*insects*

MDPI

*Article*

# SMRT Sequencing Technology Was Used to Construct the *Batocera horsfieldi* (Hope) Transcriptome and Reveal Its Features

**Xinju Wei †, Danping Xu †, Zhiqian Liu, Quanwei Liu and Zhihang Zhuo ***

College of Life Science, China West Normal University, Nanchong 637002, China; weixinjuxx@foxmail.com (X.W.); xudanping@cwnu.edu.cn (D.X.); qnhtvxhp319123@foxmail.com (Z.L.); quanwei66977@163.com (Q.L.)
* Correspondence: zhuozhihang@cwnu.edu.cn; Tel.: +86-1311-197-3927
† These authors contributed equally to this work.

**Simple Summary:** *Batocera horsfieldi* (Hope) is an important wood-boring pest in China. This pest primarily infests tree trunks by feeding on the woody tissue, creating a network of interconnected tunnels within the trunk. These tunnels become blocked with insect feces and wood debris, causing damage, decay, and even the death of the host plant's tissues. In this study, single-molecule real-time sequencing (SMRT) and Illumina RNA-seq technologies were employed to conduct full-length transcriptome sequencing of male and female adults of *B. horsfieldi*. A total of 20,356,793 subreads (38.26 G, clean reads) were generated, including 432,091 circular consensus sequences and 395,851 full-length non-chimera reads. Clustering and redundancy removal of the full-length non-chimera reads resulted in 39,912 consensus reads. Additionally, functional annotation was performed on a total of 84,650 transcripts in seven different databases. This study provides an important foundation for future exploration of gene regulation in the interaction between *B. horsfieldi* and host plants using RNA interference (RNAi), and it offers a scientific basis for the prevention and control of *B. horsfieldi*.

**Abstract:** *Batocera horsfieldi* (Hope) (Coleoptera: Cerambycidae) is an important forest pest in China that mainly infests timber and economic forests. This pest primarily causes plant tissue to necrotize, rot, and eventually die by feeding on the woody parts of tree trunks. To gain a deeper understanding of the genetic mechanism of *B. horsfieldi*, this study employed single-molecule real-time sequencing (SMRT) and Illumina RNA-seq technologies to conduct full-length transcriptome sequencing of the insect. Total RNA extracted from male and female adults was mixed and subjected to SMRT sequencing, generating a complete transcriptome. Transcriptome analysis, prediction of long non-coding RNA (lncRNA), coding sequences (CDs), analysis of simple sequence repeats (SSR), prediction of transcription factors, and functional annotation of transcripts were performed in this study. The collective 20,356,793 subreads (38.26 G, clean reads) were generated, including 432,091 circular consensus sequences and 395,851 full-length non-chimera reads. The full-length non-chimera reads (FLNC) were clustered and redundancies were removed, resulting in 39,912 consensus reads. SSR and ANGEL software v3.0 were used for predicting SSR and CDs. In addition, four tools were used for annotating 6058 lncRNAs, identifying 636 transcription factors. Furthermore, a total of 84,650 transcripts were functionally annotated in seven different databases. This is the first time that the full-length transcriptome of *B. horsfieldi* has been obtained using SMRT sequencing. This provides an important foundation for investigating the gene regulation underlying the interaction between *B. horsfieldi* and its host plants through gene editing in the future and provides a scientific basis for the prevention and control of *B. horsfieldi*.

**Keywords:** *Batocera horsfieldi*; full-length transcripts; functional annotation; Illumina RNA-seq; single-molecule real-time (SMRT) sequencing



**Citation:** Wei, X.; Xu, D.; Liu, Z.; Liu, Q.; Zhuo, Z. SMRT Sequencing Technology Was Used to Construct the *Batocera horsfieldi* (Hope) Transcriptome and Reveal Its Features. *Insects* **2023**, *14*, 625. https://doi.org/10.3390/insects14070625

Academic Editor: Mauro Mandrioli

Received: 27 May 2023
Revised: 28 June 2023
Accepted: 7 July 2023
Published: 11 July 2023

## 1. Introduction

*Batocera horsfieldi* (Hope) is an important forest pest in China that attacks trees by boring into the stems. It belongs to the family Cerambycidae of the order Coleoptera and is also known as *Batocera lineolata* [1]. The pest is mainly distributed abroad in Vietnam, Japan, India, Myanmar, and other countries [2], while domestically it is distributed in Hebei, Beijing, Zhejiang, Fujian, Jiangxi, Hubei, Hunan, Taiwan, Guangdong, Guizhou, Sichuan, Yunnan, and other places [3]. After hatching, the clouded leopard moth larvae begin to feed on the cambium of the host plant, then enter the xylem and continuously bore into the wood, causing damage, rot, and even the death of the host plant tissue [4]. Therefore, it greatly affects the growth and development of the host plant, leading to a decline in fruit quality [5]. Its adults appear in early summer and feed on the branches of the host plant until they reach sexual maturity [6]. *B. horsfieldi* has a wide variety of host species, including *Populus tomentosa*, *Juglans regia*, *Eucalyptus robusta*, *Fraxinus chinensis*, and *Castanea mollissima*, among others [7,8]. *B. horsfieldi* is an insect with a fairly complete protection mechanism. Its strong adaptability and high reproductive capacity cause serious harm to plant growth and the ecological environment. Due to the concealment of its larvae, it is difficult to control, and traditional chemical control methods are rarely effective [6,9]. Currently, research on *B. horsfieldi* mainly focuses on predicting its habitat distribution [3], investigating its impact on hosts [10–12], and analyzing the antennal transcriptome and olfactory-related genes of adult insects [13]. These studies have laid the foundation for exploring new control strategies for *B. horsfieldi*.

Transcriptome data can reflect information on cell responses, gene functions, evolution, and other biological processes, revealing different biological processes at the molecular level [14–16]. Transcriptome analysis is a high-throughput technology that can simultaneously determine the expression levels of a large number of genes in an organism, thereby revealing the complex mechanisms of insect gene expression and regulation. Through transcriptome analysis, we can gain a deeper understanding of the genetic characteristics and growth and development patterns of insects, providing strong evidence and references for future insect research. In addition, transcriptome analysis can help us discover new genes, reveal metabolic pathways and adaptive mechanisms, and deepen our understanding of insect taxonomy and ecology. Therefore, transcriptome analysis has significant implications for insect research and is expected to further promote the development of basic and applied research on insects, providing new ideas for further research on *B. horsfieldi*. In recent years, transcriptome sequencing has gradually been used in gene expression analysis of the different developmental stages of Coleoptera [17,18], such as *Nicrophorus orbicollis* [19], *Tribolium castaneum* (Herbst) [20], *Callosobruchus chinensis* (L.) [21], and *Lasioderma serricorne* [22], which have all used transcriptome library analysis. Transcriptome library construction has also been widely used in the transcriptome analysis of longhorn beetles, such as *Anoplophora nobilis* and *Monochamus alternatus*, to capture specific genes and conduct functional analysis [13]. In addition, transcriptome and olfactory-related gene analysis of the antennae of *B. horsfieldi* adults have enriched the gene database of Coleoptera and laid the foundation for further research on the behavioral regulation and control of *B. horsfieldi* at the molecular level.

However, there are incomplete splice transcripts in short-read transcriptome sequencing, making it difficult for current short-read sequence prediction programs to accurately predict gene structure [23]. Additionally, using low-quality transcripts obtained from short sequencing may lead to incorrect annotation [24,25]. Second-generation sequencing technology has some limitations, including shorter read lengths and the inability to cover the entire transcriptome, which presents significant challenges to the genome assembly process [26,27]. However, third-generation sequencing technologies represented by PacBio have effectively addressed these issues. Single-molecule sequencing technology, also known as SMRT (single-molecule real-time) sequencing, utilizes its advantage of ultra-long read lengths to directly obtain complete transcripts containing 5′UTR, 3′UTR, and polyA tails without the need for interruption and splicing. This enables accurate analysis

of structural information, such as alternative splicing and fusion genes in reference genome species, thus overcoming the challenges of shorter and incomplete transcript splicing in non-reference genome species. In addition, SMRT sequencing technology can also use second-generation sequencing data for transcript-specific expression analysis, thereby obtaining more comprehensive annotation information. As one of the third-generation high-flow sequencing techniques, SMRT has been widely applied in the transcript sequence and analysis of various species, such as *Rhynchophorus ferrugineus* [16], *Hyphantria cunea* (Drury) [28], *Agasicles hygrophila* (Selman and Vogt) [15], *Bactrocera dorsalis* (Hendel) [29], *Odontotermes formosanus* (Shiraki) [30], *Sogatella furcifera* (Horvath) [31], and *Rhopalosiphum padi* (L.) [32]. As far as we know, there are currently few reports of full-length transcriptome sequencing of *B. horsfieldi* for gene expression analysis at different stages of development.

