Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

PORTLAND PRESS

Research Article

Genome-wide survey and genetic characteristics of *Ophichthus evermanni* based on Illumina sequencing platform

Tianyan Yang[1], Zijun Ning[1], Yuping Liu[1], Shufei Zhang[2] and Tianxiang Gao[1]

[1]Fishery College, Zhejiang Ocean University, Zhoushan 316022, China; [2]Guangdong Provincial Key Laboratory of Fishery Ecology and Environment, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510300, China

Correspondence: Tianxiang Gao (gaotianxiang0611@163.com)

Ophichthidae fishes limit to continental shelf of all tropical and subtropical oceans and contain more than 350 species, representing the greatest specialization diversity in the order Anguiliformes. In the present study, we conducted a genome survey sequencing (GSS) analysis of *Ophichthus evermanni* by Illumina sequencing platform to briefly reveal its genomic characteristics and phylogenetic relationship. The first *de novo* assembled 1.97 Gb draft genome of *O. evermanni* was predicted based on K-mer analysis without obvious nucleotide bias. The heterozygosity ratio was 0.70%, and the sequence repeat ratio was calculated to be 43.30%. A total of 9016 putative coding genes were successfully predicted, in which 3587 unigenes were identified by gene ontology (GO) analysis and 4375 unigenes were classified into cluster of orthologous groups for enkaryotic complete genomes (KOG) functional categories. About 2,812,813 microsatellite motifs including mono-, di-, tri-, tetra-, penta- and hexanucleotide motifs were identified, with an occurrence frequency of 23.32%. The most abundant type was dinucleotide repeat motifs, accounting for 49.19% of the total repeat types. The mitochondrial genome, as a byproduct of GSS, was assembled to investigate the evolutionary relationships between *O. evermanni* and its relatives. Bayesian inference (BI) phylogenetic tree inferring from concatenated 12 protein-coding genes (PCGs) showed complicated relationships among Ophichthidae species, indicating a polyphyletic origin of the family. The results would achieve more thorough genetic information of snake eels and provide a theoretical basis and reference for further genome-wide analysis of *O. evermanni.*

Introduction

Ophichthidae is the family with the most various species in the order of Anguilliformes, which hitherto contains more than 350 valid species belonging to 62 genera all over the world [1]. These snake-shaped fishes are widely spread in tropical and subtropical inshore waters and prefer to slither in muddy substrates or coral reefs by pointed rayless tail tips or acute snouts [2]. Because of less distinguishable morphological features and various shapes of body in different growth stages, it brings great difficulties to effective species identification and phylogeny analysis of this group. The studies on ophichthid eels are limited to morphological identification and new species description [3–8]. There have been no reports on the genome of snake eels until now. The lacking of molecular genetic data has seriously restricted the further evolutionary and genomic studies of Ophichthidae fishes.

Nowadays, high-throughput next-generation sequencing (NGS) provides a more convenient approach to obtain massive genomic sequences, which can more comprehensively reveal the genetic background and phylogenetic relationships at DNA level [9]. With the accomplishment of the first fish whole genome sequencing (WGS) early in 2002 [10], more and more fish genomes have been published, ranging from

Received: 03 March 2022
Revised: 27 April 2022
Accepted: 29 April 2022

Accepted Manuscript online: 03 May 2022
Version of Record published: 27 May 2022

Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

the model fishes [11,12] to many commercial species [13–17]. The genome survey sequencing (GSS) is a convenient approach to provide fundamental information of genome. It could not only productively identify genome-wide simple sequence repeats (SSRs) but also efficiently predict putative gene functions and targeted the potential exon-intron boundaries [18].

In the present study, we selected *Ophichthus evermanni* [19], a kind of snake eel that mainly distributes in the East China Sea, the South China Sea and the coastal waters of southern Japan as a representative [20], and preliminarily revealed the genomic characterization such as genome size, GC content, heterozygosity and repeat ratio of this snake eel based on genome survey sequencing. Meanwhile, the genome annotation, microsatellite markers identification and mitochondrial genome assembly were conducted by a series of bioinformatics analyses. The information above could be helpful in species identification, adaptive evolutionary mechanisms and phylogenetic studies. Besides, these findings would also supplement the molecular biology data of *O. evermanni* and make a valuable contribution to the genome-wide studies on snake eels.

Materials and methods

Sample collecting and DNA extraction

One female specimen of Evermann's snake eel with body length 771.42 mm and body weight 571.63 g was obtained from coastal waters of Xiamen (118°34′E, 24°15′N), China in December 2020 (Supplementary Figure S1). After identifying it by morphological characteristics and DNA barcoding (mitochondrial DNA COI gene), the examined individual was preserved in −80°C ultra-low temperature freezer, and all animal experiments took place at Fisheries Ecology and Biodiversity Laboratory (FEBL) of Zhejiang Ocean University, Zhoushan, China. Experiments were conducted under the guideline and approval of the Ethics Committee for Animal Experimentation of Zhejiang Ocean University (ZJOU-ECAE20211876).

After species identification and morphological measurement, a piece of fresh muscle tissue was clipped from the base of dorsal fin and soaked in absolute ethanol. The genomic DNA was extracted by using the standard phenol-chloroform method followed by proteinase K digestion to ensure complete protein removal. The DNA integrity was first assessed by 1% agarose gel electrophoresis (5 V/cm, 20 min). And then, the quantity and purity of genomic DNA were checked by Qubit 2.0 fluorometer (Invitrogen, California, U.S.A.) and NanoDrop2000 spectrophotometer (Thermo Fisher Scientific, Delaware, U.S.A.), respectively.

Library construction and genome survey sequencing

The DNA sample was randomly fragmented into 300–500 bp using Covaris M220 Focused-ultrasonicator (Covaris, Massachusetts, U.S.A.) to construct the two paired-ends sequencing libraries, and then followed by terminal repair, adding an A base to the blunt ends and ligation to sequencing adaptors. After DNA purification and bridge PCR amplification, the prepared DNA library was sequenced based on the Illumina Hiseq 2500 platform with a read length of 2×150 bp by Origin-gene Biomedical Technology Co., Ltd., Shanghai, China (http://www.origin-gene.com/). All sequencing data were deposited in the short-read archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra/) under the accession number PRJNA807805.

Genome assembly and K-mer analysis

The clean data were obtained after removing reads containing adapters, duplicated reads and low quality reads from the raw genome survey sequence data. All the high-quality reads were assembled based on de Bruijn graph algorithm using SOAP de novo v2.04 software (https://soap.genomics.org.cn/) [21]. Jellyfish software [22] was conducted to count K-mer depth distribution of sequenced reads and then evaluated the genome size according to the formulas: Genome Size = K-mer number/average K-mer depth, Revised Genome Size = Genome Size × (1 − Error Rate). Because the distribution of K-mer frequency yields to Poisson distribution, the peak of K-mer distribution curve can be regarded as the expected depth of K-mer [23]. The heterozygous frequency of the genome of *O. evermanni* was roughly determined based on the K-mer analysis following the description of Liu et al [24]. And the repeat ratio was calculated according to the percentage of the total number of K-mer after the main peak 1.8 times of all K-mer numbers [24,25]. Moreover, the GC content was also an important parameter for measuring the sequencing bias of a genome, which was calculated by the 10 kb non-overlapping sliding windows along the assembled sequence.

