Biol.* **2011**, *31*, 468–484. [[CrossRef](#)]
9. Horn, R.L.; Cowley, D.E. Evolutionary relationships within *Triops* (Branchiopoda: Notostraca) using complete mitochondrial genomes. *J. Crustacean Biol.* **2014**, *34*, 795–800. [[CrossRef](#)]
10. Linder, F. Contributions to the morphology and taxonomy of the Branchiopoda Notostraca, with special reference to the North American species. *Proc. U. S. Natl. Mus.* **1952**, *102*, 1–69. [[CrossRef](#)]
11. Longhurst, A.R. A review of the Notostraca. *Bulletin of the British Museum (Natural History). Zoology* **1955**, *3*, 1–57.
12. Akita, M. Classification of Japanese tadpole shrimps. *Zool. Mag.* **1976**, *85*, 237–247.
13. Yoon, S.M.; Kim, W.; Kim, H.S. Re-description of *Triops longicaudatus* (LeConte, 1846) (Notostraca, Triopsidae) from Korea. *Korean J. Syst. Zool.* **1992**, *3*, 59–66.
14. Sassaman, C.; Simovich, M.; Fugate, M. Reproductive isolation and genetic differentiation in North American species of *Triops* (Crustacea: Branchiopoda: Notostraca). *Hydrobiologia* **1997**, *359*, 125–147. [[CrossRef](#)]
15. Kwon, S.J.; Jun, Y.C.; Park, J.H.; Won, D.H.; Seo, E.W.; Lee, J.E. Distribution and habitat characteristics of tadpole Shrimp. *Korean J. Limnol.* **2010**, *43*, 142–149.
16. Aloufi, A.B.A.; Obuid-Allah, A.H. New records and Redescription of the notostracan Tadpole shrimp, *Triops longicaudatus* (LeConte, 1846) from temporary water bodies in North West region (Tabuk and Madinah) in Saudi Arabia. *Int. J. Adv. Res.* **2014**, *2*, 1222–1231.
17. Kwon, S.J.; Kwon, H.J.; Jun, Y.C.; Lee, J.E.; Won, D.H. Effect of temperature on hatching rate of *Triops longicaudatus* (Triopsidae, Notostraca). *Korean J. Limnol.* **2009**, *42*, 32–38.
18. Ryu, J.S.; Hwang, U.W. Complete mitochondrial genome of the longtail tadpole shrimp *Triops longicaudatus* (Crustacea, Branchiopoda, Notostraca). *Mitochondrial DNA* **2010**, *21*, 170–172. [[CrossRef](#)] [[PubMed](#)]
19. Wu, P.; Qi, D.; Chen, L.; Zhang, H.; Zhang, X.; Qin, J.G.; Hu, S. Gene discovery from an ovary cDNA library of oriental river prawn *Macrobrachium nipponense* by ESTs annotation. *Comp. Biochem. Physiol. Part D Genom. Proteom.* **2009**, *4*, 111–120. [[CrossRef](#)] [[PubMed](#)]
20. Metzker, M.L. Sequencing technologies—The next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46. [[CrossRef](#)] [[PubMed](#)]
21. Qi, X.; Zhang, L.; Han, Y.; Ren, X.; Huang, J.; Chen, H. De novo transcriptome sequencing and analysis of *Coccinella septempunctata* L. in non-diapause, diapause and diapause-terminated states to identify diapause-associated genes. *BMC Genom.* **2015**, *16*, 1086. [[CrossRef](#)] [[PubMed](#)]
22. Jin, H.; Dong, D.; Yang, Q.; Zhu, D. Salt-responsive transcriptome profiling of *Suaeda glauca* via RNA Sequencing. *PLoS ONE* **2016**, *11*, e0150504. [[CrossRef](#)] [[PubMed](#)]
23. Patnaik, B.B.; Hwang, H.-J.; Kang, S.W.; Park, S.Y.; Wang, T.H.; Park, E.B.; Chung, J.M.; Song, D.K.; Kim, C.; Kim, S.; et al. Transcriptome Characterization for non-model endangered Lycaenids, *Protantigius superans* and *Spindasis takanosis*, using Illumina HiSeq 2500 Sequencing. *Int. J. Mol. Sci.* **2015**, *16*, 29948–29970. [[CrossRef](#)] [[PubMed](#)]
24. Morozova, O.; Marra, M.A. Applications of Next-Generation sequencing technologies in functional genomics. *Genomics* **2008**, *92*, 255–264. [[CrossRef](#)] [[PubMed](#)]
