genes

MDPI

Article

A High-Quality Chromosome-Level Genome Assembly of a Snail *Cipangopaludina cathayensis* (Gastropoda: Viviparidae)

Benhe Ma [1,2,†], Wu Jin [3,4,†], Huiyun Fu [1], Bing Sun [5], Su Yang [5], Xueyan Ma [3,4], Haibo Wen [3,4], Xiaoping Wu [2], Haihua Wang [1,*] and Xiaojuan Cao [5,*]

[1] Jiangxi Fisheries Research Institute, Nanchang 330039, China; mabenhe@126.com (B.M.); jxfuhuiyun@163.com (H.F.)
[2] College of Life Science, Nanchang University, Nanchang 330031, China; xpwu@ncu.edu.cn
[3] Key Laboratory of Integrated Rice-Fish Farming Ecology, Ministry of Agriculture and Rural Affairs, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi 214081, China; jinw@ffrc.cn (W.J.); maxy@ffrc.cn (X.M.); wenhb@ffrc.cn (H.W.)
[4] Wuxi Fisheries College, Nanjing Agricultural University, Wuxi 214128, China
[5] Engineering Research Center of Green Development for Conventional Aquatic Biological Industry in the Yangtze River Economic Belt, Ministry of Education, College of Fisheries, Huazhong Agricultural University, Wuhan 430070, China; sunbing931014@163.com (B.S.); yangsu0407@163.com (S.Y.)
* Correspondence: jxswhh@163.com (H.W.); caoxiaojuan@mail.hzau.edu.cn (X.C.)
† These authors contributed equally to this work.

Abstract: *Cipangopaludina cathayensis* (Gastropoda: Prosobranchia; Mesogastropoda; Viviparidae) is widely distributed in the freshwater habitats of China. It is an economically important snail with high edible and medicinal value. However, the genomic resources and the reference genome of this snail are lacking. In this study, we assembled the first chromosome-level genome of *C. cathayensis*. The preliminary assembly genome was 1.48 Gb in size, with a contig N50 size of 93.49 Mb. The assembled sequences were anchored to nine pseudochromosomes using Hi-C data. The final genome after Hi-C correction was 1.48 Gb, with a contig N50 of 98.49 Mb and scaffold N50 of 195.21 Mb. The anchored rate of the chromosome was 99.99%. A total of 22,702 protein-coding genes were predicted. Phylogenetic analyses indicated that *C. cathayensis* diverged with *Bellamya purificata* approximately 158.10 million years ago. There were 268 expanded and 505 contracted gene families in *C. cathayensis* when compared with its most recent common ancestor. Five putative genes under positive selection in *C. cathayensis* were identified (false discovery rate <0.05). These genome data provide a valuable resource for evolutionary studies of the family Viviparidae, and for the genetic improvement of *C. cathayensis*.

Keywords: genome assembly; *Cipangopaludina cathayensis*; comparative genomics

check for updates

Citation: Ma, B.; Jin, W.; Fu, H.; Sun, B.; Yang, S.; Ma, X.; Wen, H.; Wu, X.; Wang, H.; Cao, X. A High-Quality Chromosome-Level Genome Assembly of a Snail *Cipangopaludina cathayensis* (Gastropoda: Viviparidae). *Genes* **2023**, *14*, 1365. https://doi.org/10.3390/genes14071365

Academic Editor: Zhiqiang Han

Received: 1 June 2023
Revised: 21 June 2023
Accepted: 26 June 2023
Published: 28 June 2023

1. Introduction

Viviparidae, an almost globally distributed family of freshwater gastropods, belonging to the class Gastropoda, includes a variety of snail species [1]. In China, according to morphological characteristics, Viviparidae are classified into more than 70 species, and divided into nine genera [2]. Among them, *Bellamya* and *Cipangopaludina* are the most speciose [3]. Recently, *B. purificata*, the largest-size species of the genus *Bellamya*, has been deeply studied at the molecular level [4]. Huang et al. [5] performed a transcriptome and proteome analysis and several shell color-related genes/proteins were identified in *B. purificata*. Jin et al. [6] completed genome sequencing and the chromosome-level genome assembly of *B. purificata*. However, molecular genetics studies on the genus *Cipangopaludina* are lacking.

The mudsnail *C. cathayensis*, belonging to the family Viviparidae, order Mesogastropoda, subclass Prosobranchia, class Gastropoda, and phylum Mollusca, is a freshwater snail that is widely distributed in paddy fields, lakes, marshes, rivers, streams, and ponds in China [1,7]. The snail is an edible snail [8]. It has a high nutritional value, containing a variety of essential amino acids, carbohydrates, minerals, and vitamins [9]. It also has a high medicinal value [10]. The Compendium of Materia Medica states that "mudsnails are beneficial to relieve dampness and heat, quench thirst and sober up, facilitate defecation, and cure beriberi and jaundice". In addition, *C. cathayensis* has many bioactive substances that may be used for tumor and virus suppression [11–13].

Due to its high edible and medicinal values, *C. cathayensis* has become a very important aquatic economic animal in China [14]. In recent years, the annual economic value of the "snail rice noodle" has reached more than 10 billion CNY in China. In 2022, it reached more than 50 billion CNY. However, Jin et al. [6] reported that there was a giant gap between the demand and supply of freshwater Viviparidae snails. Therefore, to meet the needs of consumers, the aquaculture and breeding of freshwater Viviparidae snails, including *C. cathayensis*, have become very urgent.

Currently, genome resources for Viviparidae snails are significantly lacking, where only the *B. purificata* genome is available [6]. However, high-quality genome information is very useful for genome-wide selective breeding and economic trait improvement based on the genome editing of *C. cathayensis*. More recently, the rapid development of sequencing technology has made it easier for people to obtain high-quality genomic data [15]. In this study, we assembled a high-quality genome of *C. cathayensis* by using PacBio long-read sequencing and high-throughput chromosome conformation capture (Hi-C) technology. Meanwhile, a comparative genomic analysis was performed to explore the evolution of *C. cathayensis*. The results obtained here are very beneficial to the breeding, aquaculture, and evolution of *C. cathayensis*.