Gene prediction and functional annotation

The software GeneMark-ES (http://exon.gatech.edu/genemark/gmes_instructions.html) [26] was conducted to predict genes. The translated protein sequences were compared with Nr (Non-Redundant Protein Sequence), KOG

Table 1 The top ten species blasted against the nucleotide sequence database (NT)

Species	Number of reads	Percentage (%)
Cyprinus carpio	854	8.54
Larimichthys crocea	288	2.88
Oryzias latipes	90	0.90
Danio rerio	85	0.85
Gouania willdenowi	81	0.81
Echeneis naucrates	78	0.78
Denticeps clupeoides	74	0.74
Syngnathus acus	66	0.66
Mastacembelus armatus	65	0.65
Salmo trutta	59	0.59

(Cluster of Orthologous Groups for Enkaryotic Complete Genomes), KEGG (Kyoto Encyclopedia of Genes and Genomes) and GO (Gene Ontology) databases using Blast 2.2.28 + [27], respectively, so as to obtain the annotation information of the predicted genes.

Microsatellites identification and phylogenetic analysis

The Perl script MicroSatellite (MISA) was used to identify microsatellites in the genome of *O. evermanni* [28]. The settings implemented to detect the minimum numbers of SSRs for mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats were as follows: number of mono-nucleotide repeats was less than 10, number of di-nucleotide repeats was less than 6, and numbers of remaining repeats were all less than 5, respectively.

To further reveal the phylogeny of *O. evermanni*, we assembled and generated the complete mitochondrial genome by running a Perl script NOVOPlasty 4.3.1, a *de novo* assembler for organelle genomes from the whole genome data [29]. The circular mitogenome was annotated by the online tool MitoFish (http://mitofish.aori.u-tokyo.ac.jp/) and then checked and corrected the annotation results manually. The complete mitochondrial sequence of *O. evermanni* was submitted to NCBI (National Center for Biotechnology Information) database with the accession number OM421636. The nucleotide composition was calculated by Mega 11 [30]. Twenty Anguilliformes mitogenomes were downloaded from the GenBank, with *Gymnothorax formosus* (GenBank accession number: KP874184) selected as the outgroup. Twelve protein-coding genes (PCGs) excluding ND6 were concatenated for phylogenetic analysis based on Bayesian inference (BI) method inferring by MrBayes 3.2.6 [31] Four independent Markov Chain Monte Carlo (MCMC) chains (one cold chain and three heated chains) were run for 1,000,000 generations with sampling every thousand generations, and then the initial 25% of these sampled trees were discarded as burn in. And before that, assessing nucleotide substitution saturation and selecting the best-fit model of nucleotide substitution were carried out with DAMBE 5.0 [32] and Modeltest 3.7 [33], respectively.

Results

Illumina Sequencing data statistics

The average sequencing depth of the HiSeq data was 50× coverage, which yielded approximately 54.145 Gb clean bases with the error ratio 0.0282% after sequencing quality control. The values of Q20 (base quality > 20) and Q30 (base quality > 30) were 96.575% and 91.525%, respectively, which suggested that the sequencing depth and was sufficient to capture most of the genomic information. The proportions of single base were presented in Figure 1A, the GC content was 42.66% with no apparent abnormalities and obvious GC bias being observed. Ten thousand pairs of reads data were randomly selected from the filtered high-quality data, and the top ten species blasting against the NT (Nucleotide Sequence Database) from the NCBI was showed in Table 1, demonstrating that there was no obvious exogenous contamination during the library construction.

Genomic characteristics by K-mer analysis

About 364,763,910 clean reads were used to carry out *de novo* assembly based on K-mer analysis. Finally, a total length of 761,647,043 bp contigs were obtained with the contig N50 value of 1366 bp and N90 value of 469 bp, and the maximum contig was 18,910 bp in length. A 350-bp insert library data were used to construct the K-mer distribution map of $K = 17$ and the 17-mer frequency distribution curve exhibited a unique peak at depth of 24 (Figure 1B). Statistical analysis showed that the total number of K-mers was 47,338,914,261 after removing the anomalies of depth.

Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

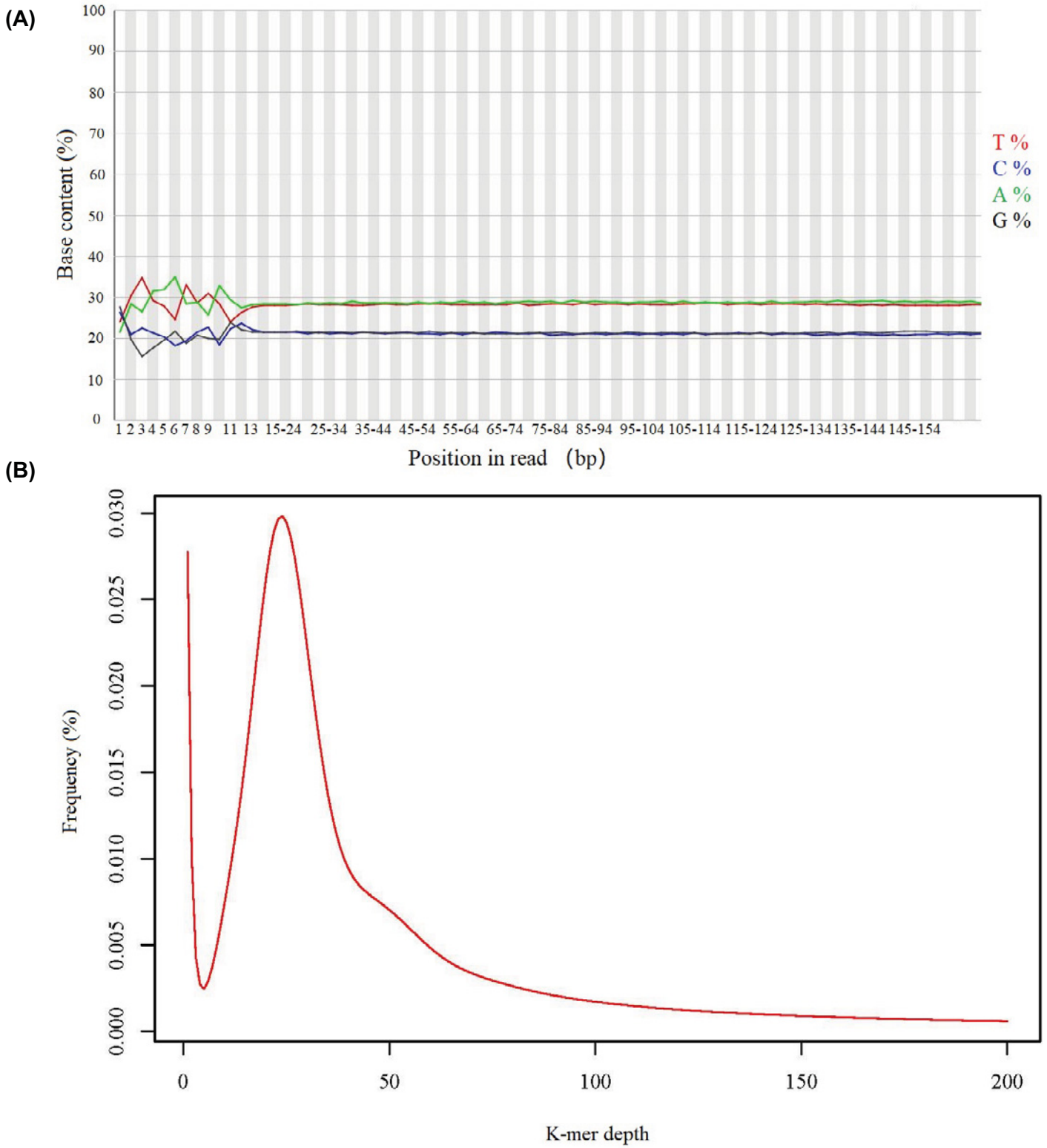

Figure 1. Sequence content across all bases and K-mer (*K* = 17) analysis for estimation of the genome size of *O. evermanni*

(**A**) The *X*-axis was the position in read and *Y*-axis was base content. (**B**) The *X*-axis represented K-mer depth and the *Y*-axis was the frequency at a given depth divided by the total frequency of all depths.

