

DATA NOTE

Draft genome of the lined seahorse, *Hippocampus erectus*

Qiang Lin^{1,†}, Ying Qiu^{2,3,†}, Ruobo Gu^{2,3,4,†}, Meng Xu^{5,†}, Jia Li^{3,†},
Chao Bian^{3,6,7,†}, Huixian Zhang¹, Geng Qin¹, Yanhong Zhang¹, Wei Luo¹,
Jieming Chen³, Xinxin You^{3,6}, Mingjun Fan³, Min Sun³, Pao Xu^{2,6},
Byrappa Venkatesh⁸, Junming Xu^{3,4,6,*}, Hongtuo Fu^{2,6,*} and Qiong Shi^{3,4,6,9,*}

¹CAS Key Laboratory of Tropical Marine Bio-resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, Guangdong 510301, China, ²Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi, Jiangsu 214081, China, ³Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Fisheries, BGI, Shenzhen, Guangdong 518083, China, ⁴BGI Zhenjiang Institute of Hydrobiology, BGI Fisheries, Zhenjiang, Jiangsu 212000, China, ⁵BGI-Shenzhen, BGI, Shenzhen, Guangdong 518083, China, ⁶BGI Research Center for Aquatic Genomics, Chinese Academy of Fishery Sciences, Shenzhen, Guangdong 518083, China, ⁷Centre of Reproduction, Development and Aging, Faculty of Health Sciences, University of Macau, Taipa, Macau, China, ⁸Institute of Molecular and Cell Biology, A*STAR, Biopolis, 138673, Singapore and ⁹Laboratory of Aquatic Genomics, College of Ecology and Evolution, School of Life Sciences, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China

*Correspondence address. Qiong Shi, Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI Fisheries, BGI, Shenzhen, Guangdong 518083, China (tel: +86 185 6627 9826; e-mail: shiqiong@genomics.cn); Hongtuo Fu, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi 214081, China (tel: +86 136 0148 0163; e-mail: fuht@ffrc.cn); Junming Xu, BGI Zhenjiang Institute of Hydrobiology, BGI Fisheries, Zhenjiang, Jiangsu 212000, China (tel: +86 139 5128 9388; e-mail: xujunming@genomics.cn)

[†]Contributed equally to this work.

Abstract

Background: The lined seahorse, *Hippocampus erectus*, is an Atlantic species and mainly inhabits shallow sea beds or coral reefs. It has become very popular in China for its wide use in traditional Chinese medicine. In order to improve the aquaculture yield of this valuable fish species, we are trying to develop genomic resources for assistant selection in genetic breeding. Here, we provide whole genome sequencing, assembly, and gene annotation of the lined seahorse, which can enrich genome resource and further application for its molecular breeding. **Findings:** A total of 174.6 Gb (Gigabase) raw DNA sequences were generated by the Illumina Hiseq2500 platform. The final assembly of the lined seahorse genome is around 458 Mb, representing 94% of the estimated genome size (489 Mb by k-mer analysis). The contig N50 and scaffold N50 reached

Received: 7 November 2016; Revised: 8 March 2017; Accepted: 17 April 2017

© The Author 2017. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

14.57 kb and 1.97 Mb, respectively. Quality of the assembled genome was assessed by BUSCO with prediction of 85% of the known vertebrate genes and evaluated using the *de novo* assembled RNA-seq transcripts to prove a high mapping ratio (more than 99% transcripts could be mapped to the assembly). Using homology-based, *de novo* and transcriptome-based prediction methods, we predicted 20 788 protein-coding genes in the generated assembly, which is less than our previously reported gene number (23 458) of the tiger tail seahorse (*H. comes*). **Conclusion:** We report a draft genome of the lined seahorse. These generated genomic data are going to enrich genome resource of this economically important fish, and also provide insights into the genetic mechanisms of its iconic morphology and male pregnancy behavior.

