Supplementary Materials

Materials and Methods

Sampling, DNA extraction, and RNA extraction

A total of 11 wild adult *Charybdis japonica* females and 9 wild adult *Charybdis japonica* males (approximately 210 g) were collected in December 2019 from Zhoushan, China. All animal procedures were approved by the Animal Ethics Committee of Yantai University. Fresh gill, muscle, heart, and intestinal tissue samples were collected using sterilized dissection scissors and a scalpel, then snap-frozen in liquid nitrogen. High-quality genomic DNA was extracted from the muscle tissue using a Blood & Cell Culture DNA Mini Kit (Qiagen, Germany), and high-quality RNA was separately extracted from each tissue using a TRIzol Reagent Kit (Invitrogen, USA), then mixed equally. Genomic DNA integrity, purity, and concentration were assessed using 1% agarose gel electrophoresis, NanoDrop 2000 analyzer (Thermo Fisher Scientific, USA), and Qubit 3.0 analyzer (Thermo Fisher Scientific, USA). RNA integrity and concentration were measured using the Agilent Bioanalyzer 2100 System (Agilent Technologies, USA) and NanoDrop 2000 analyzer (Thermo Fisher Scientific, USA).

Library construction and sequencing

Three libraries (i.e., Illumina library, SMRTbell library, and Hi-C library) were combined to construct the high-quality *C. japonica* genome. A high-quality short-insert (300–350 bp) paired-end (PE) Illumina library was constructed in accordance with the standard Illumina protocols (Illumina, USA), then sequenced on the Illumina NovaSeq-6000 platform. A high-quality SMRTbell library (20 kb fragment size) was prepared using the SMRTbell Template Preparation Kit 1.0 (PacBio, USA) according to the manufacturer's protocols to obtain long reads for promoting genome assembly. The constructed SMRTbell library was added to one SMRT cell, then transferred to the PacBio Sequel II sequencing platform for long-read genomic sequencing. A high-quality Hi-C library was constructed to obtain the chromosome-level genome assembly, then sequenced using the Illumina NovaSeq-6000 platform. A high-quality 150 bp PE RNA-sequencing (RNA-seq) library was also constructed according to the standard Illumina protocols (Illumina, USA), then sequenced on the Illumina NovaSeq-6000 platform.

In total, 20 *C. japonica* individuals (11 females and 9 males) were used for whole-genome resequencing. High-quality short-insert (300–350 bp) PE Illumina libraries were constructed in accordance with standard protocols (Illumina, USA), then sequenced on the Illumina NovaSeq-6000 platform.

Data filtering

All raw Illumina reads were filtered by removing reads that included adapter sequences, duplicated sequences, unknown nucleotides greater than 10%, and low-quality bases (quality score≤5) greater than 50%. Hi-C reads that contained adapter sequences or sequences less than 50 bp in length were removed, with only PE Hi-C reads retained. Bases with a quality score of less than 20 at both ends of the reads were eliminated. RNA-seq reads were filtered by removing reads with sequencing adaptors, unknown nucleotides (N ratio>10%), or low quality (quality score≤5).

K-mer analysis of clean Illumina reads

The remaining clean Illumina reads were used to estimate the genomic characteristics of *C. japonica* before genome assembly. In the present study, *k*-mer-based analysis was used to estimate the size, heterozygosity, and repeat sequences of the *C. japonica* genome (Liu et al., 2013). Here, 17-mer was selected for *k*-mer analysis to ensure that enough *k*-mers (4^{17}) were produced to cover the entire *C. japonica* genome.

Genome assembly and evaluation

Wtdbg2 software (v1.0; Ruan & Li, 2020) was applied to assemble the *C. japonica*

