A chromosome-level genome of Portunus trituberculatus provides insights into its evolution, salinity adaptation, and sex determination

Jianjian Lv[1], Ronghua Li[2], Zhencheng Su[3], Baoquan Gao[1], Xingbin Ti[1], Deping Yan[1], Guang-Jian Liu[3], Chunlin Wang[2], Ping Liu[1], and Jian Li[1]

[1]Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
[2]Ningbo University
[3]Novogene Bioinformatics Institute

August 3, 2020

Abstract

Portunus trituberculatus (Crustacea: Decapoda: Brachyura), commonly known as the swimming crab, is of major ecological importance, as well as being important to the fisheries industry. P. trituberculatus is also an important farmed species in China due to its rapid growth rate and high economic value. Here, we report the genome sequence of the swimming crab, which was assembled at the chromosome scale, covering ˜1.2 Gb, with 79.99% of the scaffold sequences assembled into 53 chromosomes. The contig and scaffold N50 values were 108.7 kb and 15.6 Mb, respectively, with 19,981 protein-coding genes and a high proportion of simple sequence repeats (49.43%). Based on comparative genomic analyses of crabs and shrimps, the C2H2 zinc finger protein family was found to be the only gene family expanded in crab genomes, and its members were mainly expressed in early embryonic development and during the flea-like larval stage, suggested it was closely related to the evolution of crabs. Combined with transcriptome and Bulked Segregant Analysis (BSA) providing insights into the genetic basis of salinity adaptation in P. trituberculatus, strong immunity and rapid growth of the species were also observed. In addition, the specific region of the Y chromosome was located for the first time in the genome of P. trituberculatus, and Dmrt1 was identified as a key sex determination gene in this region. Decoding the swimming crab genome not only provides a valuable genomic resource for further biological and evolutionary studies, but is also useful for molecular breeding of swimming crabs.

A chromosome-level genome of *Portunus trituberculatus* provides insights into its evolution, salinity adaptation, and sex determination

Jianjian Lv [a,c,] *, Ronghua Li[b] ,

***, Zhencheng Su [d,*], Baoquan Gao[a,c], Xingbin Ti [a], Deping Yan[a], Guangjian Liu [d], Ping Liu[a,c], Chunlin Wang [b], Jian Li[a,c]**

[a] ·Key Laboratory of Sustainable Development of Marine Fisheries, Ministry of Agriculture, P.R.China, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, 266071 Qingdao,China.

[b]· Key Laboratory of Applied Marine Biotechnology, Ministry of Education, Ningbo University, Ningbo 315211, China

[c] Function Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, No. 1 Wenhai Road, Aoshanwei Town, Jimo, Qingdao, China

[d] Novogene Bioinformatics Institute, Beijing 100016, China

[*] These authors contributed equally

Corresponding author,*lijian@ysfri.ac.cn* (Jian Li),*wangchunlin@nbu.edu.cn*(Chunlin Wang),*liuping@ysfri.ac.cn* (Ping Liu) and*liuguangjian@novogene.com*(Guangjian Liu)

Abstract

Portunus trituberculatus (Crustacea: Decapoda: Brachyura), commonly known as the swimming crab, is of major ecological importance, as well as being important to the fisheries industry. *P. trituberculatus* is also an important farmed species in China due to its rapid growth rate and high economic value. Here, we report the genome sequence of the swimming crab, which was assembled at the chromosome scale, covering ˜1.2 Gb, with 79.99% of the scaffold sequences assembled into 53 chromosomes. The contig and scaffold N50 values were 108.7 kb and 15.6 Mb, respectively, with 19,981 protein-coding genes and a high proportion of simple sequence repeats (49.43%). Based on comparative genomic analyses of crabs and shrimps, the C2H2 zinc finger protein family was found to be the only gene family expanded in crab genomes, and its members were mainly expressed in early embryonic development and during the flea-like larval stage, suggested it was closely related to the evolution of crabs. Combined with transcriptome and Bulked Segregant Analysis (BSA) providing insights into the genetic basis of salinity adaptation in *P. trituberculatus* , strong immunity and rapid growth of the species were also observed. In addition, the specific region of the Y chromosome was located for the first time in the genome of *P. trituberculatus* , and *Dmrt1*was identified as a key sex determination gene in this region. Decoding the swimming crab genome not only provides a valuable genomic resource for further biological and evolutionary studies, but is also useful for molecular breeding of swimming crabs.

Introduction

Brachyuran are generally called crabs, which are known as one of the most typical crustaceans belonging to Decapoda. In terms of numbers of species, they are one of the largest groups of the crustaceans (Warner, 1977), comprising about 6,000 species belonging to 47 families (Bowman, 1982). These species are found worldwide, largely in marine habitats, ranging from benthic to free living and planktonic or parasitic forms. Furthermore, this group includes many commercially-important species. Swimming crabs represent the most important group of crabs in fisheries and aquaculture, and have attracted considerable research attention.

In recent years, distinct morphotypes have drawn much attention and discussion within Decapoda (Haug et al., 2016). In evolutionary terms, all crabs have evolved from an ancestral macruran or “lobster” morphotype (Haug et al., 2016). Compared with the elongate bodies of shrimps and lobsters, crabs are characterised by a compact body with a depressed, short carapace, and a ventrally-folded pleon. The evolutionary transformation from a lobster-like crustacean to a crab is called ‘carcinization’, and has been interpreted as a dramatic morphological change (Scholtz, 2014). However, how crabs evolved and their underlying genetic basis have not been elucidated (McNamara & Faria, 2012).

Salinity is an important abiotic factor that influences the distribution, abundance, physiology and wellbeing of crustaceans (Romano & Zeng, 2012). The decapods have occupied and exploited virtually all habitats available on earth, and include species restricted to the marine environment, estuarine and freshwater species, and amphibious and terrestrial travelers, including desert and arboreal adventurers (McNamara & Faria, 2012). Given the innumerous challenging habitats, the mechanism of salinity adaptation of the decapod Crustacea has long aroused scientific curiosity. Osmoregulation mediated by ion transport is an important physiological process of salinity adaptation. Ion transport-related genes, including Na$^+$,K$^+$-ATPase, V-type H$^+$-ATPase, and Na$^+$,K$^+$, 2Cl $^-$cotransporters have been cloned and studied (Garcon et al., 2011; Lv, Zhang, Liu, & Li, 2016; Tsai & Lin, 2007). Some scholars also tried to study the salinity adaptation using comparative transcriptome analyses (Lv et al., 2013; Q. H. Xu & Liu, 2011), and hundreds of potential salt-tolerance-related genes were identified. However, the complex molecular mechanisms involved in salinity tolerance remain poorly understood.

Sex determination is a plastic biological developmental process, which has always had an intriguing aspect in evolutionary and developmental biology (Martins, 2002). The mechanisms of sex determination in crustaceans are remarkably diverse and are controlled by genetic and/or environmental factors (Ford, 2008). In crabs, some species (*Eriocheir japonicus* , *Hemigrapsus sanguineus* ,*Hemigrapsus penicillatus* and *Plagusia dentipes*) were believed to have an XY sex determination system based on karyotype analysis (Lecher & Noel, 1995; Niiyama, 1937, 1938, 1959); however, due to the large numbers of chromosomes and complex genomes, sex determination by karyotype analysis cannot be determined in most crustaceans (Z. Torrecilla, Martínezlage, Perina, Gonzálezortegón, & Gonzáleztizón, 2017).

The swimming crab, *Portunus trituberculatus* (Crustacea: Decapoda: Brachyura), inhabits seafloors with sand or pebbles, and is widely distributed in the coastal waters of China, Japan, and Korea[19]. This species is one of the most common edible crabs in China, and is artificially propagated and stocked. As an important crustacean species in aquaculture, the swimming crabs is famous for its rapid growth rate and large body size. Its body mass can reach up to 400-500 g in 7-8 months of artificial culture. In pond culture, crabs are often mixed with shrimps because they feed on diseased shrimps and exhibit strong disease resistance, which can effectively prevent outbreaks of shrimp diseases. Therefore, *P. trituberculatus* are also regarded as an ecologically important species. After more than 30 years of high-intensity artificial breeding (Liu et al., 2015), there are many problems in crab farming, such as significant variability in individual sizes, environmental factor stress, and disease outbreak, among others. However, the genetic mechanisms leading to rapid growth, stress resistance, and strong immunity remain poorly understood in this economically important species.

Here, we report the chromosome-scale genome assembly of *P. trituberculatus* . Genome evolution and comparative genomic analyses provided insights into the genetic basis of salinity adaptation, strong immunity, rapid growth, and sex determination in this species. This genome can serve as the genetic basis for future investigations of*P. trituberculatus* evolution and biology, and will be a valuable resource for conservation and breeding management of the swimming crab.

Materials and Methods

Sample collection and sequencing

The male individual used for genome sequencing originated from an F12 full sibling that was created by artificial insemination. Total genomic DNA was extracted from muscle tissue and stored at Novogene Bioinformatics Institute.

Pair-ended (PE) libraries with an insert size of 350 bp were constructed using the standard Illumina protocol (Mardis, 2008). Sequencing was performed using the Illumina HiSeq4000 platform. Low-quality reads with more than 10% of bases having quality scores lower than 20 (representing a 1% error rate) were removed, as were tag-contaminated sequences and duplications.

We also extracted RNA from eight *P. trituberculatus* tissues at different developmental stages*,* including the eyes (E), brain (Br), ganglia thoracalis (Tr), gill (G), blood (B), heart (H), liver (L), and muscle (M), for transcriptome sequencing on the Illumina HiSeq platform, followed by gene prediction. Four libraries for Bulked Segregant Analysis (BSA) analysis were sequenced using the Illumina Novaseq platform, and 150 bp paired-end reads were generated with an insert size of ˜350 bp.

Genome assembly

We used a K-mer frequency distribution method (R. Li et al., 2010) to estimate the genome size by the following formula: G = k-mer number/average k-mer depth, where k-mer number = total k-mers - abnormal k-mers (with too low or too high frequency). Seventeen-mers (17 bp k-mers) were extracted from the sequencing data and the frequency of each 17-mer was calculated. Finally, the k-mer depth was 41**(Fig. S1)** and the genome size of *P. trituberculatus* was estimated to be 1.2 Gb.

Libraries for single molecule real-time (SMRT) PacBio genome sequencing were constructed according to standard protocols from the Pacific Biosciences company. Briefly, high molecular-weight genomic DNA was sheared into large fragments, followed by damage repair and end repair, blunt-end adaptor ligation, and size selection. Then, the libraries were sequenced on the PacBio Sequel platform. PacBio reads were used to assemble contigs of the *P. trituberculatus* genome using SMARTdenovo, and polishing errors with Quiver (smrtlink 5.0.1) (Chin et al., 2013). Then, PacBio contigs were scaffolded using six rounds of SSPACE-LongRead (Boetzer & Pirovano, 2014), seven rounds of PBjelly (*http://sourceforge.net/projects/pb-jelly/*), and three iterations of plantanus (Kajitani et al., 2014) with default parameters for all programs. The resulting scaffolds were further connected to super-scaffolds using 10x genomics linked-read data with fragScaff (Adey et al., 2014) software. Finally, Illumina-derived short reads was used to correct the remaining errors by pilon (Walker et al., 2014), and a high-density genetic map was then constructed by chromonomer (version 1.07) (Catchen, Amores, & Bassham, 2020) to anchor the scaffolds into the chromosome-level genome.

Assessing the completeness of the assembly

To assess the completeness of the assembled *P. trituberculatus*genome, we performed Benchmarking Universal Single-Copy Orthologs (BUSCO) (*http://busco.ezlab.org/*) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) analysis by searching against the arthropod BUSCO (version 3.0). We also assessed the completeness of the *P. trituberculatus* genome using the Core Eukaryotic Genes Mapping Approach (CEGMA) (http://korflab.ucdavis.edu/datasets/cegma/) (Genis Parra, Bradnam, & Korf, 2007). Both of the analyses were performed with the default settings.

Genome prediction and annotation

The repetitive sequences in the *P. trituberculatus* genome were identified after genome assembly. Repetitive sequences included transposable elements (TEs) and tandem repeats. Two methods were used to discover TEs. RepeatMasker (version 4.0.5) (N. Chen, 2004) was used to identify TEs in an integrated repeat library, which was derived from a known repeat library (Repbase 15.02) and the *de novo* repeat library, built by RepeatModeler (http://www.repeatmasker.org) (Vision 1.0.5), RepeatScout (Price, Jones, & Pevzner, 2005), and LTR_FINDER (Z. Xu & Wang, 2007). RepeatProteinMask (Bergman & Quesneville, 2007) was used to detect TEs in the *P. trituberculatus* genome by comparing it to the TE protein database. Tandem repeats in the genome were ascertained using Tandem Repeats Finder (Benson, 1999).

Based on the repeat-masked genome, we used homology-based, ab initio, and transcriptome-based prediction methods to predict protein-coding genes in the genome. Briefly, the protein sequences of homologous species were downloaded from the NCBI database, including *Homo sapiens* , *Tetranychus urticae* , *Caenorhabditis elegans* ,*Crassostrea gigas, Drosophila melanogaster, Daphnia pulex, Ixodes scapularis, Parasteatoda tepidariorum, Penaeus vannamei, Strongylocentrotus purpuratus,* and *Tribolium castaneum,* and were used as queries to search the genome using TBLASTN (Altschul, Gish, Miller, Myers, & Lipman, 1990) (*E* -value [?] 1e-05). Homologous genome sequences were then aligned to the matching proteins using Genewise (version 2.4.0) (Birney, Clamp, & Durbin, 2004) for accurate spliced alignments. For ab initio, five tools were used for gene structures prediction, namely, Augustus (version 2.5.5) (Stanke & Morgenstern, 2005), GlimmerHMM (version 3.01) (Majoros, Pertea, & Salzberg, 2004), SNAP (Birney et al., 2004), Geneid (G. Parra, Blanco, & Guigo, 2000), and Genescan (Burge & Karlin, 1997), all with default settings. In addition, the RNA-seq data from several tissues were aligned to the genome using Tophat (version 2.0.10) (Trapnell, Pachter, & Salzberg, 2009), while gene structures were predicted using cufflinks (version 2.1.1) (Trapnell et al., 2012). Genes predicted using the above methods were then merged to a consensus gene set using EVidenceModeler (EVM) (http://evide ncemodeler.sourceforge.net/) (Haas et al., 2008).

Functional annotations of the predicted genes in the *P. trituberculatus* genome were then annotated using homology searching in several public gene databases, including NCBI-NR, KEGG (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016), and SwissProt (Apweiler et al., 2004) using BLASTP (*E* -value [?] 1e-05). We also used InterProScan (version 4.7) (Jones et al., 2014) to obtain protein domain annotations in the Interpro and Gene Ontology (GO) (The Gene Ontology Consortium, 2016) databases. Finally, functional annotations of the best alignments in each database were used as the final consensus gene annotation results.

Genome evolution

To identify gene families in the *P . trituberculatus*genome, we used the nucleotide and protein sequences from 11 other species, including *E. sinensis*(GCA_003336515.1)*, P. virginalis*(GCA_002838885.1)*, L. vannamei*(GCA_003789085.1)*, D. pulex* (GCA_000187875.1)*, D. rerio*(GCF_000002035.6)*, C. semilaevis* (GCF_000523025.1)*, O. niloticus* (GCF_001858045.2)*, C. gigas* (GCF_000297895.1)*, M. yessoensis* (GCA_002113885.2)*, C. elegans*(GCF_000002985.6), and *N. vectensis*(GCF_000209225.1). To exclude putative fragmented genes, we only retained gene models at each gene locus that encoded the longest protein sequence, and removed genes encoding protein sequences shorter than 30 amino acids. First, we performed the all-against-all BLASTP method to identify the similarities among genes in the species with an E-value cutoff of 1e-7. Then, we used OrthoMCL (L. Li, Stoeckert, & Roos, 2003) to generate orthologous and paralogous relationships among all the organisms with the parameter of "-inflation 1.5". Genes were classified into orthologues, paralogues, and single-copy orthologues (only one gene in each species). In addition, specific genes were selected by comparing *P. trituberculatus* to *E. sinensis* ,*P. virginalis* , and *L. vannamei,* and GO and KEGG enrichment analysis were then performed.

