Genomics 113 (2021) 3274–3284

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

A chromosome-level reference genome of red swamp crayfish *Procambarus clarkii* provides insights into the gene families regarding growth or development in crustaceans

Zhiqiang Xu [a,b,c,*,1], Tianheng Gao [b,c,1], Yu Xu [a,c,1], Xuguang Li [a,c], Jiajia Li [a], Hai Lin [a], Weihui Yan [a], Jianlin Pan [a,b,c], Jianqing Tang [a,c,*]

[a] Freshwater Fisheries Research Institute of Jiangsu Province, Nanjing 210017, China
[b] Institute of Marine Biology, College of Oceanography, Hohai University, Nanjing 210098, China
[c] Jiangsu Key laboratory for Conservation and Utilization of freshwater Fisheries Germplasm Resources, Nanjing 210017, China

ARTICLE INFO

Keywords:
Red swamp crayfish
Procambarus clarkii
Chromosome-level genome
Cuticle proteins
Hox genes
PSMC

ABSTRACT

Red swamp crayfish *Procambarus clarkii* is an ecologically and economically important crustacean species. Here, based on a *de novo* assembly strategy combining PacBio with Hi-C sequencing, we presented a high quality chromosome-level *P. clarkii* genome. The assembled genome is 2.75 Gb in size with a contig N50 of 216.75 kb. Transposable elements (TEs) make up the largest fraction of the genome (~79.61%), and LINEs comprise the majority of the TEs. Frequent molting and rapid growth of the red swamp crayfish may be explained by the expansion of multiple gene families regarding growth or development. Phylogenetic analysis revealed that *P. clarkii* diverged from *Portunus trituberculatus* at 278–407 million years ago (Mya). PSMC analysis identified multiple bottleneck events of the *P. clarkii* population between 2 kaBP to 14 kaBP. The obtained *P. clarkii* genome should not only facilitate us understanding the development and evolution of the crayfish species, but also contribute to the genetic improvement in future breeding selections.

1. Introduction

Crayfish belong to the largest crustacean taxon, the Decapoda, and are the only freshwater representatives of the Reptantia Macrura [1]. Crayfish have long been extensively utilized in commercially aquaculture or wild fishing for human consumptions, in aquarium trade as pets, in leisure fishery as popular fish bait, in scientific researches as model animals. Among commercially farmed crustacean species, red swamp crayfish *Procambarus clarkii* (Girard, 1852) was the second most produced species, which accounted for approximately 12% of the total productions of all cultivated crustaceans [2].

As an ecologically and economically important species, red swamp crayfish originates from the south-central United States and northeastern Mexico, which has now spread globally and has brought various impacts to ecological environment in many ways [3,4]. The use of crayfish as delicacy is in the roots of many cultural traditions, such as the Louisiana Swamp Thing & Crawfish Festival, in which families and friends gather to feast on the crayfish. There are written records of red swamp crayfish exploitation by European immigrants in Louisiana as early as the 1700s [5]. Red swamp crayfish has long been the centerpiece of Cajun cuisine in Louisiana. Records of commercial crayfish landings in Louisiana are available from the late 1800s [6]. Several thousand tons of crayfish have been harvested annually from natural wetlands in Louisiana from 1900 to the present. Commercial crayfish cultivation developed in 1950s in the USA, until now, crayfish aquaculture is practiced in most of the states in the southern USA [7]. Red swamp crayfish was first brought to Nanjing, China by the Japanese in 1930, although the reason is still unclear, and was subsequently introduced into many provinces in China in subsequent decades [3]. Because of its abundant nutrition and great taste, red swamp crayfish has now become one of the most favored leisure foods in China, especially in summer season. From 2007 to 2018, the annual aquaculture area of red swamp crayfish in China increased from 265,500 ha to 1,120,000 ha (~422% growth), producing about 1,638,700 tons of crayfish, which was worth approximately 7.2 billion USD (Fishery Bureau of Ministry of Agriculture PRC, 2018).

* Corresponding authors at: Freshwater Fisheries Research Institute of Jiangsu Province, 79 Chating East Street, Nanjing 210017, China.
E-mail addresses: zhiqiangx@163.com (Z. Xu), jstjq@163.com (J. Tang).
[1] These authors contributed equally to the study.

https://doi.org/10.1016/j.ygeno.2021.07.017
Received 17 March 2021; Received in revised form 28 May 2021; Accepted 14 July 2021
Available online 22 July 2021

In addition to its commercial values, red swamp crayfish has long been of interest to anatomists and physiologists, and have served as an excellent model organism in revealing the molecular regulatory mechanisms underlying various physiological processes, such as reproduction, sex determination, molting and growth, immunity, regeneration, and response to stress [8]. In addition, red swamp crayfish have also been considered as a perfect candidate for the research of evolution [9,10] and ecology [11,12]. As an invasive species in many countries, successful establishment of the populations might be facilitated by its high environmental tolerance, rapid population growth and superior competitive ability [13].

Whole genome sequencing of a given species would provide an important and fundamental tool to address important issues related to the protection and rational utilization. However, up to now, the limit genomic data available hinder the better explanation of complex traits regulation, as well as the studies regarding the ecological roles and evolution history of the red swamp crayfish. In recent years, although the gene-sequencing technology has progressed by leaps and bounds, due to the high degree of heterozygosity, abundant repeats and large size, chromosome-level genome assembly of a crustacean species is still challenging [14]. To date, only a draft genome of the crayfish species within the *Procambar*us genus, marbled crayfish *P. virginalis*, has been reported [15]. Here, based on a strategy combining the *de novo* PacBio assembly with Hi-C-supported scaffolding, we present a high-quality chromosome-level genome for the red swamp crayfish, and compare its genome composition with that of other crustaceans to better understand its evolutionary history leading to its massive genome expansion. The well annotated chromosome-level genome should facilitate our understanding of the development and evolution of the crayfish species, and should also contribute to the genetic improvement in the future breeding selection programs of the red swamp crayfish.

2. Materials and methods

2.1. Sampling, library construction, and sequencing

A healthy female red swamp crayfish with a body weight of 68.6 g was obtained from Hongze Lake in Jiangsu Province, China. The whole genomic DNA was extracted from the muscle of the crayfish using a DNeasyRBlood & Tissue Kit (Qiagen, Hilden, Germany). The DNA quality was measured with Qubit 3.0 (Invitrogen, Carlsbad, CA, USA) and was further checked using 1% agarose gel electrophoresis. Total RNA was isolated from 9 tissues, which included hepatopancreas, intestine, epidermis, eyestalk, gill, muscle, stomach, heart and ovary, using TRIzol extraction reagent (Thermo Fisher Scientific, USA). Equal quantities of RNA from every tissue were pooled together for the Illumina transcriptome sequencing. Another 20 healthy samples (51.8 ± 2.9 g) were randomly sampled from Hongze Lake and were used for the whole-genome re-sequencing to investigate the population history of the red swamp crayfish with Illumina HiSeq X sequencing.

Genome survey was performed by Illumina sequencing of 8 DNA libraries prepared with the genomic DNA. Transcriptome based annotation of the red swamp crayfish genome were performed by Illumina sequencing of a cDNA library prepared with the pooled RNA. Illumina libraries were constructed according to the manufacturer's standard protocol (Illumina) and sequenced on an Illumina HiSeq X Ten platform (Illumina, Inc., San Diego, CA, USA) using a paired-end 150 (PE150) strategy.

For PacBio Iso-Seq library preparation, approximately 30 μg of genomic DNA of the female red swamp crayfish above-mentioned was sheared into ~20 kb targeted size fragments with BluePippin (Sage Science, Beverly, MA, USA). The qualified libraries were prepared for single-molecule real-time (SMRT) genome sequencing using S/P2-C2 sequencing chemistry on the PacBio Sequel platform (PacBio, Pacific Biosciences, USA).

Two Hi-C libraries were prepared with the muscle of another female crayfish and sequenced to link scaffolds and anchor them to chromosomes.The Hi-C library construction included the process of cross-linking *in situ* using formaldehyde with a final concentration of 2% and homogenizing with tissue lysis by the restriction enzyme *Hin*dIII. The constructed Hi-C libraries with insert sizes of 300–700 bp were sequenced on an Illumina HiSeq X Ten platform (Illumina, SanDiego, CA, USA).

2.2. Genome estimation and de novo assembly

For the genome survey sequencing, Illumina reads were randomly selected and aligned to the Nucleotide Sequence Database (NT) using BLAST (version 2.2.31) [16] with the parameter of *E*-value = 1e-05 for pollution verification. The data were subjected to filter and correct by Fastp (version 0.19.3) [17], and then the data was processed by k-mer analysis for estimating the haploid genome size, heterozygosity, and repetitive content. The 21-mer depth distribution (k = 21) map was constructed by the data of eight Illumina libraries using Jellyfish (version 2) [18]. Genome size was estimated by the formula G = N21-mer (total number of k-mers) /D 21-mer (k-mer depth of the main peak). The repetitive content was accumulated from where the depth of k-mer was more than two times that of the main peak, and the heterozygosity were estimated at where the depth was half of the main peak.

