Dissecting the chromosome-level genome of the Asian Clam (Corbicula fluminea)

Tongqing Zhang[1], Jiawen Yin[1], Shengkai Tang[1], Daming Li[1], Xiankun Gu[1], Shengyu Zhang[2], Weiguo Suo[3], Lei Wu[4], Xiaqing Yu[5], Xiaowei Liu[1], Yanshan Liu[1], Qicheng Jiang[1], Muzi Zhao[1], Yue Yin[1], and Jianlin Pan[1]

[1]Freshwater Fisheries Research Institute of Jiangsu Province
[2]Hongze Lake Fisheries Administration Committee Office of Jiangsu Province
[3]Fisheries Management Commission of Gehu Lake
[4]Biomarker Technologies Corporation
[5]Nanjing Agricultural University

October 14, 2020

Abstract

The Asian Clam (Corbicula fluminea) is a valuable commercial and medicinal bivalve that is widely distributed in East and Southeast Asia. As a natural nutrient source, the clam is high in protein, amino acids, and microelements. In China, C. fluminea plays an important role in the diversity of freshwater ecosystems. The genome of C. fluminea has not yet been characterized, therefore, genome-assisted breeding and improvements cannot yet be implemented. In this work, we present a de novo chromosome-scale genome assembly of C. fluminea using PacBio and Hi-C sequencing technologies. The assembled genome comprised 4,728 contigs, with a contig N50 of 521.06 Kb, and 1,215 scaffolds with a scaffold N50 of 70.62 Mb. More than 1.51 Gb (99.17%) of genomic sequences were anchored to 18 chromosomes, of which 1.40 Gb (92.81%) of genomic sequences were ordered and oriented. The genome contains 38,841 coding genes, 32,591 (83.91%) of which were annotated in at least one functional database. Compared with related species, C. fluminea had 851 expanded gene families and 191 contracted gene families. The expanded genes were significantly enriched in 9 terms associated with metabolite synthesis. The phylogenetic tree showed that C. fluminea diverged from the ancestors of marine bivalves ~492.00 million years ago (Mya). Additionally, we identified two MITF genes in C. fluminea and several core genes involved in vitamin B6 metabolic pathways. The high-quality and chromosomal Asian Clam genome will be a valuable resource for a range of development and breeding studies of C. fluminea in future research.

Keywords

Corbicula fluminea , genome assembly, PacBio, Hi-C

1 Introduction

Corbicula fluminea belongs to the family Corbiculidae, genus *Corbicula* (Ishibashi et al., 2003; Korniushin, 2004). The clam has a round base and triangular double shells (Figure 1). The surface of the shells is glossy, and the shell color varies with the living environment (Alyakrinskaya, 2005). Shells are brown, yellow, green, or black and are characterized by circular growth lines (Qiu, Shi, & Komaru, 2001). There are three main teeth in the left shell, one in the front, one in the back and one in the side (Thorp & James, 1991). The clam grows rapidly and takes only 73–91 days for sexual maturation (Tao, Deng, & Li, 2016; Gu, & Wang, 2001). Due to this, it has strong reproductive capacity and diffusion ability (Sun, 1995). *C. fluminea* is native to East Asia, therefore it is also called the Asian Clam. This species is a native shellfish in China, where it

displays strong environmental adaptability. As a representative of freshwater macrobenthic invertebrates, clams are an important component of freshwater ecosystems, and this species has an important impact on the diversity of freshwater ecosystems (Ding, Deng, & Cao, 2014; Xiao, 2015). As a bivalve that has successfully undergone radiation in freshwater habitats (Sirirat, Joong, & Foighil, 2000), the clam is widely distributed in lakes and rivers in China. It is also found in Korea, Japan, and Southeast Asian countries (Jiang, & Zhao, 1997), whereas in America and Europe, it is considered as an alien species (Keogh, & Simons, 2019; Kondakov, Palatov, & Bolotov, 2018).

FIGURE 1 The Asian Clam (*Corbicula fluminea*)

The meat of *C. fluminea* is delicious and nutritious, and thus the clam is considered as a delicacy. It is rich in protein, amino acids, and microelements, and its nutritional value has been well studied (Zhao, & Liu, 2010; Zhuang, & Song, 2009). According to the Compendium of Materia Medica, the Asian Clam has medicinal applications of detumescence, dehumidification, sobering up, and benefits to the liver. Modern research has found that the extracts of the clam can protect against liver damage and reduce blood lipids (Chin, Chien, & Gow, 2010; Peng, et al., 2008). The proteins in the clam can be hydrolyzed into peptides and amino acids using proteases (Wu, & Sun, 2007). Using modern enzymolysis technology for hydrolysis, various natural products that have hepato-protective, anti-inflammatory, antitumor, antioxidant, and anti-hypertensive properties can be obtained. At present, the utilization of Asian Clam resources mainly involves fresh, dried, and canned food. As a traditional Chinese medicine, the Asian Clam is underdeveloped. For example, Japan and South Korea import tens of thousands of tons of clams to China every year for deep processing into health care drugs such as sobering agents and liver-protecting drugs (Zhang, 1996). Compared with Japan and South Korea, the deep processing ability for the Asian Clam in China is relatively backward, resulting in its economic and medicinal value not being fully exploited (Wang, & Liu, 2010).

Deciphering the genome of *C. fluminea* is the most basic step in our research program. The acquisition of a high-quality genome may provide more detailed insights into the value of *C. fluminea* . During the past decade, whole-genome sequencing has been widely performed on a number of bivalves due to the rapid development of third-generation sequencing (Sun, et al., 2017; Yan, et al., 2019). The lack of a complete genome was the motivation for a de novo genome sequencing of *C. fluminea* . In this study, we utilized PacBio and Hi-C technology to successfully assemble a chromosomal-level genome (2n = 36). The effective genome information, including chromosomes, genes, and repeat sequence distributions, is displayed in a circle diagram (Figure 2). This genome provides the foundation for a range of development and breeding studies of *C. fluminea* in future research.

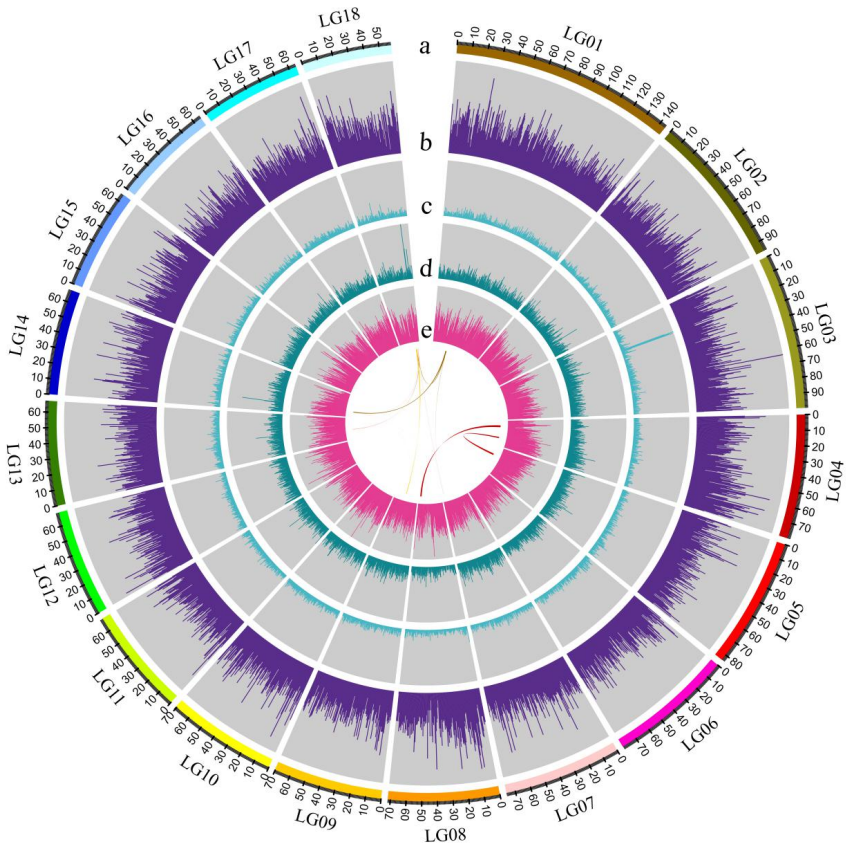

FIGURE 2 Genome landscape of *Corbicula fluminea*. From outer to inner circles: a: marker distribution on 18 chromosomes at the Mb scale; b: LARD distribution on each chromosome; c: PLE distribution on each chromosome; d: gene distribution on each chromosome; e: GC content within a 1-Mb sliding window; yellow lines in the inner circle indicate gene blocks on the forward direction strand, and red lines indicate gene blocks on the reverse direction strand.