genome with PacBio long-read sequencing, using the following parameters: best depth from input reads, 50.0; *k*-mer size, 21; readCutoff, 1k. Although PacBio long-read sequencing is reliable, a certain sequencing depth is required to ensure accuracy. Here, the PacBio long reads were first applied to polish the consensus sequence output from Wtdbg2. Specifically, pbmm2 (Chaisson & Tesler, 2012) and minimap2 (Li, 2018) were used to align the PacBio long reads to the consensus sequences, and the alignment results were then corrected using the Arrow and Racon methods (Walker et al., 2014). The clean Illumina reads were then compared with the abovementioned PacBio long read-based polished genome sequences using BWA software (v0.7.10-r789; Li & Durbin, 2009), then corrected using Pilon (v1.24; Walker et al., 2014). Finally, de-redundancy of the corrected *C. japonica* genome was performed according to the depth distribution and sequence similarity of the reads. The filtered Hi-C reads were mapped to the polished *C. japonica* genome to detect positional and directional errors in contigs during three-dimensional (3D) DNA assembly (Dudchenko et al., 2017). JuiceBox software (v1.4.3.2; Durand et al., 2016) was used to modify the order and direction of some contigs and to help in the determination of chromosome boundaries. Genomic overlap was identified based on sequence homology and long-distance interaction patterns. Finally, the chromosome-level *C. japonica* genome was obtained.

The *C. japonica* genome assembly was evaluated using three methods. First, the genome sequence was interrupted using a step length of 1 000 bp, with the interrupted sequences then compared with the nucleotide sequence (NT) database using BLAST to evaluate genome sequence accuracy. Second, BWA (v0.7.10-r789; Li & Durbin, 2009) and minimap2 (Li, 2018) were used to compare the Illumina short reads and PacBio long reads with the genome sequence, respectively. Read and genome sequence consistency was evaluated according to the comparison rate. In addition, conserved *C. japonica* gene completeness was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO v2.0; Simão et al., 2015) and the orthologous gene database of arthropods. To assess genomic integrity, RNA-seq reads were compared with the genome using HISAT2 (Vaser et al., 2017).

Genome repetitive elements, **coding genes, and non-coding RNA (ncRNA) annotation**

Annotation of the *C. japonica* genome was carried out using repeat recognition, ncRNA and gene structure prediction, and functional annotation.

First Tandem Repeat Finder (Benson, 1999) was applied to identify tandem repeats in the *C. japonica* genome. RepeatMasker and RepeatProteinMask (v4.1.0; http://www.repeatmasker.org) were used to annotate interspersed repeats (also known as transposon elements (TEs)) of the *C. japonica* genome based on the Repbase database (Jurka et al., 2005). RepeatMasker (Bedell et al., 2000) was used to compare the genome sequence to the repetitive element database acquired above to obtain a set of repetitive elements. The ultimate *C. japonica* genome repetitive elements were identified by removing the redundant repetitive elements obtained using the three methods.

Coding gene annotation includes structural prediction and functional annotation. Here, three prediction strategies, including homology, *ab initio*, and RNA-seq reads, were applied to predict the coding genes. *Eriocheir sinensis* (GCA_013436485.1), *Penaeus monodon* (GCA_015228065.1), *Penaeus vannamei* (GCA_003789085.1), and *Portunus trituberculatus* (GCA_017591435.1) were selected as they are closely related to *C. japonica*, and their protein sequences were downloaded from the National Center for Biotechnology Information (NCBI) database for the structural prediction of *C. japonica* coding genes. *Ab initio* coding gene prediction was performed using Augustus (v2.7; Stanke et al., 2006) and GenScan (v1.2; Burge & Karlin, 1997) with default settings. The filtered RNA-seq reads were mapped to the *C. japonica* genome for transcript assembly using TopHat (v2.0.0), and Cufflinks (v2.2.1) (Ghosh & Chan, 2016) was then used to predict the coding genes. MAKER2 was used to remove redundant coding genes predicted by the above methods, and the HiCESAP process

was applied to obtain more complete and accurate coding gene datasets. Predicted coding genes were then functionally annotated using the InterPro, Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG)_ALL, KEGG Orthology (KEGG_KO), Swiss-Prot, Translation of European Molecular Biology Laboratory nucleotide sequence (TrEMBL), Transcription Factor (TF), Pfam, Non-Redundant Protein Sequence (NR), and Eukaryotic Orthologous Groups (KOG) databases to determine the biological function and metabolic pathways involved in the coding genes.