Using the single-copy orthologues, we constructed a phylogenetic tree using the Maximum Likelihood model with RAxML (Stamatakis, 2006) software. Then the mcmctree program of PAML (*http://abacus.gene.ucl.ac.uk/software/paml.html*) (Yang, 2007) was used to estimate the divergence times among 12 species. Several calibration points were selected from the TimeTree (Kumar, Stecher, Suleski, & Hedges, 2017) database () to be used as normal priors to restrain the age of the nodes, such as 204-225 Mya for TMRCA of *C. semilaevis - D. rerio* ; 421-447 Mya for TMRCA of*M. yessoensis - C. gigas* ; and 256-429 Mya for TMRCA of*L. vannamei - P. virginalis* . The phylogenetic tree confirmed*N. vectensis* as the outgroup.

Gene family expansion and contraction analyses

To avoid extreme gene families, families with gene numbers [?] 200 in one species and [?] 2 in all other species were removed. The CAFE tool (version 4.0) (http://sourceforge.net/projects/cafehahnlab/) (De Bie, Cristianini, Demuth, & Hahn, 2006) was used to analyze the expansion and contraction of orthologous gene families between ancestors, and each of the 11 species using a stochastic birth and death model with a lambda parameter. This model was further used to examine the changes in gene families along each lineage on the phylogenetic tree. A probabilistic graphical model was introduced to calculate the probability of transitions in gene family size from parent to child nodes. The corresponding p-values were calculated in each lineage based on the conditional likelihood. Finally, we used the Fisher's exact test to identify overrepresented GO and KEGG pathways among the expanded and contracted genes, which were adjusted by the false discovery rate ($FDR < 0.05$).

PacBio Iso-Seq analysis

The Iso-Seq library was prepared according to the Isoform Sequencing protocol (Iso-Seq) using the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System, as described by Pacific Biosciences (PN 100-092-800-03).

Sequence data were processed using SMRTlink 5.1 software. A circular consensus sequence (CCS) was generated from subread BAM files with the following parameter settings: –minLength 50, –maxDropFraction 0.8, –minPasses 2, –minPredictedAccuracy 0.8, and –maxLength 15000. Non-full length and full-length FASTA files from the CCS were then fed into the cluster step, which performed isoform-level clustering (ICE). Finally, the Arrow polishing step was used to generate high-quality consensus FASTA files using the following parameters: –hq_quiver_min_accuracy 0.99, –bin_by_primer false, –bin_size_kb 1, –qv_trim_5p 100, and –qv_trim_3p 30.

Additional nucleotide errors in the consensus reads were corrected using the Illumina RNAseq data with LoRDEC software (Salmela & Rivals, 2014). Any redundancy in the corrected consensus reads was removed by CD-HIT (-c 0.95 -T 6 -G 0 -aL 0.00 -aS 0.99) (W. Li, Jaroszewski, & Godzik, 2002) to obtain final transcripts for subsequent analyses. The ANGEL pipeline, which is a long read implementation of ANGEL (https://github.com/PacificBiosciences/ANGEL), was used to determine the protein coding sequences from cDNAs. The following analysis mainly included structural analyses, differential enrichment analyses, and function annotation.

The structural analyses included CDS prediction, TF (transcription factors) analyses, LncRNA (long non-coding RNA) analyses, and SSR (Simple Sequence Repeat) analyses. TF analysis was performed using the animalTFDB 2.0 database (H.-M. Zhang et al., 2015). In LncRNA analyses, we used Coding-Non-Coding-Index (CNCI) (https://github.com/www-bioinfo-org/CNCI), Coding Potential Calculator (CPC) (Kong et al., 2007), Pfam-scan (Finn et al., 2015), and PLEK (A. Li, Zhang, & Zhou, 2014) tools to predict the coding potential of transcripts. CNCI profiles adjoining nucleotide triplets were performed with default parameters to distinguish protein-coding and non-coding sequences, independent of known annotations. The CPC mainly assessed the extent and quality of the ORFs, and searched the sequences in the NCBI eukaryotic protein database. For such searches, the e-value was set to 1e-10 to identify the coding and non-coding transcripts. Pfam-scan was used with default parameters to identify the occurrence of any of the known protein family domains in the Pfam database in transcripts translated in all three frames.

The PLEK SVM classifier uses an optimized K-mer approach to construct the best classifier to assess the coding potential for species that lack high-quality genomic sequences and annotations. The coding potential was predicted for all transcripts after the three tools described above were used to remove transcripts that lacked coding potential. The remaining transcripts comprised our candidate set of lncRNAs. SSR in the transcriptome were identified using *MISA*(http://pgrc.ipkgatersleben.de/misa/misa.html). MISA identifies and determines the location of perfect microsatellites, as well as compound microsatellites, which are interrupted by a specific number of nucleotides.

Differential enrichment analyses included the quantification of gene expression levels, differential expression analyses, and GO and KEGG enrichment analyses. The gene expression levels for each sample were estimated by RSEM (B. Li & Dewey, 2011), and clean data were mapped back to the transcript sequences and the read count for each transcript was obtained from the mapping results. Differential expression analysis of two conditions/groups was performed using the DESeq R package (1.18.0). DESeq provides statistical assessments to determine differential expression from digital gene expression data using a model based on a negative binomial distribution. The resulting p-values were adjusted using the Benjamini and Hochberg approach to control the false discovery rate. Genes with an adjusted p-value < 0.05 found by DESeq were assigned as differentially expressed. GO and KEGG enrichment analyses were then implemented using the Goseq R package and KOBAS (Mao, Cai, Olyarchuk, & Wei, 2005) software, respectively. Significant content with an FDR threshold of < 0.05 was selected as the results.

RNA-seq analysis

P. trituberculatus samples from different developmental stages (F, Z1-Z4, M, and J), different molting stages (PrM, InM, and PoM), the gills after 0–72 h of low salt stress (S0h, S12h, S24h, S48h and S72h), and haemocytes after 0–24 h of *V. alginolytica* infection (T0B and T24B) were used for RNA-seq analysis.

A total amount of 3 μg of RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using the NEBNext® Ultra RNA Library Prep Kit for Illumina(r) (NEB, USA), following manufacturer's recommendations. Index codes were added to assign sequences to each sample.

Clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina), according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq platform with 150 bp paired-end reads generated.

Raw data (raw reads) in FASTQ format were processed using in-house Perl scripts. During this step, clean data (clean reads) were obtained by removing reads containing adapters, containing ploy-N, or with low quality reads. At the same time, the Q20, Q30, and GC content of the clean data were calculated. All downstream analyses were based on the clean data.

The reference genome and gene model annotation files were used following genome assembly and annotation. Hisat2 (v2.0.4) software (Kim, Langmead, & Salzberg, 2015) was used to build the index of the reference genome and align clean paired-end reads to the genome. HTSeq v0.6.1 (Anders, Pyl, & Huber, 2015) was used to count the number of reads mapped to each gene. Then, the FPKM (expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced) of each gene was calculated based on the length of the gene and the read counts mapped to the same gene. This statistic simultaneously considers the effect of sequencing depth and gene length for the read counts, and is currently the most commonly used method for estimating gene expression levels (Trapnell et al., 2010).

Differential expression analysis of two conditions/groups was performed using the DESeq R package (1.18.0). The resulting p-values were adjusted using the Benjamini and Hochberg's approach to control the FDR. Genes identified by DESeq with an adjusted p-value < 0.05 were assigned as differentially expressed. GO and KEGG enrichment analyses of differentially expressed genes were processed following the same method as that used for PacBio Iso-Seq analysis.

BSA analysis

In this study, growth- and disease resistance-related genes and markers were found by BSA analysis. For growth traits, SG and BG groups composed of 20 small male individuals (68.24+-10.5g) and 20 large male individuals (142.2+-16.5g) were constructed. For disease resistance traits, 20 susceptible individuals and 20 tolerant individuals were identified by *V. parahaemolyticus* infection experiments, which were used to construct SuG and ToG groups, respectively.

A total amount of 1.5 μg of DNA per sample was collected from 20 individuals to build DNA pooling for the DNA sample preparations according to growth and disease trait, respectively. Sequencing libraries were generated using the Truseq Nano DNA HT Sample preparation Kit (Illumina USA), according to the manufacturer's recommendations. Index codes were added to assign sequences to each sample. Briefly, the DNA sample was fragmented to a size of 350 bp using sonication. Fragments were then end-polished, A-tailed, and ligated to the full-length adapter for Illumina sequencing and additional PCR amplification. Finally, PCR products were purified using an AMPure XP system and the size distribution of libraries was analyzed using the Agilent2100 Bioanalyzer and quantified using real-time PCR. Four libraries constructed as described above were sequenced using the Illumina HiSeq4000 platform, and 150 bp paired-end reads were generated with an insert size of ˜350 bp.

Raw data (raw reads) in FASTQ format were initially processed through a series of quality control (QC) procedures using in-house C scripts to ensure that reads were reliable, and without artificial bias. The QC procedures were as follows: a. removal of reads with [?] 10% unidentified nucleotides (N); b. removal of reads with $> 50\%$ of bases with a phred quality < 5; c. removal of reads with > 10 nucleotides aligned to the adapter, allowing [?] 10% mismatches; and d. removal of putative PCR duplicates generated by PCR amplification during the library construction process (i.e., read 1 and read 2 from two paired-end reads that were completely identical).

The Burrows-Wheeler Aligner (BWA) was used to align clean reads against the reference genome with the following parameters: "mem -t 4 -k 32 –M -R" and alignment files were converted to BAM files using SAMtools software with "–bS –t". In addition, potential PCR duplications were removed using the SAMtools command, "rmdup". If multiple read pairs had identical external coordinates, only the pair with the highest mapping quality was retained. Variant calling was performed for all samples using the Unified Genotyper function in GATK software (version 3.8) (McKenna et al., 2010). SNPs (single nucleotide polymorphism) were filtered using the Variant Filtration parameter in GATK (settings: –filterExpression "QD < 4.0 || FS > 60.0 || MQ < 40.0 ", -G_filter "GQ<20", –cluster WindowSize 4,). InDels were filtered using the Variant Filtration parameter (settings: –filter Expression "QD < 4.0 || FS > 200.0 || Read PosRankSum < -20.0 || Inbreeding Coeff < -0.8 ").

ANNOVAR (K. Wang, Li, & Hakonarson, 2010) (Annotate Variation), an efficient software tool to functionally annotate genetic variants, was used to annotate SNPs and InDels based on the GFF3 files for the reference genome. The homozygous SNPs and InDels between two samples were extracted from the vcf files for SNP/InDel in three comparison groups. The read depth information for homozygous SNPs and InDels described above in one sample pool was used to calculate the SNP/InDel index. We used the genotype of one sample as the reference and to statistic reads number for this sample genotype. Then calcalate the ratio of the number of different reads in total number, which is the SNP/InDel index of the base sites. We removed those sites which the SNP/InDel index in both pools less than 0.3. The sliding window method was used to present the SNP/InDel index of the whole genome. The average of all SNP/InDel indexes in each window was taken as the SNP/InDel index for the overall window. In general, we used a window size of 1 Mb and a step size of 10 Kb as the default settings. The difference in SNP/InDel index between the two pools was calculated as the delta SNP/InDel index with the following parameter settings: -fs1 0.2 -fs2 0.8 in the in-house Perl scripts.

CQ (chromosome quotient) analysis

Forty healthy males and forty healthy females were randomly selected from the core breeding population of "Huangxuan NO.1", which were reared at the Chang-Yi AquaFarming Company, Weifang, China. To improve the efficiency and accuracy of the sequencing data, DNA from individuals in two groups were pooled to generate two DNA bulks based on sexuality. Libraries with an insert size of ˜350 bp were sequenced using the Illumina HiSeq4000 platform and 150 bp paired-end reads were generated.

To obtain more reliable reads from the raw data, quality control (QC) procedures were performed similar to those for BSA analysis. The chromosome quotient (CQ) (Hall et al., 2013) method was then used to systematically discover Y chromosome genes. It uses the number of alignments from male and female sequence data to determine whether a sequence is Y-linked. First, the scaffolds were divided into fragments by masked sequences, and fragments shorter than 250 bp were removed. Then, the BWA (H. Li & Durbin, 2009) was used to align the clean reads against the split reference genome with mem -t 4 -k 32 –M -R. Alignment files were then converted to BAM files using SAMtools software (settings: –bS –t) (H. Li et al., 2009). The CQ values of those sequences were then calculated. To classify a sequence as Y-linked, a CQ value of < 0.3, with > 30 alignments from male data, and < 30 alignments from female data were filtered out. The results of the screening with CQ values equal to zero were focused on specifically.

qPCR analysis

For the genes associated with development, immunity, and sex-determination, gene expressions were analyzed by qPCR in the materials of different developmental stages (fertilized egg stage, multicellular stage, blastula stage, gastrulation stage, egg-nauplius stage, egg-zoea stage, zoea stage, megalopal stage, and juvenile crab stage), haemocytes, and hepatopancreas after *V. parahaemolyticus*infection (0–72 h). Total RNA was extracted individually using Trizol (Invitrogen, Carlsbad, CA, USA). Quantitative real-time PCR (qPCR) primers were designed using the Primer Premier 5 tool (Premier Biosoft International) **(Table S1)** . First strand cDNA was synthesized using the PrimeScript RT reagent kit (Takara, Dalian, China). qPCR was performed using a 7500 Fast Real-Time PCR System and SYBR®Premix Ex Taq Kit (Takara, Dalian, China). PCR was performed with the following parameters: 95 degC for 30 s, followed by 40 cycles of 95 degC for 15 s and 60 degC for 34 s.

Results and Discussion

Genome sequencing and assembly

A combination of three technologies, including Pacific Bioscience's single-molecule real-time sequencing, Illumina's paired-end sequencing and 10x Genomics link-reads, were used in this study. After sequencing with the PacBio SEQUEL platform at Novogene (Tianjin), a total of 120.79 Gb of long reads were generated and used for the subsequent genome assembly. Two paired-end Illumina sequencing libraries were constructed with an insert size of 350 bp, and sequencing was carried out on the Illumina HiSeq 4000 platform according to the manufacturer's instructions. A total of 111.01 Gb (92.51x coverage) of sequencing data were produced. In addition, one 10x Genomics linked-read library was constructed and sequencing was performed on the Illumina HiSeq 4000 platform, which produced 85.97 Gb (71.64x coverage) of sequencing data. The genome size of *P. trituberculatus* was estimated to be 1.2 Gb (**Fig S1**), and therefore, the average sequencing coverage was 264.81x Raw sequence data generated by the Illumina platform were filtered based on the following criteria: filtered reads with adapters, filtered reads with N bases more than 10%, and removed reads with low-quality bases ([?] 5) more than 50%. All sequence data are summarized in **Table 1** .

The genome assembly yielded a final draft *P. trituberculatus*genome with a total length of 864.45 Mb, a contig N50 of 108.68 kb, and a scaffold N50 of 15.59 Mb (**Table 2**). The assembly statistics of the crab genome are comparable to, or better than, those of previously published for shrimp genomes (contig N50: 57.65 kb and scaffold N50: 605.56kb) (X. Zhang et al., 2019).

A total of 2,396 scaffolds were attached to 53 pseudochromosomes based on the constructed genetic map, and accounted for 79.99% of the total length (691.44 Mbp). Markers between linkage groups and scaffolds exhibited fine collinearity, suggesting that our assembly was of high continuity and high accuracy **(Fig. 1).**

Completeness of the geneset and assembly

To evaluate the accuracy of the genome at the single base level, we mapped the short sequence reads generated by the Illumina platform to the assembled genome and evaluated the accuracy using BWA. We performed variant calling using SAMtools. The assembly was evaluated by mapping reads with approximately 90.72% coverage, explaining the completeness of the genome assembly. We obtained 47,482 homozygous SNPs**(Table S2)** , thus, reflecting a low homozygous rate (0.0067%) and the high accuracy of the genome assembly at the single-base level.