By combining short-read transcriptome sequencing (Illumina RNA-seq) and full-length transcriptome sequencing (PacBio Iso-seq), we have obtained for the first time the complete full-length transcriptome of *B. horsfieldi*. This provides favorable conditions for the comprehensive analysis of the transcriptome information of *B. horsfieldi*. Next, we annotated the complete full-length transcriptome, predicted coding sequences (CDS), analyzed simple sequence repeat sequences, and conducted transcription factor analysis. In addition, we analyzed lncRNAs and other splicing events. It is worth mentioning that we have provided a complete full-length transcriptome for the gene study of *B. horsfieldi*, which provides a reference for further analysis of the gene expression profile of *B. horsfieldi* and a valuable resource for future molecular biology research on *B. horsfieldi*.

## 2. Materials and Methods

### 2.1. Sample Collection and Preparation

The *B. horsfieldi* samples used in this study were collected from China West Normal University, Nanchong City, Sichuan Province. We only selected newly hatched *B. horsfieldi* adults that had not fed or reproduced and were in good physical condition. After collecting the adult insects, we separated the samples into female and male adults according to their gender. The female and male adults were then placed separately in stainless steel rearing cages (60 cm × 60 cm × 60 cm) at a room temperature of 25 ± 2 °C, with a relative humidity of 75–80%, and a light cycle of 12 L:12 D in an artificial climate chamber. They were fed with poplar leaves at room temperature for backup. Then, one male and one female adult insect were selected, and their RNA was extracted and mixed. The mixed overall sample was used for subsequent RNA sequencing, with three biological replicates conducted. All collected samples were frozen in liquid nitrogen and stored at a constant temperature of −80 °C for future experimental use.

### 2.2. Library Preparation and SMRT Sequencing

Total RNA samples were isolated using the RNeasy Plus Mini Kit (Qiagen, Valencia, CA, USA). RNA degradation and contamination were examined using 1% agarose gel electrophoresis. The purity of RNA was determined using the Nanodrop (Nanodrop Products, Rockville, MD, USA) by measuring the OD260/280 ratio. The purity of RNA was also assessed using the RNA IQ assay kit with the Qubit Fluorometer (Life Technologies, Carlsbad, CA, USA). The integrity of RNA was measured using the 2100 Agilent Bioanalyzer system (Agilent Technologies, Santa Clara, CA, USA). The purified RNA products were sent to Beijing Novogene Bioinformatics Technology Co., Ltd., Beijing, China for SMRTbellTM library preparation and sequencing. First, the full-length cDNA required for sequencing was synthesized using the SMARTerTM PCR cDNA Synthesis Kit (TaKaRa USA, Inc., Mountain View, CA, USA). Then, high-quality large-scale library amplification products were purified to 1–6 kb using the BluePippin Size Selection System. The selected full-length cDNA underwent damage repair, end repair, and SMRT dumbbell ligation to form the SMRTbell template library. Polymerase was added to the SMRTbell template library. The Iso-Seq library was prepared according to the Isoform Sequencing Protocol (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System

protocol as described by Pacific Biosciences (PN 100-092-800-03). Finally, single-molecule real-time sequencing was performed on the latest third-generation sequencing platform, PacBio RSII.

### *2.3. SMRT Sequencing Data Processing*

The sequences were aligned using SMRT software (version 5.0). The circular consensus sequence (CCS) was obtained from subread BAM files (parameters: min_length 200; max_drop_fraction 0.8; no_polish TRUE; min_zscore-9999; min_passes 1; min_predicted_accuracy 0.8; max_length 18,000) and outputted in CCS.BAM file format. The full-length reads contain a 5′ primer, a 3′ primer, and a poly(A) tail. The integrity of transcripts can be assessed based on the presence of a 5′ primer, a 3′ primer, and a poly(A) tail in CCS reads; the CCS reads are clustered into full-length and non-full-length reads. The iterative clustering for error correction (ICE) clustering analysis was performed using FLNC to obtain consensus isoforms. The quiver algorithm [33] (parameters: hq_quiver_min_accuracy 0.99; bin_by_primer false; bin_size_kb 1; qv_trim_5p 100; qv_trim_3p 30) was then used to perform accuracy correction on the consensus isoforms to identify high-quality isoforms. The full-length transcripts were corrected using LoRDEC software v0.8 [34] and Illumina RNA-seq. Finally, the CD-HIT software v4.5.88 was used to remove redundancy and similar sequences from the high-quality transcripts to obtain non-redundant transcripts.

### *2.4. Functional Annotation of Transcripts*

To obtain comprehensive functional gene information for the transcriptome of *B. horsfieldi* adult males and females, various databases were used to annotate the non-redundant genes. The BLAST software v.2.2.23 [35] was employed with an E-value threshold set at less than $1^{-10}$ to compare the obtained transcripts against the NCBI non-redundant protein (Nr) sequence database [36], NCBI non-redundant nucleotide sequence database, Protein Families Database (Pfam) [37], Clusters of Orthologous Groups of Proteins database (KOG) [38], Swiss-Prot [39], Kyoto Encyclopedia of Genes and Genomes (KEGG) [40], and Gene Ontology (GO) [41]. The Pfam database (https://www.ebi.ac.uk/interpro/, accessed on 25 April 2023) was analyzed using Hmmscan. Then, the Blast2GO v2.5 software was used to perform annotation analysis of GO based on the protein annotation results from the Pfam database.

### *2.5. CDS Prediction*

The ANGEL pipeline is capable of performing ANGEL [42] long-reads and determining the protein-coding sequences (CDS) of full-length complementary deoxyribonucleic acids (cDNAs). In this study, we tested the ANGEL pipeline using protein sequences from *B. horsfieldi* and its closely related species and then performed CDS prediction analysis on the given sequences. Ultimately, transcripts that include both the 5′- and 3′-UTRs (untranslated regions) and the complete CDSs are defined as full-length transcripts.

### *2.6. SSR Analysis*

Transcriptome SSR detection was performed using MISA [43] (version 1.0, default parameters). The minimum repeat times for each unit size were as follows: 1–10, 2–6, 3–5, 4–5, 5–5, and 6–5. For example, for 1–10, at least 10 repeats of a single nucleotide unit were required for detection; for 2–6, at least 6 repeats of a dinucleotide unit were required.

### *2.7. Transcription Factor (TF) Analysis*

Plant transcription factors were predicted using iTAK software v1.6 and animal factors were predicted using the animal TFDB 2.0 database [44]. For species that have already been collected in the database, the transcript factor is directly filtered if it is an Ensembl genid, and for genes that are not Ensembl genids, the BLASTX v2.2.31 screening is carried out through the known sequence of transcription factor proteins of the species and the

database; for species not collected in the database, the identification is based on the Pfam file of the translation factor family, using hmmsearch (Version V3.3.2).

### *2.8. lncRNAs Analysis*

Long non-coding RNAs (lncRNAs), which do not encode proteins, can be screened for coding potential using four tools: CNCI [45], CPC2 [46], Pfam-scan [37], and PLEK [47]. These tools are used to analyze the transcripts and determine whether they have coding potential. Transcripts with coding potential are filtered out, and the intersection of non-coding transcripts identified by the four analysis software programs is taken as the final predicted lncRNA result.

## 3. Results

### *3.1. Transcriptome Analysis Was Performed using Pacbio Sequencing*

A total of 462,134 polymerase readings were obtained using PacBio SMRT sequencing technology. After removing the adapter sequences form the polymerase reads, the remaining sequence is called the subreads. A total of 20,356,793 subreads were obtained from the 46.31 Gb of data, with an average length of 2275 bp. A CCS sequence is a consistency sequence obtained from the subreads in each zero-mode waveguide (ZMW) hole through its comparison correction. After removing adapters and artifacts, a total of 432,091 CCS sequences were generated, with 397,316 full-length reads. By classifying CCS, a total of 34,775 non-full-length (nFL) sequences were identified; 1465 were full-length chimera reads, and 395,851 were full-length non-chimera reads (FLNC), with an average length of 2407 bp. The proportion of the FLNC number to the CCS number is 91.61%. Details of the above data are shown in Table 1. Using the hierarchical n*log(n) algorithm, FLNCs were clustered and redundant, with a total of 39,912 consensus reads, an alkaline base number of 97,462,894, and an average length of 2442 bp. N90 is 1652 bp, and N50 is 2631 bp.

**Table 1.** Statistics of sequencing data and transcript clustering data.

| Data Type | Total Bases (bp) | Total Number | Mean Length | Min_Length | Max_Length | N50 |
|---|---|---|---|---|---|---|
| polymerase read | 47.83 G | 462,134 | 103,503 | - | - | 165,195 |
| subread | 46.31 G | 20,356,793 | 2275 | - | - | 2465 |
| CCS | - | 432,091 | 2527 | 62 | 18,215 | 2674 |
| FLNC | - | 395,851 | 2407 | 86 | 14,739 | 2555 |

In addition, the samples were subjected to replicate sequencing using Illumina Novaseq 6000, generating a total output of 41.16 G raw reads. After filtering, the total clean reads amounted to 38.26 G, as shown in Table 2. The LoRDEC software v0.8 (http://atgc.lirmm.fr/lordec, accessed on 25 April 2023) was used to correct the third-generation PacBio data using the more accurate Illumina reads. After correction, the average length was 2442 bp; N90 was 1652 bp; and N50 was 2630 bp. And, by using the CD-Hit [48] software sequence ratio for clustering and removing redundant and similar sequences, the consensus transcripts were eventually clustered into 15,233 transcript books for subsequent analysis; the comparison of data before and after redundancy removal is shown in Table 3. As shown in Table 4, 67.1% of unigenes have one allelic type, while 32.9% of unigenes have two to ten allelic types. From Figure 1, it can be observed that the length of the genes after redundancy removal is mainly distributed in the range of 1–5 k.