Using the long single molecular reads for PacBio, the pipelines of workflow for genome assemblies were as follows: Firstly, the clean data of PacBio were subjected to error correction using Canu (version 1.5) [19] with the parameter of error correct coverage = 80. Subsequently, the outputs were piped into the workflow of SMARTdenovo (version 1.0) [20] with the parameters of J = 5000, A = 1000, and r = 0.95, resulting in the genomic contigs automatically generated. Finally, the preliminary assembly was polished three times by Racon (version 1.32) [21], realizing the first correction. In view of the relatively high error rate of the third-generation sequencing, Illumina reads specifically for genome estimation were utilized for the second correction by Pilon (version 1.22) [22], and the error correction was run for three times.

2.3. Hi-C scaffolding

The contigs generated by the draft genome assembly were further assembled by filling of gaps and anchoring on the putative chromosomes with Hi-C scaffolding method. The initial contigs were piped into the Hi-C assembly workflow, and the signals of chromatin interactions were captured to construct chromosome-level genome. In summary, the putative Hi-C junctions were aligned by the unique mapped read pairs using BWA-MEM (version 0.7.10-r789) [23]. The paired reads uniquely mapped to the assembly (the valid interaction pairs) were used for the Hi-C scaffolding. The invalid reads, which included reads of self-ligation, non-ligation and dangling ends, were filtered out using HiC-Pro (version 2.10.0) [24]. After performing error correction steps, the reassembled and corrected contigs were divided into ordered, oriented and anchored groups using LACHESIS [25] to generate the putative red swamp crayfish chromosomes with parameters: CLUSTER_MIN_RE_SITES = 12; CLUSTER_MAX_LINK_DENSITY = 2; CLUSTER_NONINFORMATIVE_RATIO = 1.8; ORDER_MIN_N_RES_IN_TRUN = 15; ORDER_MIN_N_RES_IN_SHREDS = 15. Gaps between ordered contigs were filled with LR GapCloser (version 1.1) [26].

2.4. Genome quality evaluation and repeats analysis

The quality of the obtained red swamp crayfish genome was evaluated using BUSCO analysis (version 3.0) [27], CEGMA analysis (version 2.5) [28] and Illumina short-read alignments with BWA-MEM (version 0.7.10-r789) [23].

Due to the low sequence conservation of the repeated sequences among different species, a specific repeat database was construct for the repeat prediction in the red swamp crayfish genome using LTR-FINDER

(version 1.05) [29] and RepeatScout (version 1.0.5) [30]. The identified candidate repeats were further classified by PASTEClassifer (version 1.0) [31]. Then a species-specific repeats library for the red swamp crayfish genome was generated by combining the constructed repeat database and Repbase (19.06) [32]. Based on the species-specific repeats library, the repeats characteristics for the red swamp crayfish were investigated with RepeatMasker (version 4.0.6) [33].

2.5. Genome annotation

A strategy combining *de novo*-, homology-, and transcriptome-based methods was used to predict protein-coding genes. (1) Five tools included Genscan (verson3.1) [34], Augustus (version 3.1) [35], GlimmerHMM (version 3.0.4) [36], GeneID (version 1.4) [37], and SNAP (version 2006-07-28) [38], were utilized for *de novo* prediction. (2) Protein sequences from *Eriocheir sinensis*, *Penaeus vannamei*, *Litopenaeus vannamei* and *Danio rerio* were aligned to the red swamp crayfish genome to perform homology-based prediction by using GeMoMa (version 1.3.1) [39]. (3) Transcriptomic data was utilized for the functional genes prediction by Hisat (version 2.0.4) [40], Stringtie (version 1.2.3) [41], TransDecoder (version 2.0) (http://transdecoder.github.io), and GeneMarkS-T (version 5.1) [42]. The candidate protein-coding genes identified with the above strategies were further integrated into a non-redundant protein-coding gene sets with EVM (version 1.1.1) [43] and PASA (version 2.0.2) [44].

The pseudogenes in the red swamp crayfish genome were identified with Genewise [45]. The non-coding RNAs were identified by referring to the miRbase database (version 21.0) [46] and Rfam (version 13.0) [47]. Functional annotations were performed based on the databases of EuKaryotic Orthologous Groups (KOG) [48], Kyoto Encyclopedia of Genes and Genomes (KEGG) [49], TrEMBL [50], Swiss-Prot [50], Non-redundant (Nr), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene ontology (GO) databases.

2.6. Gene family identification and evolutionary analyses

Whole genome data from another 8 closely related species, which included *Penaeus vannamei, Portunus trituberculatus, Eurytemora affinis, Hyalella Azteca, Armadillidium nasatum, Armadillidium vulgare, Daphnia magna* and *Amphibalanus Amphitrite* were obtained from the NCBI genome database (https://www.ncbi.nlm.nih.gov/assembly/). The proteins of the red swamp crayfish and the other 8 related species were aligned to each other using BLAST (version 2.2.31) [16] with a maximum e-value of $1e^{-}5$. The ortholog groups for the gene families from the 9 species were generally clustered using Orthofinder (version v2.3.7) [51]. These gene families were functionally annotated on the basis of their homology using PANTHER V15 database [52].

2.7. Phylogenetic analysis and population history reconstruction

The longest transcript of each gene in each species was utilized for the identification of single-copy orthologous genes among these species. The identified single-copy orthologs shared by the 9 species (including red swamp crayfish) were aligned using MAFFT v7.205 [53]. The phylogenetic relationships among the 9 species were investigated using the identified concatenated single-copy genes with IQ-TREE v1.6.11, and the maximum likelihood (ML) method was responsible for constructing the phylogenetic tree, in which the number of bootstrap was set to 1000 [54]. Divergence times among species were analyzed with MCMCtree program of the PAML package (version 4.9i) with the approximate likelihood calculation method [55]. Molecular clock data from the TimeTree database (http://www.timetree.org/) were used as the calibration times. Based on the divergence times and phylogenetic relationships, gene family evolutions were analyzed with CAFÉ (version 4.2) [56]. The gene family expansion and contraction were analyzed by comparing the differences between the ancestor and involved species. The extended family genes for the red swamp crayfish were extracted and aligned to the functional enrichment on GO and KEGG to determine their underlying functions.

The population history of the red swamp crayfish was investigated using the Pairwise Sequentially Markovian Coalescent (PSMC) model. Single nucleotide polymorphisms (SNPs) were identified using bcftools (version 1.3.1) [57]. Population history was analyzed by PSMC (0.6.5-r67) [58] using generation time and mutation rate information. PSMC analysis was performed using the parameters "-N25 -r5 -p 4 + 25*2 + 4+6," where —N is the maximum number of iterations, −r is the initial theta/rho ratio and -p is the atomic time intervals. Results were delivered to the plot script using the parameters "-g 1 -u 4.59 e-9 -Y 100," where -g is the number of years per generation and -u is the mutation rate per nucleotide per generation [59].

3. Results

3.1. Genome sequencing assessment

A total of 310.03 Gb clean data with an insert size of 350 bp was produced by Illumina HiSeq sequencing (Genome survey sequencing), and a total of 38.57 million PacBio reads (403.31 Gb) was produced by PacBio sequencing (*de novo* whole genome sequencing). Eight libraries were constructed for the Illmina sequencing, and the data covered the depth of 112.95× for the genome of the red swamp crayfish. Ten libraries were constructed for the PacBio sequencing, and the data covered the depth of 146.93× for the genome of the red swamp crayfish. The maxium length subread for PacBio offline was 144.97 kb; the N50 and mean length of subreads were 15.46 kb and 10.46 kb, respectively. For the chromosome conformation capture, two Illumine libraries were sequenced, yielding a total of 439.15 Gb (142.23×) clean data for Hi-C scaffolding. Approximately 8 Gb transcriptomic data was further obtained and used for the transcriptome based annotation of the red swamp crayfish genome (Table S1).

3.2. Genome size estimation and assembly

More than 222.22 billion k-mers were acquired and used for the genome size estimation, giving an estimated genome size of 3.07 Gb. The genome heterozygosity rate was predicted to be about 0.71%, its repeat sequence content was about 73.10%, and its GC content was about 42.99% (Fig. 1a). The draft genome assembly of the red swamp crayfish was performed using high-quality PacBio sequencing reads, and the assembled draft genome is about 2.75 Gb in size with a contig N50 of 218.22 kb.

A total of 1.25 billion read pairs (85.18%) of Hi-C data were mapped to the draft genome assembly, and 287.94 million valid interaction pairs (19.63%) were used to build the interaction matrices (Table S2). The contigs of the draft genome (31,927 contigs) were broken and reassembled using the valid interaction pairs, yielding 31,961 corrected contigs. The final assembly presented a high-quality red swamp crayfish genome of 2.75 Gb in size, with a contig N50 of 216.75 Kb and a scaffold N50 of 17.01 Mb. The final genome comprised 24,505 scaffolds, and the longest contig and scaffold were 3.76 Mb and 37.84 Mb, respectively (Table 1). A total of 2.50 Gb of genomic sequences, which accounts for 91.22% of total sequences, were assigned to the 94 haploid chromosomes (Fig. 1b). Among the 31,961 corrected contigs, 18,843 contigs (58.96%) were anchored onto the 94 haploid chromosomes, and 1.88 Gb (75.11%) of genomic sequences were anchored with a defined order and orientation (Table S3).