NcRNAs, such as ribosomal RNA (rRNA), microRNA (miRNA), transfer RNA (tRNA), and small nuclear RNA (snRNA), do not translate proteins but have important biological functions. MiRNAs play a role in gene silencing and can degrade target genes or inhibit target gene translation into protein. Both tRNAs and rRNAs are directly involved in protein synthesis, while snRNAs are involved in the processing of RNA precursors and are the main components of RNA spliceosomes. The tRNAscan-SE (v1.3.1) program (Lowe & Eddy, 1997) was used to search for tRNA sequences in the *C. japonica* genome according to their structural characteristics. Considering the high conservation of rRNAs, BLASTN (Altschul et al., 1990) was used to search for rRNAs in the *C. japonica* genome based on the rRNA sequences of closely related species. Additionally, miRNAs and snRNAs were predicted using INFERNAL (v1.1) (Nawrocki, 2014).

Whole-genome resequencing analyses

The BWA (v0.7.10-r789) program (Li & Durbin, 2009) was used to compare high-quality whole-genome resequencing reads with the assembled *C. japonica* genome and reads with low-mapping efficiency were removed. The parameters were as follows: mem –M –t –K 10000000. Filtered reads in "SAM" format was sorted using Picard (v1.119; https://github.com/broadinstitute/picard) to remove polymerase chain reaction (PCR) duplications. Single-nucleotide polymorphisms (SNPs) and short insertions/deletions (indels) were called using a Bayesian approach implemented in SAMtools (v1.6; Li & Durbin, 2009). Finally, population differentiation index (F_{ST}) and genotype frequency were used to locate the sex-determining region in the *C. japonica* genome.

Comparative genomic analyses and testing for genomic selection

We performed an extensive orthologous gene comparison of *C. japonica* with eight other model species with genome datasets, including *Chionoecetes opilio* (GCA_016584305.1), *Hyalella azteca* (GCA_000764305.3), *Penaeus vannamei* (GCA_003789085.1), *Portunus trituberculatus* (GCA_017591435.1), *Eriocheir sinensis* (GCA_013436485.1), *Drosophila melanogaster* (GCA_003401745.1), *Amphibalanus amphitrite* (GCA_019059575.1), and *Daphnia magna* (GCA_003990815.1). We downloaded the protein sequences from the NCBI database. Subsequently, we extracted the orthologous groups using ORTHOMCL (v2.0.9) (Chen et al., 2006) and filtered the BLASTP results with default parameters. The single-copy orthologous genes shared by all nine species were further aligned using MUSCLE (v3.8.31) (Edgar, 2004), and conserved sequences were extracted from each concatenated nucleotide sequence using Gblocks (v0.91b) with the parameter −t=c. We performed 1 000 non-parametric bootstrap replicates for the optimal GTRGAMMA substitution model of all concatenated nucleotide sequences, and then constructed a phylogenetic tree of the nine species using RAxML (v8) (Stamatakis, 2014). Divergence times of the nine species were estimated using R8s (v1.7.1) software, and fossil evidence (divergence time between *D. magna* and *D. melanogaster* was 409.3–536.3 million years ago (Mya); divergence time between *H. azteca* and *A. amphitrite* was 418.7–459.0 Mya; divergence time between *P. vannamei* and *H. azteca* was 169.1–388.2 Mya; divergence time between *P. trituberculatus* and *E. sinensis* was 147.1–215.7 Mya) was used to calibrate divergence time. Furthermore, CAFE (v3.1) (De Bie et al., 2006) was applied to analyze the expansion and contraction of gene families, with $P<0.05$ used to indicate significant change. Gene enrichment analysis was

performed for the expanded and contracted gene families based on the GO and KEGG databases, respectively.

Desiccation-adaptive mechanisms of *C. japonica*

CAFE software (v3.1) (De Bie et al., 2006) was applied to analyze the expansion and contraction of gene families and thus explore the desiccation-adaptive mechanisms of *C. japonica*. Three desiccation-tolerant species (including *C. japonica*, *E. sinensis*, and *A. amphitrite*) were selected as foreground species, and multiple tree files were constructed for the nine species using all single-copy orthologous genes The branch-site model (model=2, Nsites=2) of the codeml program in PAML (v4.9) (Yang, 2007) was applied to estimate the non-synonymous/synonymous ratio (*w*) to determine positively selected genes (PSGs) in *C. japonica* and *A. amphitrite*. After Chi-square analysis, a gene was considered a PSG of *C. japonica* and *A. amphitrite* if the false discovery rate (FDR)-adjusted *P*-value was less than 0.01. Finally, gene enrichment analysis was performed for the expanded gene families and PSGs based on the GO and KEGG databases, respectively.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3): 403–410.