To assess the completeness of the assembled *P. trituberculatus*genome, we performed Benchmarking Universal Single-Copy Orthologs (BUSCO) by searching against the Arthropoda BUSCO datasets (version 3.0). Overall, 84.7% of the 1,066 Arthropoda BUSCOs were identified in the assembled genome **(Table S3)** . We also assessed the completeness of the *P. trituberculatus* genome using the Core Eukaryotic Genes Mapping Approach (CEGMA). According to CEGMA, 211 (85.08%) conserved genes were identified in the *P. trituberculatus* genome (**Table S4**). All results indicated that the genome assembly had good coverage and completeness.

Genome prediction and annotation

The *P. trituberculatus* genome encodes 19,981 protein-coding genes. The average gene length and coding sequence (CDS) length were 10,484.05 and 1,231.18 bp, respectively, which were consistent with the distributions of gene features in other arthropods **(Table S5, Fig S2)** . Among the predicted genes, 19,763 (˜98.9%) were annotated based on at least one of the NR (RefSeq non-redundant proteins), SwissProt, KEGG, and GO databases **(Table S6)** . Furthermore, ˜84.4% of complete BUSCO genes were successfully identified. The repeat content accounted for 49.46% of the assembly, which was much lower than that of *Eriocheir japonica sinensis* (61.42%) (Tang et al., 2020), but marginally higher than*Litopenaeus vannamei* (49.38%). Among the repetitive sequences, the most abundant TEs were long interspersed elements (LINE, 27.29% of the genome), followed by long terminal repeats (LTR, 17.07%) and DNA transposons (10.95%) **(Table S7, Fig S3)** .

Genome Evolution

Identifying gene families between closely related species provides important insights into the evolutionary relationship of those species. We identified 32,918 gene families altogether and 246 single-copy gene families in gene family cluster analyses performed in twelve species using OrthoMCL **(Table S8, Table S9, Fig S4)** . To investigate the phylogenetic relationship between *P. trituberculatus* and other species, we constructed a phylogenetic tree. Using single-copy orthologues, phylogenetic analyses showed that*E. sinensis* was most closely related to *P. trituberculatus,* with a divergence time around 153.0 (77.6–274.3) million years ago. The ancestor of *P. trituberculatus* and*E. sinensis* separated from the ancestor of *P. virginalis*and *L. vannamei* ~326.0 million years ago**(Fig. 2)** .

We identified 18 gene families that were expanded and seven gene families that were contracted in the crab genome, compared to the other species **(Table S10)** . KEGG (Kyoto Encyclopedia of Genes and Genomes) enrichment of the expression of the expanded genes suggested that some expanded genes played important roles in different physiological process, such as development, molting, salinity adaption, and immune stress **(Fig S5).** In addition, we also identified 1,471 unique gene families containing 3,199 genes in the crab genome (**Table S11**). Those lineage-specific gene families may contribute to traits that are specific to the species. KEGG enrichment analysis also revealed significantly enriched pathways ($p < 0.01$), which contained ribosome, protein digestion and absorption, proximal tubule bicarbonate reclamation, and metabolism of xenobiotics by cytochrome P450 **(Fig S6)** . Among which, the pathway of ribosome were the most significantly enriched, containing the largest number of genes (140 genes). Furthermore, we analyzed the expression of genes in the ribosome pathway in different stages of development, molting, under low salt stress, and upon pathogen infection **(Fig S7)** . We found that some genes were ubiquitously expressed in the above processes, but most were mainly expressed in different molting stages. The major endocrine complex of crustaceans, the X-organ/sinus gland complex (XOSG) (Hopkins), is located in the eyestalk and is known to play a role in molting (Chang & Mykles, 2011). A large number of specific ribosome-related genes are expressed in the eyestalk, and their expressions are significantly related to different molting stages, suggesting that they play an important role in the regulation of eyestalk neuroendocrine and molting.

Genetic basis of carcinization

Compared with the elongate bodies of shrimps or lobsters, crabs are characterised by a compact body organization with a depressed, short carapace and a ventrally-folded pleon (Scholtz, 2014). The evolutionary transformation from a lobster-like crustacean towards a crab is called ‘carcinization’, which has been interpreted as a dramatic morphological change (Borradaile, 2019). Comparative genomic analyses of crab and shrimp allowed us, for the first time, to infer genomic-level evolutionary processes in carbonization. Compared with the two shrimp species, one expanded and forty-seven contracted gene families were identified in two crab species using the Fisher’s exact test (**Table S12**). The only expanded gene family was the C2H2 zinc finger protein family, and previous research supports a role for the proteins in the regulation of morphogenesis (Urrutia, 2003). There are 0, 3 and 23 genes in penaeid shrimp, crayfish, and crab **(Table S13)** . Further analysis of its expression in embryonic development and larval metamorphosis showed that the gene family was mainly expressed in early embryonic development at the flea like larval stage **(Fig. 3)** ; thus, suggesting that it may have specific functions in development and metamorphosis.

Two contracted gene families deserve special attention, including the actin and WDFY families **(Table S14)** . There are 84, 8, and 4 actin genes, and 19, 1, and 0 WDFY genes in penaeid shrimp, crayfish and crab, respectively, as characterized by positive correlations between gene number and body proportion from shrimp, to crayfish, to crab. More interestingly, both genes are linked to a common human disease, rheumatoid arthritis (RA) (Steinz et al., 2019; Y. Q. Zhang, Bo, Zhang, Zhuang, & Liu, 2014). RA is a systemic autoimmune diseases caused by a failure of immune self-tolerance. RA shows some similar appearances to carcinization, such as skeletal muscle weakness, synovial inflammation and hyperplasia, and cartilage and bone destruction. Therefore, the contraction of the two gene families in crab provides clues to study the molecular mechanisms of ‘carcinization’.

The *Hox* gene family is well known for its important function in directing tissue differentiation and morphological development throughout all of the principal axes of an embryo (Yan et al., 2019). Many *Hox* genes are expressed during early embryogenesis; thus, indicating their roles in development (Hogan, 1987; McGinnis). Eight*Hox* genes were found in the crab genome, of which their expressions were detected at different ontogenic stages, identifying functions in development and metamorphosis **(Fig 4)** . The results showed that different *Hox* genes had different levels of expression at different stages, however, in general, *Lab* ,*Pb* and *Dfd* were more active before the gastrulation stage, and *Scr* , *Antp* , *Ubx* , *Abd-A* and *Abd-B*were more active after the egg-nauplius stage.

Brachyurization is one of the most important characteristics of carcinization. The development of crab from the megalopa to the first juvenile crab is the key period of brachyurization (Song et al., 2015). We found that the *Ubx* and *Abd-A* genes, which mainly control the shape of the abdominal segment of crab, were significantly down regulated from the megalopa to the first juvenile crab phase. Conversely, it should be noted that the expression of *Ubx* and*Abd-A* is maintained at a high level throughout these stages in shrimp (Xiaoqing et al., 2015). Thus, the different expression patterns of the two genes in the development of crab and shrimp suggest that they are involved in brachyurization.

Mechanism of salinity adaptation

P. trituberculatus are weak high osmoregulators (McNamara & Faria, 2012). Their internal osmotic pressure is vulnerable to external salinity changes, which indirectly affects their growth, development, and immune system. Thus, it is of great theoretical and practical importance to clarify the mechanisms relating to their osmoregulation. In the present study, we examined the complete transcriptome of *P. trituberculatus* in response to low salt stress (from 0 h to 72 h). Comparisons of gene expression showed that a total of 4,603 unigenes were significantly differentially expressed (DEG), accounting for 33.5% of DEGs, and thus, indicating that low salt stress has a substantial impact on the crab **(Table S15)** .

DEGs were clustered using the STEM Clustering Method, and were grouped into 20 profiles **(Fig S8)** . Among the profiles, Profile 1 contained the highest number of transcripts (670, 14.6%) that were down-regulated at 12 and 24 h post-salt stress, and returned to normal after 72 h. Profiles 16 (535, 11.6%) and 18 (436, 9.5%) exhibited the next highest number of DEGs upon salt stress, with transcripts that were up-regulated immediately following the salinity stress, which were then down regulated or restored to normal levels after 72 h.

To understand the mechanisms of salinity adaptation in *P. trituberculatus* , we further analyzed the expanded and unique DEGs. Eight expanded genes and 141 unique genes were identified from the DEGs**(Table S16)** . Among which, the V-type proton ATPase catalytic subunit A was involved in osmoregulation, and plays an important role in ion transport (Romano & Zeng, 2012). In addition, 14-3-3-like protein, Heat shock protein 70, and two caspase family genes were also identified, and are associated with stress resistance and apoptosis**(Kaeodee, Pongsomboon, & Tassanakajon, 2011; Q. Wang; Zhong).**The results suggest that ion transport, stress resistance, and apoptosis were involved in salinity adaptation in *P. trituberculatus* .

According to the cluster analysis of the expression patterns during the salinity stress, the expanded and unique DEGs were divided into two categories. One group showed down-regulation of genes after the salinity test, while the other group was up-regulated **(Fig 5)** . Based on the above results, we speculated that some physiological compromises or adjustments were made by *P. trituberculatus* to adapt to the low salt environment. Such adjustments were mainly through ”passive” and ”active” mechanisms. The passive pathway refers to the inhibition of some nonessential physiological processes upon salinity stress, and mainly include ribosome-mediated protein translation, and antigen processing and presentation **(Fig S9)** . The active pathway refers to the activation of some physiological processes essential for survival in low salt stress environments, including protein digestion and absorption, apoptosis, and pancreatic secretion, among others**(Fig S10)** .

Immune mechanism

Gene family expansion is probably one of the most important contributors to phenotypic diversity and evolutionary adaption in the environment (X. Zhang et al., 2019). *P. trituberculatus* has stronger resistance to some pathogens, such as WSSV and Vibrios. According to the cumulative mortality rates at 72 h, and the median lethal doses (LD50) of *V. parahaemolyticus* , the anti-disease ability of four kinds of crustacean was: *P. trituberculatus* > *E. carinicauda* > *F. chinensis* > *L. vannamei* (the data no shown). In addition, *P. trituberculatus* is not sensitive to WSSV disease compared to shrimp (Zhongfa, Zhizheng, Wenjun, Weixian, & Songye, 2008). Thus, for the first time, comparative genomic analyses were used to elucidate the immune mechanism of the crab.

Expanded gene families were significantly enriched in 27 KEGG pathways**(Fig S5)** . In particular, the expanded gene families were specifically involved in pathways related to pathogen invasion, such as the African trypanosomiasis (8.63E-08), Legionellosis (6.50E-05), and Salmonella infection (7.33E-05). Some expanded gene families were overrepresented in immune system pathways, such as phagosome (0.00418914), the NOD-like receptor signaling pathway (0.01011326), and the Hippo signaling pathway (0.017947197). In addition, the apoptosis pathway was also enriched, which has been proved to be closely related to the crab immunity (Ren, Yu, Gao, Liu, & Li, 2017). Thus, the significant expansion of these immune- or disease-related gene families may have shaped the crab's ability to resist the invasion of exogenous pathogens.

Although crabs have strong resistance to disease, their individual resistance is different. Using the materials for disease-resistance differentiation, we screened 453 markers related to disease resistance and performed BSA analysis **(Fig 6)** . A total of 427 genes were anchored in the region near the disease-resistance related markers. According to the
$^{i}ndexandsequencingdepth, somemarkersweresel ectedandverifiedbasedonwhethertheywererelatedtotheresistanceof\ V.\ par$
$squaretest, includingnineSNPsandthreeInDels($**Table S17**$).Basedonthelocationinformationofthesemarkers, threedisea$
$resistance-relatedgeneswereanchored, includingDNA-damage-inducibletranscript4(evm.model.Contig81.19), WAPfou$
$disulfidecoredomainprotein1(evm.model.Contig242.11)andproteintrappedinendoderm -$
$1(evm.model.Contig405.11).Amongthosegenes, DNA-damage-inducibletranscript4wasinvolvedinthemTORsignalingpat$
$1wasinvolvedinthecircadianentrainmentandneuroactiveligand-receptorinteractionpathways.AllsuchKEGGpathwayswe$

Regulatory mechanism of rapid growth

P. trituberculatus is an important aquaculture species in China and breeders are particularly interested in its growth traits due to their high commercial significance. The new variety of fast-growing*P. trituberculatus,* Huangxuan No. 1, was selected in 2010 after five generations of selection from a wild population. During selection, the body weight increased 20.12%, compared to the unselected population. The growth rate of *P. trituberculatus* is also significantly faster than that of *E. sinensis* , and this study sought to determine the regulatory mechanism contributing to such rapid growth.

A total of 562 growth-related SNP/InDels were found by BSA strategy**(Table S18 and Fig 8)** . Based on the locations of these markers in genome, 676 candidate growth-related genes were identified, which were significantly enriched (p<0.01) in Type II diabetes mellitus, pancreatic secretion, protein digestion and absorption, insulin resistance and salivary secretion KEGG pathways. **(Fig S11)** . The results suggest that these pathways may be associated with growth traits. Among these candidate growth-related genes, 42 genes were expanded or unique to *P. trituberculatus* **(Table 3)** . Some of the genes played key roles in pathways for carbohydrate metabolism, amino acid metabolism, and the endocrine system, which are known to be related to growth, including acidic mammalian chitinase production, and carbohydrate sulfotransferase and neural-cadherin synthesis.

Unlike *E. sinensis* and *L. vannamei* , *P. trituberculatus* is a typical carnivorous crab. There are often a large number of shellfishes and small crustacean shells found in their stomachs, which contain substantial amounts of chitin. Chitin is a linear polymer of β-1,4-linked N-acetyl-D-glucosamine (GlcNAc), and is the second most abundant natural polysaccharide in nature. Moreover, chitin functions as a major structural component in fungi, crustaceans, and insects (Koch, Stougaard, & Spaink, 2015). Chitinases are enzymes that hydrolyze chitin. In the mammalian digestive system, chitin has long been considered as a source of dietary fiber that is not digested. However, recent studies confirmed that acidic mammalian chitinase (AMCase) can function as a major protease-resistant glycosidase under stomach and intestinal conditions and can degrade chitin substrates to produce $(\text{GlcNAc})_2$, which is a source of carbon, nitrogen, and energy (Misa et al.). A specific AMCase gene was found in the *P. trituberculatus* genome, thus, suggesting the crab can efficiently digest food containing chitin to produce essential energy and substances that support rapid growth.

Rapid growth is inseparable from myogenesis. Cell alignment and fusion in myotubes are key steps during myogenesis. Cell alignment requires a different structural organization of the extracellular matrix, such as lumican, which is composed of keratan sulfates (Chakravarti & S.). Carbohydrate sulfotransferase 5 (*Chst5*) is involved in keratan sulfate biosynthesis, and contributes to the remodeling of the extracellular matrix during myogenic progression in skeletal murine satellite cells (MSC). Decreased expression of *Chst5* delayed cell fusion in myotubes (Grassot et al.). In addition, N-cadherin plays an important role in myogenesis and its expression was highest in prefusion myoblasts, which declined thereafter (Maccalman, Bardeesy, Holland, & Blaschuk). Based on comparative genomics analyses, one specific*Chst5* gene and one expanding N-cadherin family were found in the crab's genome, which indicated that there may be a specific and rapid mechanism of myogenesis in *P. trituberculatus.*

Insulin signalling is important for the regulation of glucose and lipid metabolism, which also stimulate cell growth and differentiation (Saltiel & Kahn, 2001). Growth retardation, obesity and type 2 diabetes is associated with insulin resistance (Kahn & Flier, 2000; Phillips, 1995). Protein-tyrosine phosphatase (LAR) plays an essential role in the regulation of reversible tyrosine phosphorylation of cellular proteins related to insulin resistance (P. M. Li, Zhang, & Goldstein, 1996), and the knock-down of LAR induces insulin resistance (Mander, Hodgkinson, & Sale). The result of BSA analyses showed that one SNP located between two tandem LAR genes was associated with growth traits, and one of which was unique to *P. trituberculatus* . Those data suggested that rapid growth and the differentiation of growth traits in *P. trituberculatus* could be regulated by insulin signaling.