2. Materials and Methods

2.1. Sample Collection and Sequencing

We performed an initial genome assembly to obtain a preliminary estimate of genome size, heterozygosity, and complexity. The genome DNA from the foot of a female specimen of *C. cathayensis* (Figure 1A), collected from the experimental base (28.7298° N, 115.9689° E) of Jiangxi Fisheries Research Institute, was extracted for genome sequencing using a Cetyltrimethylammonium Bromide (CTAB) method. The Blood & Cell Culture DNA Midi Kit (QIAGEN Q13343) was applied for gDNA extraction and purification. The integrity and concentration of the genomic DNA were examined using 1% agarose gel electrophoresis and a Pultton DNA/Protein Analyzer (Plextech).

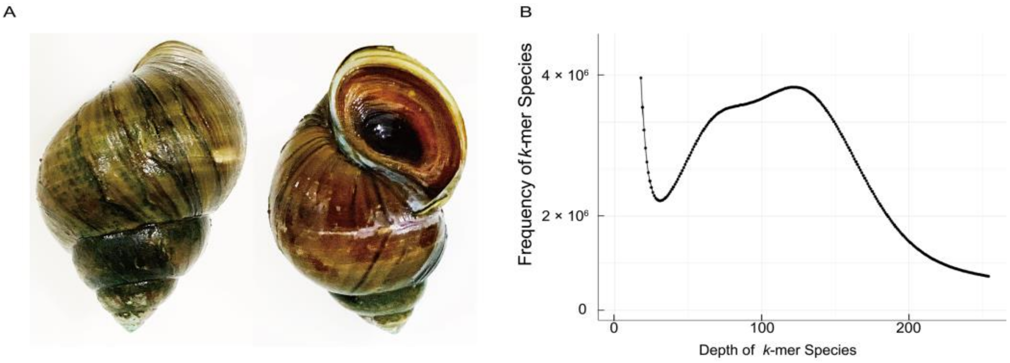

Figure 1. Physical image and genome survey analysis of *Cipangopaludina cathayensis*. (**A**) Front (left) and back (right) photographs of *C. cathayensis*. (**B**) Frequency distribution of k-mer depth and k-mer species.

In the genome survey, the qualified DNA was cut short to 300–500 bp fragments, then terminal repair, addition base A, addition sequence adapter, purification, and PCR amplification were implemented to complete the 350 bp library preparation. Next, we used pair-end sequencing on the MGI DNBSEQ-T7 platform, and fastp with 134.72× coverage (Supplementary Table S1) was used to filter out low-quality reads [16].

Genomic DNA was extracted from the foot of *C. cathayensis* using the QIAamp DNA Mini Kit (QIAGEN). The Agilent 4200 Bioanalyzer (Agilent Technologies, Palo Alto, California) was used to determine the integrity of DNA. Subsequently, g-Tubes (Covaris) and AMPure PB magnetic beads were used for genomic DNA shearing and concentration, respectively. Then, following the method of the Pacific Biosciences SMRT bell template prep kit 1.0, we constructed the SMRT bell libraries. A molecule size of 14–17 kb^2 was selected for each library by Sage ELF, followed by primer annealing and the binding of SMRT bell templates to polymerases with the DNA Polymerase Binding Kit. The Pacific Bioscience Sequel II platform was selected for sequencing with 43.8× coverage (Supplementary Table S1).

To construct the Hi-C library, we collected the foot of *C. cathayensis* and then treated it with 1–3% formaldehyde at room temperature for 20 min. Subsequently, the gDNA was extracted by the modified CTAB method, and then the restriction enzyme of *Mbo* I and biotinylated residues were used to digest the gDNA and repair the 5′ overhang, respectively. One PE library was constructed with a 300 bp insert size using the Hi-C library preparation protocol. Then, sequencing was performed on Illumina NovaSeq platform and the low-quality reads were filtered out using the fastp software (Version 0.19.5; default parameters) [16]. The Hi-C library was constructed and sequenced with 180.24× coverage (Supplementary Table S1) on the illumina Novaseq 6000 platform to accomplish a chromosome-level genome assembly following the method in [17]. The library was sequenced on the illumina Novaseq 6000 platform.

2.2. Genome Size Estimation and Genome Assembly

A k-mer-based method, by making use of Illumina short reads (134.72× coverage) (Supplementary Table S1), was applied to estimate the genome size, heterozygosity, and repeat content in *C. cathayensis* [18]. The PacBio Sequel II sequencing platform was selected for generating long reads with 43.78× coverage and used for genome assembly with HiFiasm software (Version 0.16.1; default parameters) [19].

2.3. Hi-C-assisted Chromosome-Level Assembly

The reads were mapped to a non-redundant genome with a bowtie2 (Version 2.4.5) [20]. using default parameters. We removed the redundancy of the preliminarily assembled genome via purge haplotigs (Version 1.0.4) with default parameters [21]. Using Minimap2 (Version 2.23) and SAMtools (Version 1.9) software, we selected the read pairs that were uniquely aligned to the genome at both ends for genome assembly. Subsequently, 3D-DNA with default parameters was used for chromosome-level genome assembly [22] with default settings. The positions and directions of the small contigs were adjusted with Juicebox (VersionVersion 1.11.08) using default parameters [23] manually based on the degree of contig interaction to form the chromosome.

