High-quality genome assembly of Chinese shrimp (Fenneropenaeus chinensis) suggests genome contraction and adaptation to the environment

Qiong Wang[1], Xianyun Ren[1], Ping Liu[2], Jitao Li[2], Jianjian Lv[1], Jiajia Wang[1], Haien Zhang[1], Wei Wei[1], Yuxin Zhou[1], Yuying He[2], and Jian Li[2]

[1]Chinese Academy of Fishery Science Yellow Sea Fisheries Research Institute
[2]Chinese Academy of Fishery Sciences

April 9, 2021

Abstract

A high-quality reference genome is necessary to determine the molecular mechanisms underlying important biological phenomena; therefore, in the present study, a chromosome-level genome assembly of the Chinese shrimp Fenneropenaeus chinensis was performed. Muscle of a male shrimp was sequenced using PacBio platform, and assembled by Hi-C technology. The assembled F. chinensis genome was 1,465.32 Mb with contig N50 of 472.84 Kb, including 57.73% repetitive sequences, and was anchored to 43 pseudochromosomes, with scaffold N50 of 36.87 Mb. In total, 25,026 protein-coding genes were predicted. The genome size of F. chinensis showed significant contraction in comparison with that of other penaeid species, which is likely related to migration observed in this species. However, the F. chinensis genome included several expanded gene families related to cellular processes and metabolic processes, and the contracted gene families were associated with virus infection process. The findings signify the adaptation of F. chinensis to the selection pressure of migration and cold environment. Furthermore, the selection signature analysis identified genes associated with metabolism, phototransduction, and nervous system in cultured shrimps when compared with wild population, indicating targeted, artificial selection of growth, vision, and behavior during domestication. The construction of the genome of F. chinensis provided valuable information for the further genetic mechanism analysis of important biological processes, and will facilitate the research of genetic changes during evolution.

1 Introduction

The Chinese shrimp *Fenneropenaeus chinensis* is one of the most commercially important cultured shrimp species in China (Figure 1a). It is mainly distributed in the Yellow Sea and Bohai Sea of China, and west and south coast of the Korean Peninsula (Wang et al., 2017).*Fenneropenaeus chinensis* is an annual species; they migrate to warmer sea areas to overwinter after mating and swim back to the original coast for oviposition. With the development of aquaculture techniques, *F. chinensis* became the most important cultured shrimp species in China in the 1990s. However, in 1993, the production of *F. chinensis* decreased sharply owing to an outbreak of the white spot syndrome virus (WSSV) disease. Since 1997, breeding efforts have been made to increase the production and disease resistance of *F. chinensis* . After continuous artificial selection, several cultured varieties possessing excellent characteristics, such as high yield, disease resistance, and stress resistance, have been developed and cultured during the past two decades in China.

Because of their structural complexity and high heterozygosity (Yu et al., 2015), only a few crustacean genomes have been completely characterized. Recently, third-generation sequencing, characterized as long reads, has helped to ameliorate the difficulties engendered by heterozygosity and repetitive sequences in genome assembly (van Dijk, Jaszczyszyn, Naquin, & Thermes, 2018). An assembled genome of *F. chinensis* was published recently using Illumina short reads, PacBio long reads and Hi-C technology; the assembled genome covered 1.58 Gb in 8768 scaffolds, with N50 length of 28.92 Mb (Yuan et al., 2021). Furthermore, the reference genome of *Litopenaeus vannamei* , which—similar to *F. chinensis*— previously belonged to the genus *Penaeus* , has been reported also using both Illumina short reads and PacBio long reads (X. Zhang et al., 2019), covering 1.66 Gb in 4,683 scaffolds, with N50 length of 605.56 Kb, and was improved to 31.30 Mb by using Hi-C technology (Yuan et al., 2021). Two genomic resources for another *Penaeus* species, *Penaeus monodon* , was released lately (Uengwetwanit et al., 2021; Van Quyen et al., 2020); the improved genome was generated using long-read PacBio and long-range Chicago, producing a final genome assembly of 2.39 Gb, with contig N50 length of 79 Kb (Uengwetwanit et al., 2021).

A high-quality reference genome is essential for resolving the molecular mechanism of important biological processes (You, Shan, & Shi, 2020). In this research, an improved chromosome-level genome of *F. chinensis* was assembled using the PacBio sequencing platform and Hi-C technology, in an attempt to explain the genetic changes during evolution and domestication. Quality of the genome assembly could affect the accuracy of following studies (You et al., 2020). Therefore, the new version assembled genome provides a high-quality reference, and will be a valuable resource for the further investigations of biological process and mechanism in *F. chinensis* .

2 Materials and Methods

2.1 Sample collection and genome sequencing

Fenneropenaeus chinensis shrimp were obtained from the conservation base of Haifeng Aquaculture Co., Ltd. (Weifang, Shandong Province, China). Muscle of a 7-months-old male shrimp was collected and frozen in liquid nitrogen immediately. Total genomic DNA was extracted and sequenced for genome survey and genome construction. Genome construction contains PacBio sequencing (Eid et al., 2009) and Hi-C assembly (Lieberman-Aiden et al., 2009). DNA sample used for genome survey and Hi-C assembly was sequenced by Illumina HiSeq platform (Illumina, San Diego, USA), and used for PacBio sequencing was sheared to 20 Kb and sequenced by PacBio Sequel platform (Pacific Biosciences, Menlo Park, USA).

2.2 Genome survey and assembly

The genome size, repetitive sequence proportion and heterozygosity was estimated by K-mer frequency distribution method with K-mer=17. The genome size was revised by error rate: Revised size = Genome size (1-error rate), the error rate refers to the proportion of K-mer with depth of 1.