3.3. Genome assembly evaluation and repetitive genome elements

The assembled red swamp crayfish genome covered 94.18% (1004) and 91.26% (418) of the complete core genes in the BUSCO and CEGMA database respectively. The lllumina sequencing data were further used

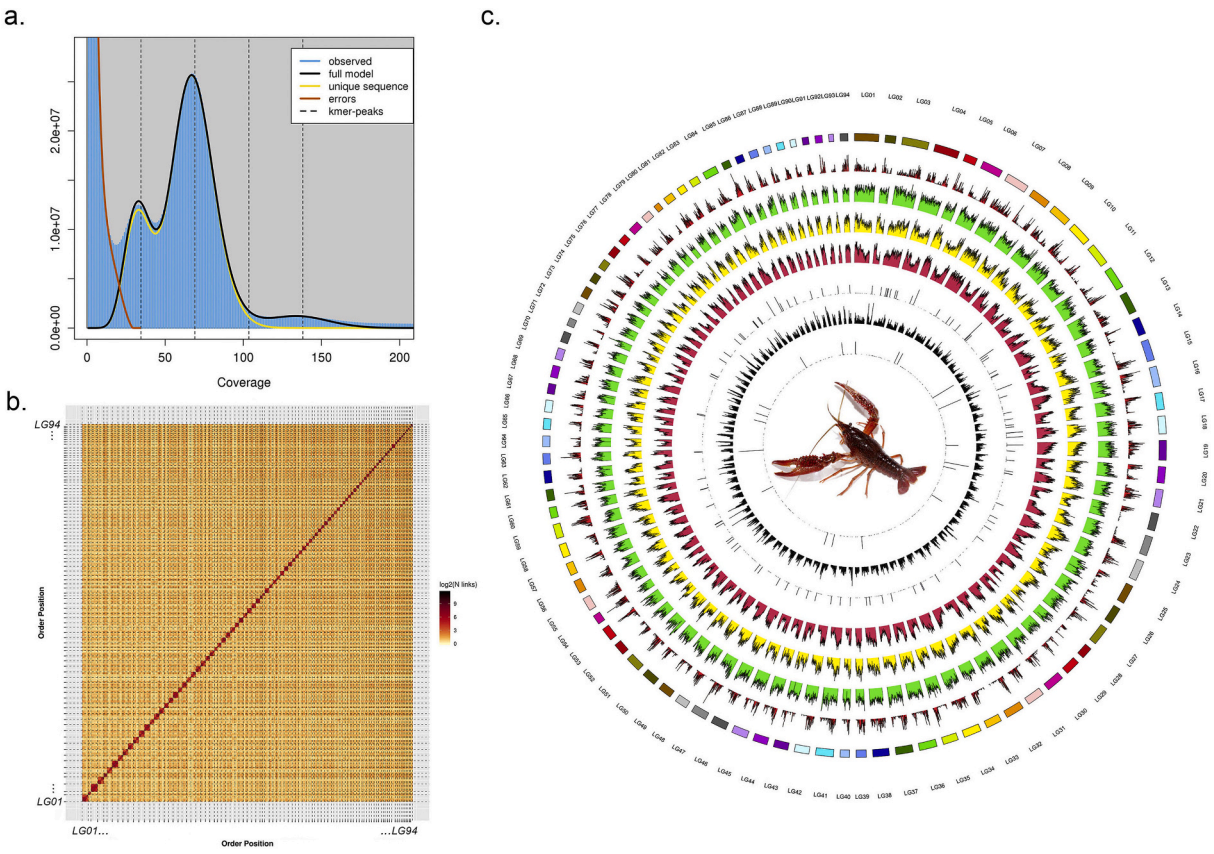

Fig. 1. The genomic characteristics of the red swamp crayfish. a. The genome survey sequencing of the red swamp crayfish. Genome size was estimated by the formula G = N21-mer (total number of k-mers) /D 21-mer (k-mer depth of the main peak). The repetitive content was accumulated from where the depth of k-mer was more than two times that of the main peak, and the heterozygosity were estimated at where the depth was half of the main peak. b. The genome-wide Hi-C heatmap of the red swamp crayfish. By using the Hi-C scaffolding, 2.50 Gb genomic sequences that accounts for 91.22% of the draft genome, were assigned to the 94 haploid chromosomes. c. Genomic characteristics of the Red swamp crayfish. Track 1 (from the outer-ring): 94 linkage groups (LGs) of the red swamp crayfish genome. Track 2: Gene density with sliding windows of 1 Mb. Higher density is shown in darker red colour. Track 3: GC content within a 1-Mb sliding window. Track 4: Distribution and density of the long tandem repeats density (LTR). Track 5: Distribution and density of the long interspersed nuclear elements (LINE).Track 6: Distribution and density of the short interspersed nuclear elements (SINE). Track 7: Distribution of the tRNAs. Track 8: Distribution of the rRNAs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Statistics and characteristics of the genome for the red swamp crayfish.

Characteristics	Number	Size	Percentage
Estimate of genome size		3.07 Gb	
Final assembly genome size		2.75 Gb	
Contig number and N50	31,961	216.75 Kb	
Maximum contig		3.76 Mb	
Scaffold number and N50	24,505	17.01 Mb	
Maximum scaffold		37.84 Mb	
Total repetitive sequences		2.26 Gb	82.42%
Total protein-coding genes	26,549	0.34 Gb	
Annotated protein-coding genes	18,436		69.44%
MicroRNA	31		
Ribosomal RNA (rRNA)	875		
Transfer RNA (tRNA)	19,583		

to assess the completeness of the red swamp crayfish genome assembly, and we found that 98.08% of the short reads could be mapped to the assembly (Table S4). These results indicate the high integrity and accuracy of the assembled red swamp crayfish genome.

More than 2.26 Gb of repeats, representing 82.42% of the total genomic sequences, were identified in the red swamp crayfish genome. Among these repeats, transposable elements (TEs) make up the largest fraction of the genome (~79.61%), which had certainly raised a challenge for the genome assembly of red swamp crayfish. Both retrotransposons and DNA transposons were identified in the red swamp crayfish genome. In the group of retrotransposons, long terminal repeat (LTR) comprises 14.47% of the total repeats in the red swamp crayfish genome. Among the non-LTR retrotransposons, long interspersed nuclear elements (LINEs) comprise the majority of the TEs, spanning 1.04 Gb of the whole genome (45.91% of total repeats) (Fig. 2a, Table S5a).

The number of substitutions to repeat consensus, which is an estimate of the relative age of the LTR, implied that the red swamp crayfish LTR has undergone a recent and apparent burst of expansion with a peak within 1 million years ago (Mya) (Fig. 2b). Insertion profile analysis showed that the TEs are mostly located in intronic and intergenic regions, however, some of them can also be found in the protein coding exons or the 5′/3′-flanking regions of the functional genes, implying the potential regulatory roles of these TEs in the red swamp crayfish genome (Fig. 2c). The neighboring genes of the LTR, such as the cuticle protein genes, chitinase genes, CHH genes and Hox genes, are frequently reported to participate in the growth or development of the crustaceans (Table S5b). Therefore, LTR elements in red swamp crayfish might possess the functions of regulating the expression of neighboring genes by coopting them into the regulatory networks in growth and development.

3.4. Genome annotation

By using a strategy combining *de novo*, homology-based, and

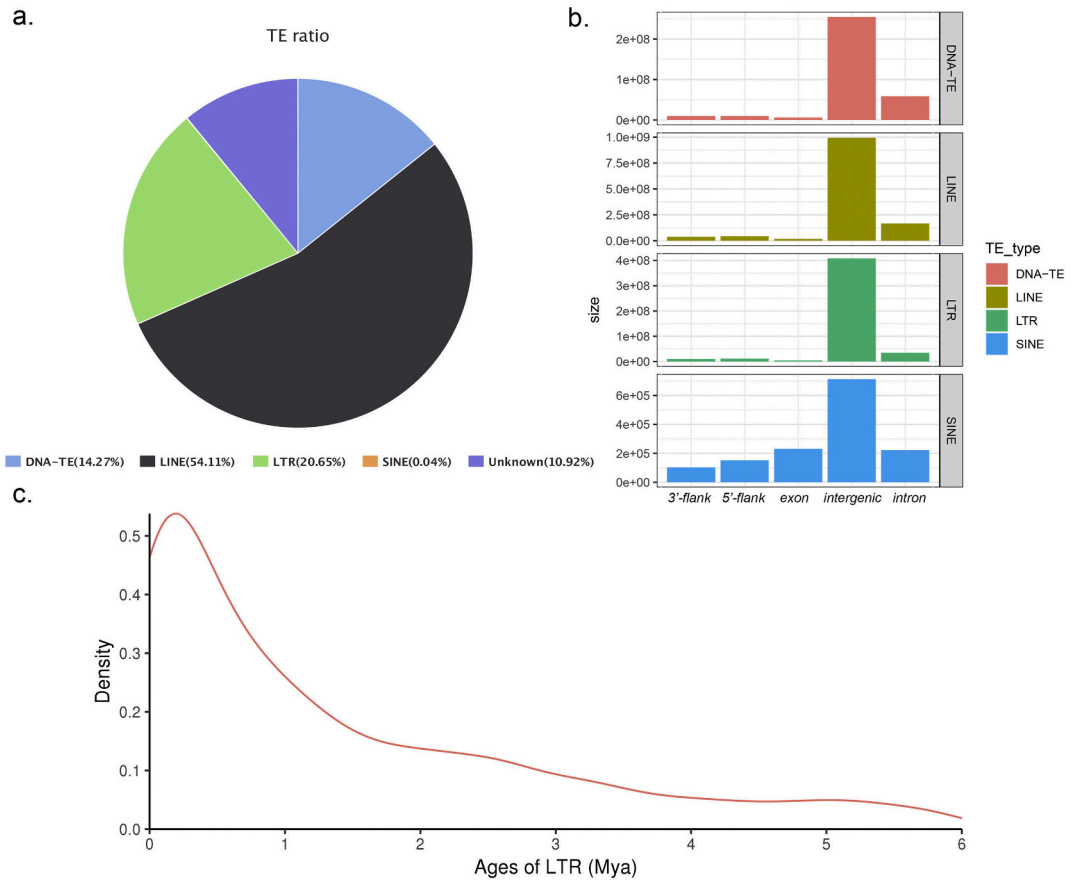

Fig. 2. Composition and evolution of the transposable elements (TEs) in the red swamp crayfish genome. a. Composition analysis of the major TEs. LTR: long terminal repeat; LINE: long interspersed nuclear elements; SINE: short interspersed nuclear elements. b. Distribution trends of the major TEs in the red swamp crayfish genome. c. Divergence time of LTR in the red swamp crayfish genome. The Kimura distance-based copy divergence analysis was calculated using a mutation rate of 4.59e-9 in *D. pulex*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