Table 2 Statistical information of predicted genes

Gene number	Gene total length (bp)	Gene average length (bp)	Gene density (genes/kb)	Intergenic region length (bp)	GC content in gene region (%)	GC content in intergenic region (%)
9,016	6,405,663	710	0.011	755,241,380	55.00	42.50

Table 3 Dominant base classes in each base repeat type in *O. evermanni*

Repeat type	Maximum repeat modify				Minimum repeat modify		
	Type	Repeat motif	Number	Proportion	Repeat motif	Number	Proportion
Mononucleotide	4	A	363,365	43.28%	G	66,406	7.91%
Dinucleotide	12	CA	373,131	26.97%	GC	2194	0.16%
Trinucleotide	60	AAT	22,229	11.24%	CGT	39	0.02%
Tetranucleotide	240	AAAT	24,328	7.93%	TCGT	10	0.003%
Pentanucleotide	966	AAAAT	2978	6.54%	–	–	–
Hexanucleotide	2361	CACACG	1227	3.09%	–	–	–

According to the calculation formula of genome size, we counted the revised genome size of the diploid species *O. evermanni* was 1.97 Gb after eliminating the effects of erroneous K-mers. The proportion of heterozygosity and repeat sequence ratio were 0.70% and 43.30%, respectively.

Gene prediction and annotation

A total of 9016 putative coding genes with the average length of 710 bp were successfully predicted by GeneMark-ES software (Table 2). The total length of genes and intergenic regions were 6,405,663 and 755,241,380 bp, with the GC content of 55.0% and 42.5%, respectively. The predicted genes were separately aligned by BLAST 2.2.28+ to the GO, KEGG, NR and KOG databases.

A total of 3587 unigenes were identified by GO analysis and further classified into the categories of molecular function, cellular component and biological process (Figure 2A). About 55.04% of them were grouped under biological processes, in which metabolic process was the most highly represented group. Second, 34.73% of the genes were grouped under cellular components, in which cell and cell part were the most significantly represented groups. Finally, 10.23% of the genes were grouped under molecular functions, in which binding represented a relatively high proportion. There were 4,375 genes were classified into KOG functional categories, the signal transduction mechanisms represented the largest group (861; 19.68%), followed by general function prediction only (562; 12.85%) and transcription (551; 12.59%) (Figure 2B).

Gene annotation analysis showed that a lot of predicted genes of *O. evermanni* genome were associated with the functional category of signal transduction mechanisms (861 genes) and immunity (950 genes). The functions of gene products in cells and their potential metabolic pathways were available in Supplementary Figure S2.

Microsatellites distribution and characteristics

Microsatellite identification tool (MISA) was used for microsatellite mining. As a result, 9,382,261 sequences with a total length of 1,214,882,177 bp were examined, and 2,812,813 SSRs were finally identified. Totally 2,187,607 SSR-containing sequences were detected accounting for 23.32% the total examined sequences. Among them, about 485,719 sequences contained more than 1 SSR and the number of SSRs present in compound formation was 425,676. The most abundant type of repeat was the dinucleotide (1,383,575; 49.19%), followed by mononucleotide (839,597; 29.85%), tetranucleotide (306,611; 10.90%), trinucleotide (197,737; 7.03%), pentanucleotide (45,529; 1.62%) and hexanucleotide (39,764; 1.41%) repeats (Figure 3A). The most and the second most common repeat types were five times repeats (451,077) and six times repeats (291,123) (Figure 3B).

In this study, the dominant repeating motifs ranging from mononucleotide to hexanucleotide were A (363,365), CA (373,131), AAT (22,229), AAAT (24,328), AAAAT (2978) and CACACG (1227) of the total SSRs (Table 3). Among the dinucleotide motifs, the most abundant repeat motif type was AC/GT, followed by AG/CT, AT/AT and CG/CG. Within the trinucleotide repeat motifs, the major repeat motifs were AAT/ATT and AAG/CTT, accounting for 44.35% and 17.77%, respectively. Percentages of different motifs in mon-, tetra-, penta- and hexa- nucleotide repeats were also showed in Figure 4.

Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

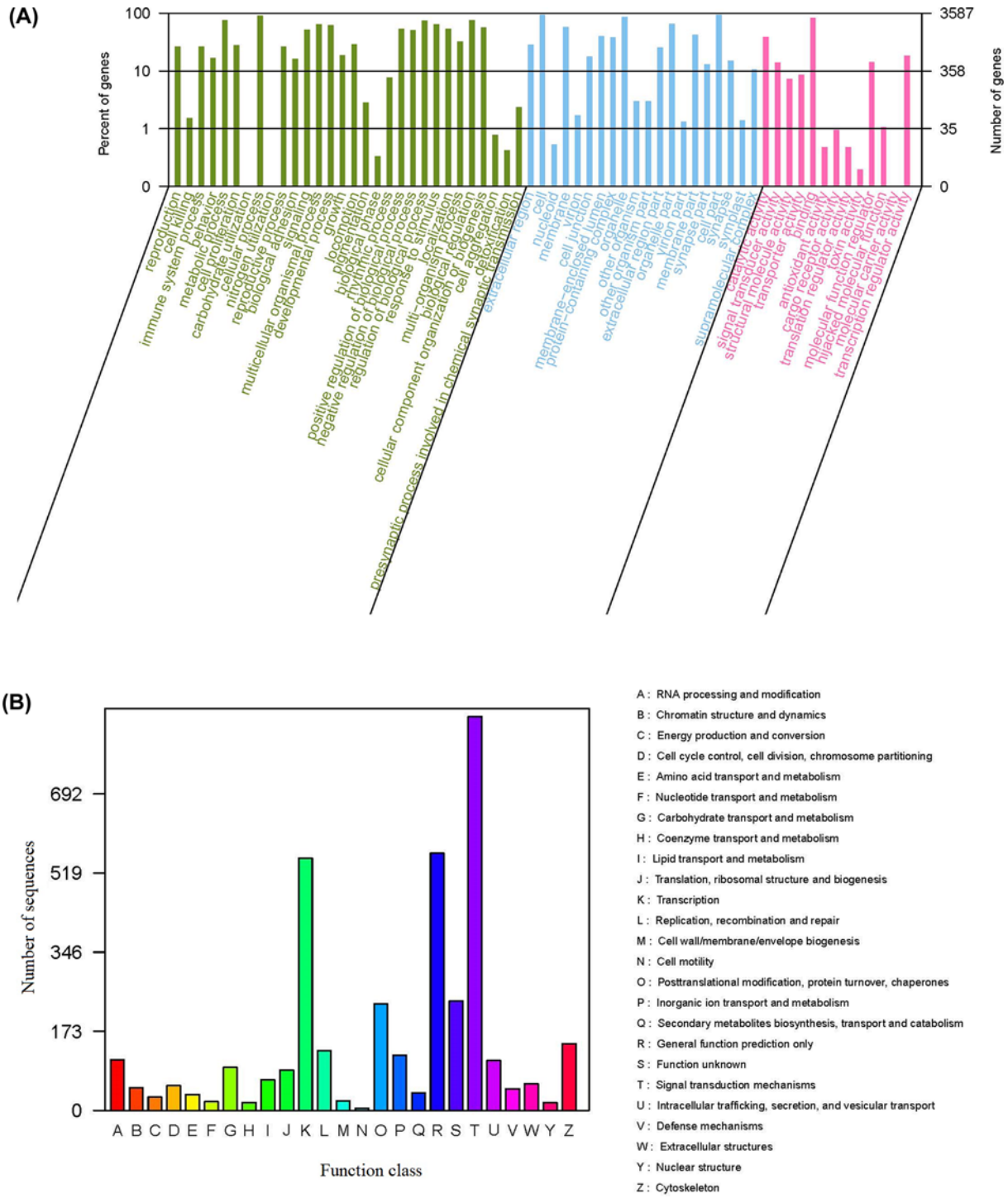

Figure 2. GO annotation and KOG function classification of putative genes in *O. evermanni*

(**A**) Genes were assigned to three categories: biological process, cellular component and molecular function. (**B**) Different color codes (A–Z) at right of the histogram represented different category.