2 Materials and methods

2.1 Sampling, library construction, and sequencing

Fresh Asian Clam (*Corbicula fluminea*) samples were collected from Hongze Lake, Jiangsu, China. Healthy, disease-free, and tender individuals of *C. fluminea* were selected for genome sequencing. After shelling and removing the intestines by hand, the remaining body tissues were immediately transferred into liquid nitrogen and sequenced by Biomarker Technologies Corporation, Beijing. High-quality genomic DNA was extracted from the Asian Clam using a DNeasyRBlood & Tissue Kit (Qiagen, Hilden, Germany). The DNA quality was measured with Qubit 3.0 (Invitrogen, Carlsbad, CA, USA) and was checked using 1% agarose gel electrophoresis.

Illumina libraries were constructed according to the manufacturer's standard PCR-free protocol (Illumina) and sequenced on an Illumina HiSeq X Ten platform (Illumina, Inc., San Diego, CA, USA) using the paired-end 150 (PE150) strategy. Approximately 30 μg of genomic DNA was used to construct PacBio libraries by shearing into ˜20 kb targeted size fragments with BluePippin (Sage Science, Beverly, MA, USA). Then, the qualified libraries were prepared for single-molecule real-time (SMRT) genome sequencing using S/P2-

C2 sequencing chemistry on the PacBio Sequel II platform (PacBio, Pacific Biosciences, USA). The Hi-C libraries were cross-linked in situ using formaldehyde with a final concentration of 2% and homogenized with tissue lysis by the restriction enzyme *Hind* III. The libraries for Hi-C with insert sizes of 300–700 bp were sequenced on an Illumina HiSeq X Ten platform (Illumina, SanDiego, CA, USA). All processes were performed at Biomarker Technologies Corporation following the standard protocols.

2.2 Genome estimation and de novo assembly

Initially, Illumina data were filtered and corrected by Fastp (version 0.19.3) (Chen, Zhou, Chen, & Gu, 2018), followed by applying the data to estimate the genomic features. Beforehand, Illumina reads were randomly selected and aligned to the Nucleotide Sequence Database (NT) using BLAST (version 2.2.31) (Altschul, Gish, Miller, Myers, & Lipman, 1990) with the parameter of E-value = $1e^{-05}$ for confirming whether contamination existed. In this study, we plotted the 21-mer depth distribution (k =21) to estimate the genome size, heterozygosity, and repeats using Jellyfish (version 2) (Marçais, & Kingsford, 2011). Genome size estimation was implemented by the formula G = N21-mer (total number of k-mers) /D 21-mer (k-mer depth of the main peak). Repetitive sequences were accumulated from where the depth of a k-mer was more than two times that of the main peak, and heterozygous sequences were estimated at where the depth was half of the main peak.

Using the long single molecular reads from PacBio, the pipelines of workflow were as follows in the genome assemblies. First, the clean data from PacBio were subjected to error correction using Canu (version 1.5) (Koren et al., 2017) with the parameter of error correct coverage = 60. Subsequently, the outputs were piped into the workflow of SMARTdenovo (version 1.0) (Schmidt, Vogel, & Denton, 2017), and the genomic contigs were automatically generated with the parameters of J=5000, A=1000, and r=0.95. Finally, the preliminary assembly was polished three times by Racon (version 1.32) (Vaser, Sović, Nagaranjan, & Šikić, 2017), resulting in the first correction being successfully realized. Recognizing the relatively high error rate of the third-generation sequencing platform, Illumina reads specifically for genome estimation had been prepared for the second correction. This was implemented by Pilon (version 1.22) (Walker et al., 2014), and the error correction was again run three times. Each of the tools used for genome assembly was well-founded for the assembly process of *C. fluminea* .

2.3 Hi-C scaffolding

The contigs generated by the preliminary genome assembly required filling of gaps and anchoring on the putative chromosomes. The initial contigs were piped into the Hi-C assembly workflow, and the signals of chromatin interactions were captured to construct chromosomes. In brief, the putative Hi-C junctions were aligned by the unique mapped read pairs using BWA-MEM (version 0.7.10-r789) (Li & Durbin, 2009). The paired reads uniquely mapped to the assembly were called the valid interaction pairs, and they were used for the Hi-C scaffolding. Other invalid reads included reads of self-ligation and non-ligation; dangling ends were filtered out using HiC-Pro (version 2.10.0) (Servant, et al., 2015).

The Hi-C reassembly broke the contigs into 50 kb fragments, and the regions that were mismatched to the initial assembly or could not be restored were listed as candidate error areas. The genome was subjected to a final round of error correction, and the gaps were filled during this round. The reassembled and corrected contigs were divided into ordered, oriented, and anchored groups by LACHESIS (Burton et al., 2013) with the parameters CLUSTER_MIN_RE_SITES = 33; CLUSTER_MAX_LINK_DENSITY = 2; CLUSTER_NONINFORMATIVE_RATIO = 2; ORDER_MIN_N_RES_IN_TRUN = 29, and ORDER_MIN_N_RES_IN_SHREDS = 29, automatically resulting in putative chromosomes. The gaps generated during the Hi-C assembly were refilled using LR GapCloser (version 1.1) (Xu et al., 2019).

2.4 Genome quality evaluation and repeats analysis

The genome of *C. fluminea* was aligned to the actinopterygii database (odb9) comprising 978 conservative core genes by BUSCO (version 3.0) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015). The eukaryotic conserved genes for the clam were searched in the database to evaluate the completeness of the

genome. The CEGMA Database comprising 458 conserved core genes of eukaryotes was searched in the same way using CEGMA (version 2.5) (Parra, Bradnam, & Korf, 2007). Additionally, another evaluation was applied to the Illumina short-read alignments to map to the assembled genome of the clam using BWA-MEM (version 0.7.10-r789) (Li & Durbin, 2009).

There are two main types of repeats, retrotransposons (Class I in our analysis) and transposons (Class II in our analysis). We constructed a specific repeats database for repeat prediction using LTR-FINDER (version 1.05) (Xu & Wang, 2007) and RepeatScout (version 1.0.5) (Price, Jones, & Pevzner, 2005) followed by the identification and classification for repeats by PASTEClassifer (version 1.0) (Hoede et al., 2014). The species-specific repeats library for the clam genome was successfully generated by aggregating our prediction and Repbase (19.06) (Bao, Kojima, & Kohany, 2015). LTR characteristics for the clam were processed by RepeatMasker (version 4.0.6) (Tarailo-Graovac & Chen, 2009).

2.5 Genome annotation

We utilized de novo-, homology-, and transcriptome-based methods to predict protein-coding genes. Five tools employed were Genscan (verson3.1) (Burge & Karlin, 1997), Augustus (version 3.1) (Stanke & Waack, 2003), GlimmerHMM (version 3.0.4) (Majoros, Pertea, & Salzberg, 2004), GeneID (version 1.4) (Blanco, Parra, & Guigó, 2007), and SNAP (version 2006-07-28) (Korf, 2004); these were used for prediction de novo. Protein sequences from four representative species (*Danio rerio* , *Crassostrea gigas* , *Crassostrea virginica* , and*Mizuhopecten yessoensis*) were aligned to the Asian Clam protein sequences to perform homology-based prediction by GeMoMa (version 1.3.1) (Keilwagen et al., 2016). Transcriptome data were mapped to the genomic sequences; Hisat (version 2.0.4) (Kim, Langmead, & Salzberg, 2015) and Stringtie (version 1.2.3) (Pertea et al., 2015) were used to assemble and dissect functional genes. TransDecoder (version 2.0) (http://transdecoder.github.io) and GeneMarkS-T (version 5.1) (Tang, Lomsadze, & Borodovsky, 2015) were used for transcriptome-based prediction. Finally, the above methods were integrated into non-redundant protein-coding gene sets by EVM (version 1.1.1) (Haas et al., 2008) and PASA (version 2.0.2) (Haas et al., 2003).

The other genome features, including pseudogenes and non-coding RNAs, were identified by referring to the miRbase database (version 21.0) (Griffiths-Jones, Grocock, Van Dongen, Bateman, & Enright, 2006) and Rfam (version 13.0) (Daub, Eberhardt, Tate, & Burge, 2015). In the process of searching for putative pseudogenes, candidates were assessed based on the premature stop codons or frameshift mutations in the gene structure using GenBlastA (version 1.0.4) (She, et al., 2011). The identification of transfer RNA (tRNA) was performed by tRNAscan-SE (version 1.3.1) (Lowe & Eddy, 1997). MicroRNA and ribosomal RNA (rRNA) were identified by Infernal (version 1.1) (Nawrocki & Eddy, 2013).