Sex determination

Due to the large number of chromosomes and complex genomes, sex determination by karyotype analysis cannot be performed in most crustaceans (Zeltia Torrecilla, Martínez-Lage, Perina, González-Ortegón, & González-Tizón, 2017). Moreover, the mechanism of sex determination in *P. trituberculatus* has not previously been determined. In our previous studies, a highly significant QTL located on LG24 (LOD > 14) was identified for the first time using a high-density linkage map. Those studies also identified heterogametic genotypes of sex-associated markers in male support the XY sex determination mechanism(Lv et al., 2018). To further analyze the mechanism of sex determination in *P. trituberculatus* , we used CQ analyses to anchor the Y chromosome. We focused on results with CQ values equal to zero and found that they were all located in the two contigs (Contig153 and Contig475)**(Fig 9)** . Interestingly, the two previously-anchored sex markers were also located on those two contigs(Lv et al., 2018); thus, proving our accuracy in locating the Y chromosome.

A total of 10 genes were predicted in the Y-linked region **(Table 4)** , and their expression patterns were analyzed during different developmental stages with known sexes of Z1 to juvenile crabs, based on the sex marker (Lv et al., 2018). Finally, eight genes were successfully analyzed, and the expression patterns of three of which, including doublesex (*Ptdmrt* , Contig475.13), myb/SANT-like (*Ptmyb* , contig153.28) and an unknown gene (Contig475.8) were significantly different in males and females during different developmental stages,**(Fig 10)** .

The doublesex gene was involved in sex determination, which is characterized by one or more highly conserved DNA-binding (DM) domains (Meng, Moore, Tang, Yuan, & Lin, 1999). In the XX/XY sex-determining system, the Y-linked *DMY/Dmrt1bY* genes of the teleost, medaka, were characterized as sex-determining genes, leading to male sexual development (Masaru et al., 2002). In chicken and flatfish, which exhibit the ZZ/ZW system, DMRT1 is located on the Z chromosome, and the gene dosage may induce male development

(S. Chen et al., 2014; Nanda et al.). In *X. laevis* , one W-linked gene, DM-W, was identified as a likely sex (ovary)-determining gene (Yoshimoto et al.). The full-length cDNA sequence (2,165 base pairs) of *Ptdmrt* was obtained by rapid amplification of cloned ends (RACE), and the open reading frame spans three exons and encodes a putative protein of 265 amino acids, including one DM domain **(Fig 11)** . *Ptdmrt* can only amplify in males (40 female and 40 male individuals were used in this study) **(Fig 12)** , and RT-PCR results also showed that the *Ptdmrt* gene was expressed in all male tissues tested, but not in females **(Fig S12),** suggesting that *Ptdmrt* is a Y-linked gene and plays an important role in sex determination. During the different stages of embryonic development, *Ptdmrt* expression was significantly upregulated at the egg-nauplius stage, reaching its peak at the egg-zoea stage (upregulated 18.54-times, compared to the multicellular stage). Expression was then decreased during the Z1 stage and reached a second peak in expression in the Z3 phase (6.20-times) (**Fig 13)** . Those results indicated that the egg-zoea and zoea periods may be critical periods of sex differentiation.

Similar to the *Ptdmrt* gene, the expression of *Ptmyb* in males was also significantly higher than that in females. The difference between the expression of the two genes was that, *Ptmyb* was mainly expressed throughout development from Z1 to juvenile crab stages. The unknown sex-related gene was different from *Ptdmrt* and*Ptmyb* , of which expression in female was significantly higher than that in male during Z1 to Z4 phase. However, in the megalopal and juvenile crab stages, the expression pattern of the unknown sex-related gene changed, and was higher in males. Neither *Ptmyb* or the unknown sex-related genes were found to be involved in sex determination in previous studies, indicating the complexity and species specificity of the sex determination mechanisms in *P. trituberculatus.*

Conclusions

We present a chromosomal-scale genome assembly of the *P. trituberculatus* , a representative crab of the brachyuran species and an important specie of aquatic crustaceans. Genomic annotation and comparative genomic analyses provided insights into the genomic structure, evolution, and mechanisms underlying the biology of complex marine ecosystems. The data shed light on the salinity adaptation of*P. trituberculatus* , as well as their strong immunity, rapid growth and sex determination. This reference genome will provide crucial resources for evolutionary and biological studies, which are also important to clarify the regulatory mechanisms of important breeding traits and MAS for this economic crab.

Acknowledgements

This research was supported by the National Key R&D Program of China (2018YFD0900303 and 2018YFD0901304), Ministry of Agriculture of China & China Agriculture Research System (CARS-48), Efficient Eco Agriculture Innovation Project of Taishan Leading Talent Project (No. LJNY2015002), National Natural Science Foundation of China (41776160).

Author contributions

Jian Li, Chunlin Wang and Ping Liu.conceived and designed the research. Guangjian Liu, and Zhencheng Su performed the genome sequencing. Jianjian Lv, Ronghua Li, Zhencheng Su, Deping Yan and Xingbin Ti performed the experiments and data analyses. Ping Liu and Baoquan Gao performed sample preparation. Jianjian Lv and Zhencheng Su wrote the manuscript. All authors reviewed the manuscript.

Data availability

All genomic sequence datasets can be found on NCBI (https://www.ncbi.nlm.nih.gov/bioproject/ PRJNA631920/)

Figure and table

Fig 1 Genomic features of the *P.trituberculatus* . From outer to inner circles: 1, distribution of across the genome; 2, distribution of positive strand genes across the genome; 3, distribution of negative strand genes across the genome; 4, GC content across the genome. 5, represents *P.trituberculatus* chromosomes. In 5,

each line joins paralogous genes at different chromosomes. 2–4 are drawn in nonoverlapping 0.5-Mb sliding windows

Fig 2 Phylogenetic tree of 12 species, which was constructed using 246 single copy orthologous genes. Gene family expansion (+) and contraction (-) are shown in green and red colors above their branches, respectively. Red points at ancestral nodes represent bootstrap values $>$ 90 in 100 experiments. In addition to*P.trituberculatus* , pictures of other species are from the website http://phylopic.org/about/#about –use.

Fig 3 Phylogenetic tree and expression of *C2H2* family at different developmental stages in *P. trituberculatus* .(A) Phylogenetic tree of genes of the expanded gene family compared with the two shrimp species. (B) Hierarchical clustering of C2H2 family genes at seventeen different developmental stages in *P.trituberculatus*including fertilized egg stage (F), zoea stage (Z1–Z4), megalopal stage (M), and juvenile crab stage (J).

Fig 4 *Hox* Genes Clusters and the Expressions at Different Developmental Stages in *P. trituberculatus.* (A) Clustering of*Hox* genes in *P. trituberculatus* (*Ptr*), *E. sinensis* (*Esi*), *L. vannamei* (*Lva*), and *P. virginalis* (Pvi). The relative position and orientation of the genes are indicated.(B) qPCR of *Hox* genes at different developmental stages in *P. trituberculatus* .

Fig 5 Cluster analysis of the expanded and unique DEG during salinity stress. Green ring represents up regulated gene and yellow ring represents down regulated gene.

Fig 6 Identification of disease resistance differentiation through BSA analysis. X-axis represents the posiotion of Contigs of *P. trituberculatus* and Y-axis represents the SNP-index or Δ(SNP-index). The color dots represent the SNP-index or Δ(SNP-index) value of every SNP locus. The red lines show the SNP-index or Δ(SNP-index) value of fitting results. a. the SNP-index graph of X group. b. the SN-index graph of XT group. c. the Δ(SNP-index) graph. The blue dashed line and purple dashed line represent the condition of screening of two groups, respectively.

Fig 7 Expression analysis of disease resistance related genes in different tissues following infection with *V. parahaemolyticus*

Fig 8 Identification of fast growth differentiation through BSA analysis. X-axis represents the posiotion of Contigs of *P. trituberculatus* and Y-axis represents the SNP-index or Δ(SNP-index). The color dots represent the SNP-index or Δ(SNP-index) value of every SNP locus. The red lines show the SNP-index or Δ(SNP-index) value of fitting results. a. the SNP-index graph of PBX group. b. the SN-index graph of PSX group. c. the Δ(SNP-index) graph. The blue dashed line and purple dashed line represent the condition of screening of two groups, respectively.

Fig 9 Reads and coverage plot of the CQ results. The alignments from male (PBX-PSX) and female (PBC-PSC) sequence data to demonstrate how Y-linked sequences can be differentiated by distinctive CQ values in the same region of the splitted genome. X-axis represrnts a region and Y-axis represents male and female individuals.

Fig 10 Expression pattern of 8 genes from Z1 to juvenile crab in male and female individuals.

Fig 11 Sequence and structural features of *Ptdmrt* genes. (A) Nucleotide and deduced amino acid sequence of *Ptdmrt.* The black box indicates the conserved DM domain and the gray labeled region is the zinc finger structure in the DM domain. (B) Alignment of cDNA sequence of *Ptdmrt* and DNA sequence of Contig 475 (Ptdmrt gene was located in Contig 475). *Ptdmrt* gene consists of three exons and two introns.

Fig 12 *Ptdmrt* amplify in male and female individuals

Fig 13 Expression of *Ptdmrt* at different developmental stages. Eleven different developmental stages from fertilized eggs to young crabs including multicellular stage (Mc), blastula stage (B), gastrulation stage (G), egg-nauplius stage (En), egg-zoea stage (Ez), zoea stage (Z1–Z4), megalopal stage (M), and juvenile crab stage (J).

Table 1 Sequencing data used for the *P. trituberculatus* genome assembly

Pair-end libraries	**Insert size**	**Total data (G)**	**Read length (bp)**	**Sequence coverage (X)**
Illumina reads	350	111.01	150	92.51
Pacbio reads	–	120.79	–	100.66
10X Genomics	–	85.97	150	71.64
Total	–	317.77	–	264.81

Table 2 Assembly statistics of *P.trituberculatus*

Sample ID	Length Contig**(bp)	Length Scaffold(bp)	number Contig**	number Scaffold
Total	846,564,517	864,453,619	12,101	2,165
Max	6,277,820	33,040,454	-	-
Number>=2000	-	-	12,077	2,165
N50	108,682	15,593,479	2,055	20
N60	82,211	12,181,407	2,951	26
N70	62,400	8,722,939	4,132	34
N80	44,666	2,431,430	5,734	51
N90	30,342	130,609	8,037	384

** Contig after scaffolding

Table 3 42 growth-related genes, also expanded or specific in *P. trituberculatus*

Gene_ID	SwissProt annotation
evm.model.Contig153.38	-
evm.model.Contig238.5	Zinc finger protein 418 OS=Homo sapiens GN=ZNF418 PE=1 SV=2
evm.model.Contig128.16	Compound eye opsin BCRH2 OS=Hemigrapsus sanguineus PE=2 SV=1
evm.model.Contig1407.2	Caspase Nc OS=Drosophila melanogaster GN=Nc PE=1 SV=1
evm.model.Contig128.21	Compound eye opsin BCRH2 OS=Hemigrapsus sanguineus PE=2 SV=1
evm.model.Contig15.40	Neural-cadherin OS=Drosophila melanogaster GN=CadN PE=1 SV=2
evm.model.Contig166.1	COMM domain-containing protein 4 OS=Homo sapiens GN=COMMD4 PE
evm.model.Contig39.49	-
evm.model.Contig6.77	Acidic mammalian chitinase OS=Mus musculus GN=Chia PE=1 SV=2
evm.model.Contig1675.1	Protein prune homolog OS=Mus musculus GN=Prune PE=2 SV=1
evm.model.Contig397.12	Gamma-aminobutyric acid receptor subunit beta-4 OS=Gallus gallus GN=G
evm.model.Contig3.172	Adhesive plaque matrix protein OS=Mytilus coruscus GN=FP1 PE=2 SV=
evm.model.Contig639.9	-
evm.model.Contig3.89	Sodium- and chloride-dependent GABA transporter ine OS=Drosophila mel
evm.model.Contig58.27	Protein CIP2A OS=Homo sapiens GN=KIAA1524 PE=1 SV=2
evm.model.Contig693.3	-
evm.model.Contig139.4	Tripartite motif-containing protein 59 OS=Homo sapiens GN=TRIM59 PE=
evm.model.Contig1679.4	-
evm.model.Contig51.19	-
evm.model.Contig35.4	Kin of IRRE-like protein 1 OS=Rattus norvegicus GN=Kirrel PE=1 SV=2
evm.model.Contig28.13	Carbohydrate sulfotransferase 5 OS=Homo sapiens GN=CHST5 PE=2 SV=
evm.model.Contig82.11	Major facilitator superfamily domain-containing protein 6 OS=Sus scrofa G
evm.model.Contig66.12	Cytochrome c oxidase assembly factor 6 homolog OS=Homo sapiens GN=C
evm.model.Contig37.39	F-box/LRR-repeat protein 6 OS=Mus musculus GN=Fbxl6 PE=2 SV=2

Gene_ID	SwissProt annotation
evm.model.Contig526.3	Iduronate 2-sulfatase OS=Homo sapiens GN=IDS PE=1 SV=1
evm.model.Contig23.12	Golgi-specific brefeldin A-resistance guanine nucleotide exchange factor 1 OS
evm.model.Contig3.60	-
evm.model.Contig23.18	-
evm.model.Contig0.31_evm.model.Contig0.32	Ras-related protein Rab-3 OS=Drosophila melanogaster GN=Rab3 PE=1 S
evm.model.Contig499.7	Neural-cadherin OS=Drosophila melanogaster GN=CadN PE=1 SV=2
evm.model.Contig15.39	Sialin OS=Homo sapiens GN=SLC17A5 PE=1 SV=2
evm.model.Contig348.16	Branched-chain-amino-acid aminotransferase, cytosolic OS=Ovis aries GN=
evm.model.Contig954.14	Leucine-rich repeat-containing protein 16A OS=Mus musculus GN=Lrrc16a
evm.model.Contig246.3	Serine/threonine-protein kinase 10 OS=Danio rerio GN=stk10 PE=2 SV=1
evm.model.Contig188.4	-
evm.model.Contig96.7	-
evm.model.Contig89.6	Rho guanine nucleotide exchange factor 11 OS=Rattus norvegicus GN=Arh
evm.model.Contig597.4	-
evm.model.Contig1137.1	Protein Dok-7 OS=Takifugu rubripes GN=dok7 PE=2 SV=1
evm.model.Contig317.10	-
evm.model.Contig180.1	Tyrosine-protein phosphatase Lar OS=Drosophila melanogaster GN=Lar P
evm.model.Contig708.7	Mitochondrial import inner membrane translocase subunit Tim10 B OS=Xe

Table 4 The information of ten genes predicted in the Y-linked region

CHROM	Start	End	Strand	Gene_ID
Contig153	1307889	1310130	+	evm.model.Contig153.28
Contig153	1514797	1515057	-	evm.model.Contig153.29
Contig153	2210235	2210559	-	evm.model.Contig153.39
Contig153	2212469	2213665	+	evm.model.Contig153.40
Contig475	369282	477231	-	evm.model.Contig475.7
Contig475	490225	497352	+	evm.model.Contig475.8
Contig475	537316	537678	-	evm.model.Contig475.9
Contig475	537720	538208	-	evm.model.Contig475.10
Contig475	538896	539131	-	evm.model.Contig475.11
Contig475	865887	866685	+	evm.model.Contig475.13

Fig S1 Distribution of 17-mer frequency in *Portunus trituberculatus* genome. The volume of K-mer is plotted against the frequency at which they occur. The left-hand peak at low frequency and high volume represented K-mer containing essentially random sequencing errors. The main volume peak of K-mer was 41. The genome size was estimated as 1.2 Gb by the formula 'Genome size=total_kmer_num/kmer_depth'

Fig S2 Comparison of gene features among *P trituberculatus* ,*Tetranychus urticae* , *Drosophila melanogaster* ,*Daphnia pulex* , *Ixodes scapularis* , *Parasteatoda tepidariorum* , *Penaeus vannamei* , *Tribolium castaneum* .