2.4. Repeat Annotation, Gene Prediction, and Gene Functional Annotation

Before gene prediction, repeat elements of *C. cathayensis* genome were annotated via homology searches and de novo predictions. RepeatMasker (Version 4.0.9) [24] and Repeat ProteinMask (Version 4.1.0) with default parameters were performed to detect and classify the known repetitive elements by comparing sequences to the Repbase database (https://www.girinst.org/repbase/, accessed on 22 December 2022) [25].

Tandem Repeat Finder (Version 4.09) [26] was executed for the tandem repeat elements detection based on sequence features. Long terminal repeat (LTR)_FINDER (Version 1.0.7) was also applied to ab initio predict the repetitive elements [27], and RepeatModeler (Version 1.0.11) [28] was used for de novo prediction based on libraries de novo constructed with LTR_FINDER (Version 1.0.7) using default parameters, PILER [29], and RepeatScout (Version 1.0.6) [30].

Based on Rfam (Version 14.0) using “cmscan—rfam—nohmmonly evalue 0.01” parameters [31] and miRbase [32] databases, Infernal (Version 1.1) [33] was applied to predict the ribosome RNAs and micro-RNAs, respectively. tRNAscan-SE (Version 1.3.1) with default parameters was used for predicting transfer RNAs [34].

For the prediction of protein-coding genes, homology-based prediction, ab initio prediction, and RNA sequencing (RNA-seq)-based prediction were used. For homology-based annotation, the protein sequences of *Aplysia californica*, *Biomphalaria glabrata*, *Haliotis rubra*, and *Plakobranchus ocellatus* were downloaded from NCBI and aligned to a genome sequence using BLASTN ($e \leq 1 \times 10^{-5}$). Homologous sequences were then accurately aligned to corresponding matching proteins using GENEWISE (Version 2.4.0) [35]. The genome sequence was also aligned to a homologous single-copy gene database of Benchmarking Universal Single-Copy Orthologs (BUSCO Version 5.3.1), which was 4analicu_odb10 [36], to find the homologous regions. Maker (Version 2.31.10) with default parameters [37] and HiCESAP (Gooalgene Co., Ltd., Wuhan, China, https://www.gooalgene.com/, https://www.girinst.org/repbase/ accessed on 23 October 2022) were employed to merge all the evidence above and the redundancies were filtered out. De novo genome prediction was performed using AUGUSTUS (Version 3.3.2) and Genscan (Version 1.0). In the meantime, homology annotation was carried out with *A. californica*, *P. ocellatus*, *B. glabrata*, and *H. rubra*, using tblastn (Version 2.11.0+) with an e-value of 0.01. To improve the results of gene prediction, a transcriptome was used for data alignment by utilizing the HISAT2 (Version 2.0.5) software. StringTie (Version 2.1.4) and TransDecoder (Version 5.5.0) were used for transcript prediction and coding region prediction, respectively. Finally, all the annotation results from the above three methods were integrated into a complete annotation file by using Maker (Version 2 2.31.10) with default parameters. The BUSCO (Version 4.1.4) analysis results indicated that it was a high-quality annotation set.

For the gene functionality annotation, some databases of NCBI, InterPro, TrEMBL, SwissProt, and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used through-Blastn and Blastx ($e \leq 1 \times 10^{-5}$), and Gene Ontology (GO) annotation was performed using Blast2GO (Version 5.2.5) [38].

2.5. Comparative Genomic Analyses and Selection Analysis

OrthoMCL (verison 2.0.9) [39], using "-I 1.5" parameters, was used to detect orthologous groups by retrieving the protein sequence of *Achatina fulica* (PRJNA511624), *B. purificata* (PRJNA818874), *B. glabrata* (ASM45736v1), *Crassostrea gigas* (GCF_902806645.1), *Elysia chlorotica* (GCA_003991915.1), *Lingula anatine* (GCF_001039355.2), *Lottia gigantean* (GCF_000327385.1), *Mytilus galloprovincialis* (GCA_900618805.1), *Patinopecten yessoensis* (GCF_002113885.1), and *Pomacea canaliculata* (GCF_003073045.1). The single-copy orthologous genes shared by all species were multiple-aligned using MUSCLE (Version 5) [40] with default parameters. A phylogenetic tree was constructed using RaxML (Version 8.2.12) [41] based on multiple sequence alignment using "-f a -N 100 m GTRGAMMA" parameters. The divergence time was evaluated using the MCMCTree program of the PAML package using "clock = 3; model = 0" parameters [42].

Phylogenetic relationships were reconstructed from 11 species by using single-copy orthologues. Moreover, gene expansion and contraction analyses were conducted using I (Version 4.0), using default parameters [43]. CAFE simulates gene gains and losses in user-specified phylogenetic trees by birth and death processes. It can calculate the transfer rate of gene family size from parent to child nodes, and infer the gene family size of ancestral species. The gene family size distributions were generated using this model. It can provide a basis for assessing the significance of the observed differences in family size between taxa. We estimated differences across the whole tree using a single birth/death parameter. The significant genes were identified by setting the cutoff *p*-value to 0.05. We performed GO and KEGG enrichment analyses for a better understanding of the biological functions of these genes and the genes of *C. cathayensis* with GO and KEGG annotations used as the background values, respectively. Terms with an enrichment-adjusted *p* value ≤ 0.05 were chosen for further analysis.

The program CODEML (Version4.9) using "branch-site model:A:model = 2, NSsites = 2, fix_omega = 1, omega = 1.0, model = 2, NSsites = 2, fix_omega = 0" parameters of PAML was used for positive selection gene (PSG) identification, and PSGs were also chosen for enrichment analysis. We used the CodeML module in PAML to detect positive selection pressures acting on protein-coding sequences. Firstly, genes were selected from the single-copy gene families. Then, multiple sequence alignment of the gene protein sequences from each single-copy gene family was performed using MAFFT software, and, after that, the results were reversed to the multiple sequence alignment results of CDS. The target species was the foreground branch, and the other species was the background branch. Based on two models, including Model A (assuming the foreground branch ω was under positive selection, $\omega > 1$) and null mode (no sites were allowed to have ω values greater than 1), the likelihood values were calculated, respectively. The likelihood ratio tests (LRTs) were performed on the above likelihood values using the chi2 program in PAML, and the significant difference results were obtained after adjusting the p value (FDR < 0.05). Based on the Bayes empirical Bayes (BEB) method, the posterior probability of a site was obtained, which was considered as being positively selected (significant positively selected genes were generally more than 0.95).