The protein-coding genes were subject to functional annotation by aligning to the EuKaryotic Orthologous Groups (KOG) (Tatusov et al., 2003), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000), TrEMBL (Boeckmann et al., 2003), Swiss-Prot (Boeckmann et al., 2003), and Non-redundant (Nr) databases (Marchler et al., 2011) using BLAST (version 2.2.31) (Altschul, Gish, Miller, Myers, & Lipman, 1990) with a maximal E-value of 1e-05. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations and Gene ontology (GO) (Consortium, 2004) terms were assigned to identify gene functions using Blast2GO (version 4.1) (Conesa et al., 2005).

2.6 Gene family identification

Protein data from *C. fluminea* and other representative species, including *Capitella teleta* , *Lingula anatina* , *Octopus vulgaris* , *Lottia gigantea* , *Crassostrea gigas* ,*Crassostrea virginica* , *Pinctada imbricata* ,*Mizuhopecten yessoensis* , *Mytilus coruscus* , and*Bathymodiolus platifrons* , were retrieved in the corresponding databases and aligned using BLAST (version 2.2.31) (Altschul, Gish, Miller, Myers, & Lipman, 1990) with a maximum e-value of 1e-5. Proteins with sequence lengths >100 amino acids were searched against the Pfam (https://pfam.xfam.org) database by Pfam scan (El-Gebali, et al., 2018). The ortholog groups for gene families were generally clustered using OrthoMCL (version 2.0.9) (Li, Stoeckert, & Roos, 2003). Four selected shellfish (*C. gigas* , *L. gigantea* , *B. platifrons* , and *C. virginica*) and *C. fluminea* were

grouped together to conduct the analysis for gene family characteristics.

2.7 Phylogenetic and gene family evolutionary analyses

The single-copy orthologs from all involved species were statistically analyzed using the longest transcripts for each gene. The single-copy orthologous genes shared by the above 11 species (including *C. fluminea*) were aligned using MUSCLE (version 3.8.31) (Edgar, 2004). The super-alignment of nucleotide sequences provided a reference tree topology using PhyML (version 3.3) (Guindon et al., 2010). The divergence times among species were roughly estimated by the MCMCTree program of the PAML package (version 4.7a) (Yang, 2007) with the approximate likelihood calculation method. We utilized molecular clock data from the TimeTree (http://www.timetree.org/) (Kumar, Stecher, Suleski, & Hedges, 2017) database as the calibration times.

According to divergence times and phylogenetic relationships, CAFE (version 4.2) (De Bie, Cristianini, Demuth, & Hahn, 2006) was used to analyze gene family evolution. The gene family expansion and contraction were analyzed by comparing the differences between the ancestor and involved species. The extended family genes for *C. fluminea* were extracted and aligned to the functional enrichment on GO and KEGG to detect their functions.

2.8 Transcriptome sequencing

Specimens of mature *Meretrix meretrix* and *Ruditapes philippinarum* were collected from coastal area in Nantong, Jiangsu. They and *Corbicula fluminea* were dissected and fixed in RNAlater. The RNA was extracted using TRIzol (Thermo Fisher, USA), and the libraries were generated using a NEBNext(r) Ultra RNA Library Prep Kit for Illumina(r) (NEB, USA) following the instruction manual. Three biological repeats were set for each species. All samples were sequenced on an Illumina HiSeq X Ten platform (Illumina, Inc., San Diego, CA, USA). The clean reads for *C. fluminea* were mapped to the Asian Clam genome using Tophat (version 2.0) (Ghosh, & Chan, 2016), and only reads with a perfect match or one mismatch were used for further analysis. For the transcriptome data of *M. meretrix* and *R. philippinarum,* a de novo transcriptome assembly was conducted by Trinity (version 2.1.1) (Haas et al., 2013), and CD-HIT (Fu, Niu, Zhu, Wu, & Li, 2012) was used to reduce sequence redundancy. Finally, we presented Kallisto (Bray, Pimentel, Melsted, & Pachter, 2016), an RNA-seq quantification program that mapped clean reads to known transcripts for quantification and standardization. Gene expression levels were estimated by TPM.

3 Results and discussion

3.1 Statistics of sequencing data

More than 252.77 Gb of clean data for survey analysis (Illumina) were generated, and the data covered the depth of 154.13X for the Asian Clam genome (Table 1). Two single-molecule real-time (SMRT) cells were processed, and approximately 15.03 million PacBio reads (293.72 Gb) were generated (Table 1). The max subread for PacBio offline was 286.39 kb; the N50 and mean length of subreads were 31.18 kb and 19.54 kb, respectively. The valid subreads were mainly distributed from 500 bp to 40,000 bp (Supporting Information Figure S1). After the Hi-C data processing by filtering of low-quality reads, we obtained approximately 780.87 million clean reads (˜233.26 Gb) from two libraries that were used for chromosomal construction (Table 1). Additionally, approximately 8 Gb clean data for transcriptome sequencing was performed for subsequent gene prediction analysis (Table 1).

TABLE 1 Statistics of the sequencing data

Types	Sequencing platform	Library size	Number of library	Clean data (Gb)	Coverage (×) +
Illumina	Illumina HiSeq X	350 bp	6	252.77	154.13
PacBio	PacBio Sequel II	20 kb	2	293.72	193.40

Hi-C	Illumina HiSeq X	300–700 bp	2	233.26	142.23
Transcriptome	Illumina HiSeq X	350 bp	1	8.18	-

3.2 K-mer analysis and genome assembly

Before k-mer analysis, Illumina reads from the survey analysis were mapped to the Nucleotide Sequence Database (NT); 86.38% of reads were successfully matched, indicating that the data were credible for further analysis. The k-mer number of 187,447,882,456 was screened out by filtering of abnormal k-mers. The k-mer depth of 115 was the main peak in the plot (Supporting Information Figure S2), and the genome size was calculated as ˜ 1.64 Gb according to the k-mer formula. The k-mer depths of 58 and 230 represented beginning locations in the computation of heterozygous and repetitive sequences, respectively. Finally, the clam genome was estimated to have a heterozygosity rate of 2.41% and a repeat ratio of 64.55%. It was deemed to be a large complex genome with high heterozygosity and a high level of repetition.

The initial filtered PacBio subreads were subjected to error correction by Canu, resulting in 15,031,088 subreads generated for subsequent assembly. The number of contigs obtained by Canu and SMARTdenovo with the polish by Racon and Pilon was 4,347. The analysis finally resulted in an Asian Clam genome of 1.52 Gb with a contig N50 of 603.64 Kb. The size of the Asian Clam genome for PacBio assembly was slightly smaller than that estimated by k-mer analysis (1.64 Gb), which was in line with the regularity. It indicates that we have captured and assembled most of the sequences of Asian Clam genome. The accuracy of the sequence needs to be verified by Hi-C technology.

3.3 Chromosome construction by Hi-C

More than 571.60 million read pairs (73.20%) of total Hi-C data were mapped to the initial genomic assembly. We utilized a total of 116.65 million valid interaction pairs for Hi-C scaffolding; invalid interaction pairs, including reads of dangling end pairs, re-ligation pairs, self-cycle pairs, and dumped pairs, were filtered out (Supporting Information Table S1).

The initial assembled contigs were broken and reassembled using unique mapped read pairs for Hi-C. The areas that could be restored as candidate areas had already been corrected. After Hi-C assembly and manual adjustment, we obtained 4,728 corrected contigs (Table 2) assigned to 18 pseudochromosomes. The final assembly presented a high-quality Asian Clam genome that reached 1.52 Gb in length and was characterized by a contig N50 of 521.06 Kb and a scaffold N50 of 70.62 Mb (Table 2). There were 1.51 Gb of genomic sequences accounting for 99.17% of total contig sequences on 18 chromosomes comprising 4,621 contigs (97.74%) (Figure 3). Additionally, 1.40 Gb (92.81%) of genomic sequences were anchored with a defined order and orientation in a Hi-C interaction heat map (Supporting Information Table S2). The scaffolding process for the Asian Clam genome showed a high level of efficiency (genomic sequences more than 99%, contigs more than 97%) and deserved to be considered as a high-quality and chromosomal-level genome.

