# Draft genome of a porcupinefish, *Diodon Holocanthus*

Mengyang Xu[1,2,3,*], Xiaoshan Su[1,2,3,*], Mengqi Zhang[1,2,3,*], Ming Li[1,4,5,*], Xiaoyun Huang[1,2,3,*], Guangyi Fan[1,2,3] , Xin Liu[1,2,3,6] , He Zhang[1,2,3,#]

[1]BGI-Qingdao, BGI-Shenzhen, Qingdao, 266555, China;

[2]BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China;

[3]China National Gene Bank, BGI-Shenzhen, Jinsha Road, Shenzhen 518120, China;

[4]College of Fisheries and Life Science, Shanghai Ocean University, Shanghai 201306, China

[5]Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences (CAFS), Key Laboratory for Sustainable Development of Marine Fisheries, Ministry of Agriculture, Qingdao 266071, China;

[6]James D. Watson Institute of Genome Science, 310008 Hangzhou, China;

*These authors contributed equally to this work

#This author supervised this work

Corresponding to zhanghe@genomics.cn (H.Z.)

## Abstract

The long-spine porcupinefish, *Diodon holocanthus* (Diodontidae, Tetraodontiformes, Actinopterygii), also known as the freckled porcupinefish, attracts great interest of ecology and economy. Its distinct characteristics including inflation reaction, spiny skin and tetradotoxin, however, have not been fully studied without a complete genome assembly.

In this study, the whole genome of a single individual was sequenced using single tube-Long Fragment Read co-barcode reads, generating 154.3 Gb of paired-end data (219.8× depth). The gap was further filled using small amount of Oxford Nanopore MinION long read dataset (11.4Gb, 15.9× depth). Taking full use of long, medium, short-range of genome assembly information, the final assembled sequences with a total length of 650.02 Mb obtained contig and scaffold N50 sizes of 2.15 Mb and 8.13 Mb, respectively, despite of high repetitive content. Benchmarking Universal Single-Copy Orthologs captured 95.7% (2,474) of core genes to assess the completeness. In addition, 206.5 Mb (32.10%) of repetitive sequences were identified, and 20,840 protein-coding genes were annotated, among which 18,281 (87.72%) proteins were assigned with possible functions.

This is the first demonstration of *de novo* genome of the porcupinefish, which will benefit downstream analysis of ontogeny, phylogeny, and evolution, and improve the exploration of its unique defensive mechanism.

## I Introduction

Pufferfish are small vertebrate fish in the order Tetraodontiformes. As indicated by their name, most pufferfish can inflate themselves, obviously expanding in size in an effort to fend off predators. This inflation ability is found in all members of the sister taxa Tetraodontidae and Diodontidae [1,2]. A small number of pufferfish species, porcupinefish, outward spines, as an additional defensive adaptation. Among them, *Diodon holocanthus* is widely distribute in the tropical and warm waters of the Pacific, the Atlantic, and the Indian Ocean. *D. holocanthus* is a demersal fish species which can reproduce several times a year [3] and is of significant economic importance.

The inflation mechanism, genome size variation, and genome evolution of pufferfish has been investigated [1,2,4,5]. The interrelationships and phylogeny of porcupinefish has been studied using single nuclear-encoded genes (e.g. *RAG1*) or mitogenomes [6-8]. However, the lack of a sequenced whole genome hampered the attempts to explore the key role in adaption and evolution as well as its unique skin character and defensive mechanism.

In this study, we sequenced *Diodon holocanthus* genome using single tube-Long Fragment Read co-barcoded (stLFR) technology, reported the assembly, repeat and gene annotation for the first porcupinefish. The final draft genome assembly was approximately 650.02 Mb with a contig N50 of 2.15 Mb and scaffold N50 of 8.13 Mb. A total of 20,840 protein-coding genes were predicted from the genome assembly. The determination of genomic resource of the porcupinefish will be of significance to improve the understanding of its unique morphological and physiological characteristics.

## II Methods

### Sample collection, library construction and genome sequencing

A single *Diodon holocanthus* fish (Figure 1) was purchased from a seafood market at Xiamen, Fujian province, southeast China. Genomic DNA was extracted from the muscle tissue using a conventional method for high molecular weight DNA [9] (liquid nitrogen grinding, followed by phenol-chloroform-Isoamyl alcohol extraction and ethanol precipitation). A stLFR library [10] was prepared by using the MGIEasy stLFR Library Prep kit (PN: 1000005622) and sequenced on BGISEQ-500 platform.