Bedell JA, Korf I, Gish W. 2000. *MaskerAid*: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**(11): 1040–1041.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**(2): 573–580.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**(1): 78–94.

Chaisson MJ, Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**: 238.

Chen F, Mackey AJ, Stoeckert Jr CJ, Roos DS. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, **34**(S1): D363–D368.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*, **22**(10): 1269–1271.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**(6333): 92–95.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, **3**(1): 99–101.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5): 1792–1797.

Ghosh S, Chan CKK. 2016. Analysis of RNA-Seq data using TopHat and cufflinks. *In*: Edwards D. Plant Bioinformatics. New York: Humana, 339–361.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, **110**(1–4): 462–467.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18): 3094–3100.

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

Liu BH, Shi YJ, Yuan JY, Hu XS, Zhang H, Li N, et al. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. arXiv*:* 1308.2012.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, **25**(5): 955–964.

Nawrocki EP. 2014. Annotating functional RNAs in genomes using infernal. *In*: Gorodkin J, Ruzzo WL. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. Totowa: Humana, 163–197.

Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, **17**(2): 155–158.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19): 3210–3212.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9): 1312–1313.

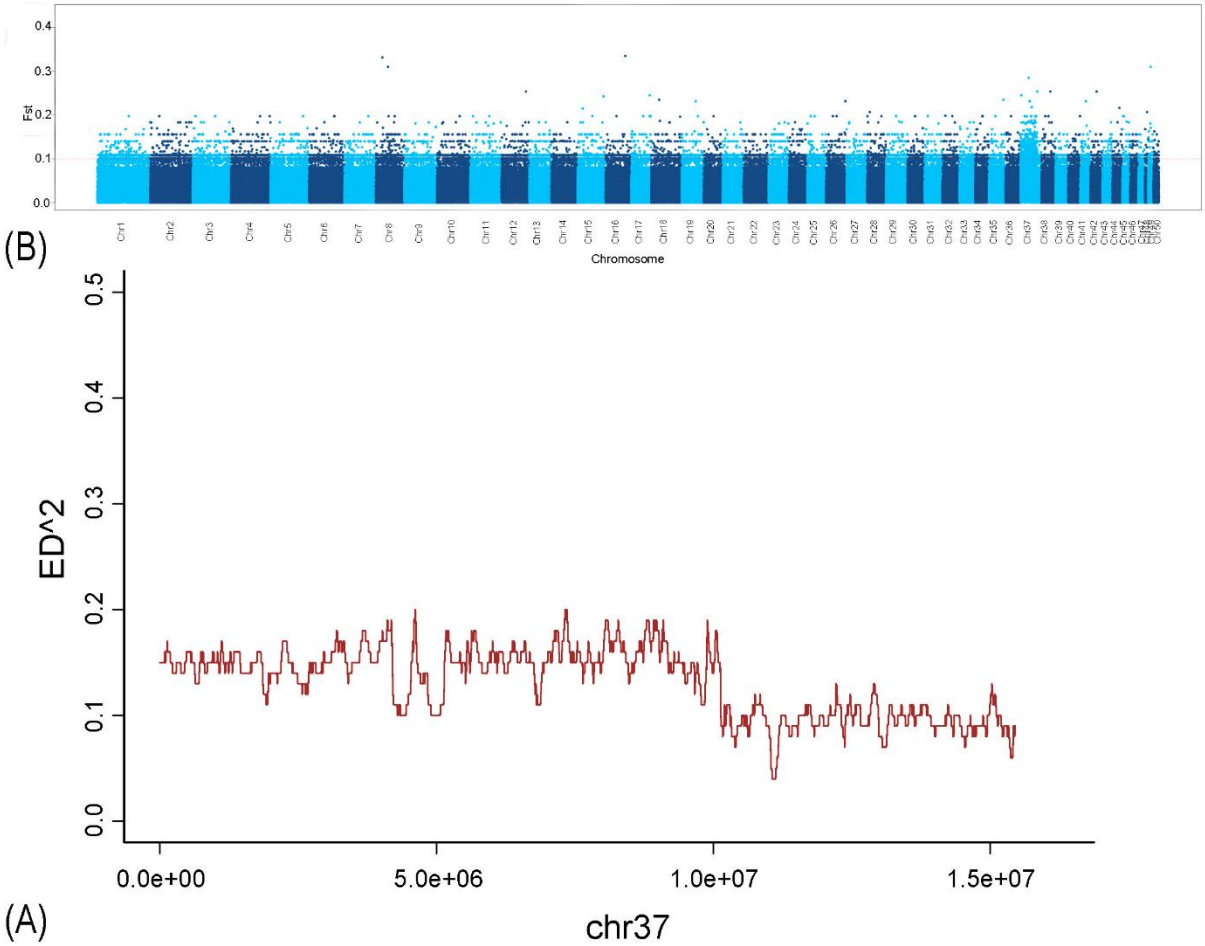
Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research*, **34**(S2): W435–W439.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, **27**(5): 737–746.