TABLE 2 Statistics and characteristics of the genome for *Corbicula fluminea*

Characteristics	Number	Size	Percentage
Estimate of genome size		1.64 Gb	
Final assembly genome size		1.52 Gb	
Contig number and N50	4,728	521.06 Kb	
Maximum contig		3.17 Mb	
Scaffold number and N50	1,215	70.62 Mb	
Maximum scaffold		144.27 Mb	
Heterozygosity rate			2.41%
Total repetitive sequences		1.06 Gb	69.66%
Total protein-coding genes	38,841	0.54 Gb	

Annotated protein-coding genes	32,591	83.91%
MicroRNA	45	
Ribosomal RNA (rRNA)	420	
Transfer RNA (tRNA)	3,707	

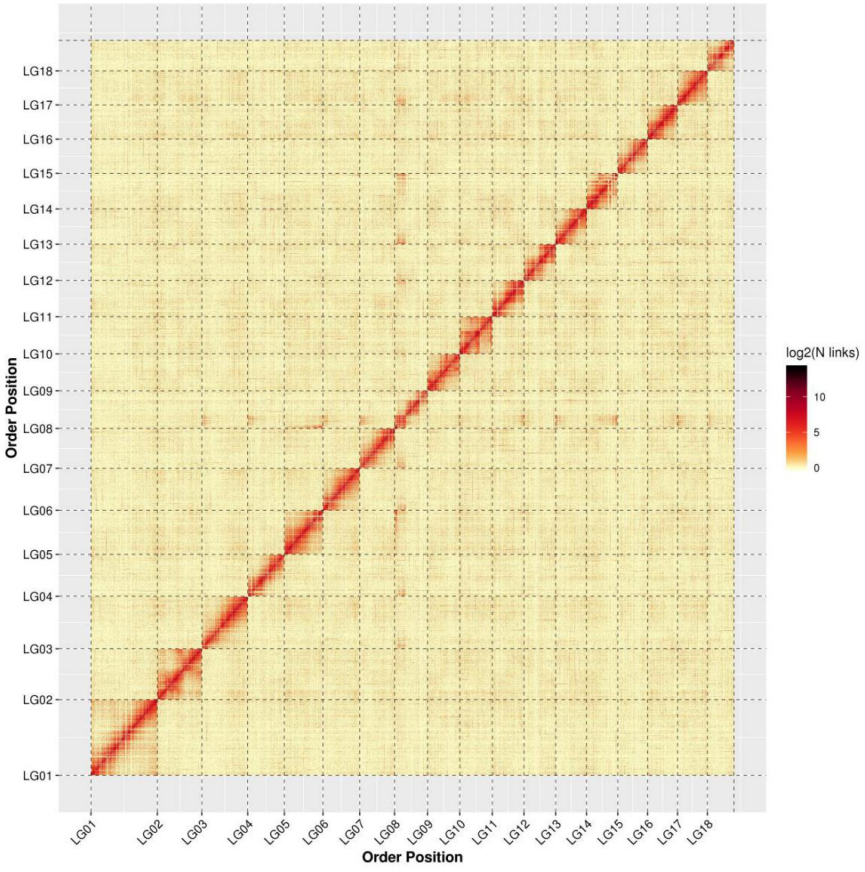

FIGURE 3 The genome-wide Hi-C heatmap of *Corbicula fluminea*. LG 1–18 are the abbreviations of Lachesis Groups 1–18 representing the 18 pseudochromosomes.

3.4 Evaluation and repeat annotation

We finally obtained a data set that contained 889 complete BUSCO groups (90.90%) and 423 (92.36%) CEGMA groups. The mapping ratio for Illumina data was up to 97.45%. The BUSCO, CEGMA, and the mapping ratio for Illumina directly supported the high-quality Asian Clam genome that we assembled. More detailed information may be found in Supporting Information Table S3. More than 1.06 Gb of genomic sequences were identified and marked as repeats, representing 69.66% of the total genomic sequences (Table 2). Approximately 608.85 Mb (57.54%) of LARDs was the predominant repeat type. Other types of repeats with high proportions were TIRs (10.46%), PLEs (12.38%), and LINEs (7.07%) (Supporting Information Table S4).

3.5 Gene prediction and gene annotation

A consensus of the results of all three methods for protein-coding genes prediction was reached, and the final number of non-redundant protein-coding genes was 38,841, with a total length of 0.54 Gb (Table 2,

Supporting Information Table S5). More than 32,591 protein-coding genes (83.91%) were annotated in at least one functional database (Table 3). All genes for each database are annotated in Supporting Information Table S6. Additionally, the Asian Clam gene sets comprised 260,971 exons, and the average gene length was ˜ 13.97 kb. The Asian Clam genome contained 3,048 pseudogenes, 45 microRNAs, 420 rRNAs, and 3,707 tRNAs (Table 2, Supporting Information Table S7).

TABLE 3 Statistics of gene annotation to different databases

Annotation database	Annotated number	Percentage (%)	100[?]Protein length<300	Protein length[?]300
GO_Annotation	7,489	19.28	2,243	5,119
KEGG_Annotation	12,757	32.84	3,466	9,144
KOG_Annotation	18,233	46.94	4,642	13,426
TrEMBL_Annotation	32,280	83.11	10,097	21,841
Nr_Annotation	32,382	83.37	10,170	21,858
All_Annotated	32,591	83.91	10,275	21,957

3.6 Gene family identification

Gene family analysis identified a total of 71,331 gene families among five kinds of shellfish (Supporting Information Table S8), and we discovered 23,063 gene families clustered by 38,841 protein-coding genes in the Asian Clam genome. We compared the *C. fluminea* genome to those of *C. gigas* , *L. gigantea* , *B. platifrons* , and*C. virginica* and discovered that 16,170 gene families were specific to *C. fluminea* (Figure 4A). Additionally, we statistically analyzed several gene features, including single-copy orthologs, multiple copy orthologs, other orthologs, and unique genes (Figure 4B). The five shellfish shared 146 single-copy orthologs, and the Asian Clam genome contained 25,878 unique genes (Supporting Information Table S9).

3.7 Phylogenetic analysis and gene family evolution

The phylogenetic position of *C. fluminea* and other representative species was estimated based on single-copy orthologs. Three time points for the most recent common ancestor (MRCA) were estimated by TimeTree. The differentiation time of *C. gigas* and *C. virginica* was 72.9 (63.2 - 82.7) million years ago (Mya) (Plazzi & Passamonti, 2010); that of *B. platifrons* and *M. coruscus* was 387 (308 - 481) Mya (Peterson, Cotton, Gehling, & Pisani, 2008); that of *C. fluminea* and six marine bivalves was 492 (472 - 516) Mya (Stöger, et al., 2013; Huang, et al., 2018). We utilized these time of MRCA to calibrate the phylogenetic tree, resulting in the phylogenetic tree constructed by seven bivalves and four other marine species (Figure 4C). As shown, all marine bivalves were clustered together, especially those belong to the same family. The phylogenetic tree showed that *C. fluminea* and the closest genetic relatives, the ancestors of marine bivalves (family Mytilidae represented by *B. platifrons*and *M. coruscus* ; family Ostreidae represented by *C. gigas*and *C. virginica;* family Pteriidae represented by *P. imbricata;* family Pectinidae represented by *M. yessoensis*) diverged ˜ 492.00 million years ago. This evidence suggests that *C. fluminea* is a kind of bivalve, which is obviously different from marine bivalves. It is suggested that bivalves were divided into two groups, and they survive in freshwater and seawater respectively for a long time (million years). Some references showed that Heterodonta living in the freshwaters originated in the Paleozoic (Moore & Raymond, 1969; Cope & Veliger, 1995) and Veneroida diversified during the Mesozoic and Cenozoic eras (Stanley, 1968). The ancestors of *C. fluminea* may have invaded and migrated to freshwater from the ocean since millions of years ago, and they have evolved to fill various freshwater habitat.