After adapter removal and filtered by minimum length of 50 bp, the subreads from PacBio platform were assembled using wtdbg2 (Ruan & Li, 2020), which uses the Fuzzy Bruijn Graph (FBG) approach. The FBG is not as sensitive to small duplications as the De Bruijn Graph. To solve the problem of high error rate, the Gapped Sequence Alignment method was used. After quality control, the high-quality Hi-C sequencing data were mapped to the draft genome by BWA software (H. Li & Durbin, 2010), and Samtools (H. Li et al., 2009) was used to remove duplicate and unmapped data to obtain high-quality data. Next, the reads near the restriction sites were extracted for assisted assembly (Burton et al., 2013).

2.3 Genome annotation

Structural annotation of the genome incorporates ab initio prediction, homology-based prediction, and RNA-Seq assisted prediction. For gene predication based on Ab initio, Augustus (v3.2.3) (Hoff & Stanke, 2019), GeneID (v1.4) (Parra, Blanco, & Guigo, 2000), Genescan (v1.0) (Aggarwal & Ramaswamy, 2002), GlimmerHMM (v3.04) (Majoros, Pertea, & Salzberg, 2004), and SNAP (2013-11-29) (Korf, 2004) were used in our automated gene prediction pipeline. Six species, *Litopenaeus vannamei* , *Hyalella azteca* , *Eurytemora affinis* ,*Daphnia pulex* , *Drosophila hydei,* and *Bombyx mori* , were used for homology-based prediction. Sequences of homologous proteins were downloaded from Ensembl and NCBI. Protein sequences were aligned to the genome using tBLASTn (v2.2.26; E-value [?] 1e^{-5}), and then the matching proteins were aligned to the homologous genome sequences for accurately spliced alignments using GeneWise (v2.4.1) software (Birney, Clamp, & Durbin, 2004). To optimize the genome annotation, the RNA-Seq reads from different tissues (NCBI BioProject: PRJNA558194) were aligned to the genome. Hierarchical indexing for spliced alignment of transcripts (HISAT; v2.0.4) (Kim, Langmead, & Salzberg, 2015) and TopHat (v2.0.11) (Cole Trapnell, Pachter, & Salzberg, 2009) were used with default parameters to identify exons and splice positions. The alignment results were then used as input for Stringtie (v1.3.3) (Pertea et al., 2015) and Cufflinks (v2.2.1) (C. Trapnell et al., 2010) with default parameters for genome-based transcriptome assembly.

Gene functions were assigned according to the best match by aligning the protein sequences to the SwissProt database using BLASTp (Altschul et al., 1997) (E-value [?] 1e^{-5}). The motifs and domains were annotated using InterProScan70 (v5.31) (Mulder & Apweiler, 2007) by searching against publicly available databases, including ProDom, PRINTS, Pfam, simple modular architecture research tool (SMART), PANTHER and PROSITE. The GO IDs for each gene were assigned according to the corresponding InterPro entry. We also mapped the gene set to the KEGG pathway database and identified the best match for each gene.

2.4 Comparative genomics analysis

The following 17 species were used for comparative genomic analysis:*F. chinensis* , *P. monodon* , *L. vannamei* ,*Portunus trituberculatus* , *Trinorchestia longiramus* ,*H. Azteca* , *Eurytemora affinis* , *Drosophila melanogaster* , *Acyrthosiphon pisum* , *Apis mellifera* ,*Bombyx mori* , *Zootermopsis nevadensis* , *Cryptotermes secundus* , *Pediculus humanus* , *Tribolium castaneum* ,*Anopheles gambiae* , and *Limulus polyphemus* . Genomic sequences were downloaded from NCBI. The gene set of each species was filtered. In brief, when a gene possessed multiply spliced transcripts, only the longest protein-coding transcripts were retained for further analysis. Furthermore, genes were excluded if the proteins encoded by them consisted of less than 30 amino acids or contained degenerate bases or termination codons. The similarity between protein sequences of all species was assessed using BLASTp (E-value [?] 1e^{-7}). The results were clustered using OrthoMCL (L. Li, Stoeckert, & Roos, 2003), with an expansion coefficient of 1.5. Single-copy and multiple-copy homologous genes were filtered by these analyses.

A phylogenetic tree was constructed using single-copy homologous genes in the 17 species. MUSCLE (Edgar, 2004) was used for sequence alignment. The final dataset was used to construct the phylogenetic tree with RAxML (Rokas, 2011) using the maximum likelihood method. The best tree was used as an input tree for divergence time estimation using MCMCTREE in the PAML package (Yang, 2007), with the following parameters: burn in = 700, sample number = 1000,000, sample frequency = 2. Fossil calibrations were used as priors for the divergence time estimation, as below:*P. monodon* and *L. vannamei* [58–108 million years ago (Mya)], *Acyrthosiphon pisum* and *Eurytemora affinis*(452–557 Mya), *Zootermopsis nevadensis* and *Cryptotermes secundus* (103–156 Mya), *Zootermopsis nevadensis* and*Pediculus humanus* (330–398 Mya), *Anopheles gambiae* and*Drosophila melanogaster* (217–301 Mya), *Drosophila melanogaster* and *Bombyx mori* (243–317 Mya), *Tribolium castaneum* and *Apis mellifera* (308–366 Mya). In the gene family expansion and contraction analysis, we filtered the gene families with the results of the clustering analysis of gene families using CAFE software (De Bie, Cristianini, Demuth, & Hahn, 2006). Protein sequences of single-copy homologous genes in *F. chinensis* , *L. vannamei* , *P. trituberculatus* and *H. azteca* were subjected to multiple alignment using MUSCLE to detect positive selection. The ratios of nonsynonymous substitution per nonsynonymous site (dN) to synonymous substitution per synonymous site (dS) were calculated using the branch-site model of the Codeml tool included in the PAML package. Likelihood ratio tests were applied to test for positive selection.