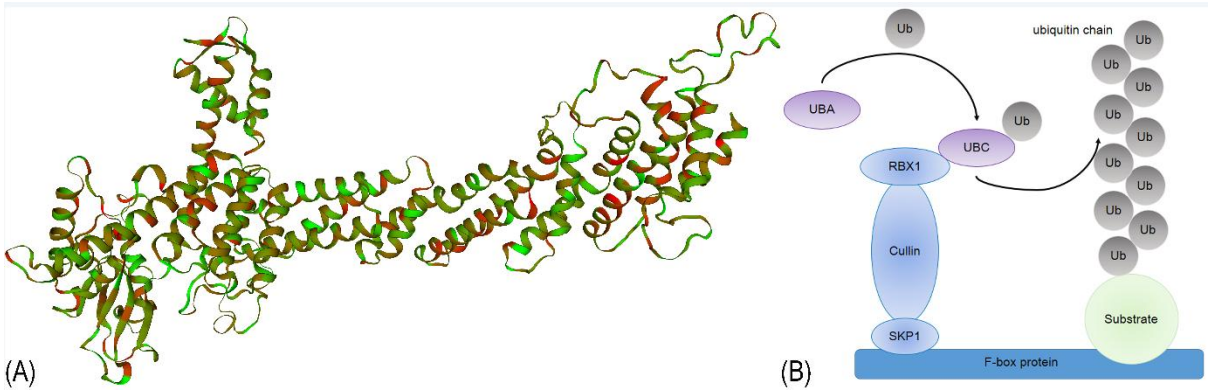
Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**(11): e112963.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8): 1586–1591.

Supplementary Figures



Supplementary Figure S1 (A). Manhattan plot of F_{ST}. (B). Euclidean distance $(ED)^2$ distribution.



Supplementary Figure S2 Overall structure (A) and schematic (B) of Skp-Cullin-F-box (SCF) ubiquitin ligase. Note: Ub: Ubiquitin; UBA: Ubiquitin-activating enzyme; UBC: Ubiquitin-conjugating enzyme; RBX1: E3 ubiquitin-protein ligase.

Supplementary Tables

Supplementary Table S1 Comparison results of Illumina short reads and PacBio long reads

Reads format	Mapping rate (%)	Average sequencing depth	Coverage (%)	Coverage at least 4× (%)	Coverage at least 10× (%)	Coverage at least 20× (%)
Short Illumina reads	94.11	87.07	130.33	95.11	89.56	85.81
Long PacBio reads	87.67	70.32	99.37	98.31	96.76	90.52

Supplementary Table S2 BUSCO assessment results

Type	Proteins	Percentage (%)
Complete BUSCOs (C)	921	86.40
Complete and single-copy BUSCOs (S)	877	82.27
Complete and duplicated BUSCOs (D)	44	4.13
Fragmented BUSCOs (F)	55	5.16
Missing BUSCOs (M)	90	8.44
Total BUSCO groups searched	1,066	100.00