Fig S3 Distribution of divergence rate of each type of TEs in the*P.trituberculatus* genome. The divergence rate was calculated between the identified TE elements in the genome and the consensus sequence in the *de novo* library we used.

Fig S4 The distribution of genes in different species

Fig S5 The statistics of KEGG pathway enrichment of expanded families

Fig S6 The statistics of KEGG pathway enrichment of identified specific families

Fig S7 the expression of genes in ribosome pathway in different stages of development, molting stage, low salt stress and pathogen infection

Fig S8 Trend of DEGs during salinity stress

Fig S9 KEGG enrichment of down-regulated expanded and specific DEGs during salt stress

Fig S10 KEGG enrichment of up-regulated expanded and specific DEGs during salt stress

Fig S11 KEGG enrichment analysis of growth-related genes

Fig S12 RT-PCR of *Ptdmrt* in male and female tissues of *P. trituberculatus*

Table S1 Primers used in this study.xlsx

Table S2 General statistics of SNP in *P.trituberculatus* genome

	Number	**Percentage(%)**
All SNP	2,586,394	0.3604
Heterozygosis SNP	2,538,522	0.3537
Homology SNP	47,872	0.0067

Table S3 BUSCO assessment of the *P.trituberculatus* genome

Species	BUSCO notation assessment results
P.trituberculatus	C:84.7%[S:82.5%,D:2.2%],F:2.9%,M:12.4%,n:1066

C:Complete BUSCOs S:Complete Single-Copy BUSCOs D:Complete Duplicated BUSCOs F:Fragmented BUSCOs M:Missing BUSCOs n:Total BUSCO groups searched

Table S4 CEGMA assessment of the *P.trituberculatus* genome

species	complete # Prots	complete %completeness	complete + partial # Prots	complete + partial %completeness
P.trituberculatus	187	75.40	211	85.08

1. Complete: core gene with a completeness >70%
2. Complete + partial: complete and partial core gene
3. #Prots: Number of core gene;
4. %completeness: the proportion core genes in the reference core gene library.

Table S5 Statistics of gene prediction

	Gene set	Number	Average transcript length(bp)	Average CDS length(bp)	Average exons per
De novo	Augustus	44,493	5,661.67	884.59	3.59
	GlimmerHMM	142,367	5,202.15	507.15	2.50
	SNAP	77,306	11,939.92	686.97	3.91
	Geneid	28,958	20,269.88	746.01	5.01
	Genscan	39,969	14,596.29	1,372.20	4.97
Homology	*Cel*	8,768	2,761.53	803.49	2.55
	Cgi	29,696	2,040.56	787.48	1.98
	Dme	10,815	4,260.04	891.58	3.24

	Gene set	Number	Average transcript length(bp)	Average CDS length(bp)	Average exons per
	Dpu	31,971	2,131.13	571.79	2.04
	Hsa	13,798	3,503.53	869.14	2.86
	Isc	47,137	1,287.92	567.46	1.58
	Ptep	23,660	2,639.54	836.84	2.27
	Pva	26,412	3,766.29	935.76	3.10
	Spu	15,073	2,878.81	1,023.85	2.39
	Tca	22,091	3,009.75	981.52	2.51
	Tur	18,806	2,754.58	914.90	2.20
RNAseq	PASA	44,920	13,768.62	1,389.51	6.28
	Cufflinks	56,571	18,303.84	3,099.64	6.63
EVM	EVM	23,993	8,599.08	1,147.41	4.49
Pasa-update*	Pasa-update*	23,395	9,475.94	1,219.26	4.76
Final set*	Final set*	19,981	10,484.05	1,231.18	5.22

注:*contains the untranslated region.

Table S6 Gene annotation in different databases.

		Number	**Percent (%)**
NR	NR	16,959	84.90
Swiss-Prot	Swiss-Prot	15,357	76.90
KEGG	KEGG	14,806	74.10
InterPro	all	19,677	98.50
	Pfam	14,492	72.50
	GO	18,257	91.40
Annotated	Annotated	19,763	98.90
Total	**Total**	19,981	-

Table S7 Categorization of repetitive sequences.

Type

Repeatmasker

(Repbase+Denovo)

Repeatmasker

(Repbase+Denovo)

TE protiens

TE protiens

Combined TEs

(all without TRF)

Combined TEs

(all without TRF)

Length (bp)

% in genome

Length (bp)

% in genome

Length (bp)

% in genome

DNA

88,177,866

10.20

7,698,813

0.89

94,695,859

10.95

LINE

205,758,594

23.80

90,240,215

10.44

235,904,361

27.29

SINE

134,806

0.02

0

0

134,806

0.02

LTR

142,449,802

16.48

15,507,775

1.79

147,540,995

17.07

Others

38

0.000004

0

0

38

0.000004

Unknown

7,689,177

0.89

0

0

7,689,177

3.42

Total

392,875,521

45.45

113,294,828

13.11

406,978,948

47.08

Table S8 Genes used for gene family clustering in each species

Species	**Name**	**Gene**
Ptr	*Portunus trituberculatus*	23694
Dre	*Danio rerio*	31951
Cse	*Cynoglossus semilaevis*	20211
Oni	*Oreochromis niloticus*	21209
Dpu	*Daphnia pulex*	30280
Esi	*Eriocheir sinensis*	7502
Lva	*Litopenaeus vannamei*	24825
Cgi	*Crassostrea gigas*	27265
Mye	*Mizuhopecten yessoensis*	23930
Pvi	*Procambarus virginalis*	20761
Cel	*Caenorhabditis elegans*	20060
Nve	*Nematostella vectensi*	24773

Table S9 The distribution of genes in different species

Species	**Single-copy orthologs**	**Multiple-copy orthologs**	**Unique**	**Other orthologs**
Cgi	869	762	5888	15655
Cse	657	1489	163	16248

Species	**Single-copy orthologs**	**Multiple-copy orthologs**	**Unique**	**Other orthologs**
Oni	621	1689	602	17458
Mye	925	667	2925	15747
Pvi	944	404	7382	10057
Dpu	889	768	10633	10968
Nve	886	755	5159	11577
Esi	1004	239	319	4551
Lva	811	958	2242	11779
Dre	484	2367	3452	23564
Ptr	854	740	2091	10517
Cel	908	564	7245	5794

Table S10 The information of expanded and contracted gene families in *P.trituberculatus*

	Gene family ID	Gene number	Classes
Expand	54	14	Growth and differentiation
	147	8	Metabolism
	287	3	Glycometabolism
	303	16	Regulation of gene expression
	378	2	Antibiosis
	410	9	Metabolism
	598	18	Regulation of gene expression, cell growth and differentiation
	967	9	Energy metabolism
	1118	10	Disease related
	1257	16	Signal transduction response regulator
	1691	12	Growth related
	2218	4	Immunity related
	3920	13	Sulfate
	4936	9	Material transportation
	4982	8	Opsin
	6221	11	Material transportation
	6224	11	Regulation of gene expression, growth and differentiation
	7258	10	Metabolism
Contract	184	1	Hydrolysis
	211	1	Immunity related

Table S11 The information of unique gene families identified.xlsx

Table S12 Expanded and contracted gene family of crab (Esi and Ptr).xlsx

Table S13 Genes of C2H2 zinc finger protein family in penaeid shrimp, crayfish and crab.xlsx

Table S14 Genes of actin and WDFY families in penaeid shrimp, crayfish and crab.xlsx

Table S15 DEG during salinity stress.xlsx

Table S16 Information of expanded and unique DEGs.xlsx

Table S17 Genotype frequency analysis of 12 markers.xlsx

Table S18 The BSA analysis result of growth trait.xlsx

References

Adey, A., Kitzman, J. O., Burton, J. N., Daza, R., Kumar, A., Christiansen, L., . . . Shendure, J. (2014). In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome research, 24* (12), 2041-2049. doi:10.1101/gr.178319.114

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:*https://doi.org/10.1016/S0022-2836(05)80360-2*

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England), 31* (2), 166-169. doi:10.1093/bioinformatics/btu638

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., . . . Yeh, L. S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic acids research, 32* (suppl_1), D115-D119. doi:10.1093/nar/gkh131

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research, 27* (2), 573-580. doi:10.1093/nar/27.2.573

Bergman, C. M., & Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics, 8* (6), 382-392. doi:10.1093/bib/bbm048

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise.*Genome research, 14* (5), 988-995. doi:10.1101/gr.1865504

Boetzer, M., & Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics, 15* , 211. Retrieved from*http://europepmc.org/abstract/MED/24950923*

https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/24950923/?tool=EBI

https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/24950923/pdf/?tool=EBI

https://doi.org/10.1186/1471-2105-15-211

https://europepmc.org/articles/PMC4076250

*https://europepmc.org/articles/PMC4076250?pdf=render*doi:10.1186/1471-2105-15-211

Borradaile, L. A. (2019). Crustacea. Part II. Porcellanopagurus: An instance of carcinization. *Zoology, 3* , 111-126.

Bowman, T. E. (1982). Classification of the recent Crustacea.*Biology of Crustacea* .

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA11Edited by F. E. Cohen. *Journal of Molecular Biology, 268* (1), 78-94. doi:*https://doi.org/10.1006/jmbi.1997.0951*

Catchen, J., Amores, A., & Bassham, S. (2020). Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. *bioRxiv* , 2020.2002.2004.934711. doi:10.1101/2020.02.04.934711

Chakravarti, & S. Lumican Regulates Collagen Fibril Assembly: Skin Fragility and Corneal Opacity in the Absence of Lumican. *Journal of Cell Biology, 141* (5), 1277-1286.

Chang, E. S., & Mykles, D. L. (2011). Regulation of crustacean molting: A review and our perspectives. *General and Comparative Endocrinology, 172* (3), 323-330. doi:10.1016/j.ygcen.2011.04.003

Chen, N. (2004). 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.bi0410s05

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., . . . Volff, J.-N. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nature Genetics, 46* (3), 253-260.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., . . . Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods, 10* (6), 563-569. doi:10.1038/nmeth.2474

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

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., . . . Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research, 44* (D1), D279-D285. doi:10.1093/nar/gkv1344

Ford, A. T. (2008). Can you feminise a crustacean? *Aquatic Toxicology, 88* (4), 316-321.

Fruman, D. A., Chiu, H., Hopkins, B. D., Bagrodia, S., & Abraham, R. T. (2017). The PI3K Pathway in Human Disease. *Cell, 170* (4), 605-635.

Garcon, D. P., Lucena, M. N., Franca, J. L., McNamara, J. C., Fontes, C. F. L., & Leone, F. A. (2011). Na+,K+-ATPase Activity in the Posterior Gills of the Blue Crab, Callinectes ornatus (Decapoda, Brachyura): Modulation of ATP Hydrolysis by the Biogenic Amines Spermidine and Spermine. *Journal of Membrane Biology, 244* (1), 9-20. doi:10.1007/s00232-011-9391-5

Grassot, V., Da Silva, A., Saliba, J., Maftah, A., Dupuy, F., & Petit, J.-M. Highlights of glycosylation and adhesion related genes involved in myogenesis. *BMC Genomics, 15* (1), 621.

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-R7. doi:10.1186/gb-2008-9-1-r7

Hall, A. B., Qi, Y., Timoshevskiy, V., Sharakhova, M. V., Sharakhov, I. V., & Tu, Z. (2013). Six novel Y chromosome genes in Anopheles mosquitoes discovered by independently sequencing males and females.*BMC Genomics, 14* , 273-273. doi:10.1186/1471-2164-14-273

Haug, J. T., Audo, D., Charbonnier, S., Palero, F., Petit, G., Saad, P. A., & Haug, C. (2016). The evolution of a key character, or how to evolve a slipper lobster. *Arthropod Structure & Development, 45* (2), 97-107. doi:10.1016/j.asd.2015.08.003

Hogan, B. L. M. (1987). Developmental and spatial patterns of expression of the mouse homeobox gene, Hox2.1. *Development, 99* (4), 603-617.

Hopkins, P. M. The eyes have it: A brief history of crustacean neuroendocrinology. *General & Comparative Endocrinology, 175* (3), 357-366.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England), 30* (9), 1236-1240. doi:10.1093/bioinformatics/btu031

Kaeodee, M., Pongsomboon, S., & Tassanakajon, A. (2011). Expression analysis and response of Penaeus monodon 14-3-3 genes to salinity stress. *Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology, 159* (4), 244-251. doi:10.1016/j.cbpb.2011.05.004

Kahn, B. B., & Flier, J. S. (2000). Obesity and insulin resistance.*Journal of Clinical Investigation, 106* (4), 473-481. doi:10.1172/jci10842

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., . . . Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research,* *24* (8), 1384-1395. doi:10.1101/gr.170720.113

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation.*Nucleic acids research, 44* (D1), D457-D462. doi:10.1093/nar/gkv1070

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

Koch, B. E. V., Stougaard, J., & Spaink, H. P. (2015). Keeping track of the growing number of biological functions of chitin and its interaction partners in biomedical research. *Glycobiology, 25* (5), 469-482.

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research, 35* (Web Server issue), W345-W349. doi:10.1093/nar/gkm391

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times.*Molecular Biology and Evolution, 34* (7), 1812-1819. doi:10.1093/molbev/msx116

Lecher, P., & Noel, D. D. (1995). Chromosomes and nuclear DNA of Crustacea. *International Journal of Invertebrate Reproduction, 27* (2), 85-114.

Li, A., Zhang, J., & Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme.*BMC bioinformatics, 15* (1), 311-311. doi:10.1186/1471-2105-15-311

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics, 12* , 323-323. doi:10.1186/1471-2105-12-323

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England), 25* (16), 2078-2079. doi:10.1093/bioinformatics/btp352

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

Li, P. M., Zhang, W. R., & Goldstein, B. J. (1996). Suppression of Insulin Receptor Activation by Overexpression of the Protein-Tyrosine Phosphatase LAR in Hepatoma Cells. *Cellular Signalling, 8* (7), 467-473.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., . . . Wang, J. (2010). The sequence and de novo assembly of the giant panda genome.*Nature, 463* (7279), 311-317. doi:10.1038/nature08696

Li, W., Jaroszewski, L., & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics, 18* (1), 77-82. doi:10.1093/bioinformatics/18.1.77

Liu, L., Li, J., Liu, P., Zhao, F., Gao, B., & Du, Y. (2015). Identification of quantitative trait loci for growth-related traits in the swimming crab Portunus trituberculatus. *Aquaculture Research, 46* (4), 850-860. doi:10.1111/are.12239

Lv, J., Liu, P., Wang, Y., Gao, B., Chen, P., & Li, J. (2013). Transcriptome Analysis of Portunus trituberculatus in Response to Salinity Stress Provides Insights into the Molecular Basis of Osmoregulation. *PLoS ONE, 8* (12), e82155. doi:10.1371/journal.pone.0082155

Lv, J., Sun, D., Huan, P., Song, L., Liu, P., & Li, J. (2018). QTL Mapping and Marker Identification for Sex-Determining: Indicating XY Sex Determination System in the Swimming Crab (Portunus trituberculatus).*Frontiers in Genetics, 9* (337). doi:10.3389/fgene.2018.00337

Lv, J., Zhang, D., Liu, P., & Li, J. (2016). Effects of salinity acclimation and eyestalk ablation on Na + , K + , 2Cl - cotransporter gene expression in the gill of Portunus trituberculatus :a molecular correlate for salt-tolerant trait. *Cell Stress & Chaperones* , 1-8.

Maccalman, C. D., Bardeesy, N., Holland, P. C., & Blaschuk, O. W. Noncoordinate developmental regulation of N-cadherin, N-CAM, integrin, and fibronectin mRNA levels during myoblast terminal differentiation.*Dev Dyn, 195* (2), 127-132.

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

Mander, A., Hodgkinson, C. P., & Sale, G. J. Knock-down of LAR protein tyrosine phosphatase induces insulin resistance. *579* (14), 0-3028.