Mitochondrial DNA structure and phylogenetic relationships

It was the first time to report the complete mitogenome for *O. evermanni* in this study. The complete mitochondrial genome was 17,759 bp in length (Figure 5), with the base composition of A (31.27%), G (16.19%), C (26.22%) and T (26.32%), respectively. The A+T content (57.59%) was greater than G+C content (42.41%), showing an obvious AT

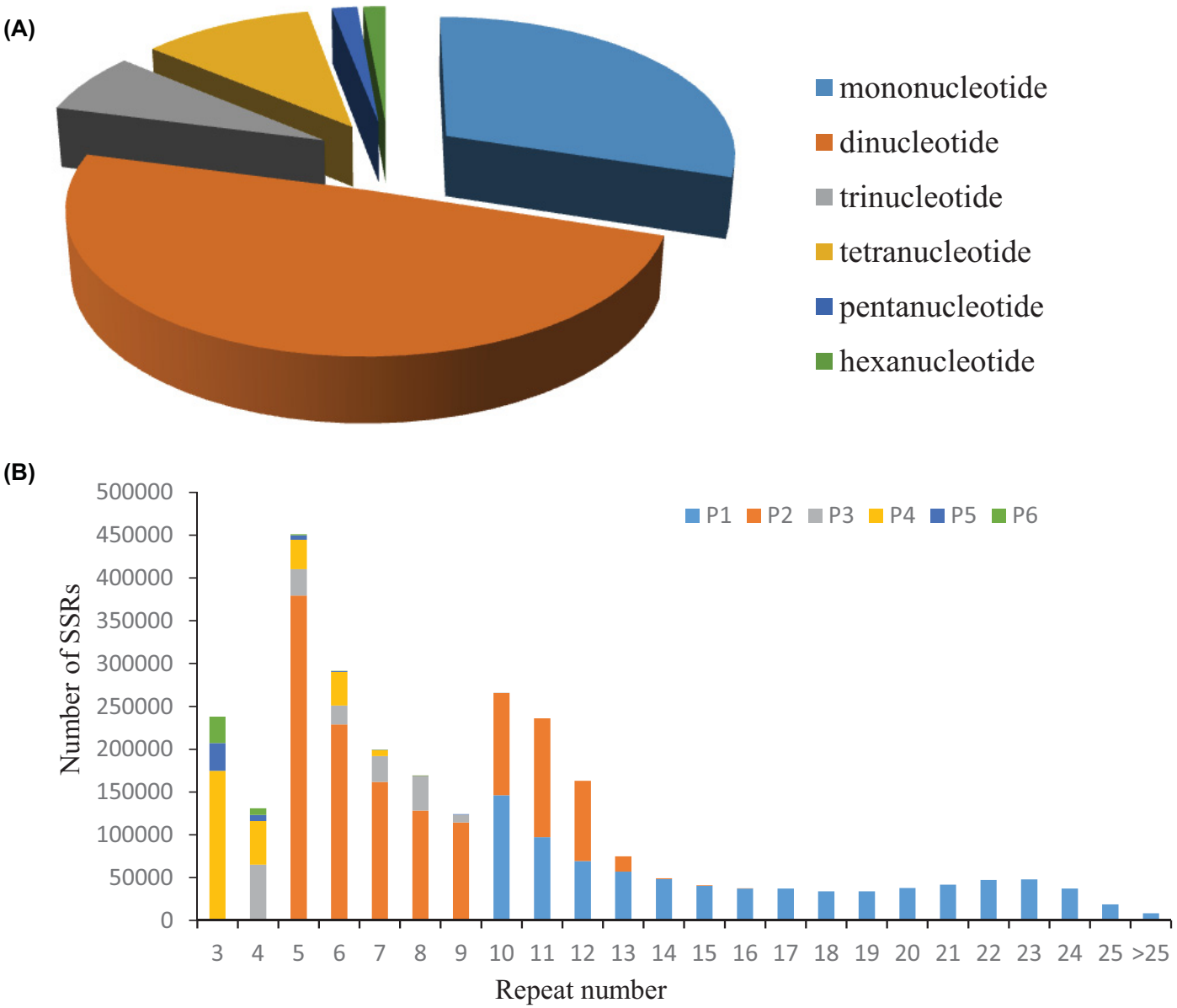

Figure 3. Distribution of SSR motifs in *O. evermanni*

(**A**) Frequency of different microsatellite motif types. (**B**) Distributions of different motif types with different repeat numbers.

bias. Unlike other typical teleosts, the gene arrangement was identified in the mitogenome of *O. evermanni*. ND6 gene and the conjoint tRNA-Glu were translocated between tRNA-Thr and tRNA-Pro, and another highly homologous D-loop region was located in the upstream of the ND6 gene. The tRNA-Gln (Q), tRNA-Ala (A), tRNA-Asn (N), tRNA-Cys (C), tRNA-Tyr (Y), tRNA-Ser^{UCA} (S1), tRNA-Glu (E), tRNA-Pro (P) and ND6 were located in the L-strand, while the rests were located in the H-strand. Except for tRNA-Ser (AGC), the remaining 21 tRNAs could fold into typical cloverleaf secondary structure.

Phylogenetic relationships were constructed based on the linked sequences of 12 PCGs (without stop codons) of 21 mitogenomes using BI method. In order to make sure that the aligned sequences were suitable for tree construction, we conducted the test of substitution saturation based on I_{ss} statistic for the dataset with DAMBE prior to phylogenetic analysis. The observed I_{ss} value (0.3013) was significantly smaller than $I_{ss.c}$ value (0.8496 assuming a symmetrical topology and 0.6444 assuming an extreme asymmetrical topology) when all three codon positions were considered as a whole. Furthermore, the plot trend-line analysis was carried out using generalized time reversible

Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

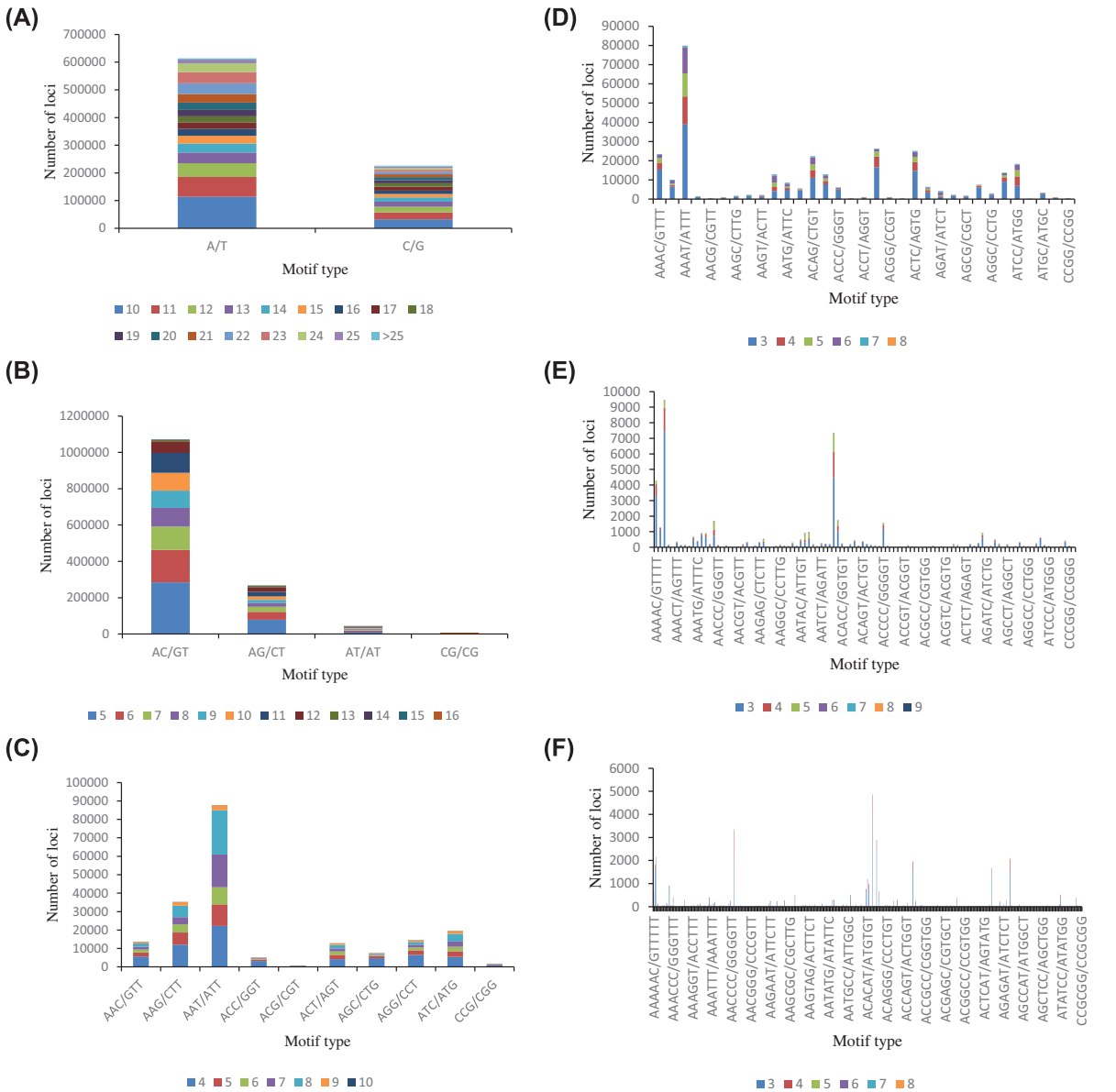

Figure 4. Type and frequency of microsatellite motifs in *O. evermanni*

(**A**) Frequency of different mononucleotide microsatellite motifs. (**B**) Frequency of different dinucleotide microsatellite motifs. (**C**) Frequency of different trinucleotide microsatellite motifs. (**D**) Frequency of different tetranucleotide microsatellite motifs. (**E**) Frequency of different pentanucleotide microsatellite motifs. (**F**) Frequency of different hexanucleotide microsatellite motifs.

(GTR) distance as abscissa and base substitution as ordinate (Figure 6). The result showed that there was an obvious linear relationship between them, indicating the sequences obviously had experienced little substitution saturation and subsequent phylogenetic analysis was feasible. GTR+G model was chosen as the appropriate model for the nucleotide sequences based on Akaike information criterion (AIC). The reconstructed BI tree was showed in Figure 7. It revealed that all Ophichthidae species gathered as one clade, and *O. evermanni* had the closest relationship with *Myrichthys maculosus*. Family Ophichthidae clustered with one group of Congridae consisting of *Conger japonicus* and *C. myriaster*. Nettastomatidae, Derichthyidae and Congridae (*Heteroconger hassi* + *Paraconger notialis*) formed another clade. While, species of Muraenesocidae located near the root of the phylogenetic tree.