2.5 Selection signature analysis

Cultured *F. chinensis* shrimps were obtained from Haifeng Aquaculture Co., Ltd, which has undergone continuous, high-intensity artificial selection for growth traits for more than ten generations. Wild *F. chinensis* were captured from the northeast region of the Huanghai Sea. We used 22 cultured and 22 wild shrimps to collect muscle tissue and extract DNA. The DNA samples were sequenced using the BGISEQ platform (BGI, Wuhan, China). Sequencing data were mapped to the reference genome, and VCFtools (Danecek et al., 2011) was used to calculate the fixation index (F_{ST}) and nucleotide diversity (π) in 40 kb sliding windows with a step size of 20 kb. Windows with the top 5% values of both F_{ST} and π were considered as outliers. Genes in these windows were picked for the subsequent Gene ontology (GO) and KEGG pathway enrichment analysis.

3 Results

3.1 Genome survey analysis

A male shrimp (*F. chinensis*) was sequenced using the Illumina HiSeq platform, and 125.99 Gb raw data were obtained. The K-mer analysis (Figure S1) revealed that the genome size was 1,384.88 Mb, with 54.79% repetitive sequences, and the genome heterozygosity was 1.04% (Table 1). The first six comparison species using 10,000 high quality reads randomly were all homologous comparisons (Table S1), which indicates an absence of exogenous contamination in our sample.

3.2 Genome assembly and assessment

A total of 350.81 Gb polymerase reads were obtained from the PacBio platform. After adapter removal and quality control, we obtained 220.53 Gb subreads (coverage depth approximately 159 X), with average read length of 16,922 bp (Table S2). The genome assembled using these subreads was 1,465.32 Mb in size and contained 9,015 contigs with an N50 length of 472,841 bp (Table 1). According to the Benchmarking Universal Single-Copy Orthologs (BUSCO) notation analysis using 1,066 lineal homologous single-copy genes, 95.3% of the genes were assembled (complete: 94%, fragmented: 1.3%, missing: 4.7%) (Table S3). Furthermore, Core Eukaryotic Genes Mapping Approach (CEGMA) analysis revealed that 234 genes were assembled from 248 Core Eukaryotic Genes (CEGs), which accounted for 94.35% (Table S4). Both analyses indicated that the genome assembly was relatively complete. The mapping rate of short reads from the Illumina platform was approximately 89.62%, and the coverage rate reached 96.15%, indicating that the short reads and the assembled genome had a good consistency (Table S5).

We obtained 203.8 Gb clean, non-duplicate data from the Illumina HiSeq platform by Hi-C technology. The contigs were anchored to 1,063 scaffolds with N50 of 36.87 Mb (Table 1, Table S6), including 43 pseudochromosomes (Figure 1b, c) and 1,020 unplaced scaffolds. The total length of the 43 pseudochromosomes was 1,451.52 Mb (Table S7), covered 99.00% of the assembly, whereas the length of unplaced scaffolds was 14.60 Mb (Table S8).

3.3 Genome annotation

There were 57.73% repetitive sequences in the *F. chinensis*genome (Table S9). A total of 25,026 genes, with average length of 11,290 bp (including untranslated regions [UTRs]), average coding sequence (CDS) length of 1,230 bp, and 5.94 exons per gene were predicted with three methods—de novo prediction, homolog prediction, and RNA-seq prediction (Table S10). Gene function of 76.7% of the predicted genes was annotated using multiple databases (Table S11). By comparing with the non-coding RNA (ncRNA) database, 72,517 ncRNA genes were annotated, including 59,026 miRNA genes, 2,592 tRNA genes, 24 rRNA genes, and 10,875 snRNA genes (Table S12).

3.4 Comparative genomics

According to the gene family clustering analysis of 17 Arthropoda species, a total of 27,512 gene families were clustered, and 443 of them were single-copy genes in all species (Figure 2a). Compared with*P. monodon* , *L. vannamei,* and *P. trituberculatus* ,*F. chinensis* possessed 593 unique gene families (Figure 2b). In general, gene families unique to a species are responsible for the biological characteristics of the species. GO enrichment analysis of the 593 unique gene families indicated that they were enriched in 47 GO terms, including structural constituent of cuticle, sodium and potassium ion transport, and chitin metabolic process (Table S13). These genes were enriched in nine KEGG pathways, including RNA polymerase, Huntington's disease, and endocrine and other factor-regulated calcium reabsorption pathways (Table S14).

Phylogenetic analysis based on the 443 single-copy homologous genes revealed that *F. chinensis* and *P. monodon* diverged approximately 44 Mya, after they diverged from *L. vannamei* 70 Mya (Figure 2c). The three penaeid shrimp species diverged from *P. trituberculatus* , which belongs to Family *Portunidae* , approximately 271.5 Mya. The *F. chinensis* genome showed 49 expanded and 51 contracted gene families in comparison with the *P. monodon* genome. The GO and KEGG enrichment analysis indicated that the expanded gene families were mostly related to cellular process and metabolic process, including chitin metabolism (Figure S2, Table S15-16), whereas the contracted gene families were mostly associated with infection with certain pathogens and phototransduction (Figure S3, Table S17-18). Compared with *L. vannamei* , *P. trituberculatus* and *H. azteca* , a total of 63 genes were subject to positive selection. These genes were mostly related to basic cellular process (Table S19). One KEGG pathway named mRNA surveillance pathway was enriched.

3.5 Selection signature analysis

A total of 447.23 Gb clean resequencing data of 42 shrimps (21 wild and 21 cultured) was obtained, which were mapped to the assembled genome. Fixation index (F_{ST}) between wild shrimp and cultured shrimp and nucleotide diversity (π) in both populations were calculated in 40 kb windows across the genome (Figure 3a). A total of 534 outlier genes were identified according to theF_{ST}-π conjoint analysis. Most of these genes were involved in metabolic process (Figure S4). In addition, the phototransduction-fly pathway and neuroactive ligand-receptor interaction pathway were enriched by the target genes (Figure 3b).

4 Discussion

4.1 Quality of this assembled genome improved

Compared to the published genome of *F. chinensis* (Yuan et al., 2021), the quality of our assembled genome improved significantly (Table S20). Primarily, the contig N50 increased from 59.00 Kb to 472.84 Kb, approximately eight times, and scaffold N50 increased from 28.92 Mb to 36.87 Mb. Secondly, the assembled complete ortholog proportion enhanced from 83.58% to 94.00% according to the BUSCO assessment, and the genome coverage enhanced from 82.46% to 96.15% according to the Illumina short read mapping test. In short, our assembled genome improved in both fragment length and completeness.