Keywords: genome; assembly; annotation; *Hippocampus erectus*

Data Description

Background

Syngnathidae, an interesting teleost family, exhibit special morphological innovations and reproductive behavior, and these phenotypes have come into being through long-term molecular evolution [1, 2]. Seahorses (*Hippocampinae*) are a group of popular and iconic species because of their unique body plan and male pregnancy. As an interesting model, seahorses could provide exceptional clues for studying evolution in virtue of their closed brood pouch, male pregnancy, and seasonal migration [3, 4]. Recently, we have reported whole genome sequence of the tiger tail seahorse (*Hippocampus comes*) [5] and provided primary insights into the genetic basis of its iconic morphology. The work also dealt with a number of fascinating areas, such as the *patristacin* subfamily of astacin metalloproteases that may be closely related to the unusual male pregnancy in this species since they were expanded and highly expressed in the male brood pouch during mid- and late pregnancy [5].

Here, we provide a draft genome of the lined seahorse (*H. erectus*; Fig. 1), which inhabits coastal waters in the Western Atlantic such as Nova Scotia, Canada, and northern Gulf of Mexico to Panama and Venezuela [6]. It has been treated as vulnerable or endangered in the Red List of Threatened Species (IUCN 2015) [7]. Moreover, the lined seahorse is easily domesticated for breeding, and it has become a popular and commercially important ingredient for traditional Chinese medicine in China [8–12]. In order to study the evolutionary history of the lined seahorse and improve its aquaculture yield, we are trying to develop genomic resources for assisted selection in genetic breeding. Hence, we performed whole genome sequencing, assembly, and gene annotation of the lined seahorse, which should facilitate further studies on species conservation and molecular breeding of this economically important fish.

Preparation and sequencing of DNA samples

Genomic DNA was extracted from a pool of four male lined seahorses (NCBI Taxonomy ID: 109281; Fishbase ID: 3283). All animal experiments conformed to the guidelines of the Animal Ethics Committee and were approved by the Institutional Review Board on Bioethics and Biosafety of BGI (approval ID: FT16091). Seven libraries, including three short-insert libraries (200, 500, and 800 bp) and four long-insert libraries (2, 5, 10, and 20 kb), were constructed based on the standard protocol of Illumina (CA, USA) and sequenced using the Illumina HiSeq2500 platform (the read length is 125 bp). Finally, we generated a total of 174.6-Gb raw sequences.

Processing of the raw sequencing reads

These raw sequences contained some sequencing errors, which may reduce the quality of genome assembly. Hence we filtered



Figure 1: Photo of a cultivated line seahorse in Shenzhen, China.

these raw sequences with the following stringent filtering processes through SOAPfilter (v. 2.2) software [13]: (i) filtered reads with 40% low-quality bases (quality scores ≤ 7); (ii) removed reads with N bases more than 10%; (iii) trimmed reads with five low-quality bases at the 5' end; (iv) discarded reads with adapter contamination and/or PCR duplicates; (v) corrected raw reads from the short-insert libraries based on the k-mer spectrum. Finally, we obtained 111.3 Gb of clean reads in total, in which 12.0, 14.5, 13.7, 18.0, 15.9, 18.9, and 18.3 Gb were kept from the seven sequencing libraries (from 200 bp to 20 kb), respectively.

Estimation of the genome size and assembly of the genome sequences

The genome size was estimated based on the k-mer spectrum [14] with the following formula: $G = \text{k-mer_number} / \text{k-mer_depth}$, where G is the genome size, k-mer_number is the total number of k-mer, and k-mer_depth means the peak frequency that was higher than any other frequencies. For the lined seahorse, the k-mer_number is 24 445 959 200 (based on 17-mer), and the k-mer_depth is 50. Therefore, the genome size was estimated to be approximately 489 Mb, which is much smaller than our estimation (695 Mb) for the tiger tail seahorse [5].

The generated clean reads were further assembled by SOAPdenovo2 (v. 2.04) [15] with optimized parameters (pregraph -K 27 -d 1; contig -M 1; scaff -b 1.5) to construct contigs and original scaffolds. Subsequently the gaps in the intra-scaffolds were filled using the reads of short-insert libraries by GapCloser 1.12 [13]. Finally, the achieved total scaffold length reached up to 457 759 912 bp with 2.8% gaps (12.8 Mb), which is smaller than that of the reported tiger tail seahorse (501 592 652 bp) [5]. The calculated scaffold N50 and contig N50 are 1.97 Mb and 14.57 kb,