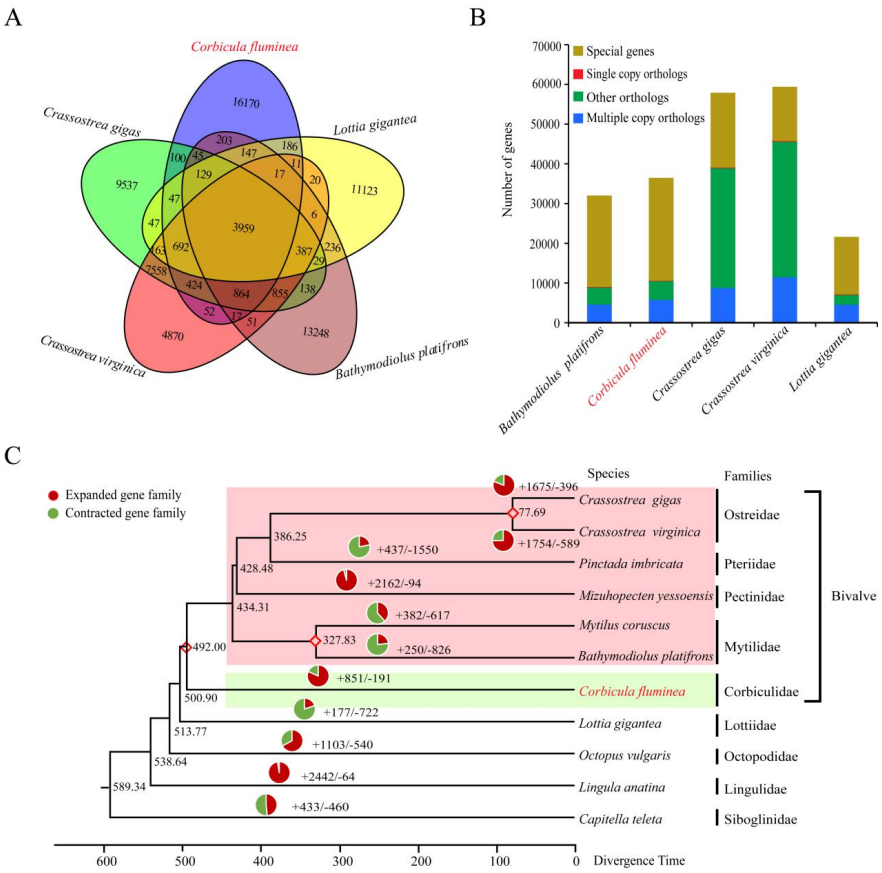

FIGURE 4 The comparative genomic analysis of *Corbicula fluminea* and other species. (A) Venn diagram of gene families between *Corbicula fluminea* and *Crassostrea gigas*, *Lottia gigantea*, *Bathymodiolus platifrons*, and *Crassostrea virginica*. (B) Distribution of multiple-copy orthologs, other orthologs, single-copy orthologs, and unique genes in *Corbicula fluminea* and the above four species. (C) Phylogenetic tree, divergence time, and profiles of gene families that underwent expansion and contraction in 11 species.

Combining the phylogenetic relationships, gene family evolution was calculated by comparing the differences between ancestors and *C. fluminea* . This analysis resulted in 851 gene families being significantly expanded ($P < 0.05$) and 191 gene families being significantly contracted ($P < 0.05$) in the Asian Clam genome (Figure 4C, Supporting Information Table S10). The 851 expanded gene families were clustered by 9,967 functional genes (Supporting Information Table S11). The functional enrichment analysis on GO and KEGG of those expanded genes identified 325 significantly enriched (q-value < 0.01) GO terms (Supporting Information Table S12) and 19 significantly enriched (q-value < 0.01) KEGG pathways (Supporting Information Table S13, Supporting Information Figure S3). Among 19 pathways, 9 pathways related to metabolite synthesis were given attention, such as taurine and hypotaurine metabolism, drug metabolism, O-Glycan biosynthesis and so on (Table 4). These results are helpful for us to study and understand the characteristics of *C. fluminea.*

TABLE 4 Significantly enriched KEGG pathways for expanded family genes in *Corbicula fluminea*

KEGG ID	KEGG pathways	Adjusted *P*-value	Number of genes	Enrichment factor

ko00430	Taurine and hypotaurine metabolism	2.58E-05	59	2.48
ko00983	Drug metabolism	9.74E-04	61	1.90
ko00512	Mucin type O-Glycan biosynthesis	0.001810	39	2.40
ko00604	Glycosphingolipid biosynthesis	0.002917	22	3.33
ko00910	Nitrogen metabolism	0.003432	26	2.86
ko00590	Arachidonic acid metabolism	0.007314	47	1.96
ko00513	Various types of N-glycan biosynthesis	0.008318	56	1.81
ko00603	Glycosphingolipid biosynthesis	0.008975	35	2.16
ko00600	Sphingolipid metabolism	0.009744	32	1.77

3.8 MITF gene family analysis

In shellfish, it has been reported that microphthalmia-associated transcription factor (MITF) plays an important role in immune defense and shell color formation (Zhang, et al., 2017; Zhang, et al., 2018). The MITF gene family consists of three subfamilies, namely TFEB, TFEC, and TFE3, in the Pfam database (Zhao, Zhao, Zhou, & Mattei, 1993). In this study, we detected two genes from the Asian Clam genome, namely EVM0008002 and EVM0031201, which were identified as MITF genes. Most species in the phylogenetic analysis possessed one or two MITF genes, while *L. gigantea* was apparently missing MITF genes (Supporting Information Table S14). This finding implies that *L. gigantea* may have lost some functions of immune defense or shell color due to the deletion of MITF genes. Additionally, there were five and seven MITF genes in *C. virginica* and *L. anatina* , respectively (Supporting Information Table S14). Combined with the gene family expansion and contraction analysis, we found the MITF gene family expanded in *C. virginica* and *L. anatina* and contracted in*L. gigantea* . Proteins from all species were aligned to a super-alignment matrix with the guidance of protein alignment, and the genic tree comprising MITF family genes from all involved species was successfully constructed using MUSCLE (Figure 5). Firstly, MITF family genes originated from the same species showed the most recent genetic relationships. Secondly, like family Mytilidae represented by *B. platifrons* and *M. coruscus* , family Ostreidae represented by*C. gigas* and *C. virginica* , those from the same families were clustered together. Additionally, the clustering relationships of MITF family genes were also similar to those shown by the phylogenetic tree of single-copy orthologs. This finding indirectly corroborates the reliability of the phylogenetic relationship analysis. Meanwhile, we observed that *C. fluminea* had a greater genetic distance from other marine bivalves via the MITF and phylogenetic tree analyses. It is probably because the divergence of *Corbiucula* from marine bivalves, and the migration from marine to freshwater is an ancient event.

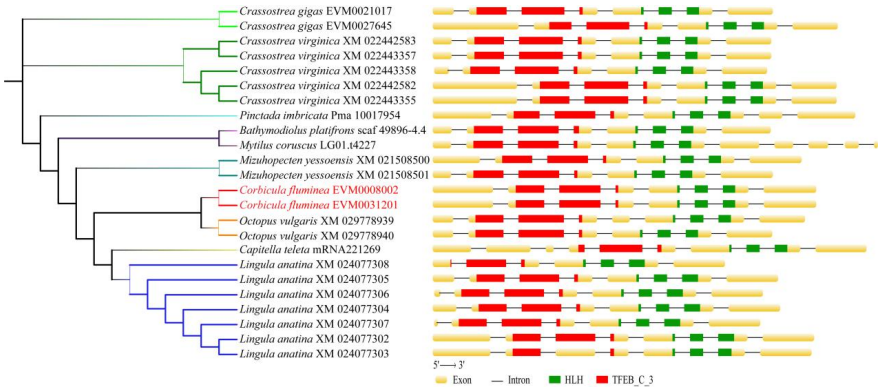

FIGURE 5 The genic tree and gene structure analysis of the MITF gene family in *Corbicula fluminea* and other species.

3.9 Vitamin B6 metabolic pathway analysis

Vitamin B6 (VB6) usually exists in food as pyridoxal, pyridoxine, or pyridoxamine, each of which is converted into pyridoxal phosphate by the mammalian liver (Ramos, et al., 2017). VB6 is a component of co-enzymes in the human body that are involved in a variety of metabolic reactions, especially those closely related to amino acid metabolism (Ueland, McCann, Midttun, & Ulvik, 2017; Bird, 2018). According to relevant studies, vitamin B1 (thiamine), vitamin B2 (riboflavin), and vitamin B3 (nicotinic acid) are found in relatively high levels in *C. fluminea* , whereas the VB6 content is very low (Wang, & Liu, 2010). Combined with the transcriptome data, we analyzed the expression of genes involved in VB6 synthesis in *C. fluminea, M. meretrix,* and*R. philippinarum.* We searched and discovered 10 existing enzymes (*PdxST* , *PNPO* , *PDXP* , *PDXK* , *PDIA1* ,*PDIA3* , *PDIA4* , *PDIA5* , *PDIA6* , and*TXNDC10*) of the VB6 pathway (Figure 6A). Ribulose 5-phosphate is converted to pyridoxal 5-phosphate by *PdxST*(K06215), and two genes (EVM0000630 and EVM0035945) in *C. fluminea* were responsible for the translation of *PdxST*.Pyridoxal 5-phosphate is converted to pyridoxal by *PDXP* (K13248), and EVM0004517 was responsible for the translation of *PDXP* . Pyridoxal can be transformed into 4-Pyridoxate or 4-Pyridoxolactone, but these pathways were blocked due to *C. fluminea* having lost the relevant enzymes or genes. For all genes on the VB6 pathways of *C. fluminea,* genes *PNPO* , *PDXK* , *PDXP* , *PDIA1* , and *PDIA4* had one copy; genes *PdxST* , *PDIA3* ,*PDIA5* , and *TXNDC10* had two copies; the *PDIA6* gene had three copies (Supporting Information Table S15). We suspect that the number of genes involved in the VB6 pathway is low in *C. fluminea* , and this may lead to the low content of VB6. Additionally, we checked the expression of homologous genes in *C. fluminea, M. meretrix,* and *R. philippinarum.* The expression levels of*PdxST* , *PDIA1* , *PDIA3* , and *PDIA4* were higher in *C. fluminea* than in *M. meretrix* and *R. philippinarum* (Figure 6B). The low expression of the *PDXP*gene may inhibit pyridoxal synthesis downstream, and it may be one of the reasons for the low expression of VB6. However, more evidence will be needed for further validation in the future.