25. Novaes, E.; Drost, D.R.; Farmerie, W.G.; Pappas, G.J.; Grattapaglia, D.; Sederoff, R.R.; Kirst, M. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genom.* **2008**, *9*. [[CrossRef](#)] [[PubMed](#)]
26. Zimmer, C.T.; Malwald, F.; Schorn, C.; Bass, C.; Ott, M.C.; Nauen, R. A de novo transcriptome of European pollen beetle populations and its analysis, with special reference to insecticide action and resistance. *Insect Mol. Biol.* **2014**, *23*, 511–526. [[CrossRef](#)] [[PubMed](#)]
27. Fullwood, M.J.; Wei, C.L.; Liu, E.T.; Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **2009**, *19*, 521–532. [[CrossRef](#)] [[PubMed](#)]
28. Li, D.J.; Deng, Z.; Qin, B.; Liu, X.H.; Men, Z.H. De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). *BMC Genom.* **2012**. [[CrossRef](#)] [[PubMed](#)]
29. Sadamoto, H.; Takahashi, H.; Okada, T.; Kenmoku, H.; Toyota, M.; Asakawa, Y. De novo Sequencing and transcriptome analysis of the Central Nervous System of Mollusc *Lymnaea stagnalis* by Deep RNA Sequencing. *PLoS ONE* **2012**, *7*, e42546. [[CrossRef](#)] [[PubMed](#)]

30. Che, R.; Sun, Y.; Wang, R.; Xu, T. Transcriptomic analysis of endangered Chinese salamander: Identification of immune, sex and reproduction-related genes and genetic markers. *PLoS ONE* **2014**, *9*, e87940. [[CrossRef](#)] [[PubMed](#)]
31. Joshi, N.A.; Fass, J.N. Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files (Version 1.33) [Software]. Available online: <https://github.com/najoshi/sickle> (accessed on 13 July 2016).
32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **2011**, *17*, 10–12. [[CrossRef](#)]
33. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. [[CrossRef](#)] [[PubMed](#)]
34. Surget-Groba, Y.; Montoya-Burgos, J.I. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.* **2010**, *20*, 1432–1440. [[CrossRef](#)] [[PubMed](#)]
35. Hao, D.C.; Ge, G.B.; Xiao, P.G.; Zhang, Y.Y.; Yang, L. The first insight into the *Taxus* genome via fosmid library construction and end sequencing. *Mol. Genet. Genom.* **2011**, *285*, 197–205. [[CrossRef](#)] [[PubMed](#)]
36. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [[CrossRef](#)] [[PubMed](#)]
37. Kang, S.W.; Patnaik, B.B.; Hwang, H.J.; Park, S.Y.; Lee, J.S.; Han, Y.S.; Lee, Y.S. PANM DB (Protosome DB) for the annotation of NGS data of mollusks. *Korean J. Malacol.* **2015**, *31*, 243–247. [[CrossRef](#)]
38. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadapoulous, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*. [[CrossRef](#)] [[PubMed](#)]
39. Conesa, A.; Gotz, S.; Garcia-Gomez, J.M.; Terol, J.; Talon, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [[CrossRef](#)] [[PubMed](#)]
40. Ye, J.; Fang, L.; Zheng, H.; Zhang, Y.; Chen, J.; Zhang, Z.; Wang, J.; Li, S.; Li, R.; Bolund, L.; et al. WEGO: A web tool for plotting GO annotations. *Nucleic Acids Res.* **2006**, *34*, W293–W297. [[CrossRef](#)] [[PubMed](#)]
41. Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **2004**, *32*, D277–D280. [[CrossRef](#)] [[PubMed](#)]
42. Chen, K.; Li, E.; Xu, Z.; Li, T.; Xu, C.; Qin, J.G.; Chen, L. Comparative Transcriptome analysis in the hepatopancreas tissue of Pacific White Shrimp *Litopenaeus vannamei* fed different lipid sources at low salinity. *PLoS ONE* **2015**, *10*, e0144889. [[CrossRef](#)] [[PubMed](#)]
43. Powell, D.; Knibb, W.; Remilton, C.; Elizur, A. De-novo transcriptome analysis of the banana shrimp (*Renneropenaeus merguensis*) and identification of genes associated with reproduction and development. *Mar. Genom.* **2015**, *22*, 71–78. [[CrossRef](#)] [[PubMed](#)]
