GigaScience

Chromosome-level genome assembly reveals adaptive evolution of the swimming crab (Portunus trituberculatus)

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00309
Full Title:	Chromosome-level genome assembly reveals adaptive evolution of the swimming crab (Portunus trituberculatus)
Article Type:	Data Note
Funding Information:	National Natural Science Foundation of China (31672267) — Dr Boping Tang National Natural Science Foundation of China (31640074) — Mr Qiuning Liu Jiangsu Agricultural Science and Technology Innovation Fund (CX(18)3027) — Dr Boping Tang Jiangsu Agricultural Science and Technology Innovation Fund (CX(18)2027) — Dr Boping Tang Natural Science Foundation of Jiangsu Province (BK20171276) — Dr Daizhen Zhang Natural Science Foundation of Jiangsu Province (BK20160444) — Mr Qiuning Liu Qinglan Project of Jiangsu Province of China (CN) (2018M642105) — Dr Daizhen Zhang
Abstract:	Background: The swimming crab, Portunus trituberculatus , is an important commercial species in China and is widely distributed in the coastal waters of Asia-Pacific countries. Despite increasing interest in swimming crab research, genomic information, including whole-genome sequencing, is lacking, with only limited transcriptome data currently available. Findings: Here, we assembled the first chromosome-level reference genome of P. trituberculatus by combining the short reads, Nanopore long reads, and Hi-C data. The genome assembly size was 1.00 Gb with a contig N50 length of 4.12 Mb. In addition, BUSCO assessment indicated that 94.7% of core eukaryotic genes were present in the genome assembly. Approximately 54.52% of the genome was identified as repetitive sequences, with a total of 16,796 annotated protein-coding genes. In addition, we anchored contigs into chromosomes and identified 50 chromosomes with a N50 length of 21.80 Mb by Hi-C technology. Conclusions: We anticipate that this chromosome-level assembly of the P. trituberculatus genome will not only promote study of basic development and evolution but also provide important resources for swimming crab reproduction.
Corresponding Author:	Yandong Ren Northwestern Polytechnical University Xi'an, Shaanxi CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Northwestern Polytechnical University
Corresponding Author's Secondary Institution:	
First Author:	Yandong Ren
First Author Secondary Information:	
Order of Authors:	Yandong Ren

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	Boping Tang
	Daizhen Zhang
	Haorong Li
	Senhao Jiang
	Fujun Xuan
	Baoming Ge
	Zhengfei Wang
	Yu Liu
	Zhongli Sha
	Yongxu Cheng
	Wei Jiang
	Hui Jiang
	Zhongkai Wang
	Kun Wang
	Chaofeng Li
	Yue Sun
	Shusheng She
	Qiang Qiu
	Wen Wang
	Xinzheng Li
	Yongxin Li
	Qiuning Liu
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources	Yes

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

1 **Chromosome-level genome assembly reveals adaptive evolution of the swimming crab**
2 **(*Portunus trituberculatus*)**
3

4 Boping Tang[1,*], Daizhen Zhang[1,*], Haorong Li[2,*], Senhao Jiang[1*], Huabin Zhang[1], Fujun
5 Xuan[1], Baoming Ge[1], Zhengfei Wang[1], Yu Liu[1], Zhongli Sha[3], Yongxu Cheng[5], Wei Jiang[3],
6 Hui Jiang[4,6], Zhongkai Wang[2], Kun Wang[2], Chaofeng Li[1], Yue Sun[1], Shusheng She[7], Qiang
7 Qiu[2], Wen Wang[2], Xinzheng Li[3†], Yongxin Li[2†], Qiuning Liu[1†], Yandong Ren[2†]
8

9 1. Jiangsu Key Laboratory for Bioresources of Saline Soils, Jiangsu Provincial Key
10 Laboratory of Coastal Wetland Bioresources and Environmental Protection, Jiangsu Synthetic
11 Innovation Center for Coastal Bio-agriculture, Yancheng Teachers University, Yancheng
12 224002, China
13 2. Center for Ecological and Environmental Sciences, Northwestern Polytechnical University,
14 Xi'an 710072, China
15 3. Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China
16 4. National Engineering Laboratory of Marine Germplasm Resources Exploration and
17 Utilization, Zhejiang Ocean University, Zhoushan 316022, China
18 5. Key Laboratory of Freshwater Aquatic Genetic Resources, Ministry of Agriculture and
19 Rural Affairs, Shanghai Ocean University, Shanghai 201306, China
20 6. National Engineering Research Center for Facilitated Marine Aquaculture, Zhejiang Ocean
21 University, Zhoushan 316022, China
22 7. China Hong Kong Ecology Consultant Company, Hong Kong, China
23

24 *These authors contributed equally to this work.
25 †Corresponding authors: Yandong Ren (renyandong90@126.com); Qiuning Liu
26 (liuqiuning@163.com); Yongxin Li (yxli28science@sina.com); Xinzheng Li
27 (lixzh@ms.qdio.ac.cn)
28
29

30 **Abstract**

31 **Background:** The swimming crab, *Portunus trituberculatus*, is an important commercial
32 species in China and is widely distributed in the coastal waters of Asia-Pacific countries.
33 Despite increasing interest in swimming crab research, genomic information, including
34 whole-genome sequencing, is lacking, with only limited transcriptome data currently
35 available. **Findings:** Here, we assembled the first chromosome-level reference genome of *P.*
36 *trituberculatus* by combining the short reads, Nanopore long reads, and Hi-C data. The
37 genome assembly size was 1.00 Gb with a contig N50 length of 4.12 Mb. In addition,
38 BUSCO assessment indicated that 94.7% of core eukaryotic genes were present in the
39 genome assembly. Approximately 54.52% of the genome was identified as repetitive
40 sequences, with a total of 16,796 annotated protein-coding genes. In addition, we anchored
41 contigs into chromosomes and identified 50 chromosomes with a N50 length of 21.80 Mb by
42 Hi-C technology. **Conclusions:** We anticipate that this chromosome-level assembly of the *P.*
43 *trituberculatus* genome will not only promote study of basic development and evolution but
44 also provide important resources for swimming crab reproduction.

45

46 **Keywords:** *Portunus trituberculatus*; genome assembly; crab; chromosome; evolution

47

48 **Introduction**

49 The swimming crab, *Portunus trituberculatus* (NCBI: txid210409), belonging to Brachyura,
50 Portunidae, Portunus, is named for its shuttle-shaped head breastplate and three verrucous
51 bumps on the back of the stomach and heart regions [1, 2]. The chelipeds of swimming crabs
52 are well developed for feeding and attacking, with the first three pairs and last pair used for
53 crawling and swimming, respectively [3, 4]. Male and female crabs are distinguished by their
54 type of abdomen, with the male having a triangular abdomen and the female having an almost
55 circular one [5]. Due to their lack of drilling ability, swimming crabs often live in soft mud or
56 sand [6] or in water grass near the shore, and also show a certain level of phototaxis, spending
57 time on the sea floor during the day and foraging at night [5]. Swimming crabs are also
58 omnivorous, feeding on shellfish, small fish, shrimp, algae, and decomposing animal and

plant carcasses [7].