**Table 2.** Quality information for sequencing data output for the samples.

| Sample | Library Name | Raw Reads | Clean Reads | Raw Base (G) | Clean Base (G) | Effective (%) | Error (%) | Q20 (%) | Q30 (%) | GC (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| BHM | FRAS220000762-4r | 48699628 | 45453468 | 7.3 | 6.82 | 93.33 | 0.03 | 97.65 | 93.02 | 36.65 |
| BHM1 | FRAS220000763-4r | 56680480 | 52076758 | 8.5 | 7.81 | 91.88 | 0.03 | 97.74 | 93.17 | 36.37 |
| BHM2 | FRAS220000764-3r | 39332558 | 36671048 | 5.9 | 5.5 | 93.23 | 0.03 | 97.46 | 92.46 | 32.88 |
| BHF | FRAS220000759-4r | 43125796 | 40408496 | 6.47 | 6.06 | 93.7 | 0.03 | 97.87 | 93.26 | 34.54 |
| BHF1 | FRAS220000760-5r | 42298740 | 39126102 | 6.34 | 5.87 | 92.5 | 0.03 | 97.77 | 93.14 | 35.24 |
| BHF2 | FRAS220000761-4r | 44239794 | 41348904 | 6.64 | 6.2 | 93.47 | 0.03 | 97.76 | 93.13 | 34.36 |

BHM stands for *B. horsfieldi* male, and BHF stands for *B. horsfieldi* female.

**Table 3.** Comparison of data before and after redundancy reduction in transcripts.

| Transcript Length Interval | <500 bp | 500–1 kbp | 1 k–2 kbp | 2 k–3 kbp | >3 kbp | Total |
|---|---|---|---|---|---|---|
| Number of transcripts | 21 | 800 | 13,054 | 16,657 | 9380 | 39,912 |
| Number of genes | 3 | 230 | 3930 | 6593 | 4477 | 15,233 |

**Table 4.** The number of genes corresponding to the transcripts. 1/2/3/4/5/6/7/8/9/10: number of genes containing the same number of transcripts.

| Isoform number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Unigene number | 10,226 | 2059 | 887 | 523 | 313 | 223 | 159 | 137 | 86 | 620 |

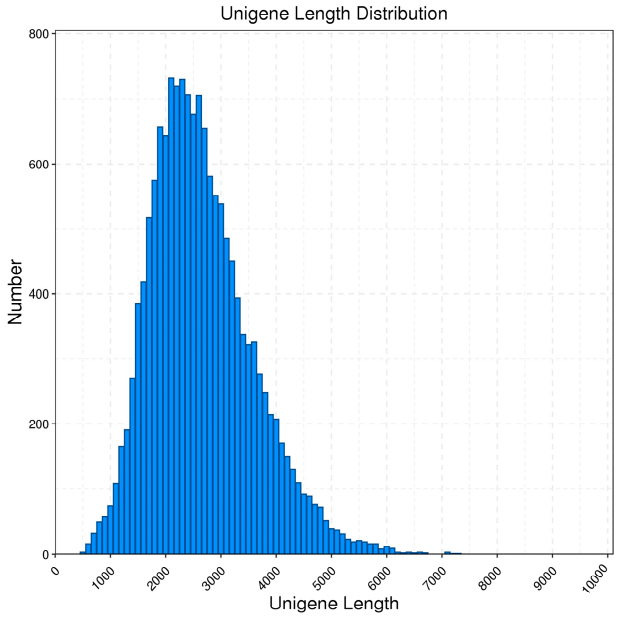


**Figure 1.** Length distribution of unigenes obtained from PacBio Iso-Seq in *B. horsfieldi*.

### *3.2. Functional Annotation of B. horsfieldi*

To comprehensively understand the gene function information of *B. horsfieldi*, we annotated 15,233 transcripts in seven databases, including Swiss-Prot, KOG, GO, NR, NT, Pfam, and KEGG. Summing up, 14,459, 12,619, 14,016, 10,788, 10,783, 11,202, and 10,783 transcripts were annotated in the NR, Swiss-Prot, KEGG, KOG, GO, NT, and Pfam databases, respectively. In addition, at least 14,791 transcripts were annotated in at least one database, and 7057 transcripts were annotated in all databases (Figure 2).

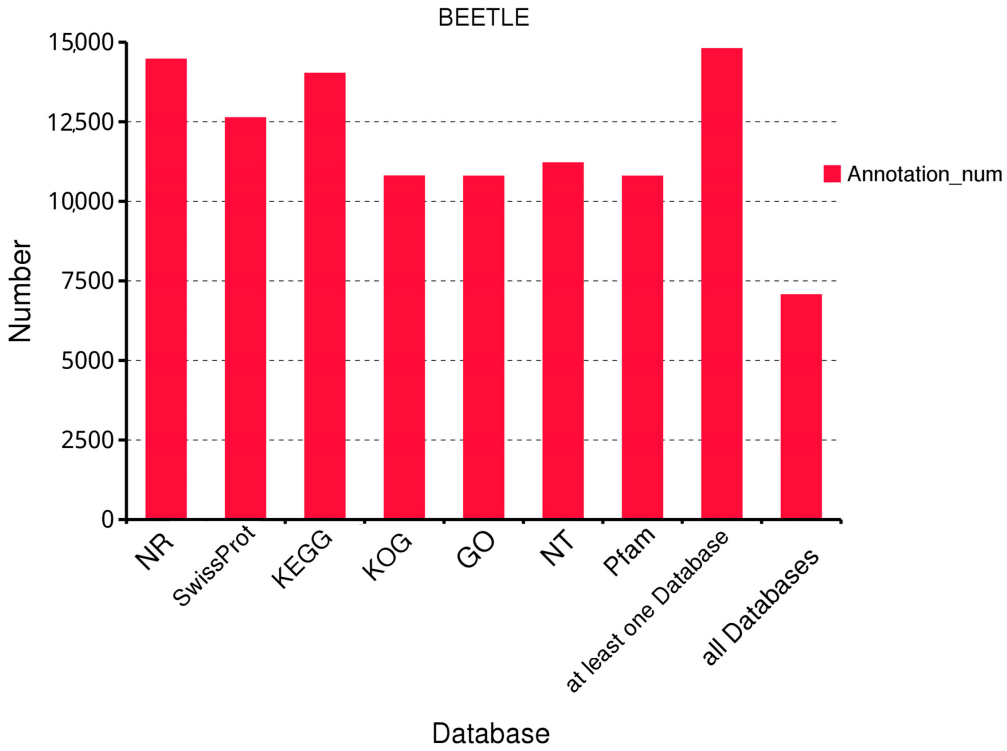


**Figure 2.** Functional annotation of *B. horsfieldi* transcripts in seven databases. NR is a non-redundant protein database. Swiss-Prot is a manually annotated and reviewed protein sequence database. KEGG stands for Kyoto Encyclopedia of Genes and Genomes. KOG is a eukaryotic orthologous gene database. GO is the Gene Ontology. NT is the NCBI's non-redundant nucleotide sequence database. Pfam is a protein family database.

The Non-Redundant Protein Database (NR) is a comprehensive protein database created and maintained by the NCBI. The annotations in the database include species information, making them useful for species classification. Comparing gene sequences with those of closely related species in the NR database can provide information on the similarity and function of genes in a given species. *B. horsfieldi* was compared to the protein sequences of closely related species in the NR database. According to the results shown in Figure 3, a total of 14,459 transcripts were annotated in the NR database. The top six species with the most annotations were Anoplophora glabripennis (83.74%), *T. castaneum* (3.08 %), *Sinocyclocheilus rhinocerous* (1.83%), *Sinocyclocheilus anshuiensis* (1.53%), *Marmota marmota* (1.29%), and *Cyprinus carpio* (1.29%). This provides important evidence for future in-depth understanding of the protein structure and function of *B. horsfieldi*.

KOG is a system for evolutionary relationships based on the complete genome-encoded proteins of bacteria, algae, and eukaryotes. To better analyze the functional aspects of the *B. horsfieldi* transcriptome, this study compared the *B. horsfieldi* transcripts with the KOG database. A total of 10,788 transcripts were successfully annotated (Figure 4), which can be categorized into 26 functional categories. The most annotated functional category was "general function prediction only" with 2149 annotations, accounting for 19.9% of the total annotated transcripts in KOG. The next two categories were "signal transduction mechanisms" with 1782 annotations, representing 16.5%, and "posttranslational modification, protein turnover, chaperones" with 1086 annotations, representing 10%. This provides a basis for analyzing the evolutionary role of this species.