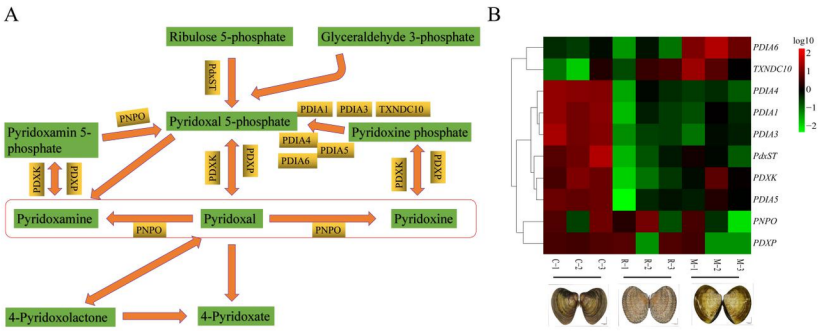

FIGURE 6 Genes involved in vitamin B6 metabolism and their expression levels in metabolic pathways. (A) Genes and metabolites involved in the vitamin B6 pathway. (B) Expression levels of genes related to vitamin B6 synthesis in *Corbicula fluminea*,*Meretrix meretrix*, and *Ruditapes philippinarum*. C represents *Corbicula fluminea*; M represents *Meretrix meretrix*; R represents *Ruditapes philippinarum*.

4 Conclusion

In this study, we assembled a chromosome-level Asian Clam genome using a combination of PacBio and Hi-C technology. The results suggested the high quality of the genome in several ways. The data filled the gap in our knowledge of *C. fluminea* and provide a reference for future research. We assembled 1.52 Gb of genome data distributed across 18 chromosomes, with a contig N50 of 521.06 Kb and a scaffold N50 of 70.62 Mb. We assigned 99.17% of contig sequences to the 18 chromosomes, thus ensuring the integrity of genetic information for each chromosome as much as possible. Additionally, the comparative genomics studies offer evidences for the evolution and characteristics of *C. fluminea* , and showed that *C. fluminea* and its closest relatives, the ancestors of marine bivalves, diverged ˜ 492.00 million years ago. MITF gene family analysis identified two genes in the Asian Clam genome. The genic tree comprising all MITF family genes from involved species showed that the relationship of MITF family genes was similar to that shown by the phylogenetic tree. All lines of evidence suggest that *C. fluminea* is clearly different from other marine bivalves, and its migration from marine to freshwater is an ancient event. In addition, we examined expanded gene families and found 9 significantly enriched pathways associated with metabolite synthesis by KEGG analysis. The vitamin B6 metabolic pathway revealed relatively few genes involved in pyridoxal, pyridoxine, or pyridoxamine synthesis, providing reference for explaining the lower content of VB6 in *C. fluminea* . The genomic information presented in our analysis will help to better understand, develop, and improve *C. fluminea* as well as establish a strong foundation for genome-assisted breeding programs in the future.

Acknowledgements

We appreciate the help from Jiangsu Marine Fisheries Research Institute during sample collection, and the bioinformatic analysis achieved by Biomarker Technologies Corporation. This work was supported by the major project of hydro bios resources in Jiangsu Province (ZYHB16-3), the monitoring of fishery resources in fishery waters of Jiangsu Province in 2018 (ZYHJ201801), the identification of *Corbicula* species and genetic conservation unit in Jiangsu inland waters (SZL201901).

Data Accessibility

Raw sequencing reads for PacBio and Illumina are available at GenBank as BioProject PRJNA657911. Raw sequencing data (Illumina, PacBio, and Hi-C data) have been deposited in the SRA (Sequence Read Archive) database as SUB7507164. The data including assembly and annotation that supported the findings of this study have been deposited in the in the FigShare database,doi.org/10.6084/m9.figshare.12805886.v1 (https://doi.org/10.6084/m9.figshare.12805886.v1).

Author Contributions

J.P., T.Z. and J.Y. designed and managed the project. T.Z. and J.Y. interpreted the data and drafted the manuscript. S.T., D.L. and X.G. prepared the materials. S.Z., W.S., X.L. and Y.L. preformed the DNA extraction, RNA extraction and libraries construction. J.Y., L.W., and X.Y. performed the bioinformatic analysis. All authors contributed to the final manuscript editing.

Conflict of interest

The authors declare no conflict of interest exists.

TABLES

TABLE 1 Statistics of the sequencing data

Types	Sequencing platform	Library size	Number of library	Clean data (Gb)	Coverage (×) +
Illumina	Illumina HiSeq X	350 bp	6	252.77	154.13
PacBio	PacBio Sequel II	20 kb	2	293.72	193.40
Hi-C	Illumina HiSeq X	300–700 bp	2	233.26	142.23
Transcriptome	Illumina HiSeq X	350 bp	1	8.18	-

TABLE 2 Statistics and characteristics of the genome for *Corbicula fluminea*

Characteristics	Number	Size	Percentage
Estimate of genome size		1.64 Gb	
Final assembly genome size		1.52 Gb	
Contig number and N50	4,728	521.06 Kb	
Maximum contig		3.17 Mb	
Scaffold number and N50	1,215	70.62 Mb	
Maximum scaffold		144.27 Mb	
Heterozygosity rate			2.41%
Total repetitive sequences		1.06 Gb	69.66%
Total protein-coding genes	38,841	0.54 Gb	
Annotated protein-coding genes	32,591		83.91%
MicroRNA	45		
Ribosomal RNA (rRNA)	420		
Transfer RNA (tRNA)	3,707		

TABLE 3 Statistics of gene annotation to different databases

Annotation database	Annotated number	Percentage (%)	100[?]Protein length<300	Protein length[?]300
GO_Annotation	7,489	19.28	2,243	5,119
KEGG_Annotation	12,757	32.84	3,466	9,144
KOG_Annotation	18,233	46.94	4,642	13,426
TrEMBL_Annotation	32,280	83.11	10,097	21,841
Nr_Annotation	32,382	83.37	10,170	21,858
All_Annotated	32,591	83.91	10,275	21,957

TABLE 4 Significantly enriched KEGG pathways for expanded family genes in *Corbicula fluminea*

KEGG ID	KEGG pathways	Adjusted *P*-value	Number of genes	Enrichment factor
ko00430	Taurine and hypotaurine metabolism	2.58E-05	59	2.48
ko00983	Drug metabolism	9.74E-04	61	1.90
ko00512	Mucin type O-Glycan biosynthesis	0.001810	39	2.40
ko00604	Glycosphingolipid biosynthesis	0.002917	22	3.33
ko00910	Nitrogen metabolism	0.003432	26	2.86
ko00590	Arachidonic acid metabolism	0.007314	47	1.96
ko00513	Various types of N-glycan biosynthesis	0.008318	56	1.81
ko00603	Glycosphingolipid biosynthesis	0.008975	35	2.16
ko00600	Sphingolipid metabolism	0.009744	32	1.77

Figure legends

FIGURE 1 The Asian Clam (*Corbicula fluminea*)

FIGURE 2 Genome landscape of *Corbicula fluminea*. From outer to inner circles: a: marker distribution on 18 chromosomes at the Mb scale; b: LARD distribution on each chromosome; c: PLE distribution on each chromosome; d: gene distribution on each chromosome; e: GC content within a 1-Mb sliding window; yellow lines in the inner circle indicate gene blocks on the forward direction strand, and red lines indicate gene blocks on the reverse direction strand.

FIGURE 3 The genome-wide Hi-C heatmap of *Corbicula fluminea*. LG 1–18 are the abbreviations of Lachesis Groups 1–18 representing the 18 pseudochromosomes.

FIGURE 4 The comparative genomic analysis of *Corbicula fluminea* and other species. (A) Venn diagram of gene families between *Corbicula fluminea* and *Crassostrea gigas*, *Lottia gigantea*, *Bathymodiolus platifrons*, and *Crassostrea virginica*. (B) Distribution of multiple-copy orthologs, other orthologs, single-copy orthologs, and unique genes in *Corbicula fluminea* and the above four species. (C) Phylogenetic tree, divergence time, and profiles of gene families that underwent expansion and contraction in 11 species.

FIGURE 5 The genic tree and gene structure analysis of the MITF gene family in *Corbicula fluminea* and other species.