The swimming crab is widely distributed in the coastal waters of Korea, Japan, China, and Southeast Asia and is one of the most valuable marine crustaceans in Asia (http://species-identification.org/species.php?species_group=crabs_of_japan&menuentry=soorten&id=1106&tab=beschrijving). It is widely found in Chinese coastal waters of the Bohai Sea, Yellow Sea, East China Sea, and South China Sea and is an important commercially cultured species [8]. Swimming crabs are considered highly nutritious, especially in regard to crab cream, and are very popular in China [9, 10]. As a result, the crab has been heavily overfished, resulting in substantial declines in its natural population [11] and initiation of artificial breeding [12, 13]. With continued research on the crab, it has become clear that morphological, physiological, but the genetic changes are poorly understood. At present, genomic research on the swimming crab has only been conducted at the transcriptome level [14-16], with the whole genome not yet described.

In the present study, we constructed a chromosome-level genome assembly of *P. trituberculatus* by combining short reads, Nanopore long reads, and Hi-C sequencing data. This chromosome-level genome will not only promote study on development and evolution, but also provide important resources for reproductive studies of *P. trituberculatus* and other crab species.

Sampling, library construction, and sequencing

A male swimming crab was collected in Bohai Bay, Hebei Province, China, for sequencing (Figure 1). To obtain sufficient high-quality DNA for the Oxford Nanopore (Oxford, UK) and BGISEQ-500 platforms (BGI, China), the swimming crab was rinsed five times with clean water and dissected immediately. Fresh muscle tissue was collected and snap-frozen in liquid nitrogen. The samples were then used to extract DNA with a Qiagen Blood & Cell Culture DNA Mini Kit and prepared for Nanopore, BGISEQ-500, and Hi-C sequencing. Muscle RNA was also extracted using TRIzol (Invitrogen) according to the manufacturer's instructions. To obtain an overview of the transcriptome, polyadenylated RNA was chosen by oligo (dT) purification and reverse-transcribed to cDNA and sequenced using the BGISEQ-500 platform.

Extracted DNA was sequenced using both the BGISEQ and Oxford Nanopore platforms. The short reads generated from the BGISEQ platform were used for estimation of genome size and error correction of the assembled genome, and the Nanopore long reads were used for genome assembly. To this end, one library with insertion lengths of ~300 bp was sequenced on the BGISEQ-500 platform, and another library with an average length of 20 kb was constructed using the Oxford Nanopore platform according to the manufacturers' protocols.

Data filtering

Three different sources of reads were used to achieve the high-quality genome assembly, i.e., Nanopore long reads, short reads, and Hi-C reads. Thus, we used different methods for filtering. For the Nanopore long reads, any reads less than 1 kb or with a mean quality value of < 7 were removed. For the short reads, any reads with more than 10% unknown reads or low-quality bases more than 50% along with its paired-end read were removed. All adaptor sequences and duplicated reads produced by polymerase chain reaction (PCR) were removed. The low-quality Hi-C reads were filtered using HiC-Pro v2.10.0 [17].

Genome characteristic estimation

All filtered BGISEQ short reads were used for estimation of genome size and other characteristics. In addition, 17-mer was chosen for k-mer analysis and the 17-mer depth frequency distribution was calculated using the k-mer method. Genome size was estimated as: Genome size = TKN17-mer / PKFD17-mer, where TKN17-mer is the total k-mer number and PKFD17-mer is the peak k-mer frequency depth of 17-mer. The estimated genome size was used to determine subsequent genome assembly results.

Genome assembly

To improve the quality of the genome and reduce the error ratio, self-error correction of all Nanopore long reads was performed using NextDenovo software (https://github.com/Nextomics/NextDenovo). The error-corrected Nanopore long reads were then used to assemble the raw genome via contig construction with WTDBG software [18]

117 and parameters: -p 0 -k 15 -AS 2 -E 1 -s 0.05 -L 5000. The assembled genomic sequences
118 were further polished by Racon v1.2.1 [19] with four iterations using the error-corrected
119 Nanopore long reads. After this, all filtered BGISEQ short reads were polished by Pilon v1.21
120 [20] at the single-base level. After completion of the error-correction steps, the Hi-C data
121 were used to obtain a chromosome-level genome assembly. All Hi-C sequencing data were
122 first filtered by Hic-Pro v2.10.0 [17] and then mapped to the polished swimming crab genome
123 to improve the connection integrity of the contigs. Finally, 3D *de novo* assembly software
124 (v180419) [21] with default parameters was used to determine contig location and direction.
125
126 **Genome assembly evaluation**
127 Three different strategies were used to evaluate the completeness and accuracy of the
128 assembled genome. First, the quality of the assembled genome and gene completeness were
129 assessed using BUSCO [22] with the core gene sets of the eukaryote and metazoan databases,
130 respectively. Second, all filtered short reads generated by BGISEQ were mapped to the
131 assembled genome using BWA-MEM v0.7.12 [23] to detect genome integrity. Third,
132 transcripts were mapped to the assembled genome using BLAT software [24].
133
134 **Repetitive element annotation**
135 Tandem repeats and transposable elements (TEs) were also annotated in the
136 chromosome-level genome. Tandem repeats were annotated using Tandem Repeat Finder
137 v4.04 [25] with default parameters. The TEs were annotated at the protein level using
138 RepeatProteinMask (RM-BLASTX) to search the protein database and at the DNA level
139 using RepeatMasker (open-4.0.7) [26] to search the *de novo* libraries and repbase. The *de*
140 *novo*-repeat libraries were constructed using RepeatModeler
141 (http://www.repeatmasker.org/RepeatModeler/), with consensus sequences used for *de novo*
142 library construction.
143
144 **Gene structure prediction and function annotation**
145 After repetitive element annotation, the repeat-masked genome was used for gene set