3. Results and Discussion

3.1. Initial Characterization of *C. cathayensis* Genome

A total of 191.31 Gb clean data were obtained for estimating the genome size of *C. cathayensis*. Using k-mers (K = 17) analysis, an estimated genome size of 1409 Mb was obtained (Figure 1B). Supplementary Table S1 shows the sequencing data information of the *C. cathayensis* genome.

3.2. Genome Assembly and Assessment

The 64.84 Gb long reads (Supplementary Table S1) were assembled using HiFiasm (Version 0.16.1) with default parameters, followed by polishing; after that, the redundancy and haplotigs were eliminated, which produced an assembly that was 1.48 Gb in size (Supplementary Table S2). The length of the genome was in accordance with the one estimated using K-mer analysis. The total number of contigs was 40, and N50 reached 98.49 Mb. The genome sizes of other gastropod species including *B. purificata*, *P. canaliculata*, *B. glabrata*, *A. fulica* and *L. gigantean* are between 359 Mb and 2.12 Gb [6,44–47], which may mean that they drive different genetic mechanisms.

Several genome pieces with a step size of 1 kb were randomly selected and mapped to the NT database (Nucleotide Sequence database), and more than 80% of these pieces could be aligned to the genomes of several shellfish. The BUSCO analysis showed that 94.65% of the complete BUSCO genes (the number was 903) were found in this assembly, including 94.03% for complete and single-copy BUSCO (the number was 897) and 0.63% for complete and duplicated BUSCO (the number was 6) (Supplementary Table S3). A Circos plot of the assessment is shown in Figure 2. These results indicate that the genome of *C. cathayensis* was assembled with high quality.

With the application of the LACHESIS software (Version 0.1.19), 266.94 Gb clean data were used and 99.99% of assembled sequences were anchored into nine pseudochromosomes (Supplementary Table S4). The nine pseudochromosomes can be clearly distinguished from the Hi-C heatmap and the internal interaction was very intense (Figure 3). Moreover, the Hi-C heatmap shows the same results as its karyotypes of *C. cathayensis* (Heude; 2n = 18), and the centromere region revealed by its C-banding is generally a high repeat region in the genome [48]. These results indicated that the nine pseudochromosomes had a high anchoring quality. However, the sex chromosomes are currently unknown, and a subsequent analysis will be conducted through resequencing methods. Finally, the final assembly resulted in high quality genome of 1.48 Gb, with a contig N50 of 98.49 Mb and a scaffold N50 of 195.21 Mb (Supplementary Table S2), which was good for annotation.

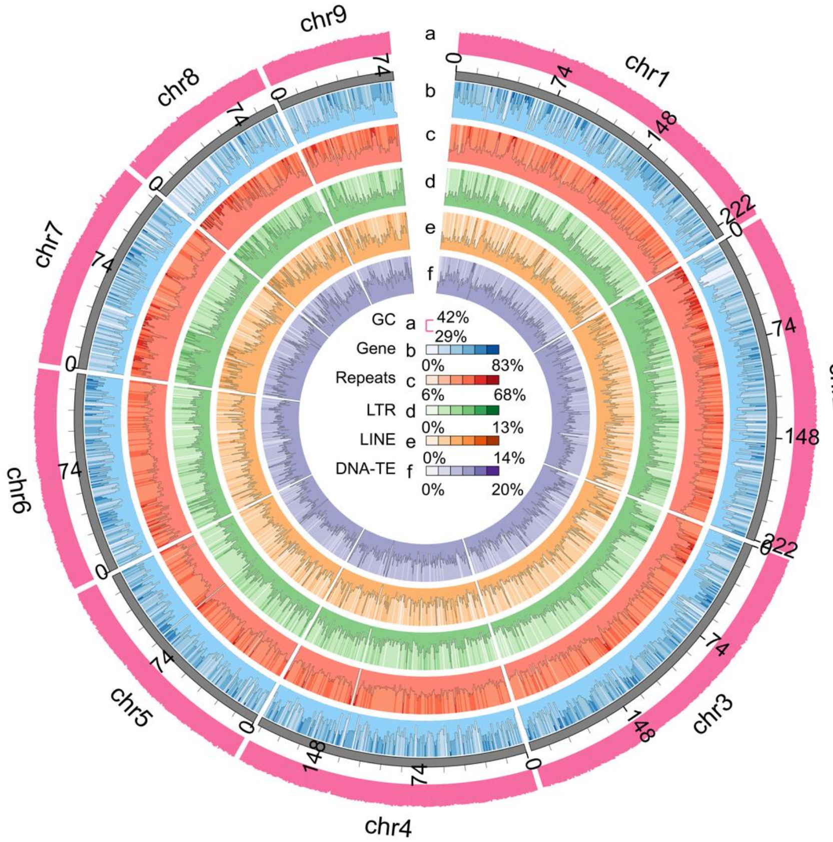

Figure 2. Genome characteristics of *Cipangopaludina cathayensis*. From the outer circle to the inner circle, (**a**) GC content of the genome, (**b**) gene distribution, (**c**) repeats, (**d**) long terminal repeat (LTI (**e**) long interspersed nuclear elements (LINE), (**f**) DNA-TE. The height of the bar is proportional to the number of items mapped to each genomic position.