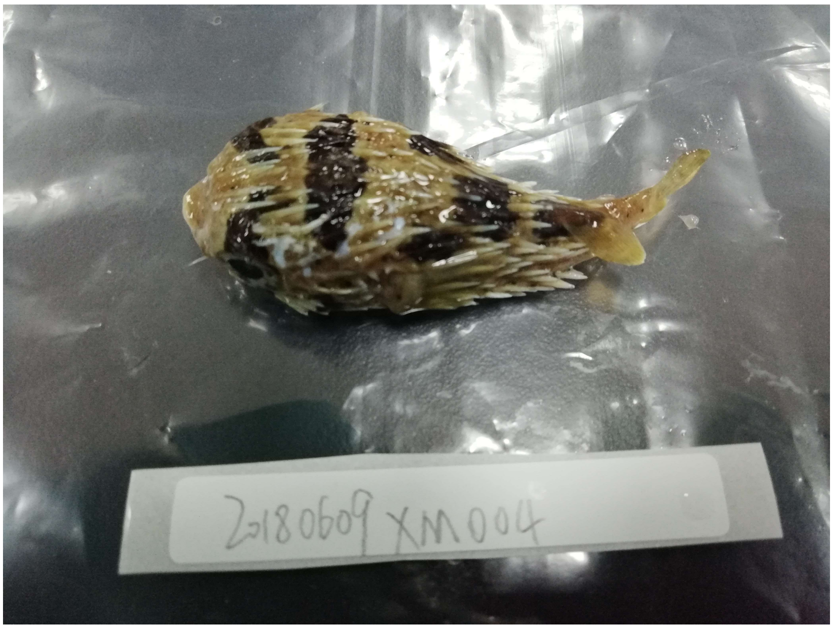

**Figure 1. Photograph of *Diodon holocanthus. (Credit: M.Z.)***

A total of 637.6 million 100bp paired-end stLFR reads were generated, representing 154.3 Gb of nucleotide sequences, with 77.4% bases ≥ Q30. Raw data was filtered using SOAPfilter software (version 2.2) [11] to obtain 93.6 Gb clean reads excluding the adaptor sequences, reads containing more than 10% unidentified nucleotides, and low-quality reads containing more than 50% bases with Phred quality score ≤10, and reads whose sequences are exactly identical, that is, PCR duplications.

A total of 11.4 Gb of sequence data, representing 15.9× theoretical coverage, were produced using the MinION (Oxford Nanopore Technologies,ONT) nanopore sequencer. The average and N50 read length are 17.3Kb and 23.6Kb, respectively.

## Genome features revealed by k-mer analysis

The high-quality clean reads from the short-insert reads (250 bp) were used to estimate the genomic information of *Diodon holocanthus* and 17-mer frequency information was generated based on *k*-mer analysis using KMERFREQ_AR (version 2.0.4) [11] for estimation of genome size, heterozygosity, and repeat content. The calculated results provided by the software GEC (version 1.0.0) [12] show that the estimated haploid genome size was 701.96 Mb, with a repeat content of 36.35% and a heterozygosity level of 0.76% represented in the first peak of the distribution (Figure 2). The high level of repeat content indicated a troublesome genome characteristic for *de novo* assembly.

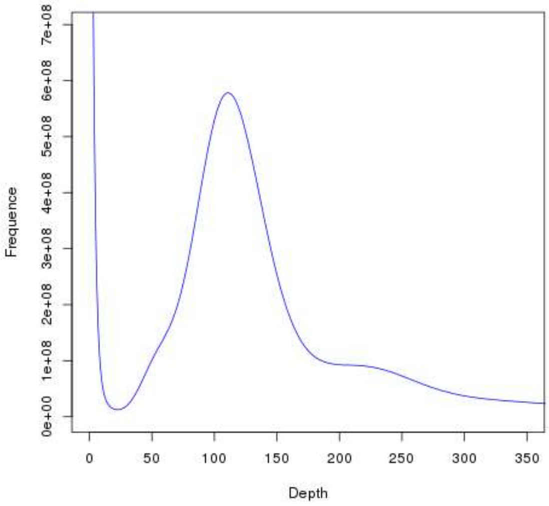

**Figure 2. Distribution of 17-mer frequency. In total 93.6 Gb of high-quality short insert reads (250 bp) were used to generate the 17-mer depth distribution curve frequency information.**