146 annotation with three different methods, i.e., *de novo* prediction, RNA-seq-based annotation,
147 and homology-based annotation. We first assembled the RNA-seq reads into transcripts using
148 Bridger r2014-12-01 [27]. The assembled genome and transcripts were then used for
149 Augustus training to obtain an accurate Augustus annotation species model. Augustus v2.5.5
150 [28] was used for *de novo* prediction of coding genes with the previous training results.
151 Second, proteins of *Bicyclus anynana* (GCF_900239965.1) [29], *Bombus terrestris*
152 (GCF_000214255.1) [30], *Drosophila melanogaster* (GCA_000001215.4) [31], *Mus*
153 *musculus* (GCF_000001635.26) [32], *Stegodyphus mimosarum* (GCA_000611955.2),
154 *Penaeus vannamei* (GCA_003789085.1), *Mesobuthus martensii* (downloaded from:
155 http://lifecenter.sgst.cn/main/en/scorpion.jsp.) [33], *Eriocheir japonica sinensis* (i.e.,
156 *Eriocheir sinensis*) (GigaDB:100186) [34-39], and *Tachypleus tridentatus*
157 (GCA_004102145.1) [40] were downloaded from the NCBI, GigaDB, or their own databases.
158 The longest transcript of each gene was selected for further annotation and phylogenetic
159 analysis. All filtered genes were searched with an e-value cutoff of 1e-5, with the *blast* results
160 then formatted and prepared for Genewise [41] prediction of the gene structure of the
161 swimming crab genome. Third, for the RNA-seq-based method, all assembled transcripts
162 were aligned against the genome using BLAT [24] (identity >90% and coverage >90%), with
163 PASA used to filter overlaps to link the spliced alignments. Finally, EvidenceModeler (EVM)
164 v1.1.1 was used to integrate the above data into an EVM-derived gene set [42].
165 Five different public protein databases were used for gene functional annotation, with
166 InterProScan v4.8 [43] used to screen proteins against the five databases (Pfam, release 27.0,
167 PRINTS, release 42.0, PROSITE, release 20.97, ProDom, 2006.1, and SMART, release 6.2)
168 to determine the number of InterPro and GO predicted protein-coding genes. In addition, the
169 Kyoto Encyclopedia of Genes and Genomes, UniProt/SwissProt, and UniProt/TrEMBL
170 databases were also used for functional annotation with BLAST v2.3.0 [44].
171
172 **Identification of orthologous genes**
173 The annotated genes in the swimming crab and six other species, including *Aedes aegypti*
174 (GCF_002204515.2), *B.anynana*, *D. melanogaster*, *S. mimosarum*, *P.vannamei*, and *E. j.*

175 *sinensis*, were used for orthologous gene identification with OrthoMCL v2.0.9 [45]. The
176 identified genes were then used to run reciprocal alignment and pairwise relationship analysis.
177 The reciprocal best similarity pairs in different species were considered as putative
178 orthologous genes and reciprocal better similarity pairs in one species were considered as
179 paralogous genes. The 1:1:1:1:1:1:1 single-copy genes in the seven species were also
180 identified for further phylogenetic and divergence time estimation analysis.
181
182 **Phylogenetic analysis and divergence time estimation**
183 Using the single-copy genes of the seven species (*P. trituberculatus*, *A. aegypti*, *B. anynana*,
184 *D. melanogaster*, *S. mimosarum*, *P. vannamei*, and *E. j. sinensis*), we connected the genes in
185 each species into one super-gene for phylogenetic tree building. Maximum likelihood-based
186 phylogenetic analysis was conducted using RAxML v8.2.10 [46] with default parameters. The
187 MCMCTREE program in the PAML package v4.8 [47] was then used to calculate divergence
188 time, with all fossil records downloaded from the TIMETREE website
189 (http://www.timetree.org) for calibration.
190
191 **Relative evolution rate**
192 The relative evolution rate of species was analyzed with LINTRE software [48] using the *tpcv*
193 model and *S .mimosarum* as an outgroup. We then evaluated the relative evolution rate
194 between the swimming crab and other related species.
195
196 **Gene family expansion and contraction**
197 Using the divergence time results calculated by MCMCTREE and the gene pairwise
198 relationships calculated by OrthoMCL [45], we determined gene family expansion and
199 contraction for each node using CAFÉ v3.1 [49]. The expansion and contraction genes of the
200 swimming crab were extracted for GO/KEGG enrichment analysis [50, 51].
201
202 **Results**
203 **Chromosome level genome assembly**

204 To obtain a high-quality chromosome-level swimming crab genome, we extracted
205 high-quality DNA from the muscle tissue and constructed libraries for genome sequencing. To
206 estimate the genome characteristics of the swimming crab, we generated 205.40 Gb of
207 BGISEQ data (Additional File: Table S1), with 17-mer analysis indicating a genome size of
208 ~918.52Mb (Additional File: Figure S1). In total, we generated 54.97 Gb (54.75-fold
209 coverage) of Nanopore long read data with N50 over 20kb (Additional File: Table S2). The
210 Nanopore long reads were assembled into contigs using WTDBG software [18] (genome size:
211 1.00 Gb; N50: 4.12 Mb) (Table 1). To further improve genome accuracy, we aligned all
212 corrected Nanopore long reads to the assembled genome and conducted error-correction using
213 Racon [19] with four iterations. The genome was subsequently corrected using all filtered
214 BGISEQ clean reads via Pilon [20] with two iterations. We then constructed the
215 chromosome-level genome with 95.95 Gb of Hi-C sequencing data (Additional File: Table S3)
216 by 3D *de novo* assembly [21]. Finally, we obtained 50 chromosomes and a mounting rate of
217 97.80% (Figure 2; Additional File: Table S4), which is the first chromosome-level crab
218 genome with N50 of 21.79 Mb (Table 1). The high mounting rate suggested successful
219 assembly of the swimming crab genome at the chromosome level.
220

221 **Genome quality evaluation**

222 We next assessed the completeness of the swimming crab genome by BUSCO [22] and
223 identified 94.7% Eukaryota and 92.9% Metazoa conserved core genes in the genome (Table
224 2). We checked the mapping rates of the BGISEQ short reads to our genome and found that
225 95.85% of reads were properly pair-mapped to the genome (Additional File: Table S5). We
226 then *de novo* assembled the transcripts using the RNA-seq data (Additional File: Table S6)
227 with Bridger software [27] and a N50 length of 2,124 bp (Additional File: Table S7). After
228 transcript mapping, we found that 97.80% of the transcripts could be mapped to the
229 swimming crab genome (Additional File: Table S8). We also analyzed the genome quality of
230 previously published high-quality genomes from closely related species and determined that
231 the quality of the assembled chromosome-level swimming crab genome was markedly higher
232 or comparable with that of other species (Additional File: Table S9). In summary, these results

233 indicated that we acquired a high-quality swimming crab genome. To investigate genome
234 characteristics, such as GC content, we analyzed the GC distribution in the genome with a
235 slide-window method. The peak value of GC content was ~41%, which agrees with the
236 average GC content in the swimming crab genome. We also found that the GC content in the
237 swimming crab was closer to that of mouse than of shrimp (Additional File: Figure S2).
238

239 **Genome annotation**

240 The repetitive sequences of the swimming crab genome were identified through four different
241 methods, resulting in 547.39 Mb of repeated sequences and accounting for 54.52% of the
242 assembled genome (Additional File: Table S10). Among the repeated sequences, 19.28%
243 (~193.56 Mb) were tandem repeats and 52.29% (~525.49 Mb) were TEs (Additional File:
244 Table S10; Table 3). The TEs could be further divided into four main types, including 0.014%
245 (~142.88kb) of short interspersed elements (SINE), 15.23% (~153.03 Mb) of long
246 interspersed elements (LINE), 14.90% (~149.71 Mb) of DNA elements, and 4.50% (~45.19
247 Mb) of long terminal repeats (LTR) (Table 3).