transcriptome-based predictions, 341.57 Mb (12.42%) sequences of the 2.75 Gb assembled genome were predicted to encode 26,549 non-redundant protein-coding genes with a mean exon number of 5.08 per gene. The average gene length was 12,865.77 bp and the average exon and intron lengths were 1790.73 and 11,075.04 bp per gene, respectively (Table S6). Among the 26,549 predicted protein-coding genes, 18,436 genes (69.44%) were annotated in at least one functional protein database (Table S7 and S8). In addition, 2185 pseudogenes, 31 miRNAs, 19,583 tRNAs and 875 rRNAs were annotated in the assembled genome, respectively (Table S9). Based on the information of genome annotation, the whole red swamp crayfish genome landscape showing the distributions of genes, repeat sequences, and GC content were recognized (Fig. 1c).

3.5. Gene family identification and expansion/contraction analysis

A total of 23,768 gene families among the nine species were identified (Table S10), among which, 10,745 gene families clustered by 26,549 protein-coding genes were identified in the red swamp crayfish genome (Table S11). Compared with other published crustacean genomes, the red swamp crayfish genome had 663 specific gene families that consisted of 2008 genes (Fig. 3a). Pathway enrichment analysis of

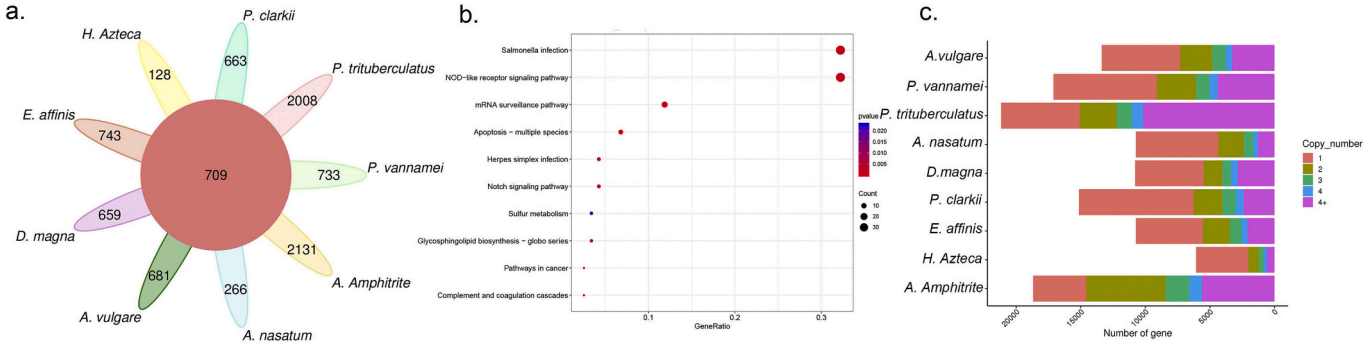

Fig. 3. Distribution and characterization of the specific gene families in the red swamp crayfish. a. The distributions of the gene families among the 9 crustacean species. b. KEGG pathway enrichment of the specific gene families in the red swamp crayfish. c. Orthologous gene numbers among the 9 crustacean species. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

these specific genes revealed 21 significantly enriched (q-value <0.05) GO terms and 9 significantly enriched (q-value <0.05) KEGG pathways. The top three most significantly enriched KEGG pathways were "NOD-like receptor signaling pathway", "Salmonella infection" and "mRNA surveillance pathway" (Fig. 3b, Table S12 and S13).

Among the gene families, the number of gene copies of 2, 3, 4, and 4+, were 2210, 1056, 608 and 2386, respectively (Fig. 3c, Table S14). Comparing with the other 8 species, red swamp crayfish possess the most abundant single-copy orthologs (8880 single-copy genes). The total number of multiple copy orthologs was less than that of single-copy orthologs, indicating the conservativeness of the red swamp crayfish genome. Gene family evolution analysis revealed 93 significantly expanded gene families and 64 significantly contracted gene families in the red swamp crayfish genome ($P < 0.05$). Underlying the long-term adaptation of red swamp crayfish is the expansion of gene families enriched in multiple basic metabolic pathways, such as cuticle protein gene families, crustacean hyperglycemic hormone (CHH) gene family, and cytochrome P450 gene family (Table S15). Functional pathway enrichment analysis of the expanded genes identified 61 significantly enriched (q-value <0.05) GO terms (Table S16) and 44 significantly enriched (q-value <0.05) KEGG pathways (Table S17). Among the enriched KEGG pathways, the top three abundant pathways were "P53 signaling pathway", "Sulfur metabolism", and "Phagosome".

3.6. Identification of key genes associated with growth or development

3.6.1. Cuticle protein genes

Two groups of cuticle protein genes, which include CPR family and CPAP family, were identified in the red swamp crayfish genome (Fig. 4 a, b, c). The most abundant family of cuticle proteins, the CPR family, which is recognized by the Rebers and Riddiford (RR) Consensus [60], a domain of about 64 amino acids that binds to chitin, is present throughout the 9 crustaceans. Red swamp crayfish CPR family can be divided into two subfamilies as RR1 and RR2. Among the RR2 family, we identified 27 CaCPs, which are a group of cuticle proteins being associated with calcification of the exoskeleton in crustaceans (Fig. 4c). Comparing with other crustaceans, red swamp crayfish has the largest number of CaCPs among the different kind of crustaceans being involved in the present study.

3.6.2. Chitinase genes

Chitinases are glycoside hydrolases (GHs) that break down the β-1,4 glycosidic bonds in amino polysaccharides such as chitin and chitooligosaccharides, which take part in multiple physiological processes, including tissue degradation and remodeling, developmental regulation, as well as immune response regulation [61]. Twenty five chitinases belonging to GH18 family were identified in the red swamp crayfish genome (Fig. 4d), which can be further divided into 6 groups (*Pc*-Cht1, *Pc*-Cht2, *Pc*-Cht3, *Pc*-Cht4, *Pc*-Cht5 and *Pc*-Cht6). Further protein structure analysis showed that each chitinases of the red swamp crayfish contain one to five Glyco_18 domains, implying their functions of catalyzing the biodegradation of β-1,4 glycosidic bond in chitins.

3.6.3. Hox genes

Hox genes encode a family of homeodomain transcription factors, which play vital roles in development, such as specifying anterior-posterior identity [62]. The body plan of the red swamp crayfish is of fairly typical of the crustacean order; therefore, Hox genes should possess critical functions for the growth or development of the red swamp crayfish. All 10 canonical Hox genes that commonly exists in the arthropod ancestor, including labial (*lab*), proboscipedia (*pb*), *Hox3*, Deformed (*Dfd*), Sex combsreduced (*Scr*), fushi tarazu (*ftz*), Antennapedia (*Antp*), Ultrabithorax(*Ubx*), abdominal-A (*abd-A*) and Abdominal-B (a*bd-B*), were identified in the red swamp crayfish. Subsequent copy number analysis revealed that red swamp crayfish has a single copy of each of these canonical Hox genes. Besides the canonical Hox genes, another 12 non-canonical Hox genes were further identified in the red swamp crayfish. The 12 non-canonical Hox genes are presented throughout the entire 9 crustacean (Fig. 5a, b, c). Among the10 canonical and 12 non-canonical Hox gene members, 14 Hox genes were assigned to the chromosome of LG 15, LG 19, LG 30, LG 33, LG 57, LG 72, and LG 93, respectively (Fig. 5d). At least six canonical Hox genes are likely to link to each other in a cluster in the same chromosome (spanning1.02 Mb), which included the *lab*, *pb*, *Dfd*, *Scr*, *ftz* and *Antp* genes. The spatial collinearity of these six Hox genes, which belong to the group of Antennapedia complex in *D. melanogaster,* is conserved with other arthropod [63]. However, the non-canonical Hox genes are not likely organized into clusters as canonical Hox genes.

3.7. Phylogenetic analysis and molecular evolution

A total of 172 single-copy genes were identified and utilized to investigate the evolutionary relationship between the red swamp crayfish and other crustaceans (Fig. 6). Red swamp crayfish*, P. trituberculatus and P. vannamei* formed a clade of Malacostraca, which diverged from *H. Azteca, a* member of the Maxillopoda, at approximately ~394–521 Mya. This divergence occurred during the late Cambrian and Ordovician, and this period was known as the "age of invertebrates" [64]. This divergence period is also consistent with the radiation of shrimp-like decapods during the period earlier than 437 Mya [65]. The phylogenetic tree suggested that red swamp crayfish and its closest relative *P. trituberculatus* diverged at ~278–407 Mya, and diverged from *P. vannamei* at an early stage of ~306–441 Mya, which should further confirm the evolutionary transition from marine to freshwater habitats.