Initially, we assembled the contigs to 44 pseudochromosomes, but the heatmap revealed a low correlation between the contigs on the last pseudochromosome (Figure S5). Therefore, only 43 pseudochromosomes were assembled in the final version of the *F. chinensis* genome. We speculate that there are only 43 pairs of chromosomes in the somatic cell of *F. chinensis* , same as that in *Marsupenaeus japonicus* (Yuan et al., 2017).

4.2 The Genome of *F. chinensis* contracted compared to other penaeid shrimps

The K-mer analysis revealed that the *F. chinensis* genome is 1.38 Gb in size. Compared to the genomes of other three penaeid shrimps,*L. vannam* ei (2.60 Gb) (X. Zhang et al., 2019), *P. monodon*(2.59 Gb) (Yuan et al., 2017), and*M. japonicus* (2.28 Gb) (Yuan et al., 2017), the genome size of*F. chinensis* contracted sharply (Table 2). Although *F. chinensis* and *L. vannam* ei genomes contain a similar number of genes (25,026 vs. 25,596, respectively), the proportion of repetitive sequences in *F. chinensis* is markedly lower (57.73% vs. 78.00%, respectively) (X. Zhang et al., 2019). The lower repetitive sequences proportion could contribute to the genome contraction, but not be the dominant factor, because the *P. monodon* , which closer to *F. chinensis* in evolution, has a similar repetitive sequences proportion with *F. chinensis* (62.5%), but the genome size is also bigger (2.59 Gb) (Uengwetwanit et al., 2021).

Research suggests that genome size is under selective pressure to contract owing to constraints placed by an elevated metabolism on cell size (Hughes & Hughes, 1995; Olmo, 1983; Szarski, 1983). A similar phenomenon is observed in vertebrates. The two living groups of flying vertebrates, birds and bats, have constricted genome sizes compared with their close relatives (Hughes & Hughes, 1995). Research shows that genome size contraction preceded flight during evolution (Organ & Shedlock, 2009). Likewise,*F. chinensis* shrimp has a special characteristic relative to *L. vannamei* , *P. monodon* , and *M. japonicus* , which is migration. Same as flight, migration requires a high metabolic intensity, which closely linked with genome size (Andrews, Mackenzie, & Gregory, 2009; Vinogradov & Anatskaya, 2006). The nucleotypic theory suggests that genome size affects nucleus size and cell size (Bennett, 1971; Gregory, 2001). The cell size may further influence housekeeping dynamics, cellular metabolism, and the rate of cell division, leading to small cells with higher metabolic rates (Kozlowski, Konarzewski, & Gawelczyk, 2003). The results of gene family expansion and contraction analysis and positive selection analysis support this conjecture. The *F. chinensis* shrimp genome showed expanded several gene families related to cellular process and metabolic process, and genes involved in cellular process had undergone positive selection during evolution.

4.3 Pathogen infection related gene families contracted in*F. chinensis*

Some of the contracted gene families in *F. chinensis* are associated with pathogen infection. Most of the contracted gene families involving in these pathways encoded actin-like protein. The actin protein is a fundamental component of the host cellular cytoskeleton, playing an important role in viral replication (Roberts & Baines, 2011; Spear & Wu, 2014). Previous research suggests that the actin gene is upregulated in *M. japonicus* after infection with *Vibrio parahaemolyticus* and WSSV (Ren, Zhang, Liu, & Li, 2019), the pathogens that cause the two most serious diseases affecting shrimp cultivation (Escobedo-Bonilla et al., 2008). The β-actin gene is also involved in WSSV infection in *L. vannamei* (J. Feng, Li, Liu, Tang, & Du, 2019). We speculate that the contraction of the actin gene family made*F. chinensis* more sensitive to viral infections because the viruses could destroy the force-generating and macromolecular scaffolding properties of the actin cytoskeleton to drive the infection process (Spear & Wu, 2014). The lower number of actin genes may make it difficult to repair damaged cell in *F. chinensis* . This finding is consistent with those of previous studies, which indicate that *F. chinensis* exhibit lower resistance for WSSV than other shrimp species (Y. Feng et al., 2017; Jiang, Yu, & Zhou, 2006). This is one of the reasons why *L. vannamei* has replaced *F. chinensis* to become the dominant cultured shrimp species in recent years.

Higher temperatures often increase the severity of disease susceptibility (Cohen & Leach, 2020). Pathogen infection needs a suitable temperature, for example, the WSSV needs an optimum temperature to successfully enter the host hematopoietic stem cells, and the virus entry is blocked at 6 °C (Korkut, Noonin, & Soderhall, 2018).*Fenneropenaeus chinensis* is mostly distributed in the northern Pacific, a colder environment relative to that of other *Penaeus*species, where pathogen infectivity is weaker. We hypothesize that, during evolution, *F. chinensis* sacrificed an aspect of immunity to preserve more energy for subsistence in the cold environment, because immunoreaction is an energy-consuming activity (Alwarawrah, Kiernan, & MacIver, 2018; Hosomi & Kunisawa, 2020; Loftus & Finlay, 2016).

4.4 Targeted artificial selection during domestication of*F. chinensis*

Domestication of animals was driven by both natural and artificial selection. Under these directed selection pressures, change in the allele frequency in specific regions may be accumulated in the animal genome after domesticated for several generations (Nielsen, Hellmann, Hubisz, Bustamante, & Clark, 2007). The cultured shrimps were derived from the wild population by continuous, high-intensity artificial selection for growth traits and is characterized by faster growth and higher body weight than wild individuals. Genes responsible for these characteristics may be detected by identifying unique selection signatures in the genome. Selection signature is characterized by the decrease in polymorphism and increase in linkage disequilibrium in certain loci. Identification of genetic variations by selection signature reveals the genetic mechanism of the formation of phenotypic trait diversity during selection (Lopez, Neira, & Yanez, 2014). In the present study, most genes with selection signatures were identified to participate in metabolic processes. Metabolism is the basis of growth and is related to growth rate (Krieger, 1978; Vahl, 1984). The changes in the allele frequency of genes related to metabolic processes suggest artificial selection on growth. However, owing to the short history of shrimp domestication, the genetic divergence between the wild and cultured shrimp populations is not significant.