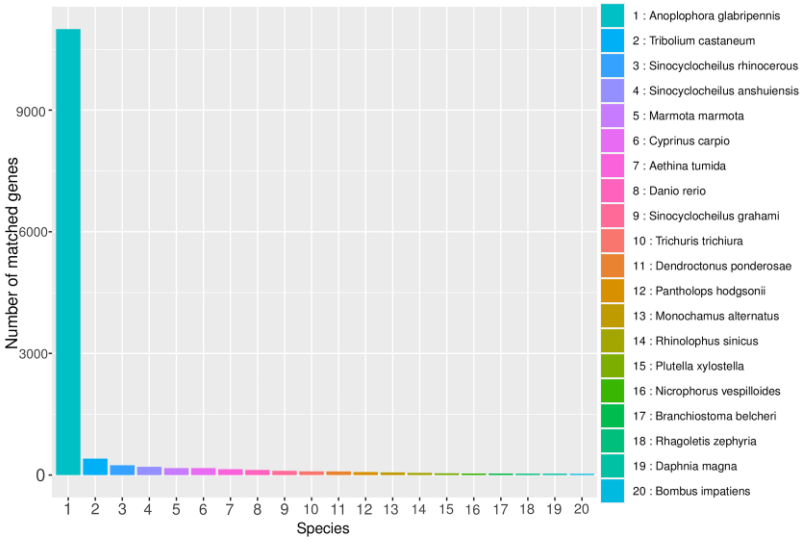


**Figure 3.** NR database annotation of *B. horsfieldi* genes. The X-axis represents the categories of closely related species, and the Y-axis represents the number of transcripts.

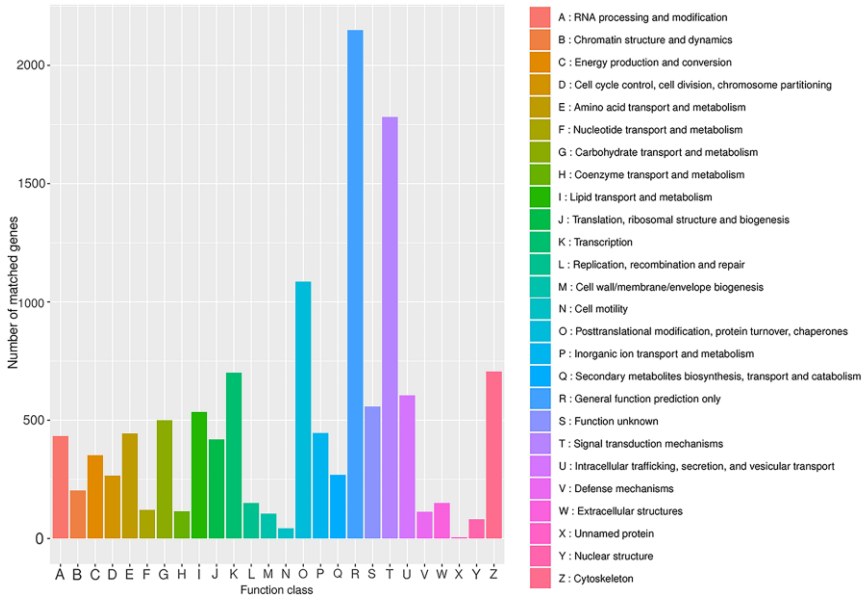


**Figure 4.** KOG annotation of *B. horsfieldi* transcripts. The x-axis represents the KOG category, and the y-axis represents the number of transcripts.

GO stands for Gene Ontology, which is an internationally standardized gene function classification system. By annotating the full-length transcripts of *B. horsfieldi* using the GO database, 10,783 transcripts were successfully classified into three major categories: Cellular Component, Molecular Function, and Biological Process (Figure 5). In the Biological Process category, the cellular process (4897) has the largest proportion, followed by the metabolic process (4588) and the single-organism process (3372). Additionally, we found that some genes were annotated as biological regulation (2037), regulation of biological process (1980), localization (1804), response to stimulus (1381), and signaling (1016) terms. In the Cellular Component category, genes involved in cell (2330), cell part (2330), organelle (1706), membrane (1633), membrane part (1545), and macromolecular complex (1142) were the most abundant. In the Molecular Function (MF) category, binding (7081) and catalytic activity (4565) are the two most abundant subcategories of the annotated transcripts.

The KEGG database is a collection of pathways that systematically analyze the metabolic pathways of genes and compounds in cells, as well as the functions of these gene products. In the

case of *B. horsfieldi*, a total of 10,783 transcripts were annotated in the KEGG metabolic pathways. These transcripts can be categorized into six major classes: Cellular Processes, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism, and Organismal Systems, which consist of 44 subcategories. As shown in Figure 6, Human Diseases, Metabolism, and Organismal Systems are the top three categories with the highest proportion. Specifically, there are a total of 4284 genes involved in the Human Diseases-related pathways, among which 510 genes are predicted to be involved in infectious disease: viral, 727 genes are predicted to be involved in cardiovascular disease, and 961 genes are predicted to be involved in cancers: an overview. Second, among the related pathways in organismal systems, the four pathways with the most abundant genes are the nervous system (368 genes), the immune system (617 genes), the endocrine system (820 genes), and the digestive system (314 genes). Moreover, the entire 2472 annotated genes were involved in the Metabolism pathway. The two most enriched pathways were lipid metabolism (337 genes) and carbohydrate metabolism (462 genes). Meanwhile, in terms of Environmental Information Processing, there are a total of 1465 genes involved in signal transduction, 112 genes involved in membrane transport, and 82 genes involved in signaling molecules and interaction, indicating a large number of genes involved in these processes. Likewise, some genes with fewer numbers were annotated in Cellular Processes and Genetic Information Processing.

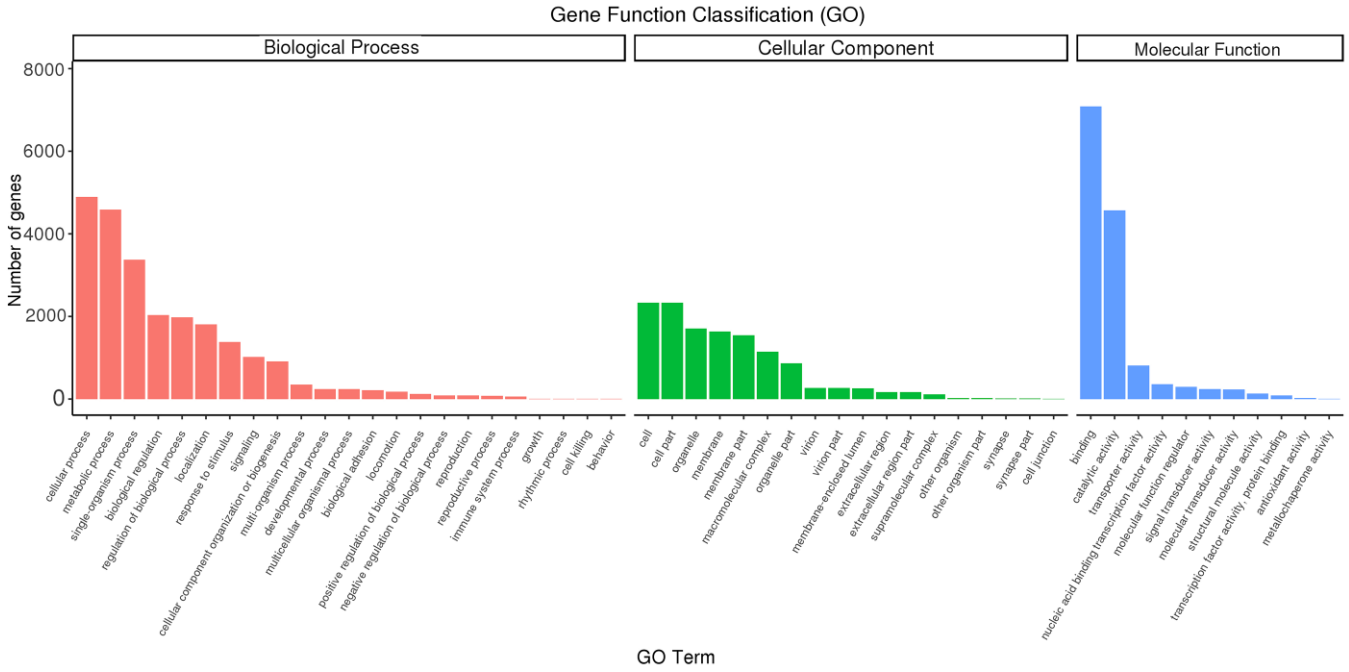


**Figure 5.** Gene Ontology (GO) annotation of *B. horsfieldi* genes. The x-axis represents the GO category, and the y-axis represents the number of transcripts.