3.8. PSMC analysis of effective population size

PSMC analysis was performed to estimate historic population dynamics of the red swamp crayfish, and multiple bottleneck events between 2 kilion anniversary before present (kaBP) to 14 kaBP were detected in the red swamp crayish population (Fig. 7). A population declining trend was detected at approximately 139 kaBP. Another three significant population declining events were detected at 14 kaBP, 4 kaBP and 2 kaBP respectively, which were concordant with the last glacial period in northern America (18 kaBP-12 kaBP) and a catastrophic worldwide drought at 4.2 kaBP [66,67]. In addition, a population expansion at 1.5 kaBP has also been detected for the red swamp crayfish.

4. Discussion

Genome assembly of non-model organisms was financially unrealistic until the advent of high-throughput next generation sequencing in recent years [68]. The fisheries genomics have made significant progress during the past decade since the first key fisheries species, Atlantic cod *Gadus morhua* genome was reported [69]. In the past few years, thanks to the rapid development and constantly falling cost of high-throughput sequencing technique, the whole genome sequencing and genome assemblies have been performed for an increasing number of aquaculture species [14]. The whole genomes data has not only provided genetic basis for the determination or evolution of phenotypic characteristics, but also contributed useful resources for the genetic improvement for these aquaculture species. However, the quality of the genomes submitted in public database varies greatly in different genome projects, which can be attributed to many factors such as the genome complexity, the sequencing and assembly strategy. Many factors intrinsic to the genome itself influence the assembly quality of a genome, particularly in the features of heterozygosity, repeat content, whole genome duplication, and ploidy [14]. In general, for a certain species, a chromosome-level genome assembly should be of more useful than a scaffold-level genome assembly in the research of evolutionary and gene regulatory mechanisms [70]. In recent years, owing to the widely application of the third generation sequencing and high-resolution chromosome

a.

Species	CPAP	CPR	CaCP	Chitinase
A.amphitrite	105	376	0	33
A.nasatum	44	19	2	14
A.vulgare	58	28	0	20
D.magna	142	276	0	34
E.affinis	116	91	0	31
H.azteca	46	46	1	24
P.clarkii	**125**	**225**	**27**	**25**
P.trituberculatus	54	66	0	17
P.vannamei	174	147	16	42

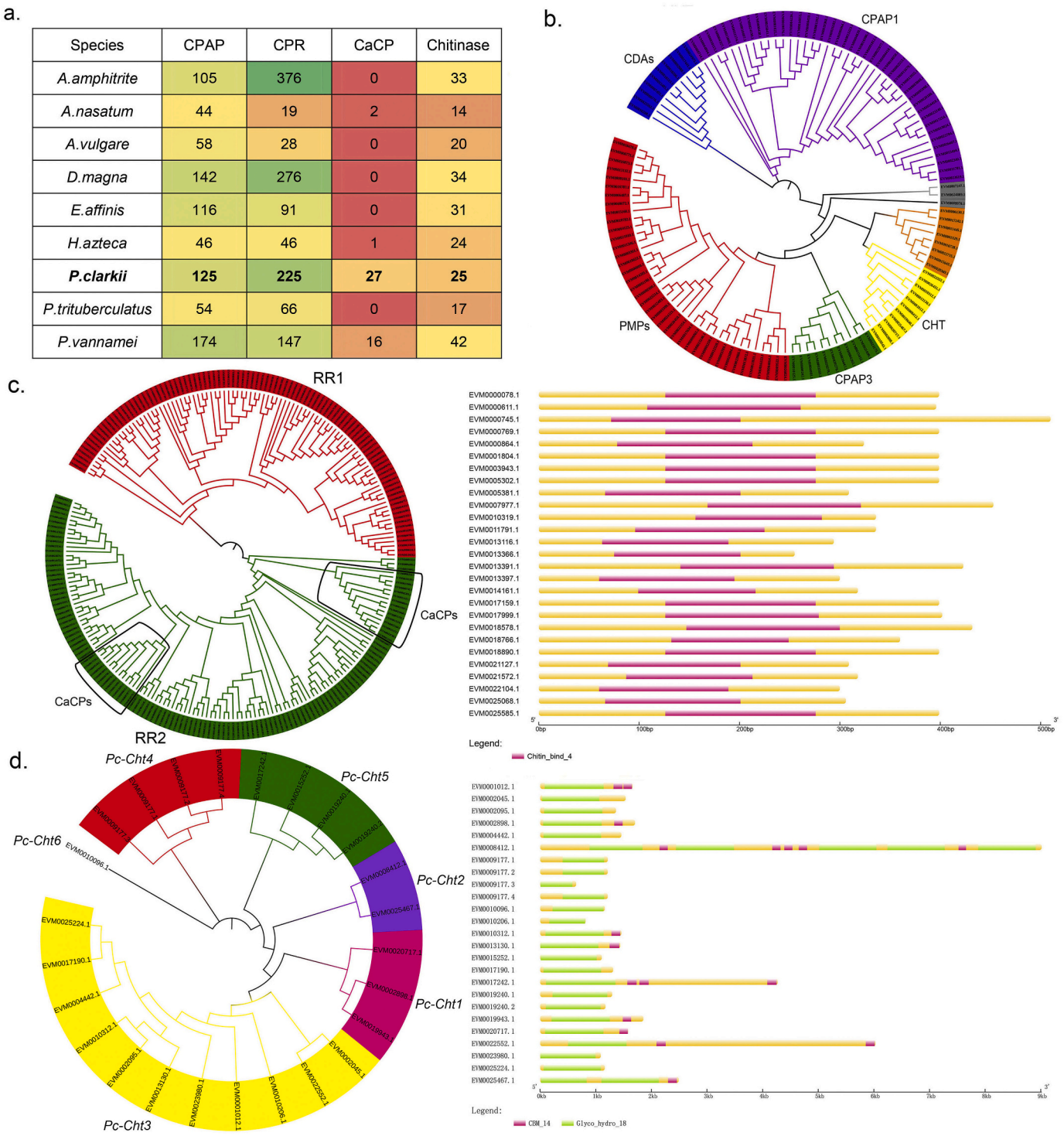

Fig. 4. Identification of molting related gene families in the red swamp crayfish. a. Distribution pattern of the molting related gene families (CPAP, CPR, CaCP and Chitinase) in red swamp crayfish and comparison with other crustaceans. b. Expansion of CPAP gene families in the red swamp crayfish. Sub-families with distinct domain structures were highlighted in different colors in the dendrogram. c. Expansion of the CPR gene families and structures of the CaCPs in the red swamp crayfish. The two subfamilies of RR1 (Rebers and Riddiford 1) and RR2 of the CPR family was highlighted with different colors in the dendrogram. Among the RR2 family, 27 CaCPs that are associated with calcification of exoskeleton in crustaceans were identified. The chitin-binding domains of each CaCPs were marked with pink in the gene structure plots. d. Expansion and structure analysis of the Chitinase gene families of the red swamp crayfish. The different subfamilies were highlighted with different colors in the dendrogram (*Pc*-Cht1, *Pc*-Cht2, *Pc*-Cht3, *Pc*-Cht4, *Pc*-Cht5 and *Pc*-Cht6). One to five Glyco_18 domains, which are responsible for the catalyzing the biodegradation of β-1,4 glycosidic bond in chitins, were highlighted with green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