Mao, X., Cai, T., Olyarchuk, J. G., & Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics, 21* (19), 3787-3793. doi:10.1093/bioinformatics/bti430

Mardis, E. R. (2008). Next-Generation DNA Sequencing Methods.*Annual Review of Genomics and Human Genetics, 9* (1), 387-402. doi:10.1146/annurev.genom.9.081307.164359

Martins, D. J. (2002). The birds and the bees - and the flowers.

Masaru, M., Yoshitaka, N., Ai, S., Tadashi, S., Chika, M., Tohru, K., . . . Nobuyoshi, S. (2002). DMY is a Y-specific DM-domain gene required for male development in the medaka fish. *Nature, 417* (6888), 559-563.

McGinnis, W. Region-specific expression of two mouse homeo box genes.*Science, 235* (4794), 1379-1382.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research, 20* (9), 1297-1303. doi:10.1101/gr.107524.110

McNamara, J. C., & Faria, S. C. (2012). Evolution of osmoregulatory patterns and gill ion transport mechanisms in the decapod Crustacea: a review. *Journal of Comparative Physiology B-Biochemical Systemic and Environmental Physiology, 182* (8), 997-1014. doi:10.1007/s00360-012-0665-8

Meng, A., Moore, B., Tang, H., Yuan, B., & Lin, S. (1999). A Drosophila doublesex-related gene, terra, is involved in somitogenesis in vertebrates. *Development, 126* (6), 1259-1268.

Misa, Ohno, Masahiro, Kimura, Haruko, Miyazaki, . . . Onuki. Acidic mammalian chitinase is a proteases-resistant glycosidase in mouse digestive system.

Nanda, I., Zend-Ajusch, E., Shan, Z., Gr, uuml, Tzner, F., . . . Goodwin, G. Conserved synteny between the chicken Z sex chromosome and human chromosome 9 includes the male regulatory gene <i>DMRT1:</i> a comparative (re)view on avian sex determination. *Cytogenetic & Genome Research, 89* (1-2), 67-78.

Niiyama, H. (1937). The Problem of Male Heterogamety in the Decapod Crustacea, with Special Reference to the Sex-Chromosomes in Plagusia dentipes de Haan and Eriocheir japonicus de Haan (With Plate VIII and 7 Textfigures). 北海道帝大理部要*, 5* (7), 1077-1081.

Niiyama, H. (1938). The X-Y chromosomes of the shore-crab, Hemigrapsus sanguineus (de HAAN). *Japanese Journal of Genetics, 14* , 34-38.

Niiyama, H. (1959). AN XX-Y SEX-MECHANISM IN THE MALE OF A DECAPOD CRUSTACEA CERVIMUNIDA PRINCEPS BENEDICT.北海道大水部研究*, 10* (2), 106-112.

Parra, G., Blanco, E., & Guigó, R. (2000). GeneID in Drosophila.*Genome research, 10* (4), 511-515. doi:10.1101/gr.10.4.511

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

Phillips, D. I. W. (1995). Relation of Fetal Growth to Adult Muscle Mass and Glucose Tolerance. *Diabetic Medicine, 12* (8), 686-690.

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

Ren, X., Yu, X., Gao, B., Liu, P., & Li, J. (2017). Characterization of three caspases and their pathogen-induced expression pattern in Portunus trituberculatus. *Fish & Shellfish Immunology, 66* , 189-197.

Romano, N., & Zeng, C. S. (2012). Osmoregulation in decapod crustaceans: implications to aquaculture productivity, methods for potential improvement and interactions with elevated ammonia exposure.*Aquaculture, 334* , 12-23. doi:10.1016/j.aquaculture.2011.12.035

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics, 30* (24), 3506-3514. doi:10.1093/bioinformatics/btu538

Saltiel, A. R., & Kahn, C. R. (2001). Insulin signalling and the regulation of glucose and lipid metabolism. *Nature, 414* (6865), 799-806. doi:10.1038/414799a

Scholtz, G. (2014). Evolution of crabs - history and deconstruction of a prime example of convergence. *Contributions to Zoology, 83* (2), 87-105.

Silva, L. M., & Jung, J. U. (2013). *Autophagy and Immunity* .

Simão, 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

Song, C., Cui, Z., Hui, M., Liu, Y., Li, Y., & Li, X. (2015). Comparative transcriptomic analysis provides insights into the molecular basis of brachyurization and adaptation to benthic lifestyle in Eriocheir sinensis. *Gene, 558* (1), 88-98. doi:10.1016/j.gene.2014.12.048

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.*Bioinformatics, 22* (21), 2688-2690. doi:10.1093/bioinformatics/btl446

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints.*Nucleic acids research, 33* (suppl_2), W465-W467. doi:10.1093/nar/gki458

Steinz, M. M., Persson, M., Aresh, B., Olsson, K., Cheng, A. J., Ahlstrand, E., . . . Lanner, J. T. (2019). Oxidative hotspots on actin promote skeletal muscle weakness in rheumatoid arthritis. *Jci Insight, 4* (9), 16. doi:10.1172/jci.insight.126347

Tang, B., Wang, Z., Liu, Q., Zhang, H., Jiang, S., Li, X., . . . Li, Y. (2020). High-Quality Genome Assembly of Eriocheir japonica sinensis Reveals Its Unique Genome Evolution. *Frontiers in Genetics, 10* , 1340.

The Gene Ontology Consortium. (2016). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research, 45* (D1), D331-D338. doi:10.1093/nar/gkw1108

Torrecilla, Z., Martínez-Lage, A., Perina, A., González-Ortegón, E., & González-Tizón, A. M. (2017). Comparative cytogenetic analysis of marine Palaemon species reveals a X 1 X 1 X 2 X 2/X 1 X 2 Y sex chromosome system in Palaemon elegans. *Frontiers in Zoology, 14* (1), 47.

Torrecilla, Z., Martínezlage, A., Perina, A., Gonzálezortegón, E., & Gonzáleztizón, A. M. (2017). Comparative cytogenetic analysis of marine Palaemon species reveals a X1X1X2X2/X1X2Y sex chromosome system in Palaemon elegans. *Frontiers in Zoology, 14* (1), 47.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25* (9), 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., . . . Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols, 7* (3), 562-578. doi:10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., . . . Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28* (5), 511-515. doi:10.1038/nbt.1621

Tsai, J.-R., & Lin, H.-C. (2007). V-type H+-ATPase and Na+,K+-ATPase in the gills of 13 euryhaline crabs during salinity acclimation.*Journal of Experimental Biology, 210* (4), 620-627. doi:10.1242/jeb.02684

Urrutia, R. (2003). KRAB-containing zinc-finger repressor proteins.*Genome Biology, 4* (10), 8. doi:10.1186/gb-2003-4-10-231

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE, 9* (11).

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.*Nucleic acids research, 38* (16), e164-e164. doi:10.1093/nar/gkq603

Wang, Q. Caspase and nm23: Apoptosis genes linked to the antibacterial response of the Chinese mitten crab. *Bioengineered Bugs, 2* (3), 174-177.

Warner, G. F. (1977). biology of crabs. *Cahiers de Biologie Marine, 19* (1), 115-115.

Xiaoqing, S., Jiankai, W., Jianbo, Y., Xiaojun, Z., Fuhua, L., & Jianhai, X. (2015). Hox Genes and Their Expression Pattern in Early Development of Litopenaeus vannamei. *Periodical of Ocean University of China, 045* (008), 52-62.

Xu, Q. H., & Liu, Y. (2011). Gene expression profiles of the swimming crab Portunus trituberculatus exposed to salinity stress. *Marine Biology, 158* (10), 2161-2172. doi:10.1007/s00227-011-1721-8

Xu, X., Ye, L., Araki, K., & Ahmed, R. (2012). mTOR, linking metabolism and immunity. *Seminars in Immunology, 24* (6), 429-435.

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., Ding, J., Li, Z., Yan, L., . . . Li, D. (2019). Clam Genome Sequence Clarifies the Molecular Basis of Its Benthic Adaptation and Extraordinary Shell Color Diversity.*iScience, 19* , 1225-1237. doi:*https://doi.org/10.1016/j.isci.2019.08.049*

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

Yoshimoto, S., Okada, E., Umemoto, H., Tamura, K., Uno, Y., Nishida-Umehara, C., . . . Ito, M. A W-linked DM-domain gene, DM-W, participates in primary ovary development in Xenopus laevis.*Proceedings of the National Academy of Sciences of the United States of America, 105* (7), 2469-2474.

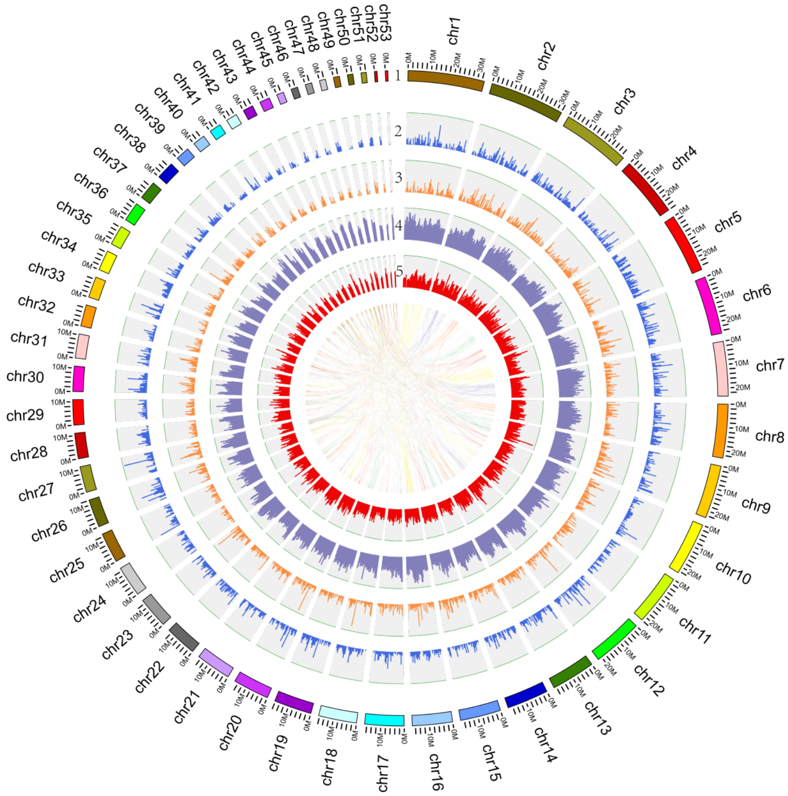
Zhang, H.-M., Liu, T., Liu, C.-J., Song, S., Zhang, X., Liu, W., . . . Guo, A.-Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic acids research, 43* (Database issue), D76-D81. doi:10.1093/nar/gku887

Zhang, X., Yuan, J., Sun, Y., Li, S., Gao, Y., Yu, Y., . . . Xiang, J. (2019). Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nature Communications, 10* (1), 356. doi:10.1038/s41467-018-08197-4

Zhang, Y. Q., Bo, L., Zhang, H., Zhuang, C., & Liu, R. P. (2014). E26 Transformation- Specific-1 (ETS1) and WDFY Family Member 4 (WDFY4) Polymorphisms in Chinese Patients with Rheumatoid Arthritis.*International Journal of Molecular Sciences, 15* (2), 2712-2721. doi:10.3390/ijms15022712

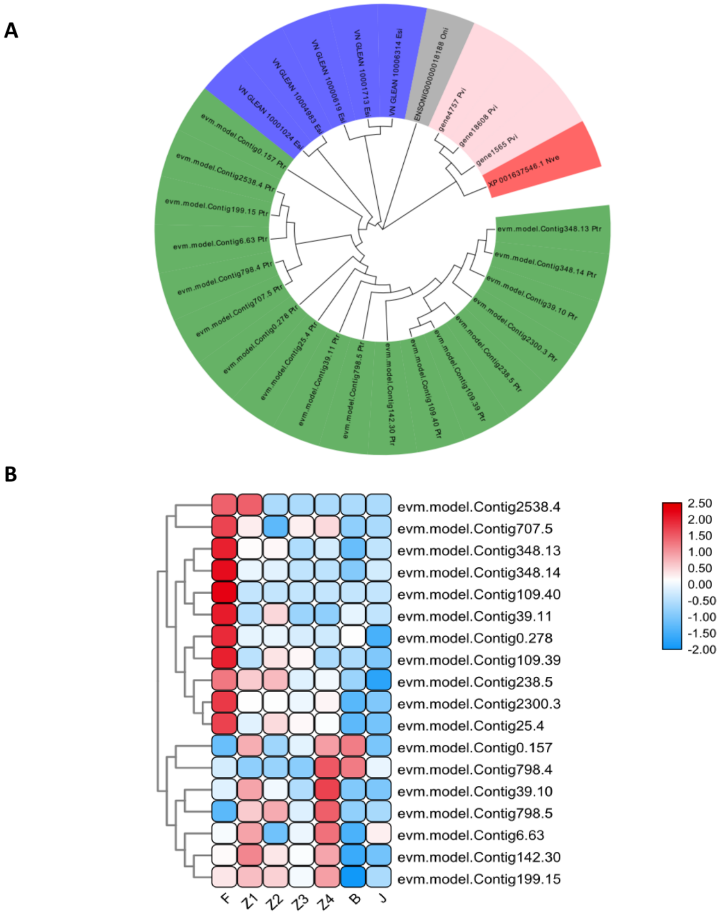
Zhong, M. Molecular cloning and expression analysis of a cytosolic heat shock protein 70 gene from mud crab Scylla serrata. *Fish & Shellfish Immunology, 34* (5), 1306-1314.

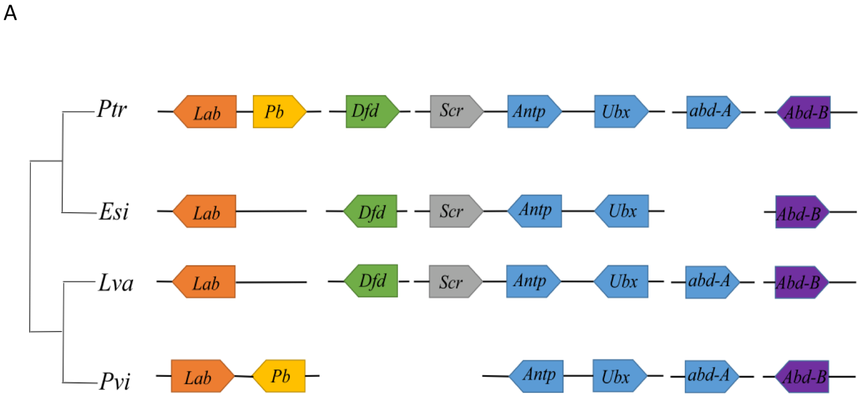
Zhongfa, W., Zhizheng, W., Wenjun, X., Weixian, H., & Songye, G. (2008). Quantitative study of lethal effect of wssv on portunu trituberculatus from mix-culture ponds of prawns and crabs.*Oceanologia et Limnologia Sinica* (2), 90-95.