### *3.3. CDS Predictions*

CDS (coding sequence) refers to the sequence that encodes a protein product. In this study, the obtained gene fragments were subjected to coding prediction using the prediction software ANGEL. The predicted CDS results are shown in Figure 7, indicating that the length distribution of CDS ranges from approximately 204 to 5820 nt, with the majority falling between 204 and 4000 nt. As the length of the transcripts increases, the number of transcripts decreases.

### *3.4. Identification of Transcription Factors*

Transcription factors (TFs) are a class of proteins that can interact with specific DNA sequences and regulate gene expression [49]. They play a crucial role in various biological processes and are an important component of the transcriptional regulatory system. Using

the existing data of *B. horsfieldi*, we predicted a total of 636 transcription factors, among which Zf-C2H2 (203, 31.92%), ZBTB (86, 13.52%), TF_bZIP (42, 6.60%), and bHLH (41, 6.45%) were the top four transcription factor families (Figure 8). These transcription factors lay the foundation for the regulatory mechanism of *B. horsfieldi*.

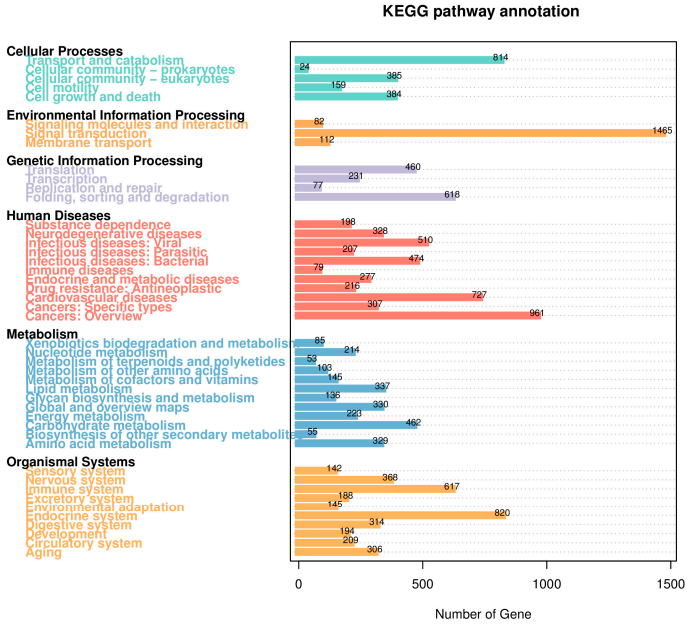


**Figure 6.** KEGG pathway classification of *B. horsfieldi* transcripts. The x-axis represents the number of transcripts, and the y-axis represents the KEGG pathway category.

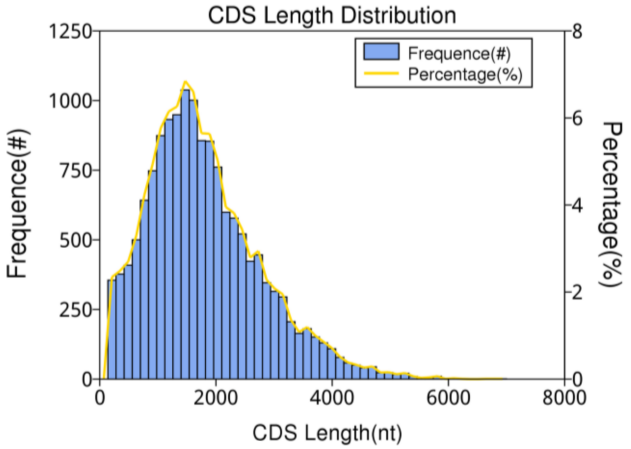


**Figure 7.** The number, percentage, and length distribution of coding sequences of *B. horsfieldi* transcripts.

### *3.5. SSR Analysis*

Simple sequence repeats (SSRs), also known as short tandem repeats or microsatellite markers, are a type of repetitive sequence widely distributed in eukaryotic genomes. They are usually composed of a few nucleotides (1–6) in repeat units with a length of several tens of nucleotides. We used MISA software (version 1.0) with default parameters. A total of 5540 SSR loci were identified, among which the most common type was mono-nucleotide motifs (3789, 68.39%), followed by tri-nucleotides (968, 17.47%), di-nucleotide motifs (748, 13.50%), tetra-nucleotides (24, 0.43%), penta-nucleotides (8, 0.18%), and hexa-nucleotide motifs (3, 0.05%) (Figure 9). The identification of SSR loci establishes the foundation for future assessments of genetic diversity in this species.

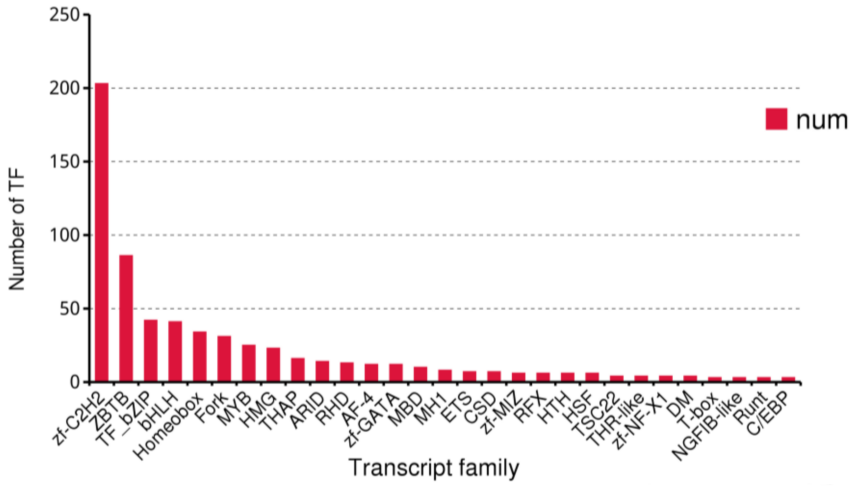


**Figure 8.** The number and family of the top 29 predicted transcription factors by SMRT.

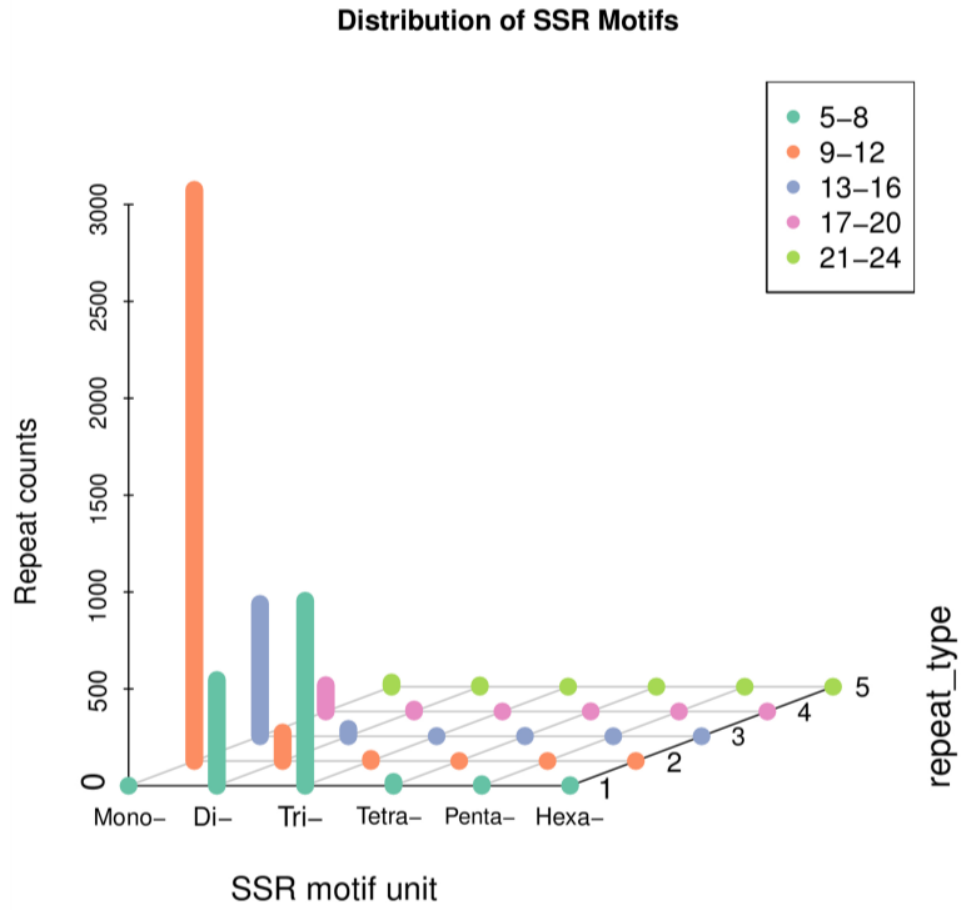


**Figure 9.** Scatter plot of simple sequence repeats in *B. horsfieldi* transcripts.

### *3.6. lncRNA Forecasts*

Long-chain noncoding RNA (lncRNA) is a class of RNA with a length greater than 200 nt that does not encode proteins. Due to the limitations of library construction principles, we could only obtain lncRNA with a polyA tail. We used four tools, CNCI [45], PLEK [47], CPC2 [46], and Pfam [37], to identify unique transcripts without protein-coding potential (i.e., lncRNAs). CNCI identified 1606 lncRNAs, PLEK identified 1155 lncRNAs, CPC2 identified 3221 lncRNAs, and Pfam identified 3836 lncRNAs (Figure 10). Meanwhile, we compared the length distribution of lncRNA and mRNA and found that the average

length of mRNA was slightly shorter than that of lncRNA (Figure 11). There is a certain correlation between lncRNA and mRNA that can be observed.