248 After masking the repeated sequences, we annotated the protein-coding genes using *de novo*
249 prediction, homology-based prediction, and transcript-based prediction. We merged the
250 results and obtained 16,791 protein-coding genes. We checked the quality of the annotated
251 genes by comparing with several closely related species. Results showed that the mRNA,
252 CDS, exon, intron length distributions of the swimming crab were similar to those of the
253 closely related species, suggesting that the swimming crab annotation results were dependable
254 (Figure 3).

255 We also performed functional annotation of the 16,791 genes with InterPro, GO, KEGG,
256 SwissProt, and TrEMBL. The highest annotation rate (74.77%) was found for SwissProt, in
257 which 12,558 genes were annotated. In total, 16,053 genes (~95.58%) were annotated,
258 indicating that most genes could be found in the public protein databases (Table 4). Thus,
259 taken together, we acquired a high-quality protein-coding gene set for the swimming crab.
260

261 **Orthologous identification and gene family analysis**

262 For comparative genomics analysis of the swimming crab, we analyzed the orthologous gene
263 relationships among several species, including *A. aegypti*, *B. anynana*, *D. melanogaster*, *S.*
264 *mimosarum*, *P. vannamei*, and *E. j. sinensis* using OrthoMCL. In total, 15,503 gene families
265 were clustered in the seven species and 1,018 one-to-one single-copy genes were identified
266 (Figure 4A). Because the swimming crab has several unique characteristics, we employed
267 gene family analysis and found 8,832 gene families shared among the seven species, with 328
268 gene families unique to the swimming crab (Figure 4B). We then employed functional
269 analysis and identified 34 enriched KEGG terms. Among them, the Notch signaling pathway
270 (*Q*-value = 0.003579101), vascular smooth muscle contraction (*Q*-value = 0.025737431), and
271 retinol metabolism (*Q*-value = 0.033319138) were significantly enriched (Additional File:
272 Table S11). As these all play important roles in development and physiological processes,
273 genes in these families may have key functions in the swimming crab.
274

275 **Phylogenetic relationships and divergence time**

276 Although the phylogenetic relationships of the swimming crab and closely related species
277 have been analyzed in previous studies, most used few nuclear and mitochondrial genes. To
278 determine the evolutionary relationship of the swimming crab, we analyzed all single-copy
279 genes using RAxML software [46], with the spider used as the outgroup species. Results
280 showed that the swimming crab has a close relationship with the Chinese mitten crab and
281 shrimp (Figure 5A). The seven species of pancrustaceans—*P. trituberculatus*, *A. aegypti*, *B.*
282 *anynana*, *D. melanogaster*, *S. mimosarum*, *P. vannamei*, and *E. j. sinensis*—formed two clades:
283 i.e., Hexapoda and Crustacea. The Hexapoda group consisted of all lepidopteran and
284 dipterous insects, whereas the second clade comprised all other crustaceans, with *P.*
285 *trituberculatus* and *E. j. sinensis* forming a Pleocyemata clade, followed by Dendrobranchiata
286 shrimp (*P. vannamei*). In addition, Hexapoda and Crustacea were both found to be
287 monophyletic (Figure 5A). To determine divergence time, we employed MCMCTREE
288 analysis in the PAML package [47] and found that the Chinese mitten crab and swimming
289 crab diverged ~183.5 million years ago (Mya), and diverged from shrimp ~428.5 Mya (Figure
290 5A).

291

292 **Relative evolution rate**

293 Species in different environments can experience different survival pressures. As such, we
294 conducted relative evolution rate analysis in LINTRE [48], with spider as the outgroup
295 species and swimming crab as the reference species. Results showed that the shrimp had the
296 slowest evolution rate among the seven species, whereas the fruit fly and butterfly exhibited
297 relatively fast evolution rates, suggesting greater survival pressures on these two species
298 (Figure 5B; Additional File: Table S12). Interestingly, the slowest evolution rates were found
299 among the Malacostraca (Figure 5B; Additional File: Table S12), suggesting that these
300 species experience relatively low survival pressure in their habitats.

301

302 **Gene family expansion and contraction**

303 We performed gene family expansion and contraction analysis of the seven species using
304 CAFÉ v4.0, and identified 148 and 25 expanded and contracted gene families ($P< 0.05$) in the
305 swimming crab, respectively. We then employed KEGG functional enrichment analysis of the
306 expanded gene families and found that the HIF-1 signaling pathway (*Q*-value = 0.000109025),
307 focal adhesion (*Q*-value = 0.000135977), Hippo signaling pathway (*Q*-value = 0.000184649),
308 and insulin signaling pathway (*Q*-value = 0.000357592) were enriched (Additional File: Table
309 S13). These biological processes are related to early development, hypoxia adaptation, and
310 other key processes, suggesting important functions of these genes in the development and
311 environmental adaptation of the unique body plan of the swimming crab.

312

313 **Conclusions**

314 Based on BGISEQ, Nanopore, and Hi-C sequencing data, we assembled a chromosome-level
315 high-quality genome of the swimming crab. Evaluation results indicated that the genome
316 quality of swimming crab was comparable with that of most high-quality model species. We
317 also successfully obtained 16,791 high-quality protein-coding genes by integrating three
318 different methods. The genome and annotation data will help researchers better understand
319 the evolution of crabs and improve their economic value. The phylogenetic results indicated

320 that the swimming crab is closely related to the Chinese mitten crab, from which it diverged
321 ~183.5 Mya. The unique and/or expanded gene family analysis provides clues to swimming
322 crab development and environmental adaptation.
323
324 **Availability of supporting data**
325 The raw sequencing data were deposited in the NCBI database under accession number
326 PRJNA555262. The genome assembly and annotation results are available via the
327 GigaScience repository GigaDB.
328
329 **Additional files**
330 Table S1: Statistics on genome sequencing data from BGISEQ platform.
331 Table S2: Statistics on sequencing reads from Oxford Nanopore platform.
332 Table S3: Statistics on Hi-C sequencing data.
333 Table S4: Statistics on assembled chromosome-level genome by 3D *de novo* assembly
334 software.
335 Table S5: Statistics on mapping ratio of the BGISEQ short reads to swimming crab genome.
336 Table S6: Statistics on RNA-seq data.
337 Table S7: Statistics on assembled transcripts by Bridger software.
338 Table S8: Statistics on transcript mapping ratio of swimming crab genome.
339 Table S9: Genome quality comparison of swimming crab with other species.
340 Table S10: Statistics on annotated repetitive sequences using different software.
341 Table S11: KEGG enrichment analysis of unique gene families in swimming crab relative to
342 six other species.
343 Table S12: Two-cluster analysis of swimming crab and other species.
344 Table S13: KEGG enrichment analysis of expanded gene families in swimming crab.
345 Figure S1: 17-mer analysis of swimming crab genome.
346 Figure S2: GC distribution in species.
347
348 **Abbreviations**