A pathway named the phototransduction-fly pathway was enriched by genes with selection signatures; of the 27 background genes of this pathway, six genes were located in the selection signature regions. This enriched pathway signifies a visual change in domesticated shrimp.*Fenneropenaeus chinensis* juveniles exhibit an intense attack behavior (P. Zhang, Zhang, Li, & Meng, 2008). We propose that the high-density culture environment required them to develop better vision to survive cannibalism.

Among the pathways enriched by the candidate genes, there was a pathway related to the nervous system, neuroactive ligand-receptor interaction pathway, which suggests that genes affecting neuronal development were also targeted during domestication. Animal behavior is regulated by the nervous system. Compared with aggressive animals, docile animals survive more easily in a culture environment (Carneiro et al., 2014; Darwin, 1860). We speculate that tame shrimps were selected for during domestication, and therefore the related genes were targeted.

4.5 Conclusion

In summary, an improved chromosome-level genome of *F. chinensis*was reported in this article. The assembled genome was 1.47 Gb in size, and was anchored to 43 pseudochromosomes, with N50 length of 36.87 Mb. The contraction of the genome size was speculated relating to the migration. Gene families related to cellular processes and metabolic processes expanded, while gene families associated with virus infection contracted to adapt to the cold environment. Furthermore, genetic variations in genes associated with metabolism, phototransduction, and nervous system were detected by selection signature, indicating targeted artificial selection on growth, vision, and behavior during domestication.

Acknowledgements

We would like to appreciate our colleagues of the Key Laboratory for Sustainable Utilization of Marine Fisheries Resources of Yellow Sea Fisheries Research Institute, for their assistance on sample collection and helpful comments on the manuscript. This work was supported by the National Key R & D Program of China (No. 2019YFD0900403), Modern Agro-industry Technology Research System (No. CARS-48), National Natural Science Foundation of China (No. 31902367) and Central Public-interest Scientific Institution Basal Research Fund of CAFS (No. 2020TD46).

References

Aggarwal, G., & Ramaswamy, R. (2002). Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J Biosci, 27* (1 Suppl 1), 7-14. doi:10.1007/BF02703679

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25* (17), 3389-3402. doi:10.1093/nar/25.17.3389

Alwarawrah, Y., Kiernan, K., & MacIver, N. J. (2018). Changes in Nutritional Status Impact Immune Cell Metabolism and Function.*Front Immunol, 9* , 1055. doi:10.3389/fimmu.2018.01055

Andrews, C. B., Mackenzie, S. A., & Gregory, T. R. (2009). Genome size and wing parameters in passerine birds. *Proc Biol Sci, 276* (1654), 55-61. doi:10.1098/rspb.2008.1012

Bennett, M. D. (1971). The duration of meiosis. *Proc. R. Soc. B., 178* , 277–299. doi:10.1098/rspb.1971.0066

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

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

Carneiro, M., Rubin, C. J., Di Palma, F., Albert, F. W., Alfoldi, J., Martinez Barrio, A., . . . Andersson, L. (2014). Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication.*Science, 345* (6200), 1074-1079. doi:10.1126/science.1253714

Cohen, S. P., & Leach, J. E. (2020). High temperature-induced plant disease susceptibility: more than the sum of its parts. *Curr Opin Plant Biol, 56* , 235-241. doi:10.1016/j.pbi.2020.02.008

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156-2158. doi:10.1093/bioinformatics/btr330

Darwin, C. (1860). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. *Br Foreign Med Chir Rev, 25* (50), 367-404.

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

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

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science, 323* (5910), 133-138. doi:10.1126/science.1162986

Escobedo-Bonilla, C. M., Alday-Sanz, V., Wille, M., Sorgeloos, P., Pensaert, M. B., & Nauwynck, H. J. (2008). A review on the morphology, molecular characterization, morphogenesis and pathogenesis of white spot syndrome virus. *J Fish Dis, 31* (1), 1-18. doi:10.1111/j.1365-2761.2007.00877.x

Feng, J., Li, D., Liu, L., Tang, Y., & Du, R. (2019). Interaction of the small GTP-binding protein (Rab7) with beta-actin in Litopenaeus vannamei and its role in white spot syndrome virus infection. *Fish Shellfish Immunol, 88* , 1-8. doi:10.1016/j.fsi.2019.02.053

Feng, Y., Kong, J., Luo, K., Luan, S., Cao, B., Liu, N., . . . Meng, X. (2017). The comparison of the sensitivity to the white spot syndrome virus between Fenneropenaeus chinensis and Litopenaeus vannamei.*Progress in Fishery Sciences, 38* (6), 78-84.

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc, 76* (1), 65-101. doi:10.1017/s1464793100005595

Hoff, K. J., & Stanke, M. (2019). Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics, 65* (1), e57. doi:10.1002/cpbi.57

Hosomi, K., & Kunisawa, J. (2020). Diversity of energy metabolism in immune responses regulated by micro-organisms and dietary nutrition.*Int Immunol, 32* (7), 447-454. doi:10.1093/intimm/dxaa020

Hughes, A. L., & Hughes, M. K. (1995). Small genomes for better flyers.*Nature, 377* (6548), 391. doi:10.1038/377391a0

Jiang, G., Yu, R., & Zhou, M. (2006). Studies on nitric oxide synthase activity in haemocytes of shrimps Fenneropenaeus chinensis and Marsupenaeus japonicus after white spot syndrome virus infection.*Nitric Oxide, 14* (3), 219-227. doi:10.1016/j.niox.2005.11.005