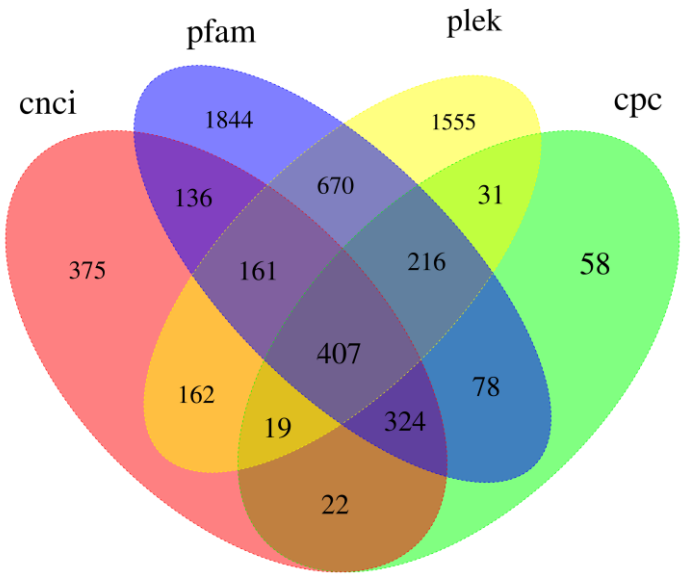


**Figure 10.** Venn diagram of identified lncRNA transcripts in *B. horsfieldi* using PLEK, CNCI, CPC, and Pfam.

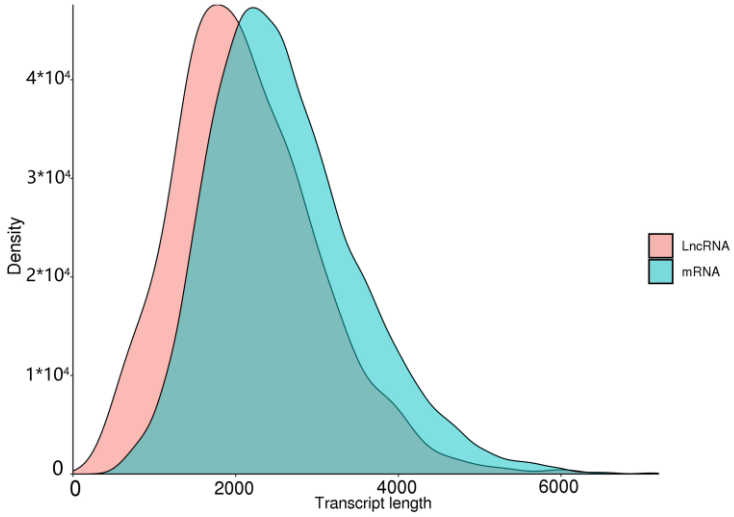


**Figure 11.** Length distribution of lncRNAs and mRNAs in *B. horsfieldi*.

## 4. Discussion

Transcriptome sequencing has become one of the primary means for investigating the mechanisms of gene expression regulation, thanks to the continual advancements and enhancements in sequencing technology of recent years. Compared with second-generation sequencing (short-read transcriptome sequencing), third-generation sequencing (full-length transcriptome sequencing) has overcome many challenges through the technology of obtaining full-length transcripts without PCR amplification through assembly [50,51]. Currently, SMRT sequencing technology has been widely applied in multiple fields, including microbial 16S rRNA gene sequencing, microbial genome assembly, transcriptome sequencing, methylation analysis, and genome resequencing [26]. In addition, third-generation sequencing has made significant contributions in multiple fields. It can accurately reflect the transcriptome information of the sequenced species, detect various alternative splicing forms, and discover more splicing sites and alternative splicing events, thereby improving the accuracy of gene function annotation. Moreover, third-generation sequencing can

discover new functional genes, supplement existing genome annotations, and promote further research. Furthermore, it can accurately analyze fusion genes, homologous genes, superfamily genes, and allelic genes, providing more possibilities for the study of variant genes and evolution. According a research paper by [52], full-length transcriptome sequencing was performed on *Oxya chinensis*, *Acrida cinerea*, *Atractomorpha sinensis*, *Manis javanica* [53], and *R. ferrugineus* [16] using the PacBio RS II platform. The results showed that the full-length transcriptome obtained higher transcript integrity and better quality compared to the transcriptome obtained through second-generation sequencing and could be used for subsequent transcriptome annotation and analysis. Although PacBio Iso-Seq, as a representative of single-molecule real-time sequencing technology, has the advantages of sequencing during synthesis and fast sequencing speed, the original sequencing data has a relatively high error rate (10–15%) [54,55], which needs to be corrected by second-generation sequencing [56].

One advantage of SMRT sequencing is that it can provide a novel understanding of full-length sequences, gene structure, and gene function. The study on common wheat has shown that SMRT sequencing has great potential for genome annotation and gene function research [57]. In the present research, PacBio Iso-Seq and Illumina RNA-Seq were used to sequence and analyze a mixed RNA sample of male and female *B. horsfieldi* adults. A total of 38.26G clean reads were obtained, including 432,091 CCS, of which 395,851 were identified as FLNC with an average length of 2442 bp, N90 of 1652 bp, and N50 of 2631 bp. After CD-Hit redundancy removal, 15,233 transcripts were obtained for subsequent analysis. In *B. horsfieldi*, the amount of transcriptome data obtained through second-generation sequencing is much higher than in other coleopteran insects at different developmental stages, such as *Hypothenemus hampei* (average length of 1609.92 bp, N50 of 2427 bp) [58] and *M. alternatus* (average length of 819 bp, N50 of 1590 bp) [59]. In addition, 97.09% of the transcripts were successfully functionally annotated, which was significantly higher than other second-generation sequencing coleopterans, such as *Holotrichia parallela* [60] and *Henosepilachna vigintioctopunctata* [61].

In gene annotation research, classifying a large number of new transcripts can help obtain gene function information. Updating and collecting homologous protein sets can be used for genome functional annotation of new sequencing data, including those for complex eukaryotes and whole genome evolution studies [38]. Genome sequencing has revealed that most genes involved in core biological functions are conserved across all eukaryotes [41]. In the SMRT sequencing results of the full-length transcriptome of *B. horsfieldi*, a total of 10,783 full-length transcripts were successfully annotated in the GO database. Among them, most of the transcripts were related to cellular processes, followed by cellular components and molecular functions. In addition, 14,016 transcripts were annotated to 40 KEGG pathways, among which signal transduction, cancers (overview), the endocrine system, and transport and catabolism had the most transcripts. Furthermore, the KOG annotation results showed that the transcripts related to "general function prediction only" and "signal transduction mechanisms" were the most abundant. The gene annotation results indicate that the new transcripts may be related to the above functions.

Some studies have indicated that the process of insect host recognition is influenced by various factors, such as the active components present in the volatile substances emitted by host plants [62,63]. These active components, when mixed in specific proportions, can regulate insect behavior. *B. horsfieldi* shows a distinct antennal response to the volatile compounds emitted by the host plants, *Viburnum awabuki* [10] and *Rosa cymosa* Tratt [12]. Although the genome sequence of *B. horsfieldi* has been published [6,13], its transcriptome features and the structures of mRNA and lncRNA have not been deeply analyzed. In the study of *B. horsfieldi*, the discovery and functional research of lncRNA can provide us with a new perspective to help us understand more about the gene regulation mechanism, especially in the aspect of mutual regulation between non-coding RNA and coding proteins. lncRNA is a class of non-protein-coding transcripts with a length of more than 200 nt that plays an important role in regulating gene expression at various levels. Many studies have

shown that non-coding RNAs also play important roles in physiological functions, such as biological growth and development regulation and abiotic stress [64,65]. According to the data in Figure 10, a total of 6058 lncRNAs were predicted, of which 1606, 1155, 3221, and 3836 lncRNA transcripts were predicted by the CNCI, CPC, PLEK, and Pfam, respectively. In addition, non-coding RNAs appear to be more species-specific and may provide more appropriate evidence for the study of biological evolution [66]. In recent years, a large number of lncRNAs have been discovered in insects and animals, such as *Agrilus zanthoxylumi*, *Portunus trituberculatus*, and *Nilaparvata lugens*. Therefore, we have reason to believe that the predicted lncRNAs in this study will help to further reveal the biological characteristics of *B. horsfieldi* [66]. Research on lncRNAs and transcription factors (TFs) can help us better understand the impact of gene regulation on the growth, development, metabolic regulation, and environmental stress of *B. horsfieldi*, which will be helpful for future research and applications in related fields.