## *De novo* genome assembly

The stLFR data were split into two parts: paired-end 100bp short reads and their corresponding barcode information. Each series of 40-base barcode sequences refers to a unique stLFR barcode ID, a combination of 3 unique tube indices.

Supernova (version 2.1.1) [13] is a well-established *de novo* assembler based on a de Bruijn graph algorithm originally designed for barcoded data from 10X Genomics Chromium Linked-Reads. To allow the use of Supernova for stLFR sequencing reads, stLFR barcodes (over 10 million) were parsed (transformed) to generate barcodes compatible with the 10X Genomics format (4.7

million). Scripts are available on GitHub (https://github.com/BGI-Qingdao/stlfr2supernova_pipeline).

We assembled the *Diodon holocanthus* genome, using stLFR clean reads with transformed barcodes, with Supernova as follows: build a 48-mer DBG based on shared *k*-mers, map barcodes to the de Bruijn graph, use barcode information to scaffold, partition the graph and make local assembly, phase based on barcode information, close gap, and reuse barcode and copy number information to further scaffold. The Supernova assembled a genome with scaffold N50 of 6,098,089 bp and contig N50 of 70,841 bp.

To make full use of the diversity of the stLFR barcode information, SLR-superscaffolder (Version 0.9.0) [14] was further applied to improve the scaffolds. First, the short reads were aligned back to the Supernova draft scaffolds using BWA-MEM (version 0.7.17) [15], in order to distribute the barcode information to scaffolds. Then, according to the shared barcodes, scaffolds were further gathered together and clustered in longer super-scaffolds. In each scaffold group, the order of scaffolds was determined based on a Mean-Spanning-Tree (MST) algorithm, while the orientation of each scaffold was chosen by the Jaccard similarity score with its nth-order neighbors. Contigs that did not belong to any draft scaffolds were also filled in the super-scaffolds based on both barcode and PE information. The scaffold N50 and N90 were improved from 6,098,089 and 976,586 bp to 8,075,516 and 1,206,047 bp, respectively. Note that the re-scaffolding procedure had no effect on contig sequences, only orientation. GapCloser (Version 1.12) [11] were applied to reduce gap regions. Considering all contigs, this step enhanced the contig N50 and N90 sizes to 143,286 and 36,130 bp, respectively. Considering the high repetitive content, approximately 15.9× Oxford Nanopore MinION reads were used to overcome the gaps induced by repeats, which was impossible by short reads and kmer extension algorithm, thus further improving assembly quality by TGSGapFiller (Unpublished, available on GitHub https://github.com/BGI-Qingdao/TGSGapFiller). Using such low coverage of long reads, the final assembly of the porcupinefish genome reached a total length of 650.02 Mb, which is similar to the estimated size, with contig N50 of 2,149,931bp, and scaffold N50 of 8,129,349 bp.

Compared with other recently published tetraodontiforme genomes, the total size is almost twice of them, and even larger than two relative species, which is consistent with previous work [5]. Using stLFR dataset, along with only 15.9× ONT long reads, the contig N50 is comparable to that of those using massive expensive third-generation long reads although the scaffold N50 is relatively smaller due to the lack of BioNano or HiC long-range information [16]. Both contig and scaffold N50 are drastically larger than that of genomes previously assembled by short reads [17,18].