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

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

Korkut, G. G., Noonin, C., & Soderhall, K. (2018). The effect of temperature on white spot disease progression in a crustacean, Pacifastacus leniusculus. *Dev Comp Immunol, 89* , 7-13. doi:10.1016/j.dci.2018.07.026

Kozlowski, J., Konarzewski, M., & Gawelczyk, A. T. (2003). Cell size as a link between noncoding DNA and metabolic rate scaling. *Proc Natl Acad Sci U S A, 100* (24), 14080-14085. doi:10.1073/pnas.2334605100

Krieger, I. (1978). Relation of specific dynamic action of food (SDA) to growth in rats. *Am J Clin Nutr, 31* (5), 764-768. doi:10.1093/ajcn/31.5.764

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26* (5), 589-595. doi:10.1093/bioinformatics/btp698

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, 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 Res, 13* (9), 2178-2189. doi:10.1101/gr.1224503

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326* (5950), 289-293. doi:10.1126/science.1181369

Loftus, R. M., & Finlay, D. K. (2016). Immunometabolism: Cellular Metabolism Turns Immune Regulator. *J Biol Chem, 291* (1), 1-10. doi:10.1074/jbc.R115.693903

Lopez, M. E., Neira, R., & Yanez, J. M. (2014). Applications in the search for genomic selection signatures in fish. *Front Genet, 5* , 458. doi:10.3389/fgene.2014.00458

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

Mulder, N., & Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol, 396* , 59-70. doi:10.1007/978-1-59745-515-2_5

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet, 8* (11), 857-868. doi:10.1038/nrg2187

Olmo, E. (1983). Nucleotype and cell size in vertebrates: a review.*Basic Appl Histochem, 27* (4), 227-256.

Organ, C. L., & Shedlock, A. M. (2009). Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. *Biol Lett, 5* (1), 47-50. doi:10.1098/rsbl.2008.0491

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

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

Ren, X., Zhang, Y., Liu, P., & Li, J. (2019). Comparative proteomic investigation of Marsupenaeus japonicus hepatopancreas challenged with Vibrio parahaemolyticus and white spot syndrome virus. *Fish Shellfish Immunol, 93* , 851-862. doi:10.1016/j.fsi.2019.08.039

Roberts, K. L., & Baines, J. D. (2011). Actin in herpesvirus infection.*Viruses, 3* (4), 336-346. doi:10.3390/v3040336

Rokas, A. (2011). Phylogenetic analysis of protein sequence data using the Randomized Axelerated Maximum Likelihood (RAXML) Program. *Curr Protoc Mol Biol, Chapter 19* , Unit19 11. doi:10.1002/0471142727.mb1911s96

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat Methods, 17* (2), 155-158. doi:10.1038/s41592-019-0669-3

Spear, M., & Wu, Y. (2014). Viral exploitation of actin: force-generation and scaffolding functions in viral infection.*Virol Sin, 29* (3), 139-147. doi:10.1007/s12250-014-3476-0

Szarski, H. (1983). Cell size and the concept of wasteful and frugal evolutionary strategies. *J Theor Biol, 105* (2), 201-209. doi:10.1016/s0022-5193(83)80002-2

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

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. *Nat Biotechnol, 28* (5), 511-515. doi:10.1038/nbt.1621

Uengwetwanit, T., Pootakham, W., Nookaew, I., Sonthirod, C., Angthong, P., Sittikankaew, K., . . . Karoonuthaisiri, N. (2021). A chromosome-level assembly of the black tiger shrimp (Penaeus monodon) genome facilitates the identification of growth-associated genes.*Mol Ecol Resour* . doi:10.1111/1755-0998.13357

Vahl, O. (1984). The relationship between specific dynamic action (SDA) and growth in the common starfish, Asterias rubens L. *Oecologia, 61* (1), 122-125. doi:10.1007/BF00379097

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends Genet, 34* (9), 666-681. doi:10.1016/j.tig.2018.05.008

Van Quyen, D., Gan, H. M., Lee, Y. P., Nguyen, D. D., Nguyen, T. H., Tran, X. T., . . . Austin, C. M. (2020). Improved genomic resources for the black tiger prawn (Penaeus monodon). *Mar Genomics, 52* , 100751. doi:10.1016/j.margen.2020.100751

Vinogradov, A. E., & Anatskaya, O. V. (2006). Genome size and metabolic intensity in tetrapods: a tale of two lines. *Proc Biol Sci, 273* (1582), 27-32. doi:10.1098/rspb.2005.3266

Wang, M., Kong, J., Meng, X., Luan, S., Luo, K., Sui, J., . . . Shi, X. (2017). Evaluation of genetic parameters for growth and cold tolerance traits in Fenneropenaeus chinensis juveniles. *PLoS One, 12* (8), e0183801. doi:10.1371/journal.pone.0183801

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood.*Mol Biol Evol, 24* (8), 1586-1591. doi:10.1093/molbev/msm088

You, X., Shan, X., & Shi, Q. (2020). Research advances in the genomics and applications for molecular breeding of aquaculture animals.*Aquaculture, 526* (735357).

Yu, Y., Zhang, X., Yuan, J., Li, F., Chen, X., Zhao, Y., . . . Xiang, J. (2015). Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific White Shrimp Litopenaeus vannamei. *Sci Rep, 5* , 15612. doi:10.1038/srep15612

Yuan, J., Zhang, X., Liu, C., Yu, Y., Wei, J., Li, F., & Xiang, J. (2017). Genomic resources and comparative analyses of two economical penaeid shrimp species, Marsupenaeus japonicus and Penaeus monodon.*Marine Genomics* , 22-25.

Yuan, J., Zhang, X., Wang, M., Sun, Y., Liu, C., Li, S., . . . Li, F. (2021). Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp. *Commun Biol, 4* (1), 186. doi:10.1038/s42003-021-01716-y

Zhang, P., Zhang, X., Li, J., & Meng, Q. (2008). Observation of behavior in Fenneropenaeus chinensis and Litopenaeus vannamei postlarvae. *JOURNAL OF FISHERIES OF CHINA, 32* (2), 223-228.