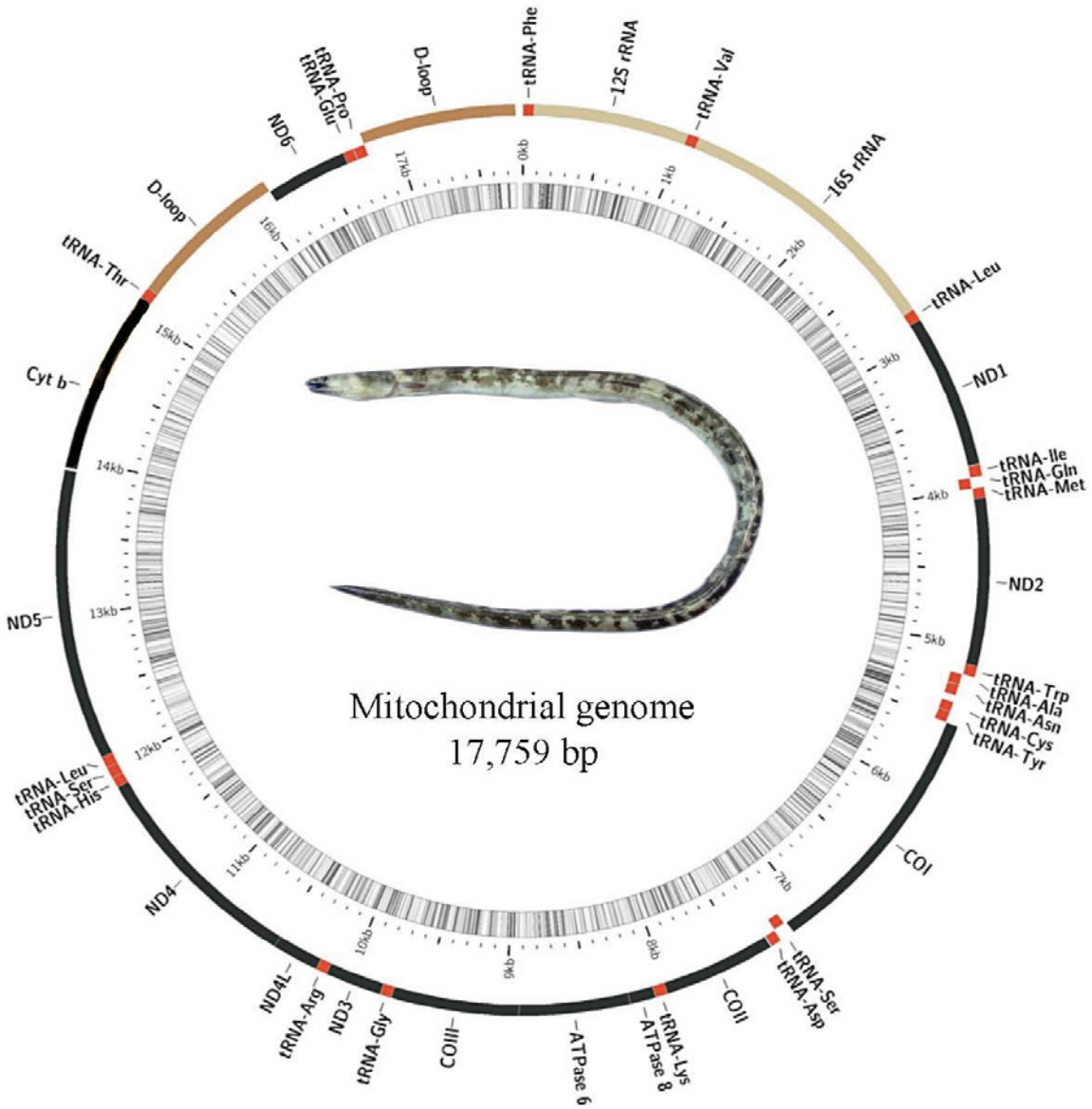

Figure 5. The mitochondrial genome structure of *O. evermanni*

Discussion

The family Ophichthidae, the most divergent group within the order Anguilliformes, comprises two subfamilies, the Myrophinae and the Ophichthinae, the latter of which is characterized by a hard, pointed and finless tail tip [3,4]. The genus *Ophichthus* includes the highest numbers of species (285 valid species worldwide) among all of the 47 recognized genera in the subfamily Ophichthinae, and there have been some new species to be constantly discovered and reported since the last two decades [1]. Therefore, the classification and identification of the snake eels are always in the utmost confusion [34]. Fourteen-four species of *Ophichthus* were recorded and described in offshore China, which mainly distributed in the sea waters south of the Yangtze River Estuary [2]. However, very limited molecular genetic researches have been focused on a certain snake eel both at home and abroad.

In the present study, the Illumina paired-end sequencing technique was applied to preliminarily unravel the genomic background of *O. evermanni*, a representative species of this group. Genome size refers to the amount of DNA contained in a haploid genome, and it serves as an important basis for comparative and evolutionary genomics [35]. Genome size and its variability were influenced by the mutational pressure of chromosomal, transposon activity and relative occurrence rates of segmental duplications and deletions to some extent [36]. The draft genome size of *O.*

Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

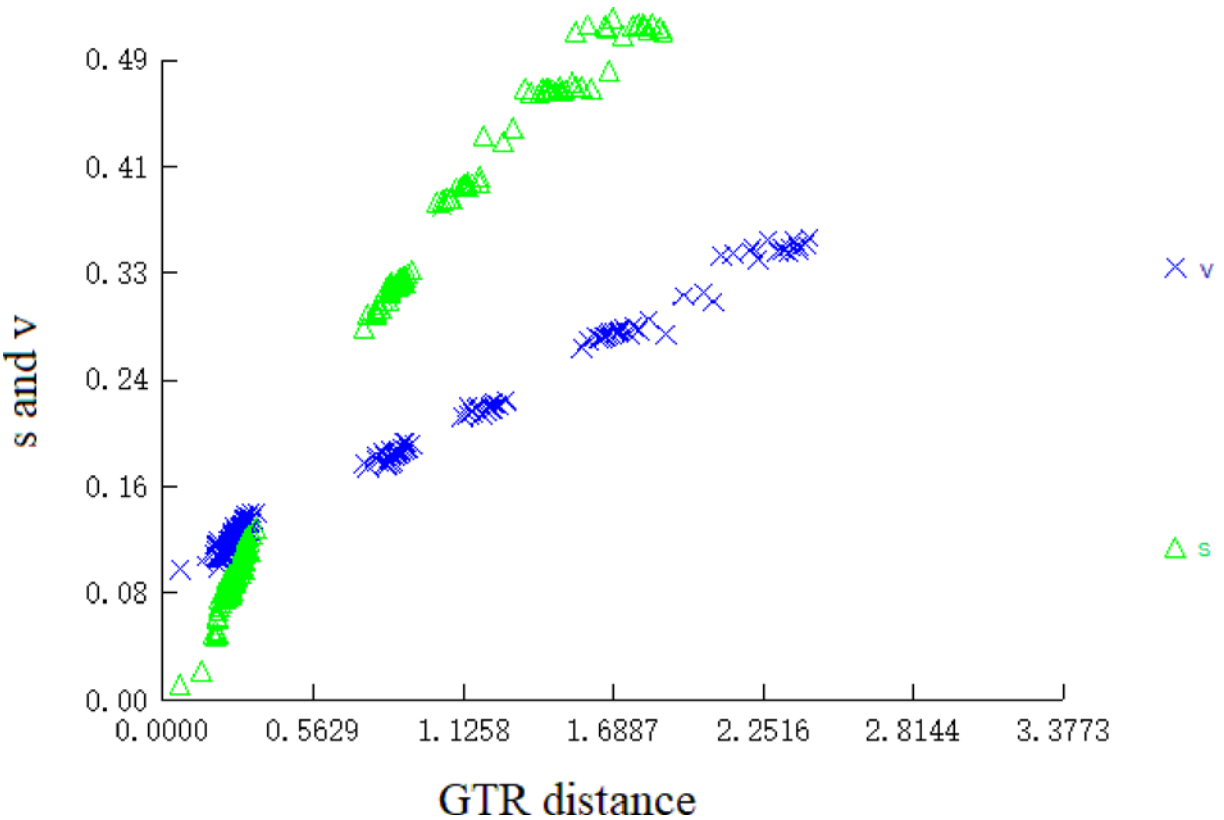

Figure 6. Nucleotide substitution saturation analysis of 12 PCGs sequences without ending codons

evermanni was relatively larger than those of most marine teleosts, such as *Fugu rubripes* (322.5 Mb) [10], *Gadus morhua* (830 Mb) [37], *Larimichthys crocea* (728 Mb) [13], *Sillago sinica* (534 Mb) and [38]. Previous researches indicated larger genomes had relatively higher mutational liability to undergoing natural selection in evolutionary process [39], and lungfish was a good case in point [40]. Our result implied that larger genomes of snake eels might accumulate more mutations and have strong ability to adapting to the benthic and burrowing living habits in sandy shores or muddy estuaries.

Genome size is determined not only by the number of genes in the genome but also by the amount of repetitive DNA. The repeat ratio (43.30%) of *O. evermanni* genome was present at a high level in the known fish genomes. It confirmed that larger genomes tended to be ones in which the copy numbers of the repeat sequences were highest [41]. Heterozygosity is important for determining the appropriate genomic splicing strategy and subsequence data processing. The genome-scale de novo assembly will become difficult when the heterozygosity exceeds 0.5% [22]. According to the criteria, the higher proportion of heterozygosity (0.70%) reflected the complexity of *O. evermanni* genome, and also inferred higher genetic diversity in *O. evermanni.* Low (<25%) or high (>65%) GC content may cause sequencing bias of Illumina platform and seriously affect the quality of genome assembly and subsequent analysis [42]. In the study, the moderate GC content was detected and the percentage of A vs. G and C vs. T were almost equal to each other, indicating the sequencing quality was good and suitable for further analysis.

As cave-dwelling fish species, the visual system structure and function of the snake eels have degenerated dramatically, by contrast, the olfactory organs and lateral line canals are well developed [43,44]. In the present study, some signal transduction pathways (MAPK signaling pathway, olfactory transduction, taste transduction, neuroactive ligand–receptor interaction etc.) were detected and therefore environmental messages are received from the sensory organs and then abundant nerve fibers can transmit external stimulus to the brain. In addition, some signaling pathways related to immune system were also founded, such as intestinal immune network for B-cell receptor signaling pathway, T-cell receptor signaling pathway, Jak-STAT signaling pathway, NOD-like receptor signaling pathway and Toll-like receptor signaling pathway. In coastal areas of Guangdong and Fujian, China, local residents regard it as healthy tonic for strengthening body and improving immunity.