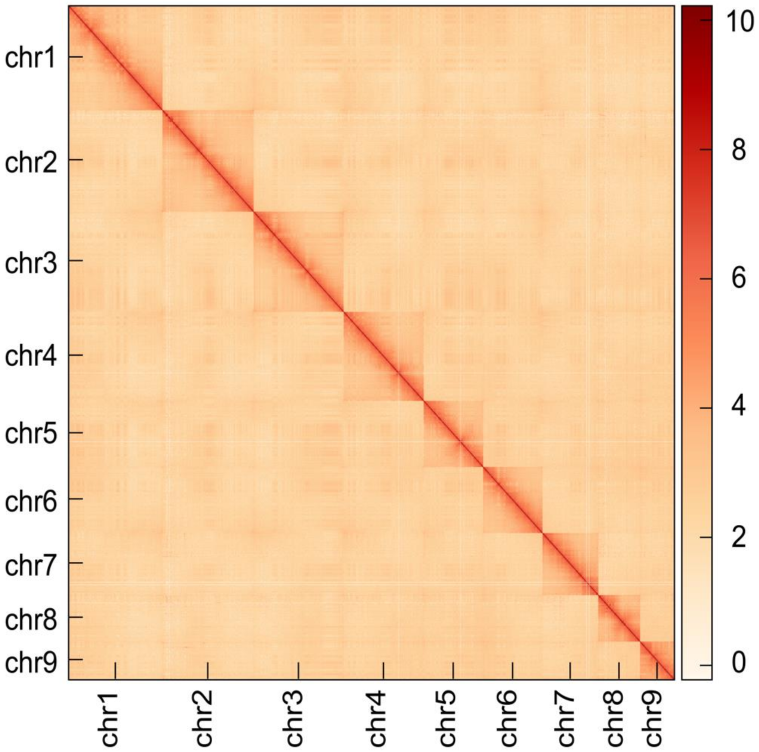

Figure 3. Genome-wide Hi-C heatmap of *Cipangopaludina cathayensis*. The blocks represent the nine pseudochromosomes. The color bar illuminated the contact density from white (**low**) to red (**high**).

3.3. Gene Structure and Function Annotations

A total of 797,589,608 bp repeat sequences were annotated, accounting for 53.83% of the total genome sequence (Supplementary Table S5). The proportion was approximately equal to that of the genome survey, but higher than that of *B. purificata* (47.93%). The predominant repeat elements were repeat DNAs (8.83%), LTR (4.45%), and LINE (4.07%) (Supplementary Table S6). Overall, 22, 702 protein-coding genes were predicted (Supplementary Table S7). The average gene length was 25, 375 bp. A total of 18,576 genes, which accounted for 81.83% of all predicted genes, were annotated (Supplementary Table S8). We successfully annotated 68 miRNAs, 208 tRNAs, 135 rRNAs, and 128 snRNAs for non-coding RNA predictions (Supplementary Table S9). A BUSCO evaluation of the predicted genome annotation revealed that 908 orthologus genes (accounting for 95.2%) were matched, including 904 (94.8%) complete single-copy BUSCOs and 4 (0.4%) complete duplicated BUSCOs (Supplementary Table S10).

3.4. Comparative Genomics

A comparative genomic analysis of *C. cathayensis* and ten other mollusk species revealed a total of 15,083 gene families and 92 single-copy genes. In the genome of *C. cathayensis*, a total of 22,702 genes were clustered into 19,316 gene families, including 191 unique families. The average gene number per family in *C. cathayensis* was 1.281. For the genomes of other species, the average gene number per family ranged from 1.205 (*B. purificata*) to 3.994 (*M. galloprovincialis*) (Supplementary Table S11).

The phylogenetic relationships reconstructed from 11 species by using 92 single-copy orthologues confirmed that *B. purificata* is the closest sister lineage of *C. cathayensis* (Figure 4A). Compared to the results of phylogenetic relationships, constructed from Jin et al. [6], we further clarified the evolutionary relationships and positions among the three species of snails (*C. cathayensis*, *B. purificata*, and *P. canaliculata*). The divergence time was estimated using the MCMCTree program, which indicated that the divergence time between *C. cathayensis* and *B. purificata* was estimated to be 158.10 million years ago (Ma), while the divergence time between the former two and *P. canaliculata* was estimated to be 348.0 million years ago (Ma). So, the Viviparidae family was more closely related to the *P. canaliculate* and the viviparous snails may have evolved from oviparous snails (Figure 4A). So far, research on the phylogenetic relationships among snails (Caenogastropoda) remains rare. The complete mitochondrial genomes of eight viviparid snails including *C. ussuriensis*, *C. dianchiensis*, *C. chinensis*, *Viviparus chui*, *Margarya melanioides*, *Margarya monodi*, *B. aeruginosa*, and *B. quadrata* were sequenced to help explore the phylogenetic relationships of caenogastropod snails [3]. It was revealed that some *Cipangopaludina* species (*cathayensis* and *dianchiensis*) should be renamed to be in the genus *Margarya*, and that the *Cipangopaludina* was more closely related to *Margarya* and *Bellamya*. Similarly, our results confirmed the near origin of *Cipangopaludina* and *Bellamya*. Supplementary Table S12 shows that the genome size and GC content of *C. cathayensis* and *B. purificata* were similar.

By comparing the genome of *C. cathayensis* with its most recent common ancestor, we found 268 expanded and 505 contracted gene families in *C. cathayensis* (Figure 4B). Supplementary Table S13 shows the KEGG enrichment analysis results of the expanded genes. These genes (including *slc4a2*, *fgfr, fgfr2*, *fgfr3*, *matk*, *lcp2*, *ctsk*, *sele*, *gstm2* and *gstm3*) were mainly enriched in pathways of salivary secretion, pancreatic secretion, gastric acid secretion, cancer, autophagy-animal, cell adhesion molecules, bile secretion, toll-like receptors, osteoclast differentiation, and platinum drug resistance.