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. *Nat Commun, 10* (1), 356. doi:10.1038/s41467-018-08197-4

Data accessibility

Genomic sequences, PacBio data used for genome assembly and Illumina data used for genomic survey and Hi-C were deposited in NCBI under accession number: PRJNA691453. The RNA-seq data used for genome annotation in this study were deposited in NCBI under accession number: PRJNA558194 and PRJNA694186. The resequencing data used for selection signature analysis were deposited in NCBI under accession number: PRJNA628752.

Author contributions

QW, JW and HZ collected samples. WW and YZ extracted DNA and RNA for sequencing. QW, XR and YH carried out genome assembly and assessment, and performed gene prediction and annotation. PL, JTL and JJL carried out comparative genome analysis. QW and JL performed selection signature analysis. JL and YH conceived this project. QW and XR wrote the manuscript. All authors contributed to the final manuscript editing.

Tables and Figures

Tables

Table 1 General information regarding *Fenneropenaeus chinensis* genome assembly

Genome survey	
Genome size (Mb)	1384.88
Heterozygosity	1.04%
Repetive sequence proportion	54.79%
PacBio assembly	
Total length (Mb)	1,465.32
GC content	37.53%
Contig number	9,015
Max length (bp)	3,793,399
N50 (bp)	472,841
N90 (bp)	77,864
Hi-C assembly	
Total length (Mb)	1,466.12
Scaffold number	1,063
Max length (bp)	48,777,264
N50 (bp)	36,870,704
N90 (bp)	24,342,607
Pseudochromosome number (2n)	86
unplaced scaffold (Mb)	14.60
Place rate	99.00%
Annotation	
Repetitive sequence proportion	57.73%
Total gene number	25,026
Average length (bp)	11,290
Annotated gene number	19,192

Table 2 Genome comparison among four penaeid shrimp species

Species	Genome size (K-mer analysis)	Chromosome number (2n)	Protein-coding gene number	Repetitive seque
*L. vannam*ei	2.60 Gb	88	25,596	78% (K-mer); 4
P. monodon	2.59 Gb	88	30,038	50.92% (K-mer)
M. japonicas	2.28 Gb	86	\	47.73%
F. chinensis	1.38 Gb	86	25,026	54.79% (K-mer)

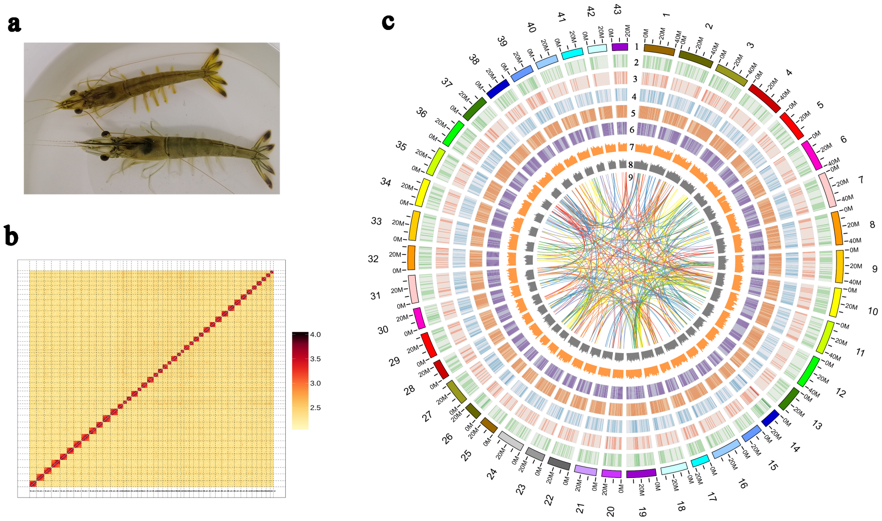
Figure Legends

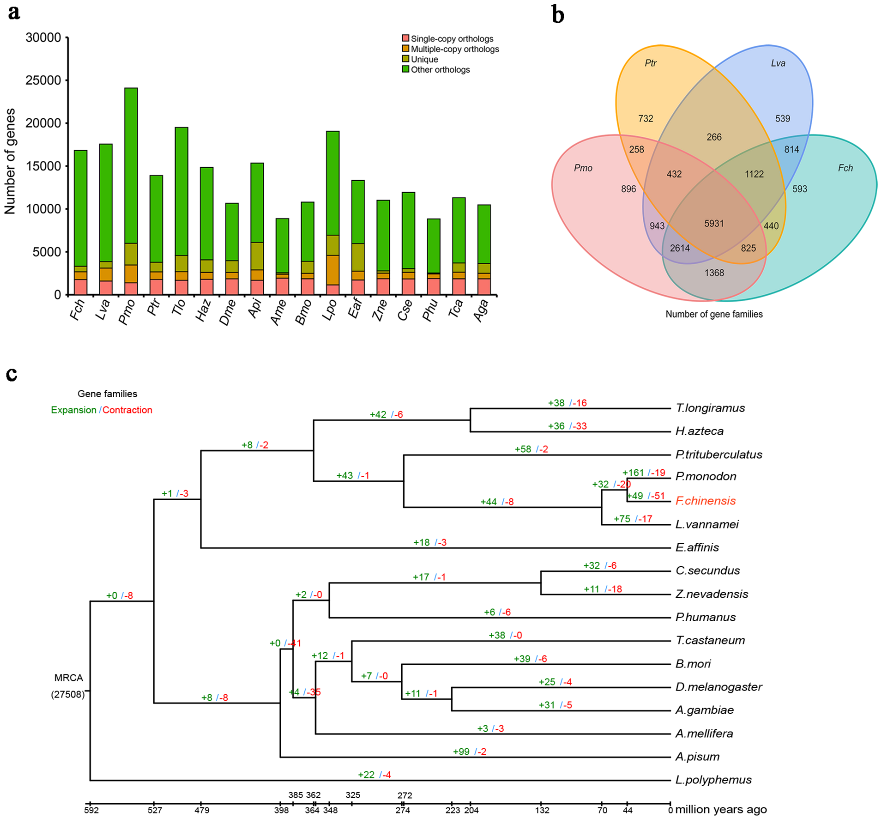
Figure 1. Basic information of *Fenneropenaeus chinensis.*(a) *F. chinensis.* Male above and female below, both are sexual maturity. (b) Heatmap of anchored chromosomes (c) Genomic characteristics of *F. chinensis* . Track 1 (from the outer-ring): 43 pseudochromosomes; Track 2: Distribution of gene density with sliding windows of 1 Mb. Higher density is indicated by darker color (the same below). Track 3: Distribution of genes on the forward strand. Track 4: Distribution of genes on the reverse strand. Track 5: Distribution of single nucleotide polymorphism (SNP) density (based on resequencing of 42 shrimps).