Table 1: Comparison of genome assembly and annotation between the lined seahorse and the reported tiger tail seahorse

| Genome Assembly | Lined Seahorse | Tiger Tail Seahorse |
|------------------------------------|-----------------|---------------------|
| Contig N50 size (kb) | 14.57 | 34.67 |
| Scaffold N50 size (Mb) | 1.97 | 1.87 |
| Estimated genome size (Mb) | 489 | 695 |
| Assembled genome size (Mb) | 457.76 | 501.59 |
| Genome coverage (×) | 243.05 | 192.05 |
| Longest scaffold (bp) | 7 855 128 | 9 810 584 |
| Genome annotation | | |
| Protein-coding gene number | 20 788 | 23 458 |
| Annotated functional gene number | 18 776 (90.32%) | 22 245 (94.83%) |
| Unannotated functional gene number | 2012 (9.68%) | 1213 (5.17%) |
| Transposable elements content | 28.1% | 24.8% |

Table 2: Assessment of the completeness of the lined seahorse genome using transcriptome data

| Dataset | Number | Total Length (bp) | Base Covered by Assembly (%) | Sequence Covered by Assembly (%) | With >90% Sequence in 1 Scaffold | | With >50% Sequence in 1 Scaffold | |
|----------|--------|-------------------|------------------------------|----------------------------------|----------------------------------|---------|----------------------------------|---------|
| | | | | | Number | Percent | Number | Percent |
| All | 71 765 | 52 877 091 | 98.22 | 99.52 | 68 292 | 95.16 | 71 255 | 99.29 |
| >200 bp | 71 765 | 52 877 091 | 98.22 | 99.52 | 68 292 | 95.16 | 71 255 | 99.29 |
| >500 bp | 29 811 | 40 111 717 | 98.12 | 99.68 | 27 902 | 93.60 | 29 640 | 99.43 |
| >1000 bp | 14 780 | 29 612 539 | 97.92 | 99.70 | 13 561 | 91.75 | 14 686 | 99.36 |

respectively (Table 1), which are comparable with the values from the tiger tail seahorse (see more details about the comparison in Table 1)[5].

Assessment of genome completeness

Benchmarking Universal Single-Copy Orthologs (BUSCO) [16] is a software that can be used to evaluate the completeness of a genome assembly by genes selected from appropriate lineage-specific orthologous groups. For the lined seahorse, the analysis data proved that our assembly contains 73% complete and 12% partial sequences of vertebrate BUSCO orthologues (3023 genes in total).

Simultaneous completeness of the lined seahorse genome was also evaluated using the *de novo* assembled RNA-seq transcripts from different developmental stages of the lined seahorse (downloaded from our recent paper [5]) to map the lined seahorse genome assembly with Blat [17]. All the results showed that more than 99% of transcripts could be mapped to the assembly (Table 2), suggesting that our assembly is of high quality.

Repeat analysis

Tandem repeats were searched in the generated genome assembly by utilizing Tandem Repeats Finder (v. 4.04) [18]. Transposable elements (TEs) were identified with an approach that combined both homology-based and *de novo* predictions. First, RepeatMask (v. 3.3.0) [19] was employed to detect known TEs based on a homologous search against the Repbase TE library (release 17.01) [20]. RepeatProteinMask (v. 3.3.0) [19], an updated software included in the RepeatMasker package, was used to identify the TE relevant proteins. Subsequently, LTR.FINDER [21] and RepeatModeler (v. 1.05) [22] were used with the default parameters to construct the *de novo* repeat library. Then we used RepeatMask [19] to identify and classify novel TEs against this *de novo* repeat library. All the repeats were finally combined to-

gether with a filtering of those redundant repetitive sequences. In total, the lined seahorse genome comprises approximately 30.43% repetitive sequences, in which 28.12% are TEs. Interestingly, the most abundant type of TE is class II DNA transposon, which covered around 15% of the genome. Our data are similar to the report of the tiger tail seahorse [5], in which 24.82% are TEs, with class II DNA transposon as the most abundant.

Gene annotation

De novo prediction

Repetitive regions in the genome sequence were replaced with “N” to reduce the ratio of pseudogene annotations. Then we chose 1000 full-length but randomly selected genes from zebrafish homology gene set to train the model parameters for AUGUSTUS. We subsequently employed AUGUSTUS 3.0.1 [23] and GenScan 1.0 [24] for *de novo* prediction of repeat-masked genome sequences. Short genes (less than 150 bp) and premature or frame-shifted genes were removed.