Figure 4C,D show the GO and KEGG enrichment of the positive selective genes, respectively. Five putative genes that appeared to be positively selected in *C. cathayensis* were identified (false discovery rate (FDR) <0.05, Supplementary Table S14). The positively selected genes (include *gart*, *jarid2*, *rpe*, *kmt5a* and *b3galt6*) were mainly enriched in pathways of purine metabolism, carbon metabolism, regulation of the pluripotency of stem cells, biosynthesis of amino acids, antifolate resistance, lysine degradation, glycosaminoglycan biosynthesis (chondroitin sulfate/dermatan sulfate), pentose and glucuronate interconversions, and pentose phosphate and glycosaminoglycan biosynthesis (heparan sulfate).

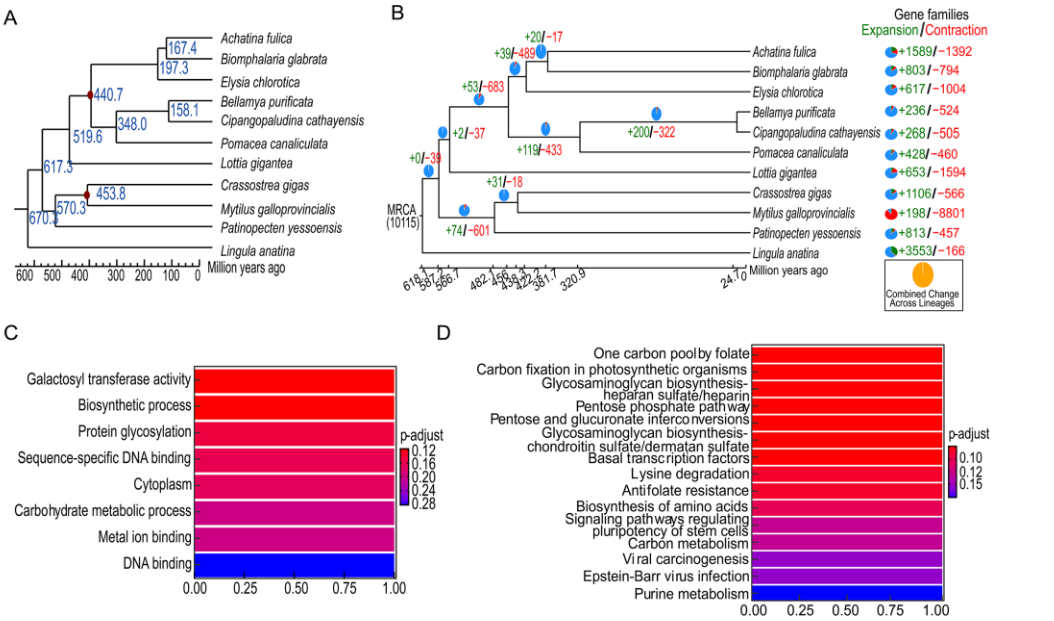

Figure 4. Comparative genomic analysis. (**A**) Estimates of species divergence times. The number of node positions represents the divergence time of a species or its ancestors. The red dot is the calibration point. (**B**) Numbers of gene families for expansion and contraction in *C. cathayensis*. The green numbers and the red numbers indicate the number of gene family members that have expanded and contracted during the evolution of the species, respectively. MRCA: Most Recent Common Ancestor. (**C**) GO enrichment of positively selected genes. (**D**) KEGG enrichment of positively selected genes.

In *C. cathayensis*, we found an expansion gene *slc4a2* and a positive gene *b3galt6*, which were, respectively, related to digestion [49] and bone development [50]. These findings may suggest that *C. cathayensis* has strong digestive and shell formation abilities in order to adapt well to various environments.

4. Conclusions

In conclusion, we assembled the first high-quality, chromosome-level genome of *C. cathayensis*. The assembled genome was 1.48 Gb, including nine chromosomes, with a contig N50 of 98.49 Mb and a scaffold N50 of 195.21 Mb. The genome assembly and annotation are of great importance for the genetic improvement of *C. cathayensis* and lay a strong foundation for evolutionary studies of the family Viviparidae.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/genes14071365/s1, Table S1: Statistics for the sequencing data of *C. cathayensis* genome. Table S2: Genome assembly results of *C. cathayensis*. Table S3: BUSCO analysis results of *C. cathayensis* genome. Table S4: Statistics of Hi-C assembly results of *C. cathayensis*. Table S5: Statistics of repetitive sequences in *C. cathayensis* genome. Table S6: Statistics of transposable elements for *C. cathayensis* genome. Table S7: Statistics of gene predictions in *C. cathayensis* genome. Table S8: Summary of functional annotations for predicted genes. Table S9: Statistics of none-coding RNA annotation of *C. cathayensis* genome. Table S10: BUSCO analysis results of *C. cathayensis* genome annotation. Table S11: Statistical results of gene family clustering. Table S12: Comparison of the sequencing data between *C. cathayensis* and *B. purificata*. Table S13: All DEGs were mapped to 83 KEGG pathways, and the number of unigenes in different pathways ranged from 1 to 42. Table S14: List of positive selective genes in *C. cathayensis* (FDR < 0.05).

Author Contributions: B.M. and W.J. conceived and managed the project. H.F., B.S., S.Y., X.M. and H.W. (Haibo Wen) collected the sequencing samples. H.W. (Haihua Wang), X.C. and X.W. contributed to the revision of the manuscript.All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key Research and Development Program of China (No. 2022YFD2400700), Joint Breeding and Industrialization Promotion and Application Project of *C. cathayensis* in Jiangxi Province (No. 2022yyzygg-05), and the Central Public-interest Scientific Institution Basal Research Fund, Freshwater Fisheries Research Center, CAFS (NO. 2021JBFM07). We thank Qi Liu at the Wuhan Onemore-tech Co., Ltd. for his assistance with genome sequencing.