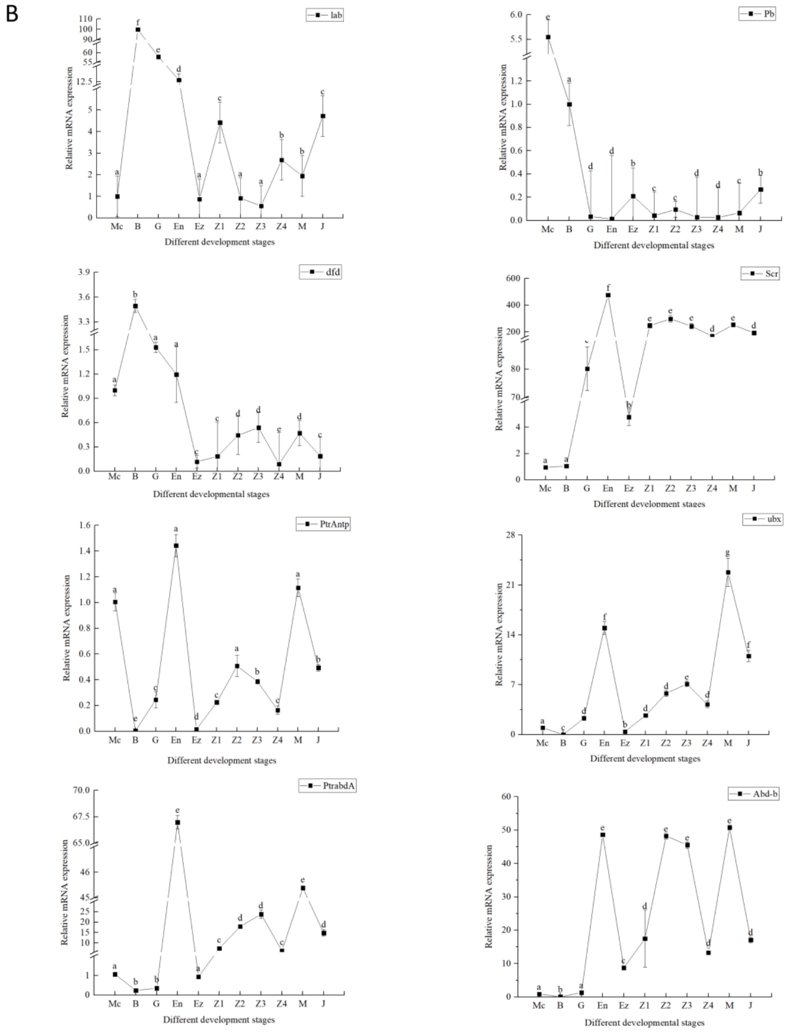


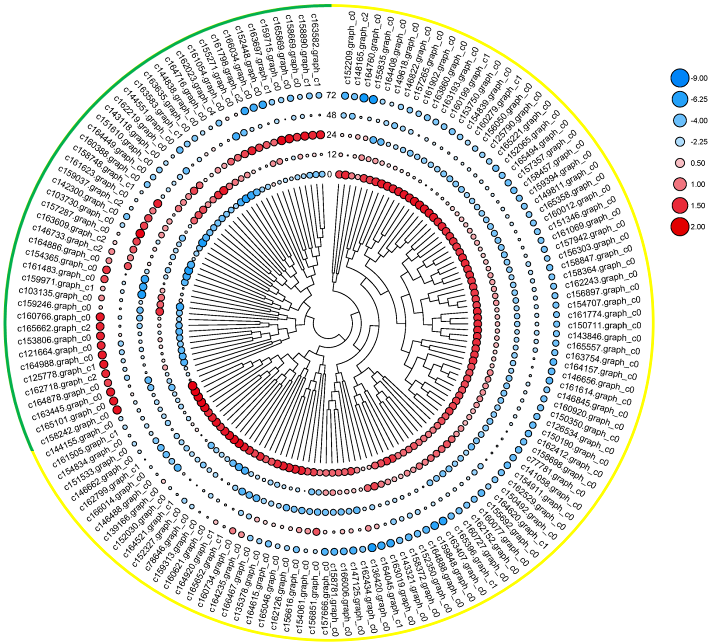
Hosted file

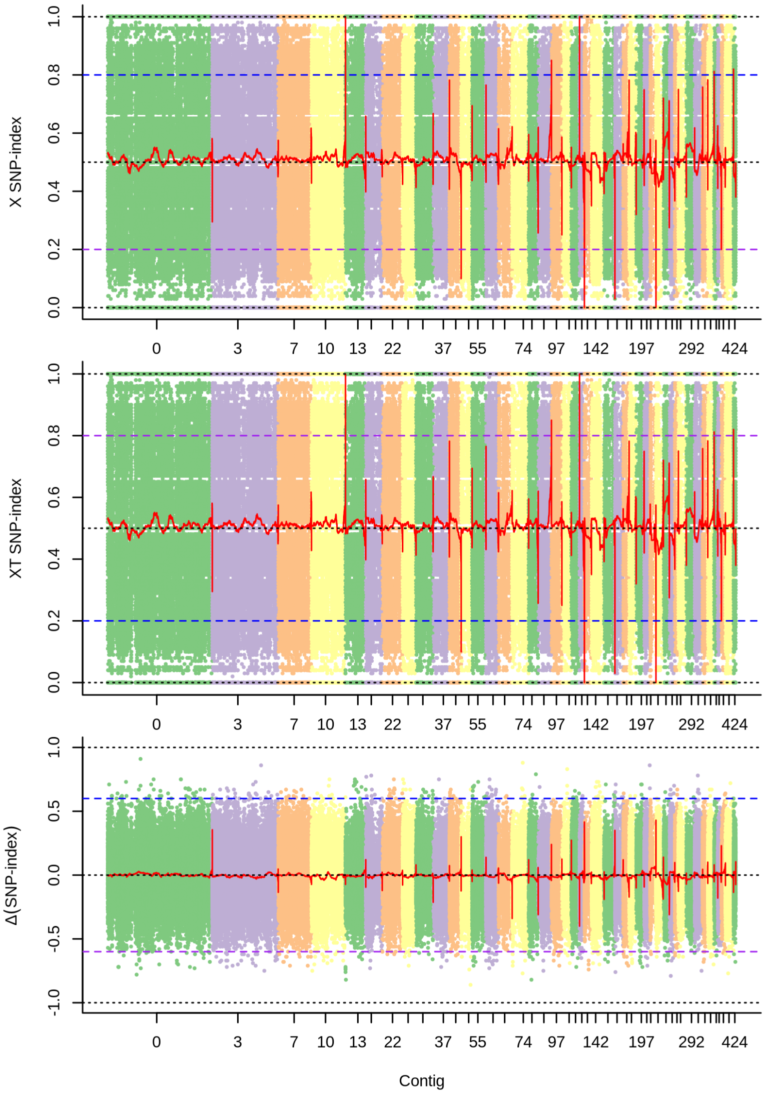
`Fig 2.pdf` available at https://authorea.com/users/348321/articles/473714-a-chromosome-level-genome-of-portunus-trituberculatus-provides-insights-into-its-evolution-salinity-adaptation-and-sex-determination

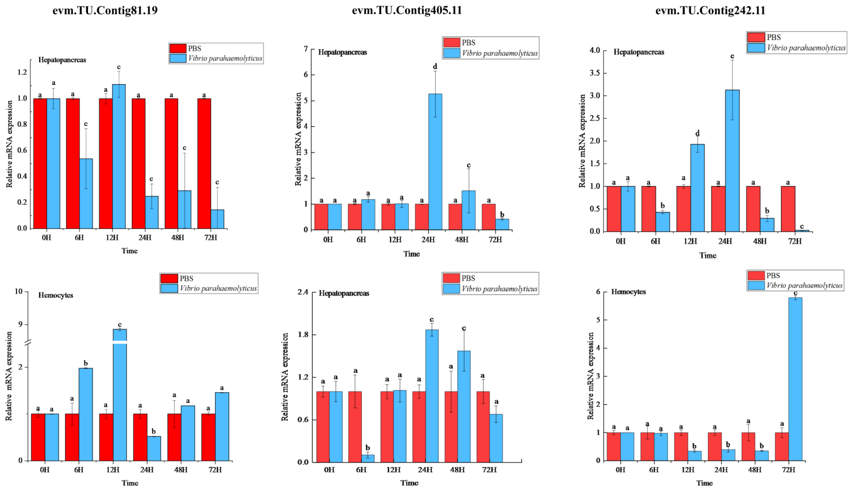

A
VN GLEAN 10001024 Esi
VN GLEAN 10004983 Esi
VN GLEAN 10000619 Esi
VN GLEAN 10001713 Esi
VN GLEAN 10006314 Esi
ENSONIG00000018188 Oni
gene4757 Pvi
gene18608 Pvi
gene1565 Pvi
XP 001637546.1 Nve
evm.model.Contig348.13 Ptr
evm.model.Contig348.14 Ptr
evm.model.Contig39.10 Ptr
evm.model.Contig2300.3 Ptr
evm.model.Contig238.5 Ptr
evm.model.Contig109.39 Ptr
evm.model.Contig109.40 Ptr
evm.model.Contig142.30 Ptr
evm.model.Contig798.5 Ptr
evm.model.Contig39.11 Ptr
evm.model.Contig25.4 Ptr
evm.model.Contig0.278 Ptr
evm.model.Contig707.5 Ptr
evm.model.Contig798.4 Ptr
evm.model.Contig6.63 Ptr
evm.model.Contig199.15 Ptr
evm.model.Contig2538.4 Ptr
evm.model.Contig0.157 Ptr
B
evm.model.Contig2538.4
evm.model.Contig707.5
evm.model.Contig348.13
evm.model.Contig348.14
evm.model.Contig109.40
evm.model.Contig39.11
evm.model.Contig0.278
evm.model.Contig109.39
evm.model.Contig238.5
evm.model.Contig2300.3
evm.model.Contig25.4
evm.model.Contig0.157
evm.model.Contig798.4
evm.model.Contig39.10
evm.model.Contig798.5
evm.model.Contig6.63
evm.model.Contig142.30
evm.model.Contig199.15
F
Z1
Z2
Z3
Z4
B
J
2.50
2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
-2.00

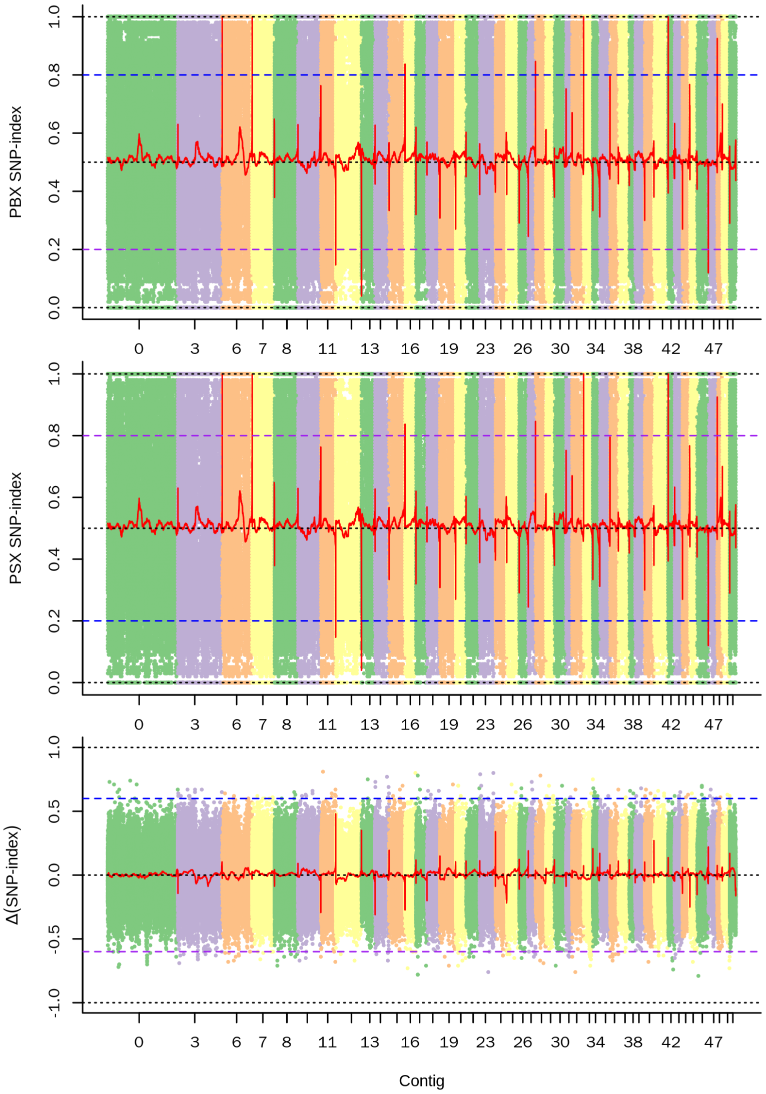

A
Ptr
Lab
Pb
Dfd
Scr
Antp
Ubx
abd-A
Abd-B
Esi
Lab
Dfd
Scr
Antp
Ubx
Abd-B
Lva
Lab
Dfd
Scr
Antp
Ubx
abd-A
Abd-B
Pvi
Lab
Pb
Antp
Ubx
abd-A
Abd-B

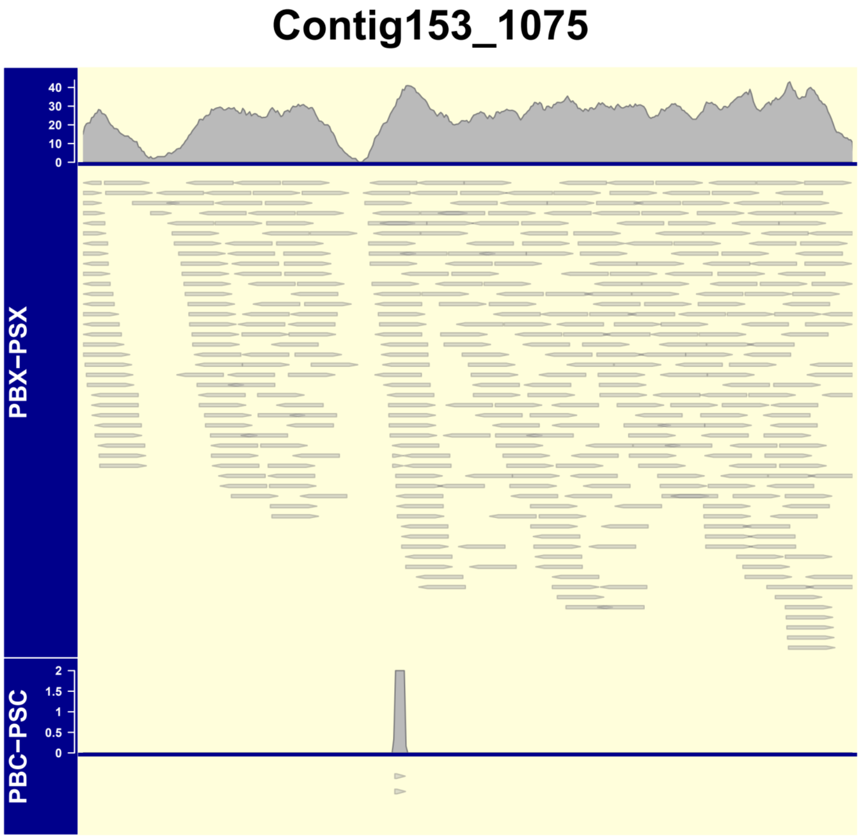

B
lab
Relative mRNA expression
Different development stages
Pb
Different developmental stages
dfd
Different developmental stages
Scr
Different development stages
PtrAntp
Different development stages
ubx
Different development stages
PtrabdA
Different development stages
Abd-b
Different developmental stages
Mc B G En Ez Z1 Z2 Z3 Z4 M J


c152209.graph_c0
c148165.graph_c2
c164760.graph_c0
c155835.graph_c0
c164408.graph_c0
c149618.graph_c0
c146822.graph_c0
c157265.graph_c0
c161902.graph_c0
c163860.graph_c0
c163193.graph_c0
c160199.graph_c1
c153750.graph_c0
c154839.graph_c0
c160279.graph_c1
c156050.graph_c0
c125790.graph_c0
c165221.graph_c0
c152065.graph_c0
c165494.graph_c0
c157357.graph_c0
c158457.graph_c0
c159394.graph_c0
c149811.graph_c0
c165358.graph_c0
c160012.graph_c0
c151346.graph_c0
c161069.graph_c0
c157942.graph_c0
c156303.graph_c0
c158847.graph_c0
c158364.graph_c0
c162243.graph_c0
c156897.graph_c0
c154707.graph_c0
c161774.graph_c0
c150711.graph_c0
c143846.graph_c0
c165557.graph_c0
c163754.graph_c0
c164157.graph_c0
c146656.graph_c0
c161614.graph_c0
c146845.graph_c0
c160920.graph_c0
c150350.graph_c0
c126534.graph_c0
c150190.graph_c0
c162412.graph_c0
c158698.graph_c0
c77781.graph_c0
c141058.graph_c0
c154911.graph_c0
c162529.graph_c0
c150492.graph_c0
c164620.graph_c1
c155692.graph_c0
c160077.graph_c0
c162152.graph_c0
c160727.graph_c0
c165396.graph_c1
c163407.graph_c0
c159848.graph_c0
c164888.graph_c0
c152350.graph_c0
c158372.graph_c0
c143321.graph_c0
c163019.graph_c1
c164045.graph_c0
c136420.graph_c0
c162434.graph_c0
c147125.graph_c0
c166006.graph_c0
c158781.graph_c0
c157666.graph_c0
c156851.graph_c0
c154061.graph_c0
c156616.graph_c0
c162126.graph_c0
c165046.graph_c0
c164615.graph_c0
c135378.graph_c0
c166467.graph_c0
c164235.graph_c0
c160734.graph_c0
c165652.graph_c1
c164920.graph_c1
c160621.graph_c0
c159313.graph_c0
c78646.graph_c0
c152327.graph_c0
c164521.graph_c1
c152030.graph_c0
c139166.graph_c0
c146488.graph_c0
c166014.graph_c0
c162799.graph_c0
c146662.graph_c1
c151533.graph_c0
c154834.graph_c0
c161505.graph_c0
c144155.graph_c1
c158242.graph_c0
c165101.graph_c0
c163445.graph_c0
c164878.graph_c0
c162718.graph_c2
c125778.graph_c1
c164988.graph_c0
c121664.graph_c0
c153806.graph_c0
c165662.graph_c2
c160766.graph_c0
c159246.graph_c0
c103135.graph_c0
c159971.graph_c1
c161483.graph_c0
c154365.graph_c0
c164886.graph_c0
c146733.graph_c2
c163609.graph_c2
c157287.graph_c0
c103730.graph_c0
c142300.graph_c0
c159037.graph_c2
c161623.graph_c0
c158748.graph_c1
c160388.graph_c0
c164449.graph_c0
c151610.graph_c0
c143118.graph_c0
c162219.graph_c0
c144551.graph_c1
c163583.graph_c0
c163635.graph_c0
c144838.graph_c0
c164716.graph_c0
c162023.graph_c0
c161054.graph_c4
c155271.graph_c0
c161798.graph_c2
c166034.graph_c0
c152448.graph_c0
c163697.graph_c1
c159715.graph_c0
c165869.graph_c0
c158669.graph_c0
c158890.graph_c1
c163582.graph_c1
72
48
24
12
0
-9.00
-6.25
-4.00
-2.25
0.50
1.00
1.50
2.00