349 Hi-C: High-throughput chromosome conformation capture; BUSCO: Benchmarking
350 Universal Single-Copy Orthologs; CDS: Coding DNA Sequence; DNA: Deoxyribonucleic
351 Acid; RNA: Ribonucleic Acid; RNA-seq: RNA sequencing; BLAST: Basic Local Alignment
352 Search Tool; KEGG: Kyoto Encyclopedia of Genes and Genomes; NCBI: National Center for
353 Biotechnology Information.
354
355 **Conflicts of interest**
356 The authors declare that they have no competing interests.
357
358 **Funding**
359 This study was supported by the National Natural Science Foundation of China (31672267,
360 31640074), Jiangsu Agriculture Science and Technology Innovation Fund (CX(18)3027,
361 CX(18)2027), Natural Science Foundation of Jiangsu Province (BK20171276, BK20160444),
362 "Qing Lan Project" of Daizhen Zhang and China Postdoctoral Science Foundation
363 (2018M642105).
364
365 **Author contributions**
366 Y.R., Q.L., Yongxin L., and X.L. conceived the project. B.T., D.Z., S.S., H.Z., Yu L., and S.J.
367 collected and dissected the samples. H.L., Zhongkai W., K.W., Y.S., Q.Q., C.L., and Yongxin
368 L. estimated genome size. F.X., Y.C., W.J., and H.J. assembled the genome. B.G., Zhengfei W.,
369 Z.S., and B.T. performed genome assembly, genome annotation, and evolution analysis. Y.L.,
370 B.T., Q.Q., and W.W. wrote the manuscript. Y.R. and W.W. revised the manuscript.
371
372 **REFERENCES**
373 1. Spiridonov VA, Neretina TV and Schepetov D. Morphological characterization and
374 molecular phylogeny of *Portunoidea Rafinesque*, 1815 (Crustacea Brachyura):
375 Implications for understanding evolution of swimming capacity and revision of the
376 family-level classification. Zoologischer Anzeiger - A Journal of Comparative
377 Zoology. 2014;253 5:404-29.

378 2. Dai A, Yang S and Song Y. Marine crabs in China Sea. Beijing: Marine Publishing
379 Company; 1986.
380 3. Okamoto K. Malformed regeneration of partly cut swimming leg as a marker for
381 swimming crab *Portunus trituberculatus*. Fisheries Science. 2010;72 5:1121-3.
382 4. Hazlett and A. B. Interspecific Fighting in Three Species of Brachyuran Crabs From
383 Hawaii1). Crustaceana.
384 5. Xue J, Du N, Wei L and Wu H. A review of studies on *Portunus trituberculatus* in
385 China. Dongahi Marineence. 1997.
386 6. Sakai T. Crabs of Japan and the Adjacent Seas. Crabs of Japan & the Adjacent Seas.
387 1976.
388 7. Wu RSS and Shin PKS. Food segregation in three species of portunid crabs.
389 Hydrobiologia. 1997;362 1-3:107-13.
390 8. Qi JB, Gu XL, Ma LB, Qiao ZG and Chen K. The research progress on food
391 organism culture and technology utilization in crab seed production in ponds in China.
392 Agricultural Sciences. 2013;4 10:563-9.
393 9. Su XR, Li TW, Ding MJ and Chien PK. Evalution on nutritive value of *Portunus*
394 *trituberculatus*. Chinese Journal of Oceanology & Limnology. 1997;15 2:168-72.
395 10. Wang Z, Sun L, Guan W, Zhou C, Tang B, Cheng Y, et al. *De novo* transcriptome
396 sequencing and analysis of male and female swimming crab (*Portunus trituberculatus*)
397 reproductive systems during mating embrace (stage II). Bmc Genetics. 2018;19 1:3.
398 11. Dan S, Oshiro M, Ashidate M and Hamasaki K. Starvation of Artemia in larval
399 rearing water affects post-larval survival and morphology of the swimming crab,
400 *Portunus trituberculatus* (Brachyura, Portunidae). Aquaculture. 2016:
401 S0044848615300442.
402 12. Hamasaki K, Obata Y, Dan S and Kitada S. A review of seed production and stock
403 enhancement for commercially important portunid crabs in Japan. Aquaculture
404 International. 2011;19 2:217-35.
405 13. Liu S, Sun J and Hurtado LA. Genetic differentiation of *Portunus trituberculatus*, the
406 world\"s largest crab fishery, among its three main fishing areas. Fisheries Research.

407 2013;148 148:38-46.

408 14. Lv J, Ping L, Yu W, Baoquan G, Ping C, Jian L, et al. Transcriptome Analysis of
409 *Portunus trituberculatus* in Response to Salinity Stress Provides Insights into the
410 Molecular Basis of Osmoregulation. Plos One. 2013;8 12:e82155.

411 15. Yang Y, Wang J, Han T, Liu T, Wang C, Xiao J, et al. Ovarian Transcriptome Analysis
412 of *Portunus trituberculatus* Provides Insights into Genes Expressed during Phase III
413 and IV Development. Plos One. 2015;10 10:e0138862.

414 16. Meng X, Liu P, Jia F, Li J and Gao BQ. Correction: *De novo* Transcriptome Analysis
415 of *Portunus trituberculatus* Ovary and Testis by RNA-Seq: Identification of Genes
416 Involved in Gonadal Development. Plos One. 2015;10 7:e0133659.

417 17. Servant N, Varoquaux N, Lajoie B, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an
418 optimized and flexible pipeline for Hi-C data processing. Genome Biology. 2015;16
419 1:259.

420 18. Ruan J and Li H. Fast and accurate long-read assembly with wtdbg2. 2019.

421 19. Talay AC and Altilar DT. RACON: a routing protocol for mobile cognitive radio
422 networks. In: *Acm Workshop on Cognitive Radio Networks* 2009.

423 20. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
424 Integrated Tool for Comprehensive Microbial Variant Detection and Genome
425 Assembly Improvement. Plos One. 2014;9 11:e112963.

426 21. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De*
427 *novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length
428 scaffolds. Science. 2017;356 6333:92-5.

429 22. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO:
430 assessing genome assembly and annotation completeness with single-copy orthologs.
431 Bioinformatics. 2015;31 19:3210-2.

432 23. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
433 transform. Bioinformatics. 2009;25 14:1754-60. doi:10.1093/bioinformatics/btp324.

434 24. Kent WJ. BLAT - The BLAST-like alignment tool. Genome Research. 2002;12
435 4:656-64.