## 5. Conclusions

In this study, we used second-generation sequencing (Illumina RNA-seq) to correct the third-generation sequencing (PacBio Iso-Seq) for full-length transcriptome sequencing of *B. horsfieldi* and analyzed the related transcripts. After filtering low-quality sequencing reads, self-correction, and redundancy removal, a total of 15,233 full-length transcripts were successfully generated. We performed gene annotation, CDS prediction, SSR analysis, and TF and lncRNA prediction. These results not only contribute to the improvement of the genome annotation information of *B. horsfieldi* but also provide a valuable foundation for the study of its gene function and for the growth and development of other Coleoptera insects in the future.

**Author Contributions:** Conceptualization, Z.Z.; methodology, X.W. and D.X.; formal analysis, X.W. and Z.Z.; investigation, Z.L.; data curation, Q.L.; writing—original draft preparation, X.W.; writing—review and editing, D.X. and Z.Z.; supervision, Z.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Sichuan Province Science and Technology Support Program (2022NSFSC0986), China West Normal University Support Program (20A007 and 21E040).

**Data Availability Statement:** The data supporting the results are available in a public repository at https://doi.org/10.6084/m9.figshare.22689796.v1, accessed on 25 April 2023.

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

## References

1. Li, J. Biocontrol of *Batocera horsfieldi* (Coleoptera: Cerambycidae) with Parasitoid *Dastarcus helophoroides* (Coleoptera: Bothrideridae). Ph.D. Thesis, Northwest A&F University, Xianyang, China, 2009.
2. Yang, H.; Yang, M.; Yang, W.; Yang, C.; Dong, J.; Shen, Y. A Study on the Spatial Distribution Pattern and the Living-inhabiting Tunnel of the Larvae of *Batocera horsfieldi* (Hope). *J. Sichuan Agr. Univ.* **2010**, *28*, 148–152.
3. Li, A.; Wang, J.; Wang, R.; Yang, H.; Yang, W.; Yang, C.; Jin, Z. MaxEnt modeling to predict current and future distributions of *Batocera lineolata* (Coleoptera: Cerambycidae) under climate change in China. *Écoscience* **2020**, *27*, 23–31. [CrossRef]
4. Wang, S.; Wang, H.; Xia, M.; Dong, S. Regularity of occurrence and control of *Batocera horsfieldi* in walnut trees. *China Fruits* **2004**, *2*, 11–13.
5. Mei, A.; Chen, J.; Wu, G.; Du, X.; Luo, F. Investigation on Yang Tree Pests, Their Occurrence Reasons, and Main Pest Control Measures in the Jianghan Plain. *For. Pest Dis.* **1998**, *02*, 36–38.
6. Yang, H.; Cai, Y.; Zhuo, Z.; Yang, W.; Yang, C.; Zhang, J.; Yang, Y.; Wang, B.; Guan, F. Transcriptome analysis in different developmental stages of *Batocera horsfieldi* (Coleoptera: Cerambycidae) and comparison of candidate olfactory genes. *PLoS ONE* **2018**, *13*, e192730. [CrossRef] [PubMed]
7. Luo, Y. Theory and Techniques of Ecological Regulation of Poplar Longhorned Beetle Disaster in Shelter-Forest. Ph.D. Thesis, Beijing Forestry University, Beijing, China, 2005.
8. Zheng, K.; Wu, S.; Zhang, D.; Wu, J.; Du, Y.; Fan, J. Differences in feeding and oviposition behavior of different populations of *Batocera horsfieldi*. *J. Zhejiang AF Univ.* **2022**, *39*, 159–165.
9. Wu, Y. Identification and Binding Characteristics of Odorant-Binding Proteins in *Batocera horsfieldi*. Master's Thesis, Huazhong Agricultural University, Wuhan, China, 2022.

10. Yang, H.; Yang, W.; Yang, C.; Zhu, T.; Huang, Q.; Han, S.; Xiao, J. Electrophysiological and behavioral responses of the white-striped longhorned beetle, *Batocera lineolata*, to the diurnal rhythm of host plant volatiles of holly, *Viburnum awabuki*. *J. Insect Sci.* **2013**, *13*, 85. [CrossRef]
11. Zhuge, P.P.; Luo, S.L.; Wang, M.Q.; Zhang, G. Electrophysiological responses of *Batocera horsfieldi* (Hope) adults to plant volatiles. *J. Appl. Entomol.* **2010**, *134*, 600–607. [CrossRef]
12. Zhuo, Z.; Jin, Y.; Xu, D.; Liao, W. Electroantennogram responses of *Batocera horsfieldi* (Hope) to the selected volatile components of host plants, *Rosa cymosa* Tratt. and *Rosa multiflora* Thunb. *Glob. Ecol. Conserv.* **2022**, *33*, e1986. [CrossRef]
13. Hu, J.; Xu, D.; Zhuo, Z.; Yang, W.; Yang, H.; Zheng, Y. Analysis of antennal transcriptome and olfaction-related genes of adult *Batocera horsfieldi* (Hope). *Chin. J. Appl. Entomol.* **2019**, *56*, 1037–1047.
14. Hittinger, C.T.; Johnston, M.; Tossberg, J.T.; Rokas, A. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 1476–1481. [CrossRef] [PubMed]
15. Jia, D.; Wang, Y.; Liu, Y.; Hu, J.; Guo, Y.; Gao, L.; Ma, R. SMRT sequencing of full-length transcriptome of flea beetle *Agasicles hygrophila* (Selman and Vogt). *Sci. Rep.* **2018**, *8*, 2197. [CrossRef] [PubMed]
16. Yang, H.; Xu, D.; Zhuo, Z.; Hu, J.; Lu, B. SMRT sequencing of the full-length transcriptome of the *Rhynchophorus ferrugineus* (Coleoptera: Curculionidae). *PeerJ* **2020**, *8*, e9133. [CrossRef]
17. Djebali, S.; Davis, C.A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; et al. Landscape of transcription in human cells. *Nature* **2012**, *489*, 101–108. [CrossRef]
18. Ekblom, R.; Galindo, J. Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* **2011**, *107*, 1–15. [CrossRef] [PubMed]
19. Won, H.I.; Schulze, T.T.; Clement, E.J.; Watson, G.F.; Watson, S.M.; Warner, R.C.; Ramler, E.A.M.; Witte, E.J.; Schoenbeck, M.A.; Rauter, C.M.; et al. De novo Assembly of the Burying Beetle *Nicrophorus orbicollis* (Coleoptera: Silphidae) Transcriptome Across Developmental Stages with Identification of Key Immune Transcripts. *J. Genom.* **2018**, *6*, 41–52. [CrossRef] [PubMed]
20. Guo, Y.; Lü, J.; Bai, C.; Gu, C.; Guo, C. Transcriptome analysis reveals adaptation mechanism of *Tribolium castaneum* (Herbst) (Coleoptera: Tenebrionidae) adults to benzoquinone stress. *J. Stored Prod. Res.* **2023**, *101*, 102083. [CrossRef]
21. Zhang, C.; Wang, H.; Zhuang, G.; Zheng, H.; Zhang, X. Comparative transcriptome analysis of *Callosobruchus chinensis* (L.) (Coleoptera: Chrysomelidae-Bruchinae) after heat and cold stress exposure. *J. Therm. Biol.* **2023**, *112*, 103479. [CrossRef]
22. Wang, G.; Chang, Y.; Guo, J.; Xi, J.; Liang, T.; Zhang, S.; Yang, M.; Hu, L.; Mu, W.; Song, J. Identification and Expression Profiles of Putative Soluble Chemoreception Proteins from *Lasioderma serricorne* (Coleoptera: Anobiidae) Antennal Transcriptome. *Environ. Entomol.* **2022**, *51*, 700–709. [CrossRef]
23. Coghlan, A.; Fiedler, T.; Mckay, S.; Flicek, P.; Harris, T.; Blasiar, D.; Stein, L. nGASP-the nematode genome annotation assessment project. *BMC Bioinform.* **2008**, *9*, 549. [CrossRef]
24. Li, Y.; Fang, C.; Fu, Y.; Hu, A.; Li, C.; Zou, C.; Li, X.; Zhao, S.; Zhang, C.; Li, C. A survey of transcriptome complexity in Sus scrofa using single-molecule long-read sequencing. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **2018**, *25*, 421–437. [CrossRef] [PubMed]
25. Lin, H.; Lin, X.; Zhu, J.; Yu, X.; Xia, X.; Yao, F.; Yang, G.; You, M. Characterization and expression profiling of serine protease inhibitors in the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). *BMC Genom.* **2017**, *18*, 162. [CrossRef] [PubMed]
26. Han, Y.; Yang, Q.; Wang, Q.; An, X.; Liu, Y.; Li, L. Application of single molecule real time sequencing in environmental microorganisms research. *Microbiol. China* **2019**, *46*, 3140–3147.
27. Koren, S.; Schatz, M.C.; Walenz, B.P.; Martin, J.; Howard, J.T.; Ganapathy, G.; Wang, Z.; Rasko, D.A.; Mccombie, W.R.; Jarvis, E.D.; et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **2012**, *30*, 693–700. [CrossRef] [PubMed]
28. Zhang, L.; Tang, X.; Wang, Z.; Tang, F. The transcriptomic response of *Hyphantria cunea* (Drury) to the infection of *Serratia marcescens* Bizio based on full-length SMRT transcriptome sequencing. *Front. Cell. Infect. Microbiol.* **2023**, *13*, 1093432. [CrossRef] [PubMed]
29. Ouyang, H.; Wang, X.; Zheng, X.; Lu, W.; Qin, F.; Chen, C. Full-Length SMRT Transcriptome Sequencing and SSR Analysis of *Bactrocera dorsalis* (Hendel). *Insects* **2021**, *12*, 938. [CrossRef]
30. Kai, F.; Xiaoyu, L.; Jian, L.; Fang, T. SMRT sequencing of the full-length transcriptome of *Odontotermes formosanus* (Shiraki) under *Serratia marcescens* treatment. *Sci. Rep.* **2020**, *10*, 15909.
31. Chen, J.; Yu, Y.; Kang, K.; Zhang, D. SMRT sequencing of the full-length transcriptome of the white-backed planthopper *Sogatella furcifera*. *PeerJ* **2020**, *8*, e9320. [CrossRef]
32. Wang, X.; Xu, X.; Ullah, F.; Ding, Q.; Gao, X.; Desneux, N.; Song, D. Comparison of full-length transcriptomes of different imidacloprid-resistant strains of *Rhopalosiphum padi* (L.). *Entomol. Gen.* **2020**, *41*, 289–304. [CrossRef]
33. Chin, C.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E.; et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **2013**, *10*, 563–569. [CrossRef]
34. Salmela, L.; Rivals, E. LoRDEC: Accurate and efficient long read error correction. *Bioinformatics* **2014**, *30*, 3506–3514. [CrossRef]
35. Pirooznia, M.; Perkins, E.; Deng, Y. Batch Blast Extractor: An automated blastx parser application. *BMC Genom.* **2008**, *9* (Suppl. 2), S10. [CrossRef]