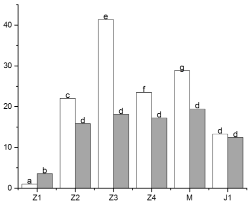

X SNP-index
1.0
0.8
0.6
0.4
0.2
0.0
0
3
7
10
13
22
37
55
74
97
142
197
292
424
XT SNP-index
1.0
0.8
0.6
0.4
0.2
0.0
0
3
7
10
13
22
37
55
74
97
142
197
292
424
Δ(SNP-index)
1.0
0.5
0.0
-0.5
-1.0
0
3
7
10
13
22
37
55
74
97
142
197
292
424
Contig


evm.TU.Contig81.19
evm.TU.Contig405.11
evm.TU.Contig242.11
Hepatopancreas
Hemocytes
PBS
Vibrio parahaemolyticus
Relative mRNA expression
Time
0H
6H
12H
24H
48H
72H
a
b
c
d

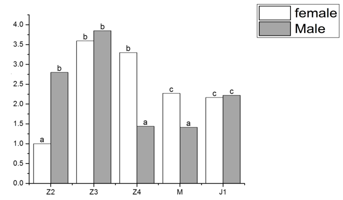

PBX SNP-index
0.0
0.2
0.4
0.6
0.8
1.0
0
3
6
7
8
11
13
16
19
23
26
30
34
38
42
47
PSX SNP-index
0.0
0.2
0.4
0.6
0.8
1.0
0
3
6
7
8
11
13
16
19
23
26
30
34
38
42
47
Δ(SNP-index)
-1.0
-0.5
0.0
0.5
1.0
0
3
6
7
8
11
13
16
19
23
26
30
34
38
42
47
Contig


Contig153_1075
PBX–PSX
40
30
20
10
0
PBC–PSC
2
1.5
1
0.5
0

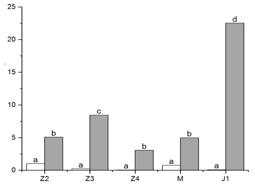
contig153.39



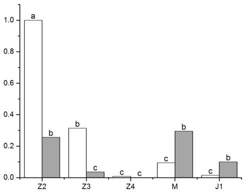
contig153.40



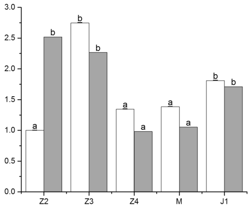
***contig153.28**



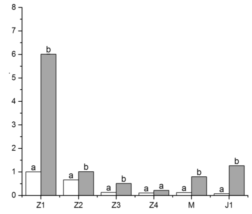
***Contig475.8**



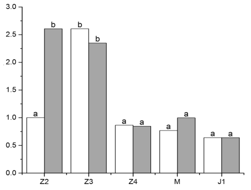
Contig475.9



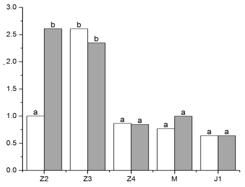
***Contig475.13**



Contig475.11

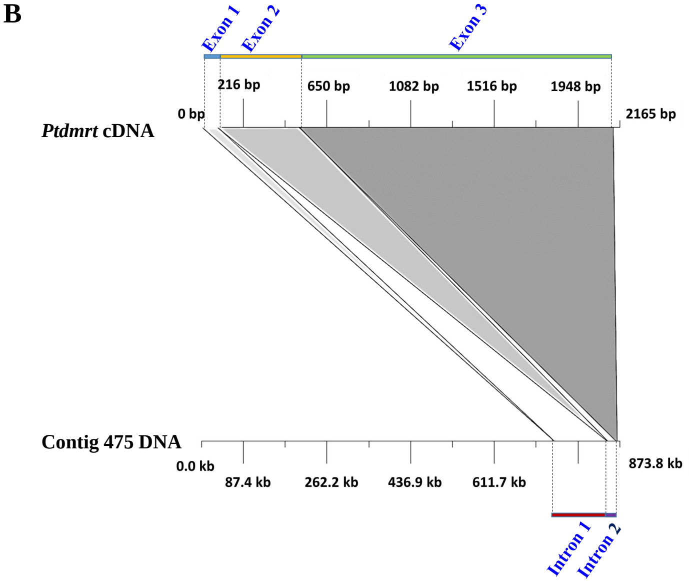


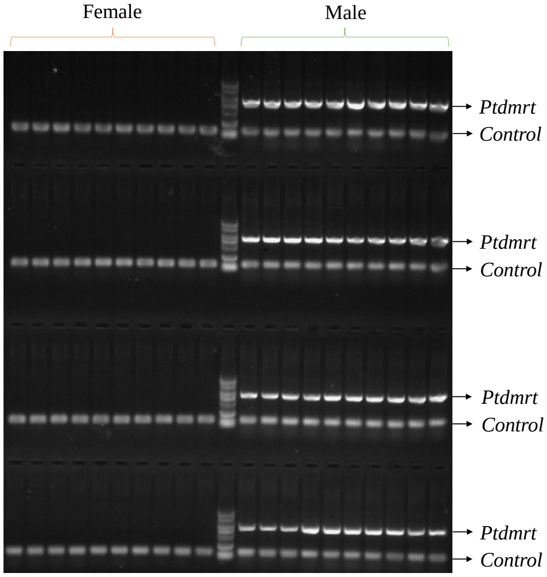
Contig475.10

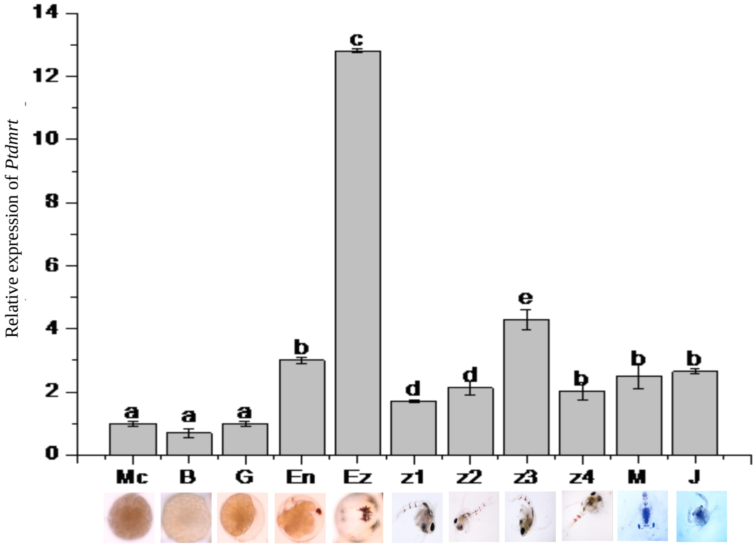


A

```
1 CTACTTCTCCTAAAGCACCGTGTCACTGTTGCCATATGGAGCATGTGGTGATGTCGAACGCTGCTGCTCCACCTATTTCGTAAGACGGGCTGGCTTTCTTGGTGATGACATCTCTGATCA
121 CCCCATCGTGCCCTGCCGTCAGTGTCCTGTCACCCTTAATATAAAATAACTTACCTAGTCAGAAATTTATCTCGCATCAGTTGTATGTATAGTGACAATCCACCAGGAATACACCAATCC
241 ACCGTCTACTTATCAGCTGCTACAACCATAACTATTTTGATAATGTGTTTTGGTAGTGATCGATGAGTGTGTCAGACAAGTGACTTCTTTACATGTTAATTGTTGGCATAGTGACTCTGC
                                                         M   A   S   A   P   K   K   R   D   F   F   N   G   S   Q   R   L
361 CCACCATCTCTTCATTTCCGTGTACTCGACAGGTGAGGTAAGGTTGTCAGC ATG GCG TCG GCG CCT AAG AAG AGA GAC TTT TTT AAT GGA TCA CAA AGA CTA
     N   L   T   E   T   Q   E   E   S   M   V   P   E   Q   N   V   T   S   G   G   R   N   G   F   S   S   S   Q   D   F
463 AAC TTA ACG GAG ACA CAG GAA GAG AGT ATG GTT CCT GAG CAA AAT GTC ACC AGT GGA GGA AGA AAT GGC TTC TCT TCC TCC CAG GAT TTC
     L   G   H   S   S   G   D   C   P   F   P   S   F   S   N   R   Q   S   S   P   V   Y   N   C   Q   D   S   C   P   S
553 CTT GGG CAC TCT TCA GGC GAC TGT CCT TTC CCA TCT TTC TCC AAT AGA CAA AGT TCT CCT GTG TAC AAT TGT CAA GAC AGC TGC CCA TCC
     L   S   S   R   Y   S   I   R   T   K   R   D   E   Y   E   N   S   K   G   T   S   Y   E   S   N   S   V   P   E   Y
643 CTC TCT TCA CGA TAT AGT ATC CGA ACG AAA AGG GAC GAA TAC GAA AAC TCC AAA GGC ACT TCA TAC GAG TCT AAT TCA GTT CCA GAA TAC
     L   S   S   A   G   S   A   E   V   A   S   L   A   K   H   D   G   N   Y   E   R   K   K   K   V   E   K   R   K   Q
733 TTA TCT TCG GCG GGC TCC GCT GAA GTT GCT TCC TTA GCC AAG CAC GAT GGA AAC TAC GAA AGG AAG AAG AAA GTG GAA AAG AGA AAA CAG
     R   C   R   L   C   A   N   H   G   K   Y   E   E   I   K   G   H   K   W   Y   C   E   Y   R   K   P   Q   H   K   C
823 AGG TGT CGA TTG TGC GCC AAT CAC GGC AAG TAC GAA GAG ATT AAA GGT CAC AAG TGG TAC TGC GAA TAC AGG AAA CCA CAA CAC AAG TGT
     S   L   C   E   I   T   H   K   K   R   L   F   L   P   K   N   E   I   R   R   K   Q   H   D   Q   E   Q   Q   L   Q
913 TCC CTG TGT GAA ATC ACC CAC AAG AAA CGA CTC TTC CTA CCG AAA AAT GAG ATC AGA CGC AAA CAG CAT GAC CAA GAG CAG CAG CTA CAA
      Q   Q   L   N   V   Q   N   R   S   T   D   E   P   W   L   D   G   Y   G   R   V   G   L   P   P   S   P   T   T   E
1003 CAA CAA TTA AAC GTG CAG AAC CGA TCG ACA GAT GAA CCA TGG CTG GAT GGT TAT GGT CGG GTC GGC TTA CCA CCC TCG CCA ACT ACC GAA
      H   I   D   F   P   R   L   Q   E   L   V   E   E   T   A   S   I   L   D   D   E   D   L   F   R   Q   I   N   E   R
1093 CAC ATA GAC TTC CCC CGG CTA CAG GAA CTA GTA GAA GAG ACA GCA TCT ATC TTG GAC GAC GAG GAT TTG TTC CGC CAA ATC AAT GAG CGC
      I   P   L   N   V   L   Q   H   *
1183 ATC CCG TTG AAT GTC CTT CAA CAT TGA TTACCTACTGTCCTCCCTAATCTACCTACCTGGTCCAGCGGAAGAATCGATGCTTGTAAAGGGATGCAGCTATTTC CACGTGCC
1294 AAATTCCGCACACAGTCTTCATCTGCTTGCAGGAAACTGGTAACGCGATAAATTGCCTTGCGTTCAAACTATGTAACGCTACATATGATTGAATATGTTTTGACAAGGTTTGTAGAGTAA
1414 TCCTTGCAATGTTTAGATAAAGAAACAACATAATCAGATATATAAATGTTGACTGAAAATGCTCAGTTAGTAAGAATACAATGATAACGTCTACCTACGAATAACCCTCGCTAATCGGAT
1534 TTTTCCTCATAGAGACAACCACTCTTATAATAAGTAAACAATAGGTCTTTAAAGAGATAAAACTAAAAAAAAAAAAAATATATCAACCGGGATGCTACCTTCAGGAATTTATGCTGATTA
1654 GATGATCGTTCATTCTATGAGCTAAAGCCAAATTGAATTTCTTGATGTCAGATATCCAATTTGAAAATAGAACAGCACCCCTTACCCAGCGAGCAACTGTGTGGGTTAATGTTTGTGTCG
1774 GGGTGTGCAGGATTACAATTATCTATTGCAATGGAGCGAGGCGCCAGTTAAGCGACCACCTGCATTACATAGATTTGATATAAATCTTATTCACAACAATTACTTAAACAATAGTGGCAG
1894 CCACGTGCGCATCCATACTAACAAAGATGATTATTATATGCCACACCGAGGGTCTTCTCTGAGACTGAAGTATTATTTTTACCAGTGATTAATTCTCATAACTACTAAATCTTAAAGTGT
2014 CCAAGACGCATGTAAGATGCCGTACATTTAGCAGCTGACACCACTTACCGCTGGACTAGGGAAGGCATACAGGCTGTGAAAAGTGGCACTAGGATTATACTCCATTCGCCGATATCCAAT
2134 ATAACAACAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATGATAATAATAATAATAATAATAATAATAATAATAATGATAATAATAATAATAATAATAATAACA
2254 ATAATAACAATGATGATGATACTACTACTATGTTAGCATGTCCTTTATTTTCCCATACTCCTTTTCTGTCTGACGTCACGTCTTTCCCCTCAGTCCTCGCTCGCTGCCGCCCGAGGTTGA
2374 CACGTACGCCTCCTGCCTCTTGTTTGCTCGGTTTACACTTGTTGTGTAGTGTGCTACCGTAAATACACTTTCATAATACCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```




Female
Male
Ptdmrt
Control
Ptdmrt
Control
Ptdmrt
Control
Ptdmrt
Control


Relative expression of Ptdmrt
0
2
4
6
8
10
12
14
a
a
a
b
c
d
d
e
b
b
b
Mc
B
G
En
Ez
z1
z2
z3
z4
M
J