436 25. Benson and G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic
437 Acids Research. 1999;27 2:573-80.
438 26. Bedell JA, Korf I and Gish W. MaskerAid : a performance enhancement to
439 RepeatMasker. Bioinformatics. 2000;16 11:1040.
440 27. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: a new framework for
441 de novo transcriptome assembly using RNA-seq data. Genome Biology. 2015;16
442 1:30.
443 28. Stanke M and Waack S. Gene prediction with a hidden Markov model and a new
444 intron submodel. Bioinformatics. 2003;19 suppl_2:215--25.
445 29. Nowell RW, Elsworth B, Oostra V, Zwaan BJ and Blaxter M. A high-coverage draft
446 genome of the mycalesine butterfly *Bicyclus anynana*. GigaScience. 2017;6 7:1-7.
447 30. Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, et al. The
448 genomes of two key bumblebee species with primitive eusocial organization. Genome
449 Biology. 2015;16 1:76. doi:10.1186/s13059-015-0623-3.
450 31. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al.
451 The Genome Sequence of *Drosophila melanogaster*. Science. 2000;287 5461:2185.
452 32. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial
453 sequencing and comparative analysis of the mouse genome. Nature. 2002;420
454 6915:520-62. doi:10.1038/nature01262.
455 33. Ma J and Shi YB. The *Mesobuthus martensii* genome reveals the molecular diversity
456 of scorpion toxins. Cell & bioscience. 2014;4 1:1. doi:10.1186/2045-3701-4-1.
457 34. Linsheng S, Chao B, Yongju L, Lingling W, Xinxin Y, Jia L, et al. Draft genome of
458 the Chinese mitten crab, *Eriocheir sinensis*. GigaScience. 2016;5 1:5.
459 35. Liu R. Introduction to the classification of recent crustacean. In: Transactions of the
460 Chinese crustacean society No. 4. The Chinese crustacean Society, Science Press,
461 Beijing. 2003;No. 4:76-86.
462 36. Tang B, Zhou K, Song D, Yang G and Dai A. Molecular systematics of the Asian
463 mitten crabs, genus Eriocheir (Crustacea: Brachyura). Molecular Phylogenetics &
464 Evolution. 2003;29 2:309-16.

465 37. Song D. X., S. L. Yang. The Crustacea fauna of Hebei. Hebei Science and
466 Technology Publishing House, Shijiazhuang. 2009:1-772.
467 38. Tang BP, Xin ZZ, Liu Y, Zhang DZ, Wang ZF, Zhang HB, et al. The complete
468 mitochondrial genome of *Sesarmops sinensis* reveals gene rearrangements and
469 phylogenetic relationships in Brachyura. 2017;12 6:e0179800.
470 doi:10.1371/journal.pone.0179800.
471 39. Xuan F WX, Liu N, et al. Reproductive potential of individual male Chinese mitten
472 crabs *Eriocheir japonica sinensis* in a local pond- reared broodstock: Implications for
473 parent crab selection and sex ratio optimization. Aquac Res. 2018;49:3498–507.
474 40. Gong L, Fan G, Ren Y, Chen Y, Qiu Q, Liu L, et al. Chromosomal level reference
475 genome of *Tachypleus tridentatus* provides insights into evolution and adaptation of
476 horseshoe crabs. 2019;19 3:744-56. doi:10.1111/1755-0998.12988.
477 41. Birney E and Durbin R. Using GeneWise in the Drosophila Annotation Experiment.
478 Genome Research. 2000;10 4:547-8.
479 42. Haas BJ, Salzberg SL, Zhu W and Pertea… M. Automated eukaryotic gene structure
480 annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.
481 Genome Biology. 2008;9 1:R7.
482 43. Zdobnov EM and Apweiler R. InterProScan - an integration platform for the
483 signature-recognition methods in InterPro. Bioinformatics. 2001;17 9:847-8.
484 44. Altschul SF. Basic local alignment search tool (BLAST). Journal of Molecular
485 Biology. 2012;215 3:403-10.
486 45. Li and L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes.
487 Genome Research. 2003;13 9:2178-89.
488 46. Stamatakis and A. RAxML version 8: a tool for phylogenetic analysis and
489 post-analysis of large phylogenies. Bioinformatics. 2014;30 9:1312-3.
490 47. Yang and Ziheng. PAML: a program package for phylogenetic analysis by maximum
491 likelihood. Computer Applications in the Biosciences Cabios. 1997;13 5:555.
492 48. None. Phylogenetic test of the molecular clock and linearized trees. Molecular
493 Biology & Evolution. 1995;12 5:823-33.

494 49. De Bie T, Cristianini N, Demuth JP and Hahn MW. CAFE: a computational tool for
495 the study of gene family evolution. Bioinformatics. 2006;22 10:1269-71.
496 50. Beissbarth T and Speed TP. GOstat: find statistically overrepresented Gene
497 Ontologies within a group of genes. Bioinformatics. 2004;20 9:1464-5.
498 51. Huang DW, Sherman BT and Lempicki RA. Bioinformatics enrichment tools: paths
499 toward the comprehensive functional analysis of large gene lists. Nucleic Acids
500 Research. 2009;37 1:1.
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522

523 **Table 1: Assembly of swimming crab genome.**

Term	Contig phase		Hi-C phase	
	Size (bp)	Number	Size (bp)	Number
N90	439,683	334	11,273,125	41
N80	1,225,551	203	14,151,211	33
N70	2,035,154	141	16,942,622	27
N60	2,950,146	100	19,786,189	21
N50	4,121,416	71	21,793,880	17
Max length	17,984,318	-	42,710,960	-
Total length	1,004,084,521	-	1,005,046,021	-
Number>=100bp	-	2446	-	523
Number>=10kb	-	1756	-	314

524 Note: Contig phase represents results assembled by WTDBG software, and Hi-C phase
525 represents scaffold statistics of genome after chromosome assembly.
526
527 **Table 2: Quality evaluation of assembled swimming crab genome by BUSCO.**

Library	Eukaryota	Metazoa
Complete BUSCO (C)	287	909
Complete and single-copy BUSCO (S)	283	903
Complete and duplicated BUSCO (D)	4	6
Fragmented BUSCO (F)	2	19
Missing BUSCO (M)	14	50
Total BUSCO groups searched	303	978
Summary	94.7%	92.9%

528
529 **Table 3: Statistics on transposable elements (TEs) in swimming crab genome.**

Type	RepbaseTEs		TE proteins		*De novo*		Combined TEs	
	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	131,799,733	13.11%	2,434,533	0.24%	19,288,080	1.92%	149,711,951	14.90%
LINE	16,171,649	1.61%	75,759,827	7.54%	131,530,457	13.09%	153,027,744	15.23%
SINE	142,878	0.01%	0	0	0	0	142,878	0.014%
LTR	26,546,055	2.64%	10,195,324	1.01%	18,421,957	1.83%	45,189,365	4.50%
Other	89,969,319	8.95%	0	0	211,157,523	21.01%	230,116,216	22.90%
Unknown	34,752	0.0035%	0	0	90,989,908	9.05%	91,007,921	9.06%
Total	213,558,503	21.25%	88,375,336	8.79%	464,908,824	46.26%	525,492,271	52.29%