Institutional Review Board Statement: This study was carried out according to the regulations of the Jiangxi Fisheries Research Institute, China (2022AA1005020008).

Data Availability Statement: The raw sequencing reads of *C. cathayensis* genome were submitted to NCBI and GSA under PRJNA913660 and CRA009296, respectively. The other data can be obtained by contacting the corresponding authors.

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

References

1. Yang, H.; Zhang, J.E.; Luo, H.; Luo, M.; Guo, J.; Deng, Z.; Zhao, B. The complete mitochondrial genome of the mudsnail *Cipangopaludina cathayensis* (Gastropoda: Viviparidae). *Mitochondrial DNA Part A DNA Mapp. Seq. Anal.* **2016**, *27*, 1892–1894. [CrossRef]
2. Liu, Y.; Zhang, W.; Wang, Y.; Wang, E. *Economic Fauna Sinica of China, Freshwater Mollusks*; Science Press: Beijing, China, 1979.
3. Wang, J.G.; Zhang, D.; Jakovlić, I.; Wang, W.M. Sequencing of the complete mitochondrial genomes of eight freshwater snail species exposes pervasive paraphyly within the Viviparidae family (Caenogastropoda). *PLoS ONE* **2017**, *12*, e0181699. [CrossRef] [PubMed]
4. Jin, W.; Cao, J.Y.; Ma, C.; Ma, X.Y.; Lv, G.H.; Wen, H.B. Genetic diversity and genetic differentiation analysis of *Bellamya purificata* in eleven populations based on the microsatellite makers. *Freshw. Fish.* **2022**, *52*, 16–21. (In Chinese) [CrossRef]
5. Huang, S.Q.; Jiang, H.J.; Zhang, L.; Gu, Q.H.; Wang, W.M.; Wen, Y.H.; Luo, F.; Jin, W.; Cao, X. Integrated proteomic and transcriptomic analysis reveals that polymorphic shell colors vary with melanin synthesis in *Bellamya purificata* snail. *J. Proteom.* **2021**, *230*, 103950. [CrossRef] [PubMed]
6. Jin, W.; Cao, X.J.; Ma, X.Y.; Lv, G.H.; Xu, G.C.; Xu, P.; Sun, B.; Xu, D.P.; Wen, H.B. Chromosome-level genome assembly of the freshwater snail *Bellamya purificata* (Caenogastropoda). *Zool. Res.* **2022**, *43*, 683–686. [CrossRef]
7. Jiang, J.Y.; Li, W.H.; Wu, Y.Y.; Cheng, C.X.; Ye, Q.Q.; Feng, J.X.; Xie, Z.X. Effects of cadmium exposure on intestinal microflora of *Cipangopaludina cathayensis*. *Front Microbiol.* **2022**, *13*, 984757. [CrossRef]
8. Zhang, G.; Yin, D.; He, T.; Xu, Y.; Ran, S.; Zhou, X.; Tian, X.; Wang, Y. Mercury Bioaccumulation in Freshwater Snails as Influenced by Soil Composition. *Bull. Environ. Contam. Toxicol.* **2021**, *106*, 153–159. [CrossRef]
9. Guo, Z.F.; Ai, H. Research progress of Chinese Cipangopaludina cathayensis and its bioactive constituent. *Food Res. Dev.* **2015**, *36*, 132–134. (In Chinese) [CrossRef]
10. Bhattacharya, S.; Chakraborty, M.; Bose, M.; Mukherjee, D.; Roychoudhury, A.; Dhar, P.; Mishra, R. Indian freshwater edible snail Bellamya bengalensis lipid extract prevents T cell mediated hypersensitivity and inhibits LPS induced macrophage activation. *J. Ethnopharmacol.* **2014**, *157*, 320–329. [CrossRef]
11. Wang, C.; Liu, J.; Huang, Y.; Zhang, X. In vitro polysaccharide extraction from *Cipangopaludina cathayensis* and its pharmacological potential. *J. Environ. Biol.* **2016**, *37*, 1069–1072.
12. Zhao, T.; Xiong, J.; Chen, W.; Xu, A.; Zhu, D.; Liu, J. Purification and Characterization of a Novel Fibrinolytic Enzyme from *Cipangopaludina Cahayensis*. *Iran. J. Biotechnol.* **2021**, *19*, e2805. [CrossRef]
13. Dhiman, V.; Pant, D. Human health and snails. *J. Immunoass. Immunochem.* **2021**, *42*, 211–235. [CrossRef]
14. Wu, Y.Y.; Cheng, C.X.; Yang, L.; Ye, Q.Q.; Li, W.H.; Jiang, J.Y. Characterization of Gut Microbiome in the Mud Snail *Cipangopaludina cathayensis* in Response to High-Temperature Stress. *Animals* **2022**, *12*, 2361. [CrossRef]
15. Li, K.; Jiang, W.; Hui, Y.; Kong, M.; Feng, L.Y.; Gao, L.Z.; Li, P.; Lu, S. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. *Mol. Plant* **2021**, *14*, 1745–1756. [CrossRef]
16. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
17. Rao, S.S.; Huntley, M.H.; Durand, N.C.; Stamenova, E.K.; Bochkov, I.D.; Robinson, J.T.; Sanborn, A.L.; Machol, I.; Omer, A.D.; Lander, E.S.; et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **2014**, *159*, 1665–1680. [CrossRef]
18. Liu, B.; Shi, Y.; Yuan, J.; Hu, X.; Zhang, H.; Li, N.; Li, Z.; Chen, Y.; Mu, D.; Fan, W. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv* **2013**, arXiv:1308.2012.
19. Cheng, H.; Concepcion, G.T.; Feng, X.; Zhang, H.; Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **2021**, *18*, 170–175. [CrossRef]
20. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef]
21. Roach, M.J.; Schmidt, S.A.; Borneman, A.R. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **2018**, *19*, 460. [CrossRef]