44. Valenzuela-Miranda, D.; Gallardo-Escarate, C.; Valenzuela-Munoz, V.; Farlora, R.; Gajardo, G. Sex-dependent transcriptome analysis and single nucleotide polymorphism (SNP) discovery in the brine shrimp *Artemia franciscana*. *Mar. Genom.* **2014**, *18 PB*, 151–154. [[CrossRef](#)] [[PubMed](#)]
45. Verbruggen, B.; Bickley, L.K.; Santos, E.M.; Tyler, C.R.; Stentiford, G.D.; Bateman, K.S.; van Aerle, R. De novo assembly of the *Carcinus maenus* transcriptome and characterization of innate immune system pathways. *BMC Genom.* **2015**, *16*. [[CrossRef](#)] [[PubMed](#)]
46. Ma, H.; Ma, C.; Li, S.; Jiang, W.; Li, X.; Liu, Y.; Ma, L. Transcriptome analysis of the Mud Crab (*Scylla paramamosain*) by 454 deep sequencing: Assembly, Annotation and Marker Discovery. *PLoS ONE* **2014**, *9*, e102668. [[CrossRef](#)] [[PubMed](#)]
47. Lv, J.; Liu, P.; Gao, B.; Wang, Y.; Wang, Z.; Chen, P.; Li, J. Transcriptome analysis of the *Portunus trituberculatus*: De novo assembly, growth-related gene identification and marker discovery. *PLoS ONE* **2014**, *9*, e94055. [[CrossRef](#)] [[PubMed](#)]
48. Guo, H.; Ye, C.X.; Wang, A.L.; Xian, J.A.; Liao, S.A.; Miao, Y.T.; Zhang, S.P. Transcriptome analysis of the Pacific white shrimp *Litopenaeus vannamei* exposed to nitrite by RNA-seq. *Fish Shellfish Immunol.* **2013**, *35*, 2008–2016. [[CrossRef](#)] [[PubMed](#)]

49. Ali, M.Y.; Pavasovic, A.; Mather, P.B.; Prentis, P.J. Transcriptome analysis and characterization of gill-expressed carbonic anhydrase and other key osmoregulatory genes in freshwater crayfish *Cherax quadricarinatus*. *Data Brief* **2015**, *5*, 187–193. [[CrossRef](#)] [[PubMed](#)]
50. Ryouka, K.-M.; Wada, K.; Azuma, N.; Chiba, S. Expression profiling without genome sequence information in a non-model species, Pandalid shrimp (*Pandalus latirostris*), by next-generation sequencing. *PLoS ONE* **2011**, *6*, e26043.
51. Zeng, V.; Villaneuva, K.E.; Ewen-Campen, B.S.; Alwes, F.; Browne, W.E.; Extavour, C.G. De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean *Parhyale hawaiiensis*. *BMC Genom.* **2011**. [[CrossRef](#)] [[PubMed](#)]
52. Asselman, J.; Pfrender, M.E.; Lopez, J.A.; De Coninck, D.I.; Janssen, C.R.; Shaw, J.R.; De Schamphelaere, K.A. Conserved transcriptional responses to cyanobacterial stressors are mediated by alternate regulation of paralogous genes in *Daphnia*. *Mol. Ecol.* **2015**, *24*, 1844–1855. [[CrossRef](#)] [[PubMed](#)]
53. Kang, S.W.; Hwang, H.J.; Park, S.Y.; Wang, T.H.; Park, E.B.; Lee, T.H.; Hwang, U.W.; Lee, J.S.; Park, H.S.; Han, Y.S.; et al. Mollusks sequence database: Version II. *Korean J. Malacol.* **2014**, *30*, 429–431. [[CrossRef](#)]
54. Hwang, H.-J.; Patnaik, B.B.; Kang, S.W.; Park, S.Y.; Wang, T.H.; Park, E.B.; Chung, J.M.; Song, D.K.; Patnaik, H.H.; Kim, C.; et al. RNA sequencing, de novo assembly, and functional annotation of an endangered Nymphalid butterfly, *Fabriciana nerippe* Felder, 1862. *Entomol. Res.* **2016**, *2*, 148–161. [[CrossRef](#)]
55. Rhee, S.Y.; Wood, V.; Dolinski, K.; Draghici, S. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* **2008**, *9*, 509–515. [[CrossRef](#)] [[PubMed](#)]
56. Patnaik, B.B.; Wang, T.H.; Kang, S.W.; Hwang, H.-J.; Park, S.Y.; Park, E.B.; Chung, J.M.; Song, D.K.; Kim, C.; Kim, S.; et al. Sequencing, de novo assembly, and annotation of the transcriptome of the endangered freshwater pearl bivalve, *Cristaria plicata*, provides novel insights into functional genes and marker discovery. *PLoS ONE* **2016**, *11*, e0148622. [[CrossRef](#)] [[PubMed](#)]