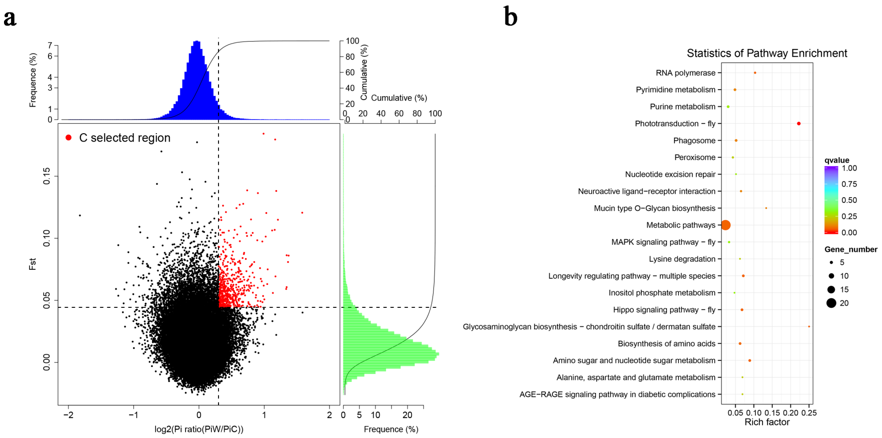
Track 6: Distribution of simple sequence repeats (SSRs). Track 7: Distribution of GC content (only values between 20% and 50% are displayed). Track 8: Distribution of proportion of repetitive sequences. Track 9: Schematic representation of interchromosomal relationships in the genome.

Figure 2. Comparative genomic analysis of *Fenneropenaeus chinensis* and homologous species. (a) The distribution of genes in different species. The horizontal axis represents 17 species, and the vertical axis represents the number of genes. (b) Common and unique gene families among four homologous species. (c) Phylogenetic tree of 17 species. Green numbers indicate the number of expanded gene families, and red numbers indicate the number of contracted gene families.

Figure 3. Selective sweeps on the genome of domestic*Fenneropenaeus chinensis*. (a) Conjoint analysis ofF_{ST} and $\vartheta\pi$· the red dots indicate the top 5% of rank level values. (b) Results of the KEGG enrichment analysis of the genes under artificial selection pressure in *F. chinensis* .




a
Number of genes
30000
25000
20000
15000
10000
5000
0
🟥 Single-copy orthologs
🟧 Multiple-copy orthologs
🟨 Unique
🟩 Other orthologs
Fch
Lva
Pmo
Ptr
Tlo
Haz
Dme
Api
Ame
Bmo
Lpo
Eaf
Zne
Cse
Phu
Tca
Aga
b
Ptr
Lva
Pmo
Fch
732
539
266
258
814
432
1122
896
593
943
5931
440
2614
825
1368
Number of gene families
c
Gene families
Expansion /Contraction
+38 /-16
T.longiramus
+36 /-33
H.azteca
+42 /-6
+8 /-2
+58 /-2
P.trituberculatus
+43 /-1
+161 /-19
P.monodon
+32 /-20
+49 /-51
F.chinensis
+44 /-8
+75 /-17
L.vannamei
+1 /-3
+18 /-3
E.affinis
+32 /-6
C.secundus
+17 /-1
+11 /-18
Z.nevadensis
+0 /-8
+2 /-0
+6 /-6
P.humanus
+38 /-0
T.castaneum
+0 /-41
+12 /-1
+39 /-6
B.mori
+7 /-0
+25 /-4
D.melanogaster
MRCA
(27508)
+8 /-8
+4 /-35
+11 /-1
+31 /-5
A.gambiae
+3 /-3
A.mellifera
+99 /-2
A.pisum
+22 /-4
L.polyphemus
385 362
325
272
592
527
479
398
364 348
274
223 204
132
70
44
0 million years ago


a
Frequence (%)
7
6
4
3
1
0
100
80
60
40
20
0
Cumulative (%)
Cumulative (%)
0 20 40 60 80 100
● C selected region
Fst
0.15
0.10
0.05
0.00
−2
−1
0
1
2
log2(Pi ratio(PiW/PiC))
0 5 10 20
Frequence (%)
b
Statistics of Pathway Enrichment
RNA polymerase
Pyrimidine metabolism
Purine metabolism
Phototransduction − fly
Phagosome
Peroxisome
Nucleotide excision repair
Neuroactive ligand−receptor interaction
Mucin type O−Glycan biosynthesis
Metabolic pathways
MAPK signaling pathway − fly
Lysine degradation
Longevity regulating pathway − multiple species
Inositol phosphate metabolism
Hippo signaling pathway − fly
Glycosaminoglycan biosynthesis − chondroitin sulfate / dermatan sulfate
Biosynthesis of amino acids
Amino sugar and nucleotide sugar metabolism
Alanine, aspartate and glutamate metabolism
AGE−RAGE signaling pathway in diabetic complications
0.05 0.10 0.15 0.20 0.25
Rich factor
qvalue
1.00
0.75
0.50
0.25
0.00
Gene_number
5
10
15
20