Homology-based annotation

Protein sequences of zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), fugu (*Takifugu rubripes*), stickleback (*Gasterosteus aculeatus*), and Nile tilapia (*Oreochromis niloticus*) were downloaded from Ensembl (release 83) [25]. Protein sequences of the tiger tail seahorse (*H. comes*) were downloaded from our recently published genome data (Bioproject ID: PRJNA314292) [5]. Protein sets of these species were mapped to the assembled lined seahorse genome using tBlastn (v. 2.2.19) [26] with E-value $\leq 1e-5$. GeneWise (v. 2.2.0) [27] was applied to refine the potential gene models of all alignments. Ultimately, we filtered short genes (less than 150 bp) and premature or frame-shifted genes.

Transcriptome-based prediction

We downloaded the transcriptome data of the lined seahorse from our previous work [10]. The raw reads were mapped onto

the genome using TopHat (v. 2.0) [28] with the default parameters and assembled into transcripts using Cufflinks [29].

Gene set integration and optimization

The gene models based on *de novo* prediction, homology-based annotation, and transcriptome-based prediction were merged to form a comprehensive and non-redundant gene set using GLEAN [30]. Finally, we obtained a gene set containing 20 788 genes, which is less than the reported gene number (23 458) of the tiger tail seahorse [5].

Annotation of *patristacin* gene family

The *patristacin* subfamily of the astacin metalloprotease family may be closely related to the unusual male pregnancy in seahorses since we identified six *patristacin* genes in the tiger tail seahorse and confirmed their expansion and high expression in the male brood pouch [5]. We also analyzed *patristacin* in the lined seahorse genome. Related *patristacin* protein sequences were downloaded from the tiger tail seahorse genome data [5] and used for homology searches against the lined seahorse genome using tBlastn (v. 2.2.19) [26]. We chose alignments with coverage >50% and identity >50% and then used Genewise (v. 2.2.0) [27] to predict the gene structures. We also downloaded the RNA-seq data at the pregnancy stage of the male lined seahorse from our recently published paper [10] to confirm the existence of the six *patristacin* genes in the lined seahorse. The RNA-seq reads were mapped by TopHat [28], and gene expression levels were measured by RPKM (reads per kilobases per million reads). Finally, we observed that all the six *patristacin* genes were expressed during pregnancy in the male lined seahorse.

Functional assignment

The protein sequences predicted from the lined seahorse genome were aligned to the Swiss-Prot and TrEMBL databases [31] using BlastP at E-value $\leq 1e-5$. The motifs and domains were annotated using InterProScan [32] by searching publicly available databases including Pfam [33], ProDom [34], SMART [35], PRINTS [36], and PANTHER [37], and then we retrieved Gene Ontology (GO) [38] annotation from the results of InterProScan. The gene pathways were assigned based on the best blast hit against the KEGG database [39]. In summary, approximately 90.32% of the genes are supported by at least one related function from the searched databases (Swiss-Prot, Interpro, TrEMBL, and KEGG).

Construction of gene families

Protein sequences of seven ray-fin fishes, including zebrafish, medaka, fugu, stickleback, Nile tilapia, platyfish (*Xiphophorus maculatus*), and spotted gar (*Lepisosteus oculatus*), were downloaded from Ensembl (release 83) [25]. Protein sequences of the tiger tail seahorse (*H. comes*) were downloaded from our recently published genome data [5]. Protein sequences of Gulf pipefish (*Syngnathus scovelli*) were downloaded from the Cresko Lab web server (<http://creskolab.uoregon.edu>) [40]. The consensus proteome set of the above nine species and the lined seahorse were composed of a final data set of 209 747 protein sequences. Finally, we used OrthoMCL [41] to cluster gene families and obtained 19 053 OrthoMCL families with all-to-all BLASTP strategy (E-value $\leq 1e-5$) and a Markov Chain Clustering (MCL) default inflation parameter.

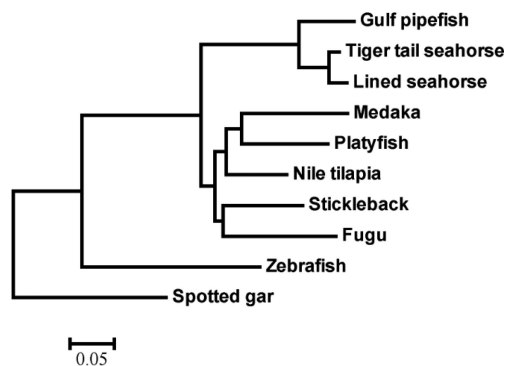


Figure 2: Phylogeny of ray-finned fishes. The Spotted gar was used as the out-group species. See more details of the protein sequence sources in the main context.