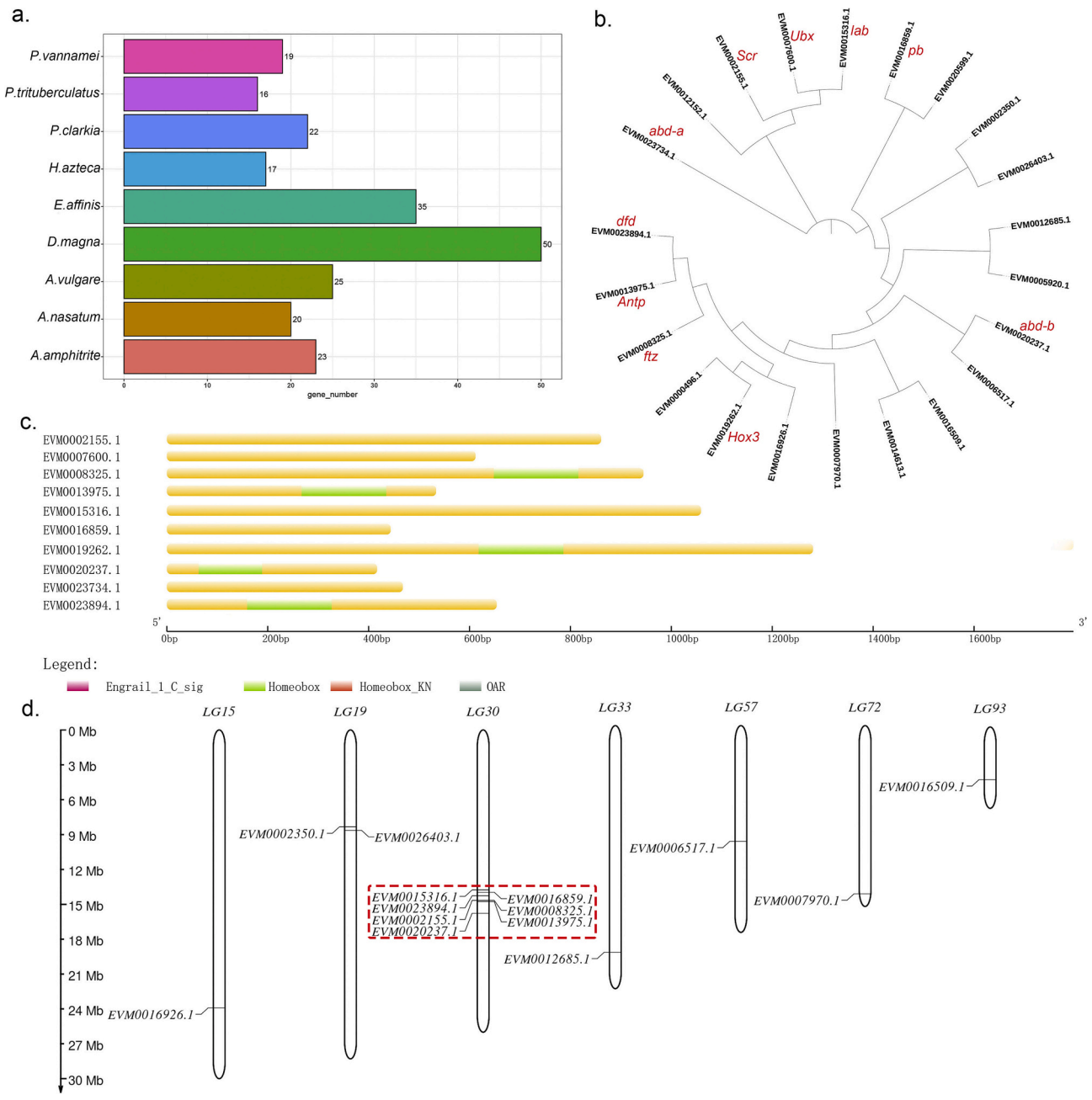

Fig. 5. Identification of Hox genes in the red swamp crayfish. a. Distribution of the Hox genes in red swamp crayfish and other related crustaceans. b. Phylogenetic relationship of the Hox genes of the red swamp crayfish. Annotations of the 10 canonical Hox gene members were highlighted in red in the dendrogram, which included labial (*lab*), proboscipedia (*pb*), *Hox3*, Deformed (*Dfd*), Sex combsreduced (*Scr*), fushi tarazu (*ftz*), Antennapedia (*Antp*), Ultrabithorax(*Ubx*), abdominal-A (*abd-A*) and Abdominal-B (*abd-B*). c. Genetic structural characteristics of the 10 canonical Hox genes. d. Assignment of the Hox genes to the chromosome of red swamp crayfish. Among the 22 Hox gene members,14 Hox genes were assigned to the chromosome of LG 15, LG 19, LG 30, LG 33, LG 57, LG 72, and LG 93, respectively. The six canonical Hox genes in the same chromosome (spanning1.02 Mb) were highlighted in a red rectangular box. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conformation capture techniques, a great number of chromosome-level genomes of aquaculture animals have been assembled and reported, which included the Pacific white shrimp *Litopenaeus vannamei* and the swimming crab *Portunus trituberculatus* [71,72]. The sequencing and assembly of the red swamp crayfish genome were challenging due to the big size, high heterozygosity and abundant repeats as revealed by genome survey analysis. We therefore tried to apply a sequencing and assembly strategy combining a *de novo* PacBio assembly with Hi-C-supported scaffolding to acquire a chromosome-level genome of the red swamp crayfish. We have finally obtained a high-quality chromosome-level genome sequence assembly of 2.75 Gb in size with a contig N50 of 216.75 kb and a scaffold N50 of 17.01 Mb. It has been well documented that TEs amplification might contribute to the genome expansion in various organisms [73]. In the present study, comparing with other crustaceans, the highest proportion of TEs (~79.61%) was identified in the red swamp crayfish genome. Previous studies have showed that TEs might possess critical biological significance owing to their interaction with the host genome [74]. In red swamp crayfish, insertion profiles revealed that a fairly large number of TEs were located in the 5′/3′-flanking sequences of coding genes with various functions, implying critical functions of these TEs in generating genetic variation underlying adaptive evolution, which are worthy of further and in-depth

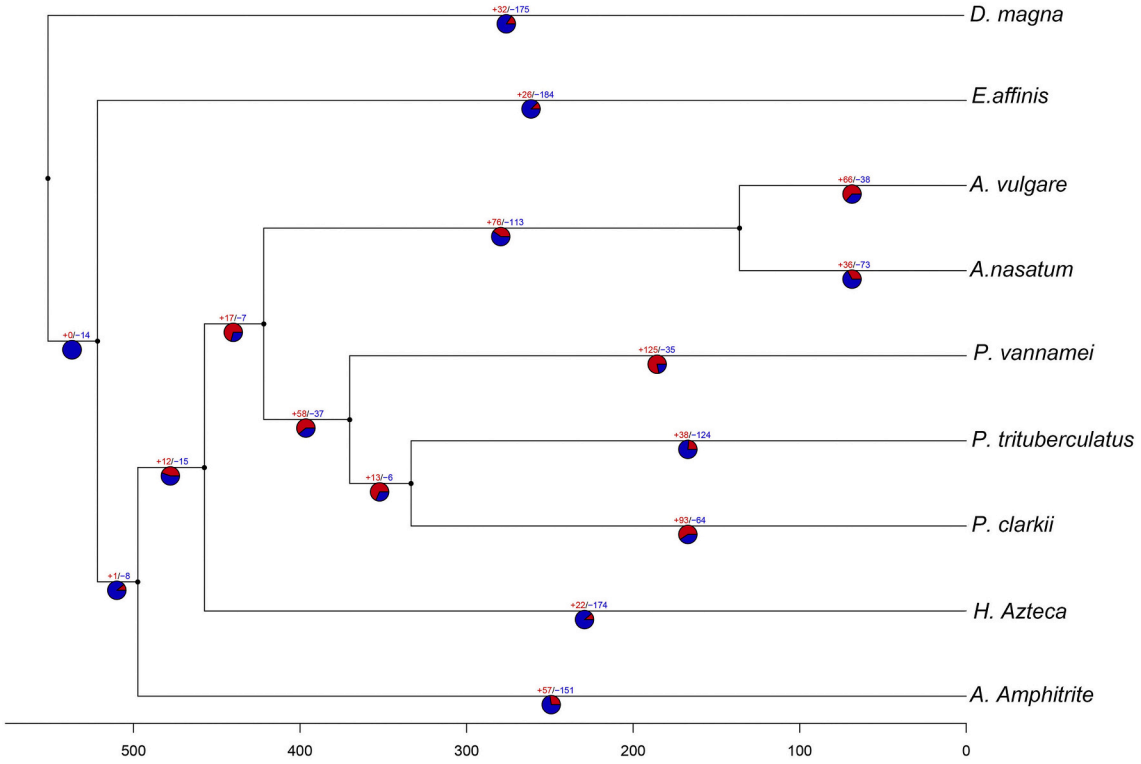

Fig. 6. Phylogenetic relationship and the probable divergence time of the red swamp crayfish. Numbers on branches indicate the number of gene gains (+) or losses (−). The estimated divergence times are displayed below the phylogenetic tree. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

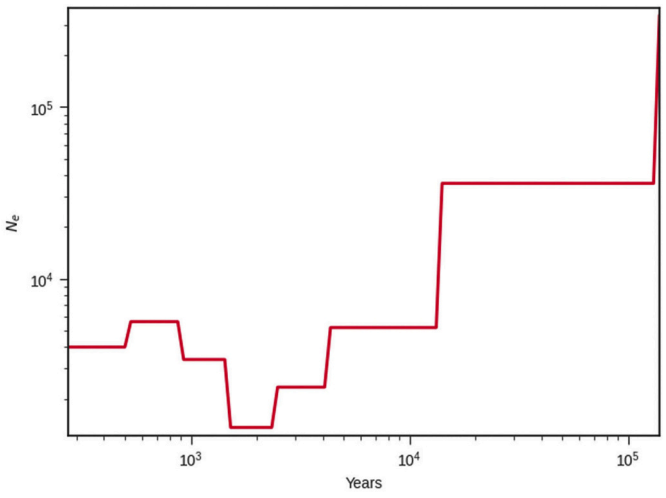

Fig. 7. Inferred historical population sizes of the red swamp crayfish by pairwise sequential Markovian coalescent (PSMC) analysis. The generation time of 1 years and the mutation rate per nucleotide per generation of 4.59e-9 were used in the PSMC analysis (the mutation rate was derived from a closely related crustacean *D. pulex*). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

investigation in future studies.

Based on the comparative genomics analysis with the reported crustacean genomes, significant expansion or contraction in the size of some particular gene families (such as the cuticle-related genes), were detected in the red swamp crayfish. The cuticle of crustaceans covers the entire outer surfaces and the regions of the gastrointestinal tract, which provides a physical barrier against external injuries or pathogenic infections [75].Cuticle proteins are main structural proteins responsible for the constructing the cuticle in crustaceans. As a vital process in crustaceans, molting is tightly linked to growth, metamorphosis and reproduction. During the molting cycles, to allow for the growth in body size or development, the chitin in the cuticle is degraded prior to molting by chitinase [76]. Therefore, cuticle protein and chitinase directly participate in the periodic molting process, which should be closely associated with multiple growth- and development-related traits in crustaceans. Identification and elucidating the regulatory mechanisms of the cuticle protein synthesis should provide useful clues for genetic improvement of the red swamp crayfish in future breeding selection programs focusing on the rapid growth traits.