To assess the quality of the genome assembly, we performed the Benchmarking Universal Single-Copy Orthologs (BUSCO) software (Version 3.0.2) [19]. The genome was queried (options ‘−m geno −sp zebrafish’) against the “metazoa.odb9” lineage set containing 65 eukaryotic organisms to assess the coverage of core eukaryotic genes. Using the vertebrata_odb9 database, BUSCO analysis revealed that 95.7% (91.5% single-copy genes and 4.2% duplicates) of the expected vertebrate genes were complete.

| | *Diodon* | *Takifugu* | *Takifugu* | *Takifugu* | *Cynoglossus* | *Paralichthys* |
|---|---|---|---|---|---|---|
| | *Holocanthus* | *rubripes* | *bimaculatus* | *flavidus* | *semilaevis* | *olivaceus* |

| **Sequencing Technology** | stLFR+ONT | Pacbio+Illumina+BioNano | Pacbio+Illumina+HiC | SOLiD | Illumina | Illumina |
|---|---|---|---|---|---|---|
| **Published Year** | 2019 | 2019 | 2019 | 2014 | 2014 | 2017 |
| **Assembly Size** | 650 Mb | 384 Mb | 372 Mb | 378 Mb | 477 Mb | 546 Mb |
| **Sequencing Coverage** | 235× | 204× | 327× | 131× | 211× | 118× |
| **Contig N50 (Kb)** | 2,150 | 3,136 | 1,398 | 8 | 26.5 | 30 |
| **Scaffold N50 (Kb)** | 8,129 | 16,705 | 16,786 | 315 | 867 | 3,817 |
| **GC Content** | 41.8% | 45.8% | 45.4% | 45.6% | 41.2% | 42.4% |
| **Gene Number** | 20,840 | 23,164 | / | 29,192 | 21,516 | 21,787 |
| **References** | / | GenBank accession: GCA_901000725.2 | GenBank accession: GCA_004026145.1 | [16] | [17] | [18] |

**Table 1. Summary statistics of recently published Tetraodontiforme genomes and relative species.**

## III Results and Discussions

### Repeat Content

Repetitive sequences usually refer to two major types of repetitive sequences: tandem repeats and interspersed repeats. For the repeat annotation of the porcupinefish genome, both homology-based predictions and *de novo* methods were employed. In the homolog-based methods, RepeatMasker and ProteinMask (version 3.2.9) [20] were utilized to detect interspersed repeats searching, against the published RepBase 16.02 [21] sequences. In the *de novo* method, the interspersed repeats in the genome were identified using RepeatMasker and RepeatModeler (version 1.1.0.4) [22]. Tandem Repeats Finder (TRF version 4.04) [23] was subsequently used to search for tandem repeats. A total of 211 Mb of non-redundant repetitive sequences in were identified in the porcupinefish genome, accounting for 32,91% of the whole genome (Table 2). The most predominant elements were long interspersed nuclear elements (LINEs), which accounted for 15.17% (97.6 Mb) of the genome (Table 3).

| Type | Repeat size (bp) | % of genome |
|---|---|---|
| **TRF** | 10,251,761 | 1.59 |
| **Repeatmasker** | 97,485,124 | 15.15 |
| **Proteinmask** | 49,879,868 | 7.75 |
| ***De novo*** | 192,737,809 | 29.95 |
| **Total** | **211,735,883** | **32.91** |

**Table 2. Prediction of repeat elements in the *Diodon Holocanthus* genome.**

| | **RepBase TEs** | | **TE proteins** | | ***De novo*** | | **Combined TEs** | |
|---|---|---|---|---|---|---|---|---|
| **Type** | Length (bp) | % in | Length (bp) | % in | Length (bp) | % in | Length (bp) | % in |

| | | genome | | genome | | genome | | genome |
|---|---|---|---|---|---|---|---|---|
| **DNA** | 36,599,586 | 5.69 | 5,920,922 | 0.92 | 58,987,243 | 9.17 | 78,826,558 | 12.25 |
| **LINE** | 42,087,139 | 6.54 | 35,005,111 | 5.44 | 79,625,667 | 12.37 | 97,621,545 | 15.17 |
| **SINE** | 17,253,616 | 2.68 | - | 0.00 | 10,266,445 | 1.60 | 25,398,653 | 3.95 |
| **LTR** | 11,380,434 | 1.77 | 8,990,419 | 1.40 | 67,612,186 | 10.51 | 72,182,179 | 11.22 |
| **Other** | 11,748 | 0.00 | - | 0.00 | - | 0.00 | 11,748 | 0.00 |
| **Unknown** | - | 0.00 | - | 0.00 | 7,598,160 | 1.18 | 7,598,160 | 1.18 |
| **Total** | **97,485,124** | **15.15** | **9,879,868** | **7.75** | **191,374,885** | **29.74** | **206,528,174** | **32.10** |