FIGURE 6 Genes involved in vitamin B6 metabolism and their expression levels in metabolic pathways. (A) Genes and metabolites involved in the vitamin B6 pathway. (B) Expression levels of genes related to vitamin B6 synthesis in *Corbicula fluminea*,*Meretrix meretrix*, and *Ruditapes philippinarum*. C represents *Corbicula fluminea*; M represents *Meretrix meretrix*; R represents *Ruditapes philippinarum*.

Additional files

Supporting Information Figure S1 The number of subreads distributed in different lengths.

Supporting Information Figure S2 Frequency distribution of the 21-mer graph analysis used to estimate the features of *Corbicula fluminea* genome.

Supporting Information Figure S3 KEGG enrichment analysis for expanded family genes in *Corbicula fluminea.*

Supporting Information Table S1 Statistics of the different types of Hi-C reads.

Supporting Information Table S2 Summary of the Hi-C assembly.

Supporting Information Table S3 Summary of the assessment of genome assembly.

Supporting Information Table S4 Statistics of the repeated sequences.

Supporting Information Table S5 Summary of the gene prediction results.

Supporting Information Table S6 Integrated lists of gene annotation for *Corbicula fluminea* genome.

Supporting Information Table S7 The family consisting of miRNAs, rRNAs and tRNAs in *Corbicula fluminea.*

Supporting Information Table S8 The identification of gene families among five shellfish.

Supporting Information Table S9 Summary of gene features in five shellfish.

Supporting Information Table S10 The expanded and contracted gene families in *Corbicula fluminea.*

Supporting Information Table S11 The expanded family genes in *Corbicula fluminea.*

Supporting Information Table S12 Significantly enriched GO terms for expanded family genes in *Corbicula fluminea.*

Supporting Information Table S13 Significantly enriched KEGG pathways for expanded family genes in *Corbicula fluminea.*

Supporting Information Table S14 MITF genes between *Corbicula fluminea* and other representative species.

Supporting Information Table S15 Genes involved in vitamin B6 synthesis pathway in *Corbicula fluminea*, *Meretrix meretrix* and *Ruditapes philippinarum.*

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology* , *215* (3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Alyakrinskaya, I. O. (2005). Functional significance and weight properties of the shell in some mollusks. *Biology Bulletin* ,*32* (4), 397-418. doi:10.1007/s10525-005-0118-y

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* ,*6* (1), 11. doi:10.1186/s13100-015-0041-9

Bird, R. P. (2018). The Emerging Role of Vitamin B6 in Inflammation and Carcinogenesis. *Adv Food Nutr Res, 3* (83), 151-194. doi:10.1016/bs.afnr.2017.11.004

Blanco, E., Parra, G., & Guigó, R. (2007). Using geneid to identify genes. *Current protocols in bioinformatics* , *18* (1), 4.3.1-4.3.28. doi:10.1002/0471250953.bi0403s00

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., . . . Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* , *31* (1), 365-370. doi:10.1093/nar/gkg095

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol, 34* (5), 525-527. doi:10.1038/nbt.3519

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* ,*268* (1), 78-94. doi:10.1006/jmbi.1997.0951

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology* ,*31* (12), 1119-1125. doi:10.1038/nbt.2727

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34* (17), 884-890. doi:10.1093/bioinformatics/bty560

Chin, L. H., Chien, C. H., & Gow, C. Y. (2010). Hepatoprotection by freshwater clam extract against ccl4-induced hepatic damage in rats.*American Journal of Chinese Medicine* , *38* (5), 881-894. doi:10.1142/S0192415X10008329

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.*Bioinformatics* , *21* (18), 3674-3676. doi:10.1093/bioinformatics/bti610

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* , *32* (Database issue), D258-D261. doi:10.1093/nar/gkh036

Cope, J., Veliger, C. (1995). The early evolution of the Bivalvia.*Origin and, evolutionary, radiation of the Mollusca* ,*123* (34), 342-335.

Daniel, L., Graf, K. S., & Cummings. (2006). Palaeoheterodont diversity (mollusca: trigonioida + unionoida): what we know and what we wish we knew about freshwater mussel evolution. *Zoological Journal of the Linnean Society, 12* (6), 245-251. doi:10.1111/j.1096-3642.2006.00259.x.

Daub, J., Eberhardt, R. Y., Tate, J. G., & Burge, S. W. (2015). Rfam: annotating families of non-coding RNA sequences. In Picardi E. (Eds.), RNA Bioinformatics. Methods in Molecular Biology (pp. 349-363).*New York* , NY: Humana Press.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution.*Bioinformatics* , *22* (10), 1269-1271. doi:10.1093/bioinformatics/btl097

Ding, L. Y., Deng, Y. H, & Cao, Y. H. (2014). Ecological environment indicator function of *Corbicula fluminea* . *Contemporary fisheries* , *8* (1), 78-79. (in Chinese)

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* ,*32* (5), 1792-1797. doi:10.1093/nar/gkh340

El-Gebali, S., Mistry, J., Beteman, A., Eddy, S. R., Luciani, A., Potter, S. C., . . . Tosatto, S. C. E. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research* , *47* (1), 427-432. doi:10.1093/nar/gky995

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data.*Bioinformatics* , *28* (23), 3150-3152. doi:10.1093/bioinformatics/bts565

Ghosh, S., & Chan, C. K. (2016). Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods Mol Biol, 13* (74), 339-361. doi:10.1007/978-1-4939-3167-5_18

Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research* , *34* (Database issue), D140-D144. doi:10.1093/nar/gkj112

Gu, M. Q., & Wang, Z. (2001). Embryonic development observation and staging of *Corbicula fluminea* (müller). *Fisheries information and strategy* , *5* (10), 28-29. (in Chinese)

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.*Systematic biology* , *59* (3), 307-321. doi:10.1093/sysbio/syq010

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., . . . White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.*Nucleic acids research* , *31* (19), 5654-5666. doi:10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., . . . Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* , *8* (8), 1494-1512. doi:10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., . . . Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* , *9* (1), R7. doi:10.1186/gb-2008-9-1-r7

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014). PASTEC: an automatic transposable element classification tool. *Plos One* , *9* (5), e91929. doi:10.1371/journal.pone.0091929

Huang, X. C., Wu, R. W., An, C. T., Xie, G. L., Su, J. H. . . . Wu, X. P. (2018). Reclassification of Lamprotula rochechouartii as Margaritifera rochechouartiicomb. nov. (Bivalvia: Margaritiferidae) revealed by time-calibrated multi-locus phylogenetic analyses and mitochondrial phylogenomics of *Unionoida. Mol Phylogenet Evol* ,*62* (120), 297-306. doi:10.1016/j.ympev.2017.12.017.

Ishibashi, R., Ookubo, K., . . . Kawamura, K. (2003). Androgenetic Reproduction in a Freshwater Diploid Clam *Corbicula fluminea*(Bivalvia: Corbiculidae). *Zoolog* , *20* (6), 727-732. doi:10.2108/zsj.20.727

Jiang, H. T., & Zhao, W. P. (1997). Exploitation and culture technology of *Corbicula fluminea* . *Anhui agriculture* , *8* (10), 26-27. (in Chinese)

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* , *28* (1), 27-30. doi:10.1093/nar/28.1.27

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic acids research* ,*44* (9), e89. doi:10.1093/nar/gkw092

Keogh, S. M., & Simons, A. M. (2019). Molecules and morphology reveal 'new' widespread north american freshwater mussel species (bivalvia: unionidae). *Molecular Phylogenetics and Evolution* , *138* (1), 182-192. doi:10.1016/j.ympev.2019.05.029

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* ,*12* (4), 357-360. doi:10.1038/nmeth.3317

Kondakov, A. V., Palatov, D. M., . . . Bolotov, I. N. (2018). DNA analysis of a non-native lineage of *Sinanodonta woodiana* species complex (Bivalvia: Unionidae) from Middle Asia supports the Chinese origin of the European invaders. *Zootaxa, 4462* (4), 511-522. doi:10.11646/zootaxa.4462.4.4

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* , *27* (5), 722-736. doi:10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics* , *5* (1), 59. doi:10.1186/1471-2105-5-59

Korniushin, A. V. (2004). A revision of some Asian and African freshwater clams assigned to *Corbicula fluminalis* (Müller, 1774) (Mollusca: Bivalvia: Corbiculidae), with a review of anatomical characters and reproductive features based on museum collections.*Hydrobiologia* . *529* (3), 251-270. doi:10.1007/s10750-004-9322-x

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times.*Molecular Biology & Evolution* , *34* (7), 1812-1819. doi: 10.1093/molbev/msx116

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25* (14), 1754–1760. doi:10.1093/bioinformatics/btp324

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* , *13* (9), 2178-2189. doi:10.1101/gr.1224503

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* , *25* (5), 955-964. doi:10.1093/nar/25.5.955

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.*Bioinformatics* , *20* (16), 2878-2879. doi:10.1093/bioinformatics/bth315

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.*Bioinformatics* , *27* (6), 764-770. doi:10.1093/bioinformatics/btr011.