Supplementary Table S3 Classification results of repeat elements

Type	RepBase TEs		TE proteins		*De novo*		Combined TEs	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA transposable element	222,547,342	15.57	5,550,462	0.39	112,219,133	7.85	318,621,782	22.29
Long interspersed nuclear element	133,138,998	9.31	109,275,692	7.64	129,540,697	9.06	236,309,645	16.53
Short interspersed nuclear element	5,915,427	0.41	0	0.00	2,656,934	0.19	8,240,344	0.58
Long terminal repeat	89,499,509	6.26	24,164,373	1.69	99,477,496	6.96	179,303,759	12.54
Satellite	70,788,128	4.95	0	0.00	1,257,831	0.09	71,793,940	5.02
Simple repeat	0	0.00	0	0.00	328,443	0.02	328,443	0.02
Other	89,220	0.01	933	0.00	0	0.00	90,153	0.01
Unknown	7,505,046	0.53	6,318	0.00	193,983,018	13.57	201,152,647	14.07
Total	378,962,007	26.51	139,024,401	9.73	512,974,229	35.89	824,020,605	57.65

Supplementary Table S4 Coding gene prediction results

Gene set	Protein coding gene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo/Genscan	65,791	10,108	1,319	4.10	321.74	2,836
De novo/Augustus	112,067	5,711	866.06	3.97	218.10	1,631
Homo/*E. sinensis*	84,409	5,187	463.97	1.83	252.90	5,658
Homo/*P. monodon*	130,995	6,568	698.14	2.18	319.64	4,957
Homo/*P. vanamei*	153,053	6,949	673.77	2.07	326.06	5,885
Homo/*P. trituberrculatus*	646,952	3,574	375.90	1.48	253.66	6,636
Tans.orf/RNA	7,327	21,251	1,339	7.02	449.02	3,005
BUSCO	1,020	14,665	1,475	9.84	149.94	1,492

MAKER2	52,283	9,669	1,179	4.91	278.62	2,125
HiCESAP	30,900	11,027	1,386	5.12	341.12	2,255

Supplementary Table S5 Annotation information of ncRNAs

Type		Copy	Average length (bp)	Total length (bp)	% in genome
miRNA		474	121	57,561	0.004027
tRNA		15,570	73	1,135,923	0.079470
rRNA	18S	8	1,571	12,566	0.000879
	28S	7	142	997	0.000070
	5.8S	57	152	8,678	0.000607
	5S	237	116	27,554	0.001928
	Total	15,570	73	1,135,923	0.079470
snRNA	CD-box	24	127	3,059	0.000214
	HACA-box	34	308	10,477	0.000733
	Splicing	98	147	14,441	0.001010
	scaRNA	1	133	133	0.000009
	Total	157	179	28,110	0.001967

Supplementary Table S6 SNP statistical results

Type	Number
Downstream	96,328
Exonic | nonsynonymous SNV	8,606
Exonic | stopgain	300
Exonic | stoploss	13
Exonic | synonymous SNV	10,885
Exonic | unknown	14,520
Intergenic	1,511,526
Upstream	98,572
Upstream; Downstream	70,864

Supplementary Table S7 Indel statistical results

Type	Number
Downstream	92,524
Exonic | frameshift deletion	2,295
Exonic | frameshift insertion	1,762
Exonic | nonframeshift deletion	1,240
Exonic | nonframeshift insertion	694
Exonic | stopgain	108
Exonic | stoploss	8
Exonic | unknown	4,736
Intergenic	1,405,288
Upstream	92,190
Upstream; Downstream	71,136
UTR5	14

Supplementary Table S8 Gene family clustering results

Species	Genes number	Unclustered genes	Genes in families	Family number	Unique families	Unique families genes	Common families	Common families genes	Single copy genes	Average genes per family
C. japonica	30,900	7,546	23,354	12,963	832	3,271	2,593	4,047	340	1.802
E. sinensis	28,033	8,928	19,105	11,627	1,140	3,492	2,593	3,908	340	1.643
P. trituberculatus	17,292	2,156	15,136	12,023	196	669	2,593	3,253	340	1.259
A. amphitrite	27,357	1,831	25,526	9,984	2,472	7,439	2,593	5,947	340	2.557
C. opilio	21,739	3,637	18,102	8,336	627	4,165	2,593	3,229	340	2.172
D. magna	16,891	2,290	14,601	7,803	794	3,922	2,593	3,498	340	1.871
D. melanogaster	13,968	4,481	9,487	6,794	550	1,868	2,593	3,200	340	1.396
H. azteca	18,608	4,612	13,996	9,964	509	2,014	2,593	3,374	340	1.405
P. vannamei	24,974	6,423	18,551	11,376	682	2,603	2,593	3,828	340	1.631