**Table 3. Statistics of repeat elements in the *Diodon Holocanthus* genome.**

## Genome annotation

Structural annotation of genes. Two methods (homology-based and*ab initio* predictions) were used to predict spiny porcupinefish genes. In the homology-based method, the protein repertoires of *Astatotilapia calliptera*, *Maylandia zebra*, *Oreochromis niloticus* and *Pundamilia nyererei* were downloaded from the NCBI database and mapped onto the spiny porcupinefish genome using BLAT (version 0.36) [24] and GeneWise (version 2.4.1) [25] to define gene models. For the *ab initio* prediction approach, Augustus [26], GlimmerHMM [27], and GENSCAN [28] were applied to predict the coding regions of genes using *Danio rerio* as the model species of Hidden Markov Model method. In total, 20,840 non-redundant protein-coding genes were annotated in the spiny porcupinefish genome by combining the different evidences using EVidenceModeler (version 1.1.1) [29] (Table 4). The average length was 13,736.38 bp, with an average of 9.71 exons. The average length of coding sequences, exons and introns were 1,677.23bp, 172.80 bp and 1385.12 bp, respectively, similar to that of the other released fish genomes, such as *Astatotilapia calliptera*, *Maylandia zebra*, *Oreochromis niloticus* and *Pundamilia*.

| Gene set | # of genes | CDS+intron length (avg) | CDS length (avg) | Exon length (avg) | Intron length (avg) | Exons per gene (avg) |
|---|---|---|---|---|---|---|
| *Diodon Holocanthus* | 20,840 | 13736.38 | 1677.23 | 172.8 | 1385.12 | 9.71 |
| *Danio_rerio* | 25,619 | 25207.59 | 1642.64 | 174.39 | 2798.97 | 9.42 |
| *Oryzias_latipes* | 19,699 | 12145.58 | 1515.82 | 147.82 | 1148.61 | 10.25 |
| *Takifugu_rubripes* | 18,523 | 7492.75 | 1693.53 | 152.61 | 574.33 | 11.1 |

**Table 4. Comparison on gene structures of annotated gene models of the *Diodon Holocanthus* with other species.**

Several databases including TrEMBL [30], SWISS-PROT [31] and InterPro [32] were used to search homologs to detect functions of the annotated protein-coding genes. 18,214 (87.40%), 17,015 (81.65%), 4,812 (23.09%) of protein-coding genes were found to present their homologous alignments in the three databases, respectively. We note that the remaining 2,559 (12.28%) protein-coding genes which cannot be identified and functionally annotated by existing database might be related to the specific characters of the *Diodon Holocanthus* genome, deserving further investigation.

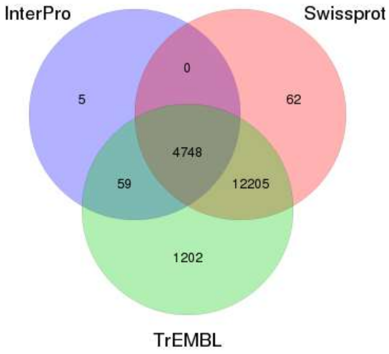

**Figure 3. Venn diagram of the number of genes with functional annotation using multiple public databases.**

## Data Records

Raw reads from BGISEQ-500 sequencing are deposited in the CNGB Nucleotide Sequence Archive (CNSA) with accession number CNP0000691 (https://db.cngb.org/cnsa). Data Citation 1: CNGB Nucleotide Sequence Archive CNP0000691).

## Author contributions

XXX

## Competing interests

The authors declare no competing interests.

## Acknowledgements

This work was supported by the special funding of "Blue granary" scientific and technological innovation of China (2018YFD0900301-05). We also appreciate the support and help from the Fish10K project. The work also received the technical support from China National Gene Bank.

## References

1 Wainwright, P. C. & Turingan, R. G. Evolution of Pufferfish Inflation Behavior. *Evolution* **51**, 506-518, doi:10.1111/j.1558-5646.1997.tb02438.x (1997).