57. Park, S.Y.; Patnaik, B.B.; Kang, S.W.; Hwang, H.-J.; Chung, J.M.; Song, D.K.; Sang, M.K.; Patnaik, H.H.; Lee, J.B.; Noh, M.Y.; et al. Transcriptome analysis of the endangered neritid species *Clithon retropictus*: De novo Assembly, Functional Annotation, and Marker Discovery. *Genes* **2016**, *7*. [[CrossRef](#)] [[PubMed](#)]
58. Chen, K.; Li, E.; Li, T.; Xu, C.; Wang, X.; Lin, H.; Qin, J.G.; Chen, L. Transcriptome and pathway analysis of the Hepatopancreas in the Pacific White Shrimp *Litopenaeus vannamei* under Chronic Low-Salinity stress. *PLoS ONE* **2015**, *10*, e0131503. [[CrossRef](#)] [[PubMed](#)]
59. Pathak, E.; Atri, N.; Mishra, R. Role of highly central residues of P-loop and its flanking region in preserving the archetypal conformation of Walker A motif of diverse P-loop NTPases. *Bioinformation* **2013**, *9*, 23–28. [[CrossRef](#)] [[PubMed](#)]
60. Razin, S.V.; Borunova, V.V.; Maksimenko, O.G.; Kantidze, O.L. Cys2His2 Zinc Finger Protein Family: Classification, Functions, and Major Members. *Biochemistry (Moscow)* **2012**, *77*, 217–226. [[CrossRef](#)] [[PubMed](#)]
61. Luo, H.; Xiao, S.; Ye, H.; Zhang, Z.; Lv, C.; Zheng, S.; Wang, Z.; Wang, X. Identification of immune related genes and development of SSR/SNP markers from the spleen transcriptome of *Schizothorax prenanti*. *PLoS ONE* **2016**, *11*, e0152572. [[CrossRef](#)] [[PubMed](#)]
62. Liang, M.; Yang, X.; Li, H.; Su, S.; Yi, H.; Chai, L.; Deng, X. De novo transcriptome assembly of pummelo and molecular marker development. *PLoS ONE* **2015**, *10*, e0120615. [[CrossRef](#)] [[PubMed](#)]
63. La Rota, M.; Kantety, R.V.; Yu, J.K.; Sorrells, M.E. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genom.* **2005**, *6*. [[CrossRef](#)] [[PubMed](#)]
64. Annadurai, R.S.; Neethiraj, R.; Jayakumar, V.; Damodaran, A.C.; Rao, S.N.; Katta, M.A.V.S.K.; Gopinathan, S.; Sarma, S.P.; Senthikumar, V.; Niranjana, V.; et al. De novo transcriptome assembly (NGS) of *Curcuma longa* L. rhizome reveals novel transcripts related to anticancer and antimalarial terpenoids. *PLoS ONE* **2013**, *8*, e56217. [[CrossRef](#)] [[PubMed](#)]
65. Chen, M.; Tan, Z.; Zeng, G.; Peng, J. Comprehensive analysis of simple sequence repeats in pre-miRNAs. *Mol. Biol. Evol.* **2010**, *27*, 2227–2232. [[CrossRef](#)] [[PubMed](#)]
66. Miller, A.D.; Good, R.T.; Coleman, R.A.; Lancaster, M.L.; Weeks, A.R. Microsatellite loci and the complete mitochondrial DNA sequence characterized through next generation sequencing and de novo assembly for the critically endangered orange-bellied parrot, *Neophema chrysogaster*. *Mol. Biol. Rep.* **2013**, *40*, 35–42. [[CrossRef](#)] [[PubMed](#)]



67. Zhang, S.H.; Luo, H.; Du, H.; Wang, D.Q.; Wei, Q.W. Isolation and characterization of twenty-six microsatellite loci for the tetraploid fish Dabry's sturgeon (*Acipenser dabryanus*). *Conserv. Genet. Res.* **2013**, *5*, 409–412. [[CrossRef](#)]
68. Andriantahina, F.; Liu, X.; Huang, H. Genetic map construction and quantitative trait locus (QTL) detection of growth-related traits in *Litopenaeus vannamei* for selective breeding applications. *PLoS ONE* **2013**, *8*, e75206. [[CrossRef](#)]
69. Shen, D.; Bo, W.; Xu, F.; Wu, R. Genetic diversity and population structure of the Tibetan poplar (*Populus szechuanica* var. *tibetica*) along an altitude gradient. *BMC Genet.* **2014**, *15*. [[CrossRef](#)]



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).