Phylogenetic analysis

We extracted 2812 one-to-one orthologous genes from the above-mentioned gene family set. The protein sequences of each selected family were aligned using MUSCLE (v. 3.8.31) [42] with the default parameters. The protein alignments were then converted to corresponding coding sequences (CDS) using an in-house Perl script. All these nucleotide sequences were concatenated into a supergene for each species, all of which were used to construct a phylogenetic tree using PhyML (Fig. 2) [43].

Conclusions

Seahorses are a fascinating teleost group with special morphological innovations and reproductive behavior. In our previous genome paper about the tiger tail seahorse [5], we paid much attention to the genetic bases of their unique morphology and reproductive system. However, besides the spectacular aspects of the phenotype, seahorses have been very popular in traditional Chinese medicine. Here we report the first draft genome assembly of the lined seahorse, an economically important aquaculture fish in China. With availability of these genomic data, we can develop genetic markers for construction of a high-density genetic linkage map, and subsequently further genetic selection and molecular breeding in the future. These works will support a significant increase of the aquaculture yield, which could produce remarkable economic benefits and realize the ecological protection of seahorses in the world. Our genome data will also facilitate the genetic mechanism study and evolutionary history analysis of the lined seahorse.

Availability of supporting data

Supporting data are available in the GigaScience database [44], and the raw data have been deposited in NCBI with the project accession PRJNA347499.

Acknowledgements

This work was supported by the Youth Foundation of National High Technology Research and Development Program (2015AA020909), the Outstanding Youth Foundation in Guangdong Province (S2013050014802), the Special Fund for Agroscientific Research in the Public Interest (201403008), the

National Natural Science Foundation of China (41576145), China National Natural Science Foundation (No. 31370047), Shenzhen Special Program for Future Industrial Development (No. JSGG20141020113728803), Special Project on the Integration of Industry, Education and Research of Guangdong Province (No. 2013B090800017), Shenzhen Science and Technology Program (No. SGLH20131010105856414 and GJHZ20160229173052805), and Shenzhen Dapeng Special Program for Industrial Development (No. KY20160307).

Conflicts of interest

The authors declare that they have no competing interests.

Author's contributions

Q.S. and Q.L. designed the project. J.C., X.Y., M.F., and M.S. collected the samples and prepared the quality control. Y.Q., M.X., J.L., and C.B. assembled and annotated the genome. Y.Z., H.Z., G.Q., and W.L. were involved in the data analysis. Y.Q., Q.S., C.B., Q.L., P.X., and R.G. wrote the manuscript. J.X., H.F., B.V., and Q.S. participated in discussions and provided advice. All authors read and approved the final manuscript.