530
531 **Table 4: Functional annotation of predicted protein-coding genes.**

Term	Gene number	Percentage (%)
GO	8,712	51.87
InterPro	11,691	69.61
KEGG	10,880	64.78
SwissProt	12,558	74.77

TrEMBL	12,256	72.97
Annotated	16,053	95.58
Unannotated	743	4.42
Total	16,796	100

532
533 **Figure 1: Swimming crab, *Portunus trituberculatus*.**
534
535 **Figure 2: Genome characteristics of swimming crab.**
536
537 **Figure 3: Annotation quality comparison of protein-coding genes.**
538
539 **Figure 4: Gene family analysis of swimming crab.** A. Orthologous genes among species. B:
540 Unique and common gene families among species.
541
542 **Figure 5: Phylogenetic relationships, divergence time, and evolution rate analysis.** A.
543 Phylogenetic relationship and divergence time of species. Red dot represents fossil record
544 used here. B. Relative evolution rate of species.
545
546
547
548

Click here to download Figure Figure 1.pdf

Figure 1

Click here to download Figure Figure 2.pdf

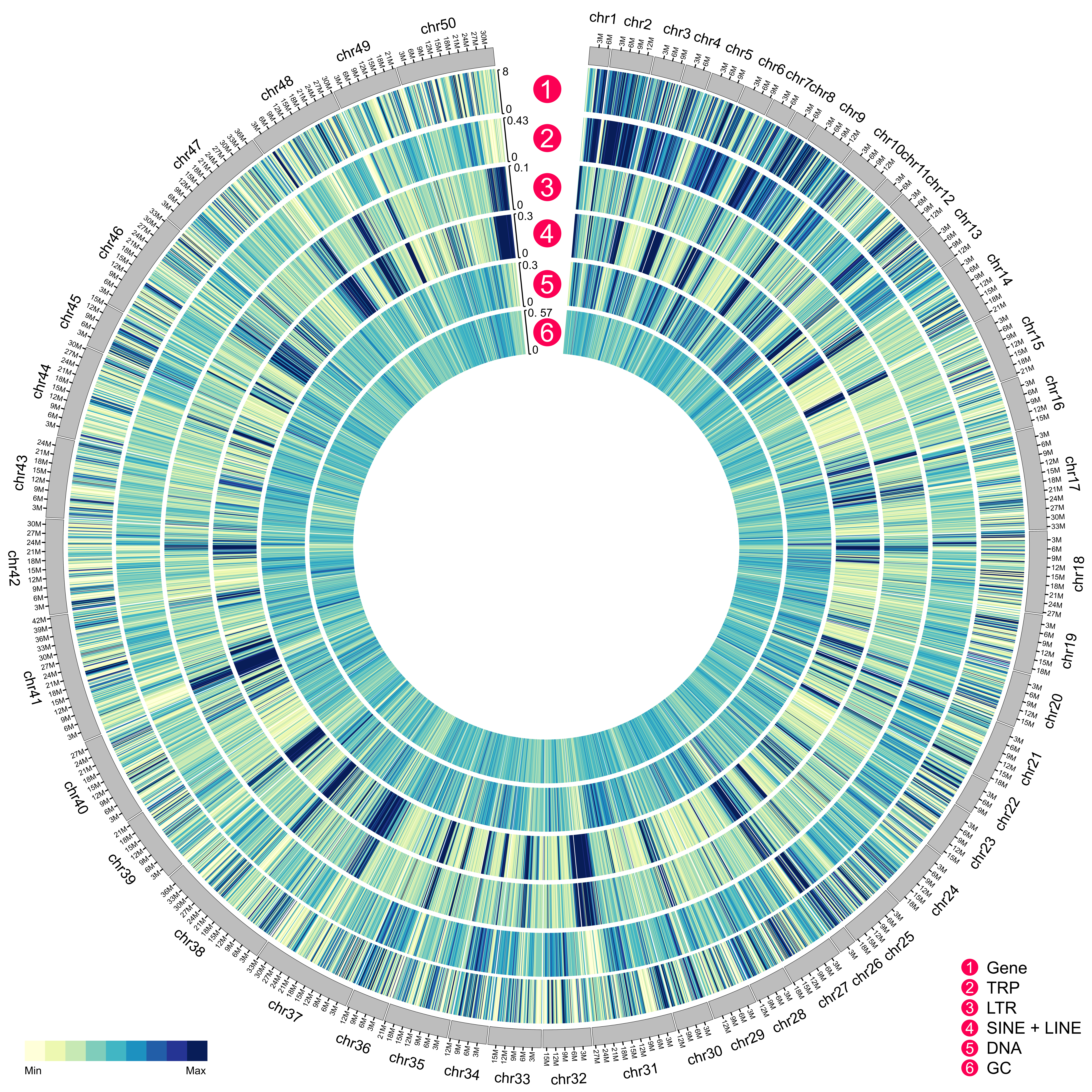
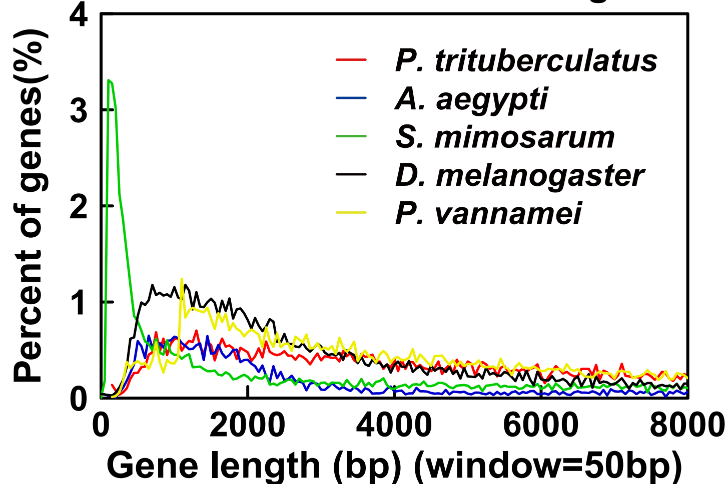
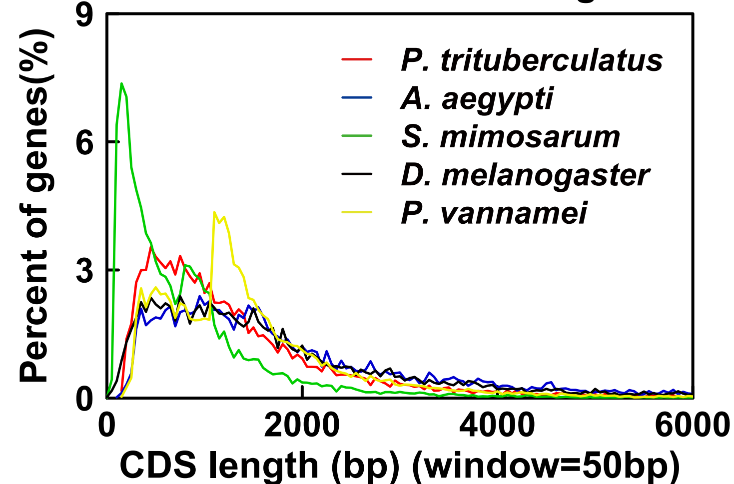
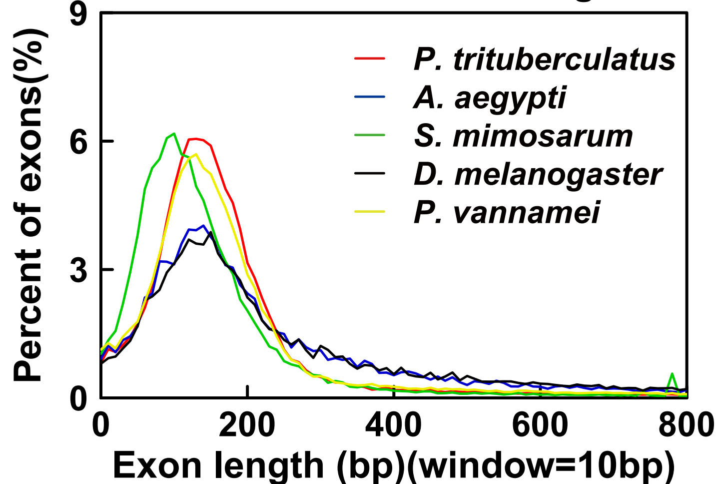