36. Li, W.; Jaroszewski, L.; Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **2002**, *18*, 77–82. [CrossRef] [PubMed]
37. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [CrossRef] [PubMed]
38. Tatusov, R.; Fedorova, N.; Jackson, J.; Jacobs, A.; Kiryutin, B.; Koonin, E.; Krylov, D.; Mazumder, R.; Mekhedov, S.; Nikolskaya, A.; et al. The COG database: An updated version includes eukaryotes. *BMC Bioinform.* **2003**, *4*, 1. [CrossRef] [PubMed]
39. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **2000**, *28*, 45–48. [CrossRef] [PubMed]
40. 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]
41. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
42. Shimizu, K.; Adachi, J.; Muraoka, Y. ANGLE: A sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *J. Bioinform. Comput. Biol.* **2006**, *4*, 649–664. [CrossRef]
43. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [CrossRef]
44. Zhang, H.; Liu, T.; Liu, C.; Song, S.; Zhang, X.; Liu, W.; Jia, H.; Xue, Y.; Guo, A. AnimalTFDB 2.0: A resource for expression, prediction and functional study of animal transcription factors. Nucleic. *Nucleic Acids Res.* **2015**, *43*, D76–D81. [CrossRef] [PubMed]
45. Sun, L.; Luo, H.; Bu, D.; Zhao, G.; Yu, K.; Zhang, C.; Liu, Y.; Chen, R.; Zhao, Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* **2013**, *41*, e166. [CrossRef] [PubMed]
46. Kang, Y.; Yang, D.; Kong, L.; Hou, M.; Meng, Y.; Wei, L.; Gao, G. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **2017**, *45*, W12–W16. [CrossRef] [PubMed]
47. Li, A.; Zhang, J.; Zhou, Z. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinform.* **2014**, *15*, 311. [CrossRef]
48. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
49. Shen, W.; Chen, S.; Gan, Z.; Zhang, Y.; Yue, T.; Chen, M.; Xue, Y.; Hu, H.; Guo, A. AnimalTFDB 4.0: A comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res.* **2022**, *51*, D39–D45. [CrossRef]
50. Schadt, E.E.; Turner, S.; Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **2010**, *19*, R227–R240. [CrossRef]
51. Yang, F.; Chen, T.; Mi, L.; Ma, L.; Xie, Y.; Li, J.; Liu, Z. Current status andprospect of biological research on transcriptome sequencing. *Anim. Husb. Vet. Med.* **2019**, *51*, 133–138.
52. Zhao, L. Analysis of Full-Length Transcriptome and Mitochondrial Transcriptome of Three Orthoptera Insects. Ph.D. Thesis, Shaanxi Normal Univiversity, Xi'an, China, 2018.
53. Jing-E, M.; Hai-Ying, J.; Lin-Miao, L.; Xiu-Juan, Z.; Hui-Ming, L.; Guan-Yu, L.; Da-Ying, M.; Jin-Ping, C. SMRT sequencing of the full-length transcriptome of the Sunda pangolin (*Manis javanica*). *Gene* **2019**, *692*, 208–216.
54. Chaisson, M.J.; Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinform.* **2012**, *13*, 1. [CrossRef]
55. Lv, Y. Variation Analysis of Single-Molecule Real-Time Sequencing Data Based on Deep Learning. Master's Thesis, Beijing University of Chemical Technology, Beijing, China, 2021.
56. Yang, H.; Hu, J.; Wang, Z.; Xu, D.; Zhuo, Z. Using PacBio Iso-Seq to determine the transcriptome of *Rhynchophorus ferrugineus*. *Chin. J. Appl. Entomol.* **2021**, *58*, 655–663.
57. Dong, L.; Liu, H.; Zhang, J.; Yang, S.; Kong, G.; Chu, J.S.C.; Chen, N.; Wang, D. Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. *BMC Genom.* **2015**, *16*, 1039. [CrossRef] [PubMed]
58. Noriega, D.D.; Arias, P.L.; Barbosa, H.R.; Arraes, F.B.M.; Ossa, G.A.; Villegas, B.; Coelho, R.R.; Albuquerque, E.V.S.; Togawa, R.C.; Grynberg, P.; et al. Transcriptome and gene expression analysis of three developmental stages of the coffee berry borer, *Hypothenemus hampei*. *Sci. Rep.* **2019**, *9*, 12804. [CrossRef] [PubMed]
59. Li, H.; Zhao, X.; Qiao, H.; He, X.; Tan, J.; Hao, D. Comparative Transcriptome Analysis of the Heat Stress Response in *Monochamus alternatus* Hope (Coleoptera: Cerambycidae). *Front. Physiol.* **2020**, *10*, 1568. [CrossRef]
60. Jian-Kun, Y.; Shuang, Y.; Shang, W.; Jun, W.; Xin-Xin, Z.; Yan, L.; Jing-Hui, X. Identification of candidate chemosensory receptors in the antennal transcriptome of the large black chafer *Holotrichia parallela* Motschulsky (Coleoptera: Scarabaeidae). *Comp. Biochem. Physiol. Part D Genom. Proteom.* **2018**, *28*, 63–71.
61. Wei, G.; Jing, L.; Mujuan, G.; Shimin, C.; Baoli, Q.; Wen, S.; Chunxiao, Y.; Youjun, Z.; Huipeng, P. De Novo Transcriptome Analysis Reveals Abundant Gonad-specific Genes in the Ovary and Testis of *Henosepilachna vigintioctopunctata*. *Int. J. Mol. Sci.* **2019**, *20*, 4084.

62. Liang, X. Preference of *Batocera horsfieldi* (Hope) for Host of Supplementary Feeding. Master's Thesis, Sichuan Agricultural University, Chengdu, China, 2007.
63. Zhuge, P. The Semiochemicals in Host Location of Longhorn Beetle *Batocera horsfieldi* (Hope). Master's Thesis, Huazhong Agricultural University, Wuhan, China, 2009.
64. Fang, F.; Yao, Y.; Ma, Z. Exploration of the Long Noncoding RNAs Involved in the Crosstalk between M2 Macrophages and Tumor Metabolism in Lung Cancer. *Genet. Res.* **2023**, *2023*, 4512820. [CrossRef]
65. Pedro, J.B.; Howard, Y.C. Long Noncoding RNAs: Cellular Address Codes in Development and Disease. *Cell* **2013**, *152*, 1298–1307.
66. Lou, F.; Han, Z. Full-length transcripts facilitates *Portunus trituberculatus* genome structure annotation. *J. Oceanol. Limnol.* **2022**, *40*, 2042–2051. [CrossRef]