Marchler, B., Lu, S. N., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., . . . Bryant, S. H. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* ,*39* (1), 225-229. doi:10.1093/nar/gkq1189

Moore. & Raymond, C. (1969). Treatise on invertebrate paleontology.*Geological Society of America* , *18* (23), 167-172.

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* , *29* (22), 2933-2935. doi:10.1093/bioinformatics/btt509

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.*Bioinformatics* , *23* (9), 1061-1067. doi:10.1093/bioinformatics/btm071

Peng, T. C., Subeq, Y. M., Lee, C. J., Lee, C. C., Tsai, C. J., & Chang, F. M. (2008). Freshwater clam extract ameliorates acute liver injury induced by hemorrhage in rats. *American Journal of Chinese Medicine* , *36* (6), 1121-1133. doi:10.1142/S0192415X08006466

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* ,*33* (3), 290-295. doi:10.1038/nbt.3122

Peterson, K. J., Cotton, J. A., Gehling, J. G., & Pisani, D. (2008). The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos Trans R Soc Lond B Biol Sci* , *363* (1496), 1435-1443. doi:10.1098/rstb.2007.2233.

Plazzi, F., & Passamonti, M. (2010). Towards a molecular phylogeny of Mollusks: bivalves' early evolution as revealed by mitochondrial genes.*Mol Phylogenet Evol, 57* (2), 641-657. doi:10.1016/j.ympev.2010.08.032.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes.*Bioinformatics* , *21* (suppl_1), i351-i358. doi:10.1093/bioinformatics/bti1018

Qiu, A. D., Shi, A. J., & Komaru, A. (2001). Yellow and brown shell color morphs of *corbicula fluminea* (bivalvia : corbiculidae) from sichuan province, china, are triploids and tetraploids. *Journal of Shellfish Research* , *20* (1), 323-328. (in Chinese)

Ramos, R. J., Pras-Raves, M. L., Gerrits, J., Willemsen, M., Prinsen, H., . . . Nanda, M. (2017). Vitamin B6 is essential for serine de novo biosynthesis. *J Inherit Metab Dis* , *40* (6), 883-891. doi:10.1007/s10545-017-0061-3

Schmidt, M. H., Vogel, A., Denton, A. K. (2017). De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell* , *29* (3), 2336-2348. doi:10.1105/tpc.17.00521

Servant, N., Varoquaux, N., Lajoie, B. R., Eric, V. & Emmanuel, B. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* , *16* (1), 259-262. doi:10.1186/s13059-015-0831-x

She, R., Chu, J. S., Uyar, B., Wang, J., Wang, K., & Chen, N. (2011). genBlastG: using BLAST searches to build homologous gene models.*Bioinformatics* , *27* (15), 2141-2143. doi:10.1093/bioinformatics/btr342

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* ,*31* (19), 3210-3212. doi:10.1093/bioinformatics/btv351

Sirirat, S., Joong, K. P., & Foighil, D. Ó. (2000). Two lineages of the introduced asian freshwater clam *corbicula* occur in north america. *Journal of Molluscan Studies, 17* (3), 275-286. doi:10.1093/mollus/66.3.423

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* ,*19* (suppl_2), ii215-ii225. doi:10.1093/bioinformatics/btg1080

Stanley, S. M. (1968). Post-Paleozoic adaptive radiation of infaunal bivalve molluscs - a consequence of mantle fusion and siphon formation.*J. Paleontol* , *3* (42), 214-229.

Stöger, I., Sigwart, J. D., Kano, Y., Knebelsberger, T., Marshall, B. A., Schwabe, E., Schrödl, M. (2013). The continuing debate on deep molluscan phylogeny: evidence for Serialia (Mollusca, Monoplacophora + Polyplacophora). *Biomed Res Int* , *2013* (40), 70-72. doi:10.1155/2013/407072.

Sun, H. (1995). Utilization and culture of *Corbicula fluminea* .*Scientific fish culture* , *34* (02), 30-31. (in Chinese)

Sun, J., Zhang, Y., . . . Xu, T. (2017). Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat Ecol Evol, 1* (5), 121-126. doi:10.1038/s41559-017-0121

Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic acids research* , *43* (12), e78. doi:10.1093/nar/gkv227

Tao, Z. Y., Deng, Y. H., & Li, C. G. (2016). Embryonic and postembryonic development of *Corbicula fluminea* . *Jiangsu Agricultural Sciences* , *44* (10), 305-307. (in Chinese)

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* , *5* (1), 4.10.11-14.10.14. doi:10.1002/0471250953.bi0410s25

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., . . . Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC bioinformatics* , *4* (1), 41. doi:10.1186/1471-2105-4-41

Thorp & James, H. (1991). Ecology and classification of north american freshwater invertebrates. *Quarterly Review of Biology* ,*39* (4), 209-212. doi:10.1021/ba-1995-0246.pr001

Ueland, P. M., McCann, A., Midttun, Ø., & Ulvik, A. (2017). Inflammation, vitamin B6 and related pathways. *Mol Aspects Med,8* (53), 10-27. doi:10.1016/j.mam.2016.08.001

Vaser, R., Sović, I., Nagaranjan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads.*Genome Research* , *27* (5), 737-746. doi:10.1101/gr.214270.11632.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.

Plos One , *9* (11), e112963. doi:10.1371/journal.pone.0112963

Wang, Y., & Liu, D. H. (2010). Research status and Prospect of functional components of *Corbicula fluminea* . *Food and fermentation industry* , *36* (6), 122-124. (in Chinese)

Wang, Y., & Liu, D. H. (2010). Research status and prospect of functional components in Corbicula fluminea. Food and fermentation industry, 1(06), 128-130. (in Chinese)

Wu, Y. T., & Sun, H. L. (2007). Enzymatic utilization of marine shellfish protein resources. *Chinese Journal of bioengineering* ,*27* (9), 120-125. (in Chinese)

Xiao, L. Z., Jiao, G., Hui, J., Xiao, Y. N., & Kuan, Y. L. (2015). Effects of *corbicula fluminea* in lake taihu on improvement of eutrophic water quality. *Journal of Lake ences* , *27* (3), 486-492. (in Chinese)

Xu, G. C., Xu, T. J., Zhu, R., Zhang, Y., & Li, G. T. (2019). LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* , *8* (1), 157-160. doi:10.1093/gigascience/giy157

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* , *35* (Web Server issue), W265-W268. doi:10.1093/nar/gkm286

Yan, X., Nie, H., Huo, Z., Yan, X., Nie, H., . . . Dong, D. L. (2019). Clam Genome Sequence Clarifies the Molecular Basis of Its Benthic Adaptation and Extraordinary Shell Color Diversity. *IScience, 19* (3), 1225-1237. doi:10.1016/j.isci.2019.08.049

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood.*Molecular Biology & Evolution* , *24* (8), 1586-1591. doi:10.1093/molbev/msm088

Zhang, P. (1996). Utilization and cultivation of *Corbicula fluminea* . *Practical science and technology information in rural areas* , *7* (3), 19-21. (in Chinese)

Zhang, S., Wang, H., Yu, J., Jiang, F., Yue, X., & Liu, B. (2018). Identification of a gene encoding microphthalmia-associated transcription factor and its association with shell color in the clam*Meretrix petechialis* . *Comp Biochem Physiol, 34* (225), 75-83. doi:10.1016/j.cbpb.2018.04.007

Zhang, S., Yue, X., Jiang, F., Wang, H., & Liu, B. (2017). Identification of an MITF gene and its polymorphisms associated with the Vibrio resistance trait in the clam Meretrix petechialis. *Fish Shellfish Immunol, 13* (68), 466-473. doi:10.1016/j.fsi.2017.07.035

Zhao L., & Liu, H. Q. (2010). Evaluation of protein nutritional value in *Corbicula fluminea* extraction. *Anhui Agricultural Science* , *23* (8), 4105-4107. (in Chinese)

Zhao, G. Q., Zhao, Q., Zhou, X., & Mattei, M. G. (1993). a basic helix-loop-helix protein, forms heterodimers with TFE3 and inhibits TFE3-dependent transcription activation. *Mol Cell Biol, 13* (8), 4505-4512. doi:10.1128/mcb.13.8.4505

Zhuang, P., Song, C., & Zhang, L. Z. (2009). Analysis and evaluation of nutritional components of *Corbicula fluminea* in the Yangtze River Estuary. *Acta nutrition Sinica* , *31* (3), 304-306. (in Chinese)