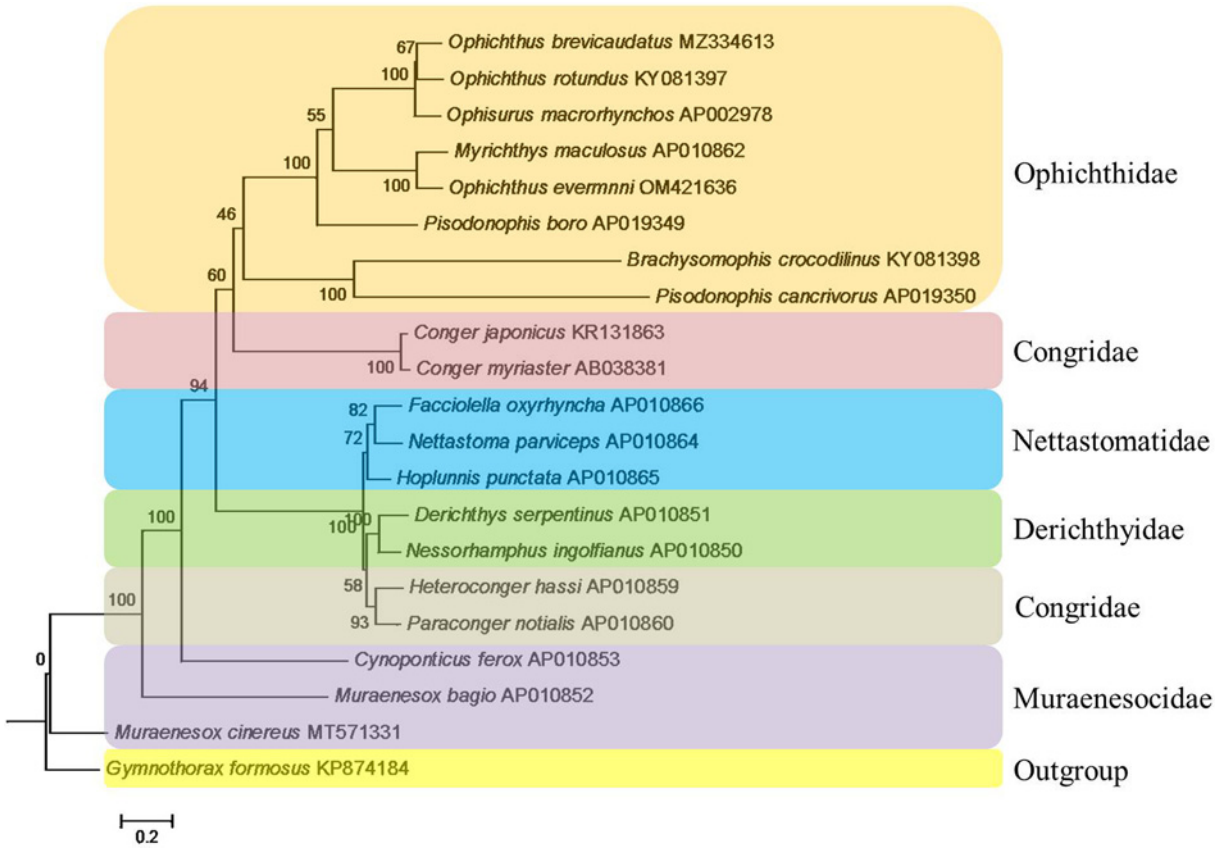

Figure 7. The phylogenetic tree inferred from the mitogenome sequences of 21 Anguilliformes fishes. Sample from this study was written in red letters

Microsatellite DNA marker offers several advantages of codominant, extensive distribution, abundant polymorphisms and convenient analysis, and has become an ideal tool in genetics and evolution studies [45]. In the present study, the dinucleotide repeats had the highest number and type of repeats, which was consistent with *Acanthogobius ommaturus* [46], *Sillago sihama* [47], *Harpadon nehereus* [48] and *Cociella crocodilus* [49]. SSR polymorphic loci are mainly distributed in dinucleotide and trinucleotide repeats [50]. Hence, the development of polymorphic SSR markers from low repetitive motifs will have great potential in population genetics research of *O. evermanni* subsequently. The complexity of repeated motif is usually related to evolutionary level and DNA mutation rate [51]. The frequency of mononucleotides to trinucleotides was amount to 86.07%, which implied that *O. evermanni* might have experienced a long evolutionary history and accumulated more genetic variation. Apart from SSRs, another important molecular marker mitochondrial DNA was also assembled to explore the systematical evolution of *O. evermanni*. The intricate clustering relationship in family Ophichthidae was presented in the topological structure of BI tree, deducing that Ophichthidae was not a monophyletic group and should be a polyphyletic group. The conclusion was identical with morphological and anatomical evidences of olfactory organs [44]. The snake eels have later divergence time on evolution comparing to other related species, and the short interval time of differentiation might cause a rapid affair of evolution radiation and species forming in this group.

Conclusions

In the present study, the genome size of *O. evermann* estimated by K-mer analysis ($K = 17$) was 1.97 Gb, with the heterozygosity and duplication rates 0.70% and 43.30%, respectively. The results showed *O. evermann* owned relatively larger genome size, higher heterozygosity and nucleotide repetition ratio in bony fishes. Besides, the gene annotation, SSR characteristics and phylogenetic relationship analyses were tentatively carried out. Our results would provide meaningful data for further genomic studies and lay a useful basis for novel molecular marker development. Because genome size based on K-mer analysis might be affected by data quality, analytical software, parameters setting

and some other confounding factors. Hence, the novel state-of-the-art genetic techniques, such as Illumina combined with PacBio and Hi-C-based assembly needs to be conducted to obtain chromosomal-level scaffolding genome in the future.

Data Availability

The data presented in this study are openly available in NCBI database.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

The study was supported by Science and Technology Planning Project of Zhoushan [grant number 2022C41022], Fund of Guangdong Provincial Key Laboratory of Fishery Ecology and Environment [grant number FEEL-2021-7] and The Province Key Research and Development Program of Zhejiang [grant number 2021C02047].

CRediT Author Contribution

Tianyan Yang: Conceptualization, Writing—original draft. **Zijun Ning:** Formal analysis, Investigation. **Yuping Liu:** Data curation, Formal analysis. **Shufei Zhang:** Resources. **Tianxiang Gao:** Project administration, Writing—review & editing.

Acknowledgements

We sincerely thank the reviewers for their constructive comments. Besides, we would like to thank Professor Zhiqiang Han for data analysis guidance.

Abbreviations

AIC, Akaike information criterion; BI, Bayesian inference; GO, gene ontology; GSS, genome survey sequencing; GTR, generalized time reversible; MCMC, Markov Chain Monte Carlo; MISA, microsatellite identification tool; NGS, next-generation sequencing; PCG, protein-coding gene.