Chitinases are glycosyl hydrolases (GH) that can break down the β-1,4 glycosidic bonds of amino polysaccharides (such as chitin) into *N*-acetylglucosamine, and are widely involved in the molting process and even immune response in various crustaceans [76]. Crustaceans are reported to possess a greater number of chitinases than other organism. Chitinases are produced during molting for the shedding of old cuticle in crustaceans. The catalysis and structural characterization of chitinase revealed in various studies demonstrated that the enzyme was indispensable to molting in crustaceans [77,78]. In our present study, comparative genomics analysis revealed that multiple chitinases families were present throughout the 9 crustaceans, implying the key functions of chitinases in life activities of various crustacean species.

In Arthropods, calcification of the cuticle seems to occur only in crustacean species. Although a calcified cuticle will add extra weight to the body, this does not appear to affect the action of aquatic crustaceans significantly. However, for most terrestrial Arthropods, especially insects, might have evolved not to keep calcium carbonate during the evolution from aquatic to terrestrial environments [60]. In the present study, comparing with other 8 crustaceans, it is interest to find that red swamp crayfish possess the largest number of CaCPs, which are a group of cuticle proteins being associated with calcification of the exoskeleton in crustaceans. Future functional studies are required to explain the expansions of cuticle protein families in the red swamp crayfish,

especially to elucidate the critical roles of CaCPs in cuticle constructing and calcification of the exoskeleton.

The highly conserved homeotic (Hox) genes play essential roles in establishing regional identity along the anterior-posterior (AP) axis in bilaterian animals, and have been involved in determination of morphological diversity throughout evolution [79]. Homologs of Hox genes have been cloned from various arthropod classes and some outgroups of the Arthropoda [80,81]. Considering the vital roles of Hox genes in establishing regional and segmental identity, various studies have provided numerous molecular, genetic, and developmental evidences for the Hox genes in generating morphological diversity during evolution [77]. In the present research, Hox genes are present throughout the 9 crustaceans through comparative genomics research, implying the vital functions of Hox genes in life activities of crustaceans. Hox genes are commonly retained in multiple copies following whole-genome duplication, for instances, 50 Hox gene members were identified in *D. magna*. In red swamp crayfish, however, we found only a single Hox complement. In most invertebrates or vertebrates, HOX genes are likely to be clustered in some specific chromosomal regions, their orders are likely to be conserved across animals, and which usually reflects where they are expressed along the anterior/ posterior axis [81]. In the red swamp genome, at least six canonical Hox genes are likely to link to each other in a cluster in the same chromosome, and the spatial collinearity of these six Hox genes is conserved with other arthropod [63]. Further investigations are necessary to further determine whether the other 4 canonical Hox genes lie together in another single cluster, as well as the regulation of the distinct spatial and temporal expression patterns of these canonical and non-canonical Hox genes in the growth or development of the red swamp crayfish.

In a word, a high-quality reference genome assembly with functional annotations is essential for the understanding of evolution and the regulation of important biological characteristics such as molting and development, as well as for the future genetic improvement programs. Our assembled red swamp crayfish genome should provide us a useful resource not only for the revealing the regulatory mechanisms of various physiological processes, but also for the genome-wide association studies and genomic selection in future breeding studies.

Data availability

Genome sequences data that support the findings of this study have been deposited in NCBI GenBank under the accession codes of PRJNA 727411.

Declaration of Competing Interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted. This manuscript has not been submitted to any other journals for publication and has been reviewed and approved by all the authors.

Acknowledgements

This work was supported in part by National Key R&D Program of China (no. 2020YFD0900304); Project from Jiangsu Province Science and Technology Agency (no. BE2020348); Jiangsu Agricultural Industry Technology System (no. JATS[2020]365, JATS[2020]366); China Agriculture Research System (no. CARS-48); Agriculture Breeding Project supported by Jiangsu Provincial Department of Agriculture and Rural Affairs (no. PZCZ201746); Jiangsu Agricultural Science and Technology Innovation Fund (no. CX(20)3005). We appreciate Biomarker Technologies Co., Ltd. (Beijing, China) for the high-throughput sequencing and bioinformatics analysis service. We declare that we have no conflict of interest other than that stated in the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.07.017.

References

[1] G. Scholtz, S. Richter, Phylogenetic systematics of the reptantian Decapoda (Crustacea, Malacostraca), Zool. J. Linnean Soc. 113 (1995) 289–328.
[2] FAO, FAO Yearbook of Fishery and Aquaculture Statistics, 2019.
[3] S.C.H. Cheung, The social life of American crayfish in Asia, in: James Farrer (Ed.), Globalization, Food and Social Identities in the Asia Pacific Region, Sophia University institute of Comparative Culture, Tokyo, 2010.
[4] M.J. Moore, R.J. DiStefano, E.R. Larson, An assessment of life-history studies for USA and Canadian crayfishes: identifying biases and knowledge gaps to improve conservation and management, Freshwater Sci. 32 (2013) 1276–1287.
[5] M. Comeaux, Atchafalaya Swamp Life: Settlement and Folk Occupations, Louisiana State University, School of Geoscience, Geoscience and Man 2, Baton Rouge, 1972.
[6] D.L. Gary, The Commercial Crawfish Industry in South Louisiana, Louisiana Sea Grant College Program, Louisiana State University, Baton Rouge, 1974.
[7] D. Holdich, Biology of Freshwater Crayfish, Blackwell Publishing Ltd, Oxford, 2002.
[8] M. Longshaw, P. Stebbing, The Management of Invasive Crayfish, 2016.
[9] W. Bode, F.X. Gomis-Rüth, R. Huber, R. Zwilling, W. StöCker, Structure of astacin and implications for activation of astacins and zinc-ligation of collagenases, Nature. 358 (6382) (1992) 164–167.
[10] K. Crandall, T. Cronin, The molecular evolution of visual pigments of freshwater crayfishes (Decapoda: Cambaridae), J. Mol. Evol. 45 (1997) 524–534.
[11] D.M. Lodge, Biological invasions-lessons for ecology, Trends Ecol. Evol. 8 (1993) 133–137.
[12] J.E. Garvey, R.A. Stein, H.M. Thomas, Assessing how fish predation and interspecific prey competition influence a crayfish assemblage, Ecology 75 (1994) 532–547.
[13] J.O. Francisco, I.S. Marta, C. Miguel, One century away from home: how the red swamp crayfish took over the world, Rev. Fish Biol. Fish. 30 (2020) 121–135.
[14] X.X. You, X.X. Shan, Q. Shia, Research advances in the genomics and applications for molecular breeding of aquaculture animals, Aquaculture. 526 (2020) 735357.
[15] J. Gutekunst, R. Andriantsoa, C. Falckenhayn, K. Hanna, W. Stein, J. Rasamy, F. Lyko, Clonal genome evolution and rapid invasive spread of the marbled crayfish, Nat. Ecol. Evol. 2 (2018) 567–573.
[16] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (3) (1990) 403–410.
[17] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (17) (2018) 884–890.
[18] G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics. 27 (6) (2011) 764–770.
[19] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation, Genome Res. 27 (5) (2017) 722–736.
[20] M.H. Schmidt, A. Vogel, A.K. Denton, De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing, Plant Cell 29 (3) (2017) 2336–2348.
[21] R. Vaser, I. Sović, N. Nagaranjan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads, Genome Res. 27 (5) (2017) 737–746.
[22] B.J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q.Z. Zeng, J. Wortman, S.K. Young, A.M. Earl, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, PLoS One 9 (11) (2014), e112963.
[23] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[24] N. Servant, N. Varoquaux, B.R. Lajoie, V. Eric, B. Emmanuel, HiC-Pro: an optimized and flexible pipeline for Hi-C data processing, Genome Biol. 16 (1) (2015) 259–262.
[25] J.N. Burton, A. Adey, R.P. Patwardhan, R. Qiu, J.O. Kitzman, J. Shendure, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, Nat. Biotechnol. 31 (12) (2013) 1119–1125.
[26] G.C. Xu, T.J. Xu, R. Zhu, Y. Zhang, G.T. Li, LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly, Gigascience 8 (1) (2019) 157–160.
[27] F.A. Simao, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (19) (2015) 3210–3212.
[28] G. Parra, K. Bradnam, I. Korf, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, Bioinformatics 23 (9) (2007) 1061–1067.
[29] Z. Xu, H. Wang, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, Nucleic Acids Res. 35 (2007) W265–W268. Web Server issue.
[30] A.L. Price, N.C. Jones, P.A. Pevzner, De novo identification of repeat families in large genomes, Bioinformatics 21 (suppl_1) (2005) i351–i358.
[31] C. Hoede, S. Arnoux, M. Moisset, T. Chaumier, O. Inizan, V. Jamilloux, H. Quesneville, PASTEC: an automatic transposable element classification tool, PLoS One 9 (5) (2014), e91929.
[32] W. Bao, K.K. Kojima, O. Kohany, Repbase update, a database of repetitive elements in eukaryotic genomes, Mob. DNA 6 (1) (2015) 11.
[33] M. Tarailo-Graovac, N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences, Curr. Protoc. Bioinformatics 5 (1) (2009) 4.10.11–14.10.14.