References

- Wilson AB, Vincent A, Ahnesjo I, et al. Male pregnancy in seahorses and pipefishes (family Syngnathidae): rapid diversification of paternal brood pouch morphology inferred from a molecular phylogeny. *J Hered* 2001;**92**(2):159–66.
- Stolting KN, Wilson AB. Male pregnancy in seahorses and pipefish: beyond the mammalian model. *Bioessays* 2007;**29**(9):884–96.
- Harlin-Cognato A, Hoffman EA, Jones AG. Gene cooption without duplication during the evolution of a male-pregnancy gene in pipefish. *Proc Natl Acad Sci U S A* 2006;**103**(51):19407–12.
- Foster S, Vincent A. Life history and ecology of seahorses: implications for conservation and management. *J Fish Biol* 2004;**65**(1):1–61.
- Lin Q, Fan S, Zhang Y et al. The seahorse genome provides insights into the evolution of its iconic body plan and male pregnancy. *Nature* 2016;**540**:395–99.
- Abe F, Akimoto H, Akopian A et al. Observation of top quark production in p p collisions with the collider detector at fermilab. *Phys Rev Lett* 1995;**74**(14):2626.
- Cardoso P, Stoev P, Georgiev T et al. Species Conservation Profiles compliant with the IUCN Red List of Threatened Species. *Biodivers Data J* 2016(4):e10356.
- Lin Q, Zhang D, Lin J. Effects of light intensity, stocking density, feeding frequency and salinity on the growth of sub-adult seahorses *Hippocampus erectus* Perry, 1810. *Aquaculture* 2009;**292**(1):111–16.
- Lin Q, Lin J, Huang L. Effects of substrate color, light intensity and temperature on survival and skin color change of juvenile seahorses, *Hippocampus erectus* Perry, 1810. *Aquaculture* 2009;**298**(1):157–61.
- Lin Q, Luo W, Wan S et al. De novo transcriptome analysis of two seahorse species (*Hippocampus erectus* and *H. mohnikei*) and the development of molecular markers for population genetics. *PLoS One* 2016;**11**(4):e0154096.
- Qin G, Zhang Y, Huang L et al. Effects of water current on swimming performance, ventilation frequency, and feeding behavior of young seahorses (*Hippocampus erectus*). *J Exp Marine Biol Ecol* 2014;**461**:337–43.
- Wang X, Zhang Y, Qin G et al. A novel pathogenic bacteria (*Vibrio fortis*) causing enteritis in cultured seahorses, *Hippocampus erectus* Perry, 1810. *J Fish Dis* 2016;**39**:765–69.
- Li R, Yu C, Li Y et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009;**25**(15):1966–67.
- Liu B, Shi Y, Yuan J et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint 2013; arXiv:1308.2012.
- Luo R, Liu B, Xie Y et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Genomics* 2012;**1**(1):18.
- Simao FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**(19):3210–12.
- Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res* 2002;**12**(4):656–64.
- Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999;**27**(2):573–80.
- Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2009; chapter 4: unit 4 10.
- Jurka J, Kapitonov VV, Pavlicek A et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005;**110**(1–4):462–67.
- Xu Z, Wang H. LTR FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* 2007;**35**(web server issue):W265–68.
- Abrusan G, Grundmann N, DeMester L et al. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 2009;**25**(10):1329–30.
- Stanke M, Keller O, Gunduz I et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 2006;**34**(web server issue):W435–39.
- Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997;**268**(1):78–94.
- Cunningham F, Amode MR, Barrell D et al. Ensembl 2015. *Nucleic Acids Res* 2015;**43**(database issue):D662–69.
- Mount DW. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc* 2007; 2007: pdb top17.
- Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res* 2004;**14**(5):988–95.
- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 2009;**25**(9):1105–11.
- Trapnell C, Williams BA, Pertea G et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**(5):511–15.
- Elsik CG, Mackey AJ, Reese JT et al. Creating a honey bee consensus gene set. *Genome Biol* 2007;**8**(1):R13.
- Boeckmann B, Bairoch A, Apweiler R et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;**31**(1):365–70.
- Hunter S, Apweiler R, Attwood TK et al. InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009;**37**(database issue):D211–15.
- Finn RD, Bateman A, Clements J et al. Pfam: the protein families database. *Nucleic Acids Res* 2014;**42**(database issue):D222–30.
- Bru C, Courcelle E, Carrere S et al. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* 2005;**33**(database issue):D212–15.

35. Letunic I, Copley RR, Schmidt S et al. SMART 4.0: towards genomic data integration. *Nucleic Acids Res* 2004;**32**(database issue):D142–44.
36. Attwood TK. The PRINTS database: a resource for identification of protein families. *Brief Bioinform* 2002;**3**(3): 252–63.
37. Thomas PD, Kejariwal A, Campbell MJ et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* 2003;**31**(1): 334–41.
38. Ashburner M, Ball CA, Blake JA et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–29.
39. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia Of Genes and Genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.
40. Small CM, Bassham S, Catchen J et al. The genome of the Gulf pipefish enables understanding of evolutionary innovations. *Genome Biol* 2016;**17**(1):258.
41. Li L, Stoeckert CJ, Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**(9):2178–89.
42. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**(5):1792–97.
43. Guindon S, Dufayard J-F, Lefort V et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010;**59**(3):307–21.
44. Lin Q, Qiu Y, Gu R et al. Supporting data for “Draft genome of the lined seahorse, *Hippocampus erectus*.” *GigaScience DB* 2017; <http://dx.doi.org/10.5524/100298>.