Figure 3

Click here to download Figure Figure 3.pdf







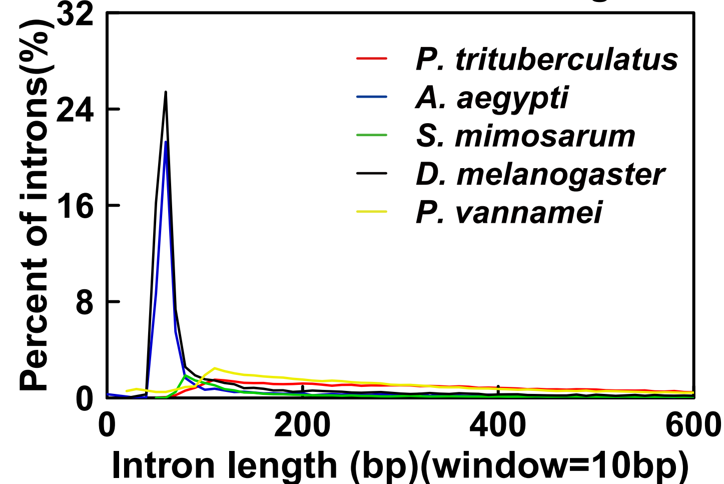


Figure 4

Click here to download Figure Figure 4.pdf

A

Gene number

0
10,000
20,000

E. j. sinensis
P. vannamei
D. melanogaster
B. anynana
P. trituberculatus
A. aegypti
S. mimosarum

Multiple copy orthologs
Other orthologs
Single copy orthologs
Unclustered genes
Unique paralogs

B

B. anynana
A. aegypti
D. melanogaster
P. vannamei
E. j. sinensis
S. mimosarum
P. trituberculatus

6,343
8,832
328

Figure 5

Click here to download Figure Figure 5.pdf

A

B. anynana
A. aegypti
D. melanogaster
P. vannamei
E. j. sinensis
P. trituberculatus
S. mimosarum

296.2(233.5-355.0)
173.7(96.2-243.9)
430.1(420.6-443.6)
428.5(418.6-441.8)
601.7(560.1-642.1)
183.5(138.0-237.7)

Lepidoptera
Diptera
Dendrobranchiata
Pleocyemata
Hexapoda
Crustacea
Pancrustacea

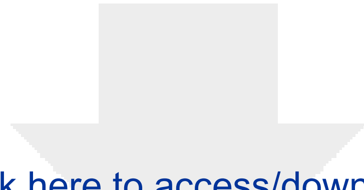
Million years ago
600 500 400 300 200 100 0

B

Branch size relative to swimming crab
0.08
0.04
0.00

outgroup
(S. mimosarum)
reference
(P. trituberculatus)

A. aegypti
B. anynana
D. melanogaster
E. j. sinensis
P. vannamei



Click here to access/download

Supplementary Material

Supplementary tables and figures.docx

Dear Editors,

We would like to submit the enclosed manuscript, entitled "**Chromosome-level genome assembly reveals adaptive evolution of the swimming crab (*Portunus trituberculatus*)**" for your consideration of publication as an original research paper in "***GigaScience***".

The swimming crab, *Portunus trituberculatus*, is an important commercial species in China and is widely distributed in the coastal waters of Asia-Pacific countries. Swimming crabs are considered highly nutritious, especially in regard to crab cream, and are very popular in China. As a result, the crab has been heavily overfished, resulting in substantial declines in its natural population and initiation of artificial breeding. With continued research on the crab, it has become clear that morphological, physiological, but the genetic changes are poorly understood. At present, genomic research on the swimming crab has only been conducted at the transcriptome level, with the whole genome not yet described.

In this study, based on BGISEQ short reads, Nanopore, and Hi-C sequencing data, we assembled the first chromosome-level high-quality genome of the swimming crab. Evaluation results indicated that the genome quality of swimming crab was comparable with that of most high-quality genomes in other species. We also successfully obtained 16,791 high-quality protein-coding genes by integrating three different methods. The genome and annotation data will help researchers better understand the evolution of crabs and improve their economic value. The phylogenetic results indicated that the swimming crab is closely related to the Chinese mitten crab, from which it diverged ~183.5 Mya. The unique and/or expanded gene family analysis provides clues to swimming crab development and environmental adaptation.

We believe these findings will be of interest to a broad audience of evolutionary and developmental biologists, as well as to researchers working to understand the molecular basis of the development in the swimming crab. In addition, these genomic resources could help scientists studying the development and ecological adaptation of the swimming crab. As a premier international journal devoted to the rapid dissemination of significant biological findings, ***GigaScience*** represents the ideal platform for sharing these results with the international research community.

We thank you for your consideration of our manuscript and look forward to hearing from you at your earliest convenience.

Yours sincerely,

Yandong Ren, Ph.D

Center for Ecological and Environmental Sciences, Northwestern Polytechnical University, Xi'an 710072, China

Email: renyandong90@126.com