[34] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1) (1997) 78–94.
[35] M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel, Bioinformatics 19 (suppl_2) (2003) ii215–ii225.
[36] W.H. Majoros, M. Pertea, S.L. Salzberg, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, Bioinformatics. 20 (16) (2004) 2878–2879.
[37] E. Blanco, G. Parra, R. Guigó, Using geneid to identify genes, Curr. Protoc. Bioinformatics 18 (1) (2007) 4.3.1–4.3.28.
[38] I. Korf, Gene finding in novel genomes, BMC Bioinformatics 5 (1) (2004) 59.
[39] J. Keilwagen, M. Wenk, J.L. Erickson, M.H. Schattat, J. Grau, F. Hartung, Using intron position conservation for homology-based gene prediction, Nucleic Acids Res. 44 (9) (2016), e89.
[40] D. Kim, B. Langmead, S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods 12 (4) (2015) 357–360.
[41] M. Pertea, G.M. Pertea, C.M. Antonescu, T.C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol. 33 (3) (2015) 290–295.
[42] S. Tang, A. Lomsadze, M. Borodovsky, Identification of protein coding regions in RNA transcripts, Nucleic Acids Res. 43 (12) (2015), e78.
[43] B.J. Haas, S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C. R. Buell, J.R. Wortman, Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments, Genome Biol. 9 (1) (2008) R7.
[44] B.J. Haas, A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, S.L. Salzberg, O. White, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, Nucleic Acids Res. 31 (19) (2003) 5654–5666.
[45] E. Birney, M. Clamp, R. Durbin, GeneWise and genomewise, Genome Res. 14 (5) (2004) 988–995.
[46] S. Griffiths-Jones, R.J. Grocock, S. Van Dongen, A. Bateman, A.J. Enright, miRBase: microRNA sequences, targets and gene nomenclature, Nucleic Acids Res. 34 (Database issue) (2006) D140–D144.
[47] J. Daub, R.Y. Eberhardt, J.G. Tate, S.W. Burge, Rfam: annotating families of non-coding RNA sequences, in: E. Picardi (Ed.), RNA Bioinformatics. Methods in Molecular Biology, Humana Press, New York, NY, 2015, pp. 349–363.
[48] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D. A. Natale, The COG database: an updated version includes eukaryotes, BMC Bioinformatics 4 (1) (2003) 41.
[49] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[50] B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (1) (2003) 365–370.
[51] D.M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics, Genome Biol. 20 (2019) 238.
[52] H. Mi, A. Muruganujan, D. Ebert, X. Huang, P.D. Thomas, ANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools, Nucleic Acids Res. 2019 (2019) 47.
[53] K. Katoh, G. Asimenos, H. Toh, Multiple alignment of DNA sequences with MAFFT, Methods Mol. Biol. 537 (2009) 39–64.
[54] L.T. Nguyen, H.A. Schmidt, A. Von Haeseler, B.Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, Mol. Biol. Evol. 32 (2015) 268–274.
[55] Z. Yang, PAML 4: a program package for phylogenetic analysis by maximum likelihood, Mol. Biol. Evol. 24 (2007) 1586–1591.
[56] T.J. De Bie, N. Cristianini, J.P. Demuth, M.W. Hahn, CAFE: a computational tool for the study of gene family evolution, Bioinformatics 22 (10) (2006) 1269–1271.
[57] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, R. Durbin, BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data, Bioinformatics 32 (11) (2016) 1749–1751.
[58] H. Li, R. Durbin, Inferenceof human population history from individual whole-genome sequences, Nature 475 (7357) (2011) 493.
[59] J.M. Flynn, F.J.J. Chain, D.J. Schoen, M.E. Cristescu, Spontaneous mutation accumulation in *Daphnia pulex* in selection-free *vs.* competitive environments, Mol. Biol. Evol. 34 (1) (2017) 160–173.
[60] N. Hiromichi, The crustacean cuticle: structure, composition and mineralization, Front. Biosci. E4 (2012) 711–720.
[61] O. Ayokunmi, Y.M. Normi, Chitinase: diversity, limitations and trends in engineering for suitable applications, Biosci. Rep. 38 (4) (2018), BSR20180323.
[62] P.A. Lawrence, G. Morata, Homeobox genes: their function in Drosophila segmentation and pattern formation, Cell 78 (1994) 181–189.
[63] G.B. James, C.W. Stephen, D.L. Anthony, A new standard for crustacean genomes: the highly contiguous, annotated genome assembly of the clam shrimp *Eulimnadia texana* reveals HOX gene order and identifies the sex chromosome, Genome Biol. Evol. 10 (1) (2018) 143–156.
[64] J.X. Fan, S.Z. Shen, D.H. Erwin, P.M. Sadler, Y.Y. Zhao, A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity, Science 367 (2020) 272–277.
[65] M.L. Porter, M. Perez-Losada, K.A. Crandall, Model-based multi-locus estimation of decapod phylogeny and divergence times, Mol. Phylogenet. Evol. 37 (2005) 355–369.
[66] R.F. Flint, Glacial and Quaternary Geology, John Wiley and Sons Inc, New York, 1971.
[67] M. Berkelhammer, A. Sinha, L. Stott, H. Cheng, F.S.R. Pausata, K. Yoshimura, An abrupt shift in the Indian monsoon 4000 years ago, Geophys. Monogr. Ser. 198 (7) (2012) 75–88.
[68] G.H. Yue, L. Wang, Current status of genome sequencing and its applications in aquaculture, Aquaculture 468 (2017) 337–347.
[69] B. Star, A. Nederbragt, S. Jentoft, K.J. Jakobsen, The genome sequence of Atlantic cod reveals a unique immune system, Nature 477 (2011) 207–210.
[70] G.R. Gong, C. Dan, S.J. Xiao, W.J. Guo, P.P. Huang, Y. Xiong, J.J. Wu, Y. He, J. C. Zhang, X.H. Li, N.S. Chen, J.F. Gui, J. Mei, Chromosomal-level assembly of yellow catfifish genome using third-generation DNA sequencing and Hi-C analysis, GigaScience 7 (2018), giy120.
[71] X. Zhang, J. Yuan, Y. Sun, S. Li, Y. Gao, Y. Yu, C. Liu, Q. Wang, X. Lv, X. Zhang, K. Y. Ma, X. Wang, W. Lin, L. Wang, X. Zhu, C. Zhang, J. Zhang, S. Jin, K. Yu, J. Kong, P. Xu, J. Chen, H. Zhang, P. Sorgeloos, A. Sagi, A. Alcivar-Warren, Z. Liu, L. Wang, J. Ruan, K.H. Chu, B. Liu, F. Li, J. Xiang, Penaeid shrimp genome provides insights into benthic adaptation and frequent molting, Nat. Commun. 10 (1) (2019) 356.
[72] B. Tang, D. Zhang, H. Li, S. Jiang, H. Zhang, F. Xuan, B. Ge, Z. Wang, Y. Liu, Z. Sha, Y. Cheng, W. Jiang, H. Jiang, Z. Wang, K. Wang, C. Li, Y. Sun, S. She, Q. Qiu, W. Wang, X. Li, Y. Li, Q. Liu, Y. Ren, Chromosome-level genome assembly reveals the unique genome evolution of the swimming crab (*Portunus trituberculatus*), Gigascience 9 (1) (2020), giz161.
[73] F. Shao, M.J. Han, Z.G. Peng, Evolution and diversity of transposable elements in fish genomes, Sci Rep. 9 (1) (2019), 15399.
[74] L.Y. Zhang, J.H. Hu, X.L. Han, J.J. Li, Y. Gao, C.M. Richards, C.X. Zhang, Y. Tian, G. M. Liu, H. Gul, D.J. Wang, Y. Tian, C.X. Yang, M.H. Meng, G.P. Yuan, G.D. Kang, Y. L. Wu, K. Wang, H.T. Zhang, D.P. Wang, P.H. Cong, A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour, Nat. Commun. 10 (1) (2019) 1494.
[75] A.F. Rowley, The immune system of crustaceans, in: Encyclopedia of Immunobiology 1, 2016, pp. 437–453.
[76] X.G. Li, Z.Q. Xu, G. Zhou, H. Lin, J. Zhou, Q.F. Zeng, Z.G. Mao, X.H. Gu, Molecular characterization and expression analysis of five chitinases associated with molting in the Chinese mitten crab, *Eriocheir sinensis*, Comp. Biochem. Physiol. B 187 (2015) 110–120.
[77] W. Chen, X. Jiang, Q. Yang, Glycoside hydrolase family 18 chitinases: the known and the unknown, Biotechnol. Adv. 43 (2020) 107553.
[78] M. Liu, C. Chen, Q.C. Wu, J.L. Chen, L.S. Dai, S.H. Chu, Q.N. Liu, Chitinase involved in immune regulation by mediated the toll pathway of crustacea *procambarus clarkii*, Fish Shellfish Immunol. 110 (2021) 67–74.
[79] M. Averof, Arthropod Hox genes: insights on the evolutionary forces that shape gene functions, Curr. Opin. Genet. Dev. 12 (2002) 386–392.
[80] J.K. Grenier, T.L. Garber, R. Warren, P.M. Whitington, S. Carroll, Evolution of the entire arthropod hox gene set predated the origin and radiation of the onychophoran/arthropod clade, Curr. Biol. 7 (8) (1997) 547–553.
[81] D. Duboule, The rise and fall of Hox gene clusters, Development 134 (14) (2007) 2549–2560.