2 Brainerd, E. L. Pufferfish inflation: Functional morphology of postcranial structures in Diodon holocanthus (Tetraodontiformes). *J Morphol* **220**, 243-261, doi:10.1002/jmor.1052200304 (1994).

3 Lucano-Ramirez, G., Pena-Perez, E., Ruiz-Ramirez, S., Rojo-Vazquez, J. & Gonzalez-Sanson, G. [Reproduction of the spiny puffer, Diodon holocanthus (Pisces: Diodontidae) in the continental shelf of Mexican Central Pacific]. *Rev Biol Trop* **59**, 217-232 (2011).

4 Guo, B., Zou, M., Gan, X. & He, S. Genome size evolution in pufferfish: an insight from BAC

clone-based Diodon holocanthus genome sequencing. *BMC Genomics* **11**, 396, doi:10.1186/1471-2164-11-396 (2010).

5 Neafsey, D. E. & Palumbi, S. R. Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. *Genome Res* **13**, 821-830, doi:10.1101/gr.841703 (2003).

6 Holcroft, N. I. A molecular analysis of the interrelationships of tetraodontiform fishes (Acanthomorpha: Tetraodontiformes). *Mol Phylogenet Evol* **34**, 525-544, doi:10.1016/j.ympev.2004.11.003 (2005).

7 Noleto, R. B., Vicari, M. R., Cipriano, R. R., Artoni, R. F. & Cestari, M. M. Physical mapping of 5S and 45S rDNA loci in pufferfishes (Tetraodontiformes). *Genetica* **130**, 133-138, doi:10.1007/s10709-006-9000-1 (2007).

8 Yamanoue, Y. *et al.* A new perspective on phylogeny and evolution of tetraodontiform fishes (Pisces: Acanthopterygii) based on whole mitochondrial genome sequences: basal ecological diversification? *BMC Evol Biol* **8**, 212, doi:10.1186/1471-2148-8-212 (2008).

9 Panova, M. *et al.* DNA Extraction Protocols for Whole-Genome Sequencing in Marine Organisms. *Methods Mol Biol* **1452**, 13-44, doi:10.1007/978-1-4939-3774-5_2 (2016).

10 Wang, O. *et al.* Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res* **29**, 798-808, doi:10.1101/gr.245126.118 (2019).

11 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18, doi:10.1186/2047-217x-1-18 (2012).

12 Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *arXiv e-prints* (2013). <https://ui.adsabs.harvard.edu/abs/2013arXiv1308.2012L>.

13 Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res* **27**, 757-767, doi:10.1101/gr.214874.116 (2017).

14 Deng, L. *et al.* SLR-superscaffolder: a <em>de novo</em> scaffolding tool for synthetic long reads using a top-to-bottom scheme. *bioRxiv*, 762385, doi:10.1101/762385 (2019).

15 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv e-prints* (2013). <https://ui.adsabs.harvard.edu/abs/2013arXiv1303.3997L>.

16 Gao, Y. *et al.* Draft sequencing and analysis of the genome of pufferfish Takifugu flavidus. *DNA Res* **21**, 627-637, doi:10.1093/dnares/dsu025 (2014).

17 Chen, S. *et al.* Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nat Genet* **46**, 253-260, doi:10.1038/ng.2890 (2014).

18 Shao, C. *et al.* The genome and transcriptome of Japanese flounder provide insights into flatfish asymmetry. *Nat Genet* **49**, 119-124, doi:10.1038/ng.3732 (2017).

19 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*, doi:10.1093/molbev/msx319 (2017).

20 Smit, A., Hubley, R. & Green, P. (1996).

21 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).

22 Smit, A. F. & Hubley, R. RepeatModeler Open-1.0. *Available fom http://www. repeatmasker. org* (2008).

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

24 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202 (2002).

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

26 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-439, doi:10.1093/nar/gkl200 (2006).

27 Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878-2879, doi:10.1093/bioinformatics/bth315 (2004).

28 Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94, doi:10.1006/jmbi.1997.0951 (1997).

29 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, doi:10.1186/gb-2008-9-1-r7 (2008).

30 Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res* **26**, 38-42, doi:10.1093/nar/26.1.38 (1998).

31 Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45-48, doi:10.1093/nar/28.1.45 (2000).

32 Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* **47**, D351-d360, doi:10.1093/nar/gky1100 (2019).