22. Burton, J.N.; Adey, A.; Patwardhan, R.P.; Qiu, R.; Kitzman, J.O.; Shendure, J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **2013**, *31*, 1119–1125. [CrossRef] [PubMed]
23. Durand, N.C.; Shamim, M.S.; Machol, I.; Rao, S.S.; Huntley, M.H.; Lander, E.S.; Aiden, E.L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **2016**, *3*, 95–98. [CrossRef] [PubMed]
24. Tarailo-Graovac, M.; Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinform.* **2009**, *4*, 4.10.1–4.10.14. [CrossRef] [PubMed]
25. Jurka, J.; Kapitonov, V.V.; Pavlicek, A.; Klonowski, P.; Kohany, O.; Walichiewicz, J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **2005**, *110*, 462–467. [CrossRef] [PubMed]
26. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, *27*, 573–580. [CrossRef]
27. Xu, Z.; Wang, H. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **2007**, *35*, W265–W268. [CrossRef]
28. Flynn, J.M.; Hubley, R.; Goubert, C.; Rosen, J.; Clark, A.G.; Feschotte, C.; Smit, A.F. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9451–9457. [CrossRef]
29. Edgar, R.C.; Myers, E.W. PILER: Identification and classification of genomic repeats. *Bioinformatics* **2005**, *21* (Suppl. 1), i152–i158. [CrossRef]
30. Price, A.L.; Jones, N.C.; Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **2005**, *21* (Suppl. 1), i351–i358. [CrossRef]
31. Griffiths-Jones, S.; Bateman, A.; Marshall, M.; Khanna, A.; Eddy, S.R. Rfam: An RNA family database. *Nucleic Acids Res.* **2003**, *31*, 439–441. [CrossRef]
32. Kozomara, A.; Griffiths-Jones, S. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **2011**, *39*, D152–D157. [CrossRef]
33. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef]
34. Lowe, T.M.; Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **1997**, *25*, 955–964. [CrossRef]
35. Doerks, T.; Copley, R.R.; Schultz, J.; Ponting, C.P.; Bork, P. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.* **2002**, *12*, 47–56. [CrossRef]
36. Manni, M.; Berkeley, M.R.; Seppey, M.; Zdobnov, E.M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **2021**, *1*, e323. [CrossRef]
37. Cantarel, B.L.; Korf, I.; Robb, S.M.; Parra, G.; Ross, E.; Moore, B.; Holt, C.; Alvarado, A.S.; Yandell, M. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **2008**, *18*, 188–196. [CrossRef]
38. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef]
39. Li, L.; Stoeckert, C.J., Jr.; Roos, D.S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **2003**, *13*, 2178–2189. [CrossRef]
40. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
41. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
42. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007**, *24*, 1586–1591. [CrossRef] [PubMed]
43. De Bie, T.; Cristianini, N.; Demuth, J.P.; Hahn, M.W. CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* **2006**, *22*, 1269–1271. [CrossRef] [PubMed]
44. Simakov, O.; Marletaz, F.; Cho, S.J.; Edsinger-Gonzales, E.; Havlak, P.; Hellsten, U.; Kuo, D.H.; Larsson, T.; Lv, J.; Arendt, D.; et al. Insights into bilaterian evolution from three spiralian genomes. *Nature* **2013**, *493*, 526–531. [CrossRef] [PubMed]
45. Adema, C.M.; Hillier, L.W.; Jones, C.S.; Loker, E.S.; Knight, M.; Minx, P.; Oliveira, G.; Raghavan, N.; Shedlock, A.; Do Amaral, L.R.; et al. Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat. Commun.* **2017**, *8*, 15451. [CrossRef]
46. Liu, C.; Liu, C.; Zhang, Y.; Ren, Y.; Wang, H.; Li, S.; Jiang, F.; Yin, L.; Qiao, X.; Zhang, G.; et al. The genome of the golden apple snail Pomacea canaliculata provides insight into stress tolerance and invasive adaptation. *GigaScience* **2018**, *7*, giy101. [CrossRef]
47. Guo, Y.; Zhang, Y.; Liu, Q.; Huang, Y.; Mao, G.; Yue, Z.; Abe, E.M.; Li, J.; Wu, Z.; Li, S.; et al. A chromosomal-level genome assembly for the giant African snail Achatina fulica. *Gigascience* **2019**, *8*, giz124. [CrossRef]
48. Zhao, D.; Zhao, M.; Wu, Z.D. The karyotype of five species of freshwater snails of the family Viviparidae. *Acta Zool. Sin.* **1988**, *4*, 364–370.
49. Calvete, O.; Reyes, J.; Valdés-Socin, H.; Martin, P.; Marazuela, M.; Barroso, A.; Escalada, J.; Castells, A.; Torres-Ruiz, R.; Rodríguez-Perales, S.; et al. Alterations in SLC4A2, SLC26A7 and SLC26A9 Drive Acid-Base Imbalance in Gastric Neuroendocrine Tumors and Uncover a Novel Mechanism for a Co-Occurring Polyautoimmune Scenario. *Cells* **2021**, *10*, 3500. [CrossRef]
50. Nikpour, M.; Noborn, F.; Nilsson, J.; Van Damme, T.; Kaye, O.; Syx, D.; Malfait, F.; Larson, G. Glycosaminoglycan linkage region of urinary bikunin as a potentially useful biomarker for β3GalT6-deficient spondylodysplastic Ehlers-Danlos syndrome. *JIMD Rep.* **2022**, *63*, 462–467. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.