References

1 Eschmeyer, W.N. and Fong, J.D. (2022) Genera/species by family/subfamily in Eschmeyer's catalog of fishes. http://researcharchive.calacademy.org/research/ichthyology/catalog/SpeciesByFamily.asp
2 Tang, W.Q. and Zhang, C.G. (2004) A taxonomic study on snake eel family Ophichthidae in China with the review of Ophichthidae (Pisces, Anguilliformes). *J. Shanghai Fish Univ.* **13**, 16–22, (In Chinese)
3 Gosline, W.A. (1951) The osteology and classification of the Ophichthid eels of the Hawaiian Islands. *Pac. Sci.* **5**, 298–320
4 McCosker, J.E. (1977) The osteology, classification and relationships of the eel family Ophichthidae. *Proc. Calif. Acad. Sci.* **41**, 1–123
5 Zhu, Y.D., Wu, H.L. and Jin, X.B. (1981) Four new species of the families Ophichthyidae and Neenchelidae. *J. Fish. Chin.* **5**, 21–27, (In Chinese)
6 McCosker, J.E. and Psomadakis, P.N. (2018) Snake eels of the genus *Ophichthus* (Anguilliformes: Ophichthidae) from Myanmar (Indian Ocean) with the description of two new species. *Zootaxa* **4526**, 71–83, https://doi.org/10.11646/zootaxa.4526.1.5
7 Mohapatra, A., Ray, D., Mohanty, S.R. and Mishra, S.S. (2018) *Ophichthus johnmccoskeri* sp. nov. (Anguilliformes: Ophichthidae): a new snake eel from Indian waters, Bay of Bengal. *Zootaxa* **4462**, 251–256, https://doi.org/10.11646/zootaxa.4462.2.7
8 McCosker, J.E., Bogorodsky, S.V., Mal, A.O. and Alpermann, T.J. (2020) Description of a new snake eel *Ophichthus olivaceus* (Teleostei: Anguilliformes, Ophichthidae) from the Red Sea. *Zootaxa* **4750**, 31–48, https://doi.org/10.11646/zootaxa.4750.1.2
9 Yang, M.Q., Athey, B.D., Arabnia, H.R., Sung, A.H., Liu, Q.Z., Yang, J.K. et al. (2009) High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. *BMC Genom.* **10**, 1–3, https://doi.org/10.1186/1471-2164-10-S1-I1
10 Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P. et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310, https://doi.org/10.1126/science.1072104
11 Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B. et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714–719, https://doi.org/10.1038/nature05846
12 Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M. et al. (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503, https://doi.org/10.1038/nature12111
13 Wu, C.W., Zhang, D., Kan, M.Y., Lv, Z.M., Zhu, A.Y., Su, Y.Q. et al. (2014) The draft genome of the large yellow croaker reveals well-developed innate immunity. *Nat. Commun.* **5**, 5227, https://doi.org/10.1038/ncomms6227
14 Chen, S.L., Zhang, G.J., Shao, C.W., Huang, Q.F., Liu, G., Zhang, P. et al. (2014) Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nat. Genet.* **46**, 253–260, https://doi.org/10.1038/ng.2890
15 Xu, P., Zhang, X.F., Wang, X.M., Li, J.T., Liu, G.M., Kuang, Y.Y. et al. (2014) Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nat. Genet.* **46**, 1212–1219, https://doi.org/10.1038/ng.3098

16 Chen, B.H., Li, Y., Peng, W.Z., Peng, W.Z., Zhou, Z.X., Shi, Y. et al. (2019) Chromosome-level assembly of the Chinese seabass (*Lateolabrax maculatus*) genome. *Front. Genet.* **10**, 275, https://doi.org/10.3389/fgene.2019.00275
17 Huang, Y.Q., Mustapha, U.F., Huang, Y., Tian, C.X., Yang, W., Chen, H.P. et al. (2021) A chromosome-level genome assembly of the spotted scat (*Scatophagus argus*). *Genome Biol. Evol.* **13**, 1–8, https://doi.org/10.1093/gbe/evab092
18 Lu, X., Luan, S., Kong, J., Hu, L.Y., Mao, Y. and Zhong, S.P. (2017) Genome-wide mining, characterization, and development of microsatellite markers in *Marsupenaeus japonicus* by genome survey sequencing. *J. Ocean. Limn.* **35**, 203–214, https://doi.org/10.1007/s00343-016-5250-7
19 Jordan, D.S. and Richardson, R.E. (1909) A catalog of the fishes of the island of Formosa, or Taiwan, based on the collections of Dr. Hans Sauter. *Mem. Carnegie Mus.* **4**, 172, https://doi.org/10.5962/p.48328
20 Chen, D.G. and Zhang, M.Z. (2016) *Marine Fishes of China*, China Ocean University Press, Qingdao, (In Chinese)
21 Luo, R.B., Liu, B.H., Xie, Y.L., Li, Z.Y., Huang, W.H., Yuan, J.Y. et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18, https://doi.org/10.1186/2047-217X-1-18
22 Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011
23 Shi, L.L., Yi, S.K. and Li, Y.H. (2018) Genome survey sequencing of red swamp crayfish *Procambarus clarkii*. *Mol. Biol. Rep.* **45**, 799–806, https://doi.org/10.1007/s11033-018-4219-3
24 Liu, B.H., Shi, Y.J., Yuan, J.Y., Hu, X.S., Zhang, H., Li, N. et al. (2013) Estimation of genomic characteristics by analyzing K-mer frequency in de novo genome projects. *Quant. Biol.* **35**, 62–67
25 Li, X. and Waterman, M.S. (2003) Estimating the repeat structure and length of DNA sequences using L-tuples. *Genome Res.* **13**, 1916–1922, https://doi.org/10.1101/gr.1251803
26 Borodovsky, M. and Lomsadze, A. (2011) *Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES*, John Wiley & Sons, Inc., New York
27 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. et al. (1997) Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402, https://doi.org/10.1093/nar/25.17.3389
28 Beier, S., Thiel, T., Münch, T., Scholz, U. and Mascher, M. (2017) MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585, https://doi.org/10.1093/bioinformatics/btx198
29 Dierckxsens, N., Mardulyn, P. and Smits, G. (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18
30 Tamura, K., Stecher, G. and Kumar, S. (2021) MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027, https://doi.org/10.1093/molbev/msab120
31 Ronquist, F., Teslenko, M., Mark, P., van der Mark, P., Ayres, D.L., Darling, A. et al. (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542, https://doi.org/10.1093/sysbio/sys029
32 Xia, X.H. (2013) DAMBE5: A Comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **30**, 1720–1728, https://doi.org/10.1093/molbev/mst064
33 Posada, D. and Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818, https://doi.org/10.1093/bioinformatics/14.9.817
34 Myers, G.S. and Storey, M.H. (1939) *Hesperomyrus fryi,* a new genus and species of echelid eels from California. *Stanford Ichthyol. Bull.* **1**, 156–159
35 Sessions, S.K. (2013) Genome Size. *Brenner's Encyclopedia of Genetics*, Second Edition, Elsevier Academic Press, Amsterdam, https://doi.org/10.1016/B978-0-12-374984-0.00639-2
36 Lynch, M. (2007) *The Origins of Genome Architecture*, Sinauer Associates, Sunderland
37 Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F. et al. (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207–210, https://doi.org/10.1038/nature10342
38 Xu, S.Y., Xiao, S.J., Zhu, S.L., Zheng, X.F., Luo, J., Liu, J.Q. et al. (2018) A draft genome assembly of the Chinese sillago (*Sillago sinica*), the first reference genome for Sillaginidae fishes. *Gigaence* **7**, 1–8, https://doi.org/10.1093/gigascience/giy108
39 Dufresne, F. and Jeffery, N. (2011) A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Res.* **19**, 925–938, https://doi.org/10.1007/s10577-011-9248-x
40 Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J.M., Irisarri, I. et al. (2021) Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **590**, 284–289, https://doi.org/10.1038/s41586-021-03198-8
41 Charles, R.C. and Cassandra, L.S. (1999) *Genomics - The Science and Technology behind the Human Genome Project*, John Wiley & Sons Inc., New York
42 Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C. et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18, https://doi.org/10.1186/gb-2011-12-2-r18
43 Zhang, Y.W. (1964) On the structure of the lateral line canal of Apodes and its significance on classification. *Acta Zool. Sin.* **16**, 653–657, (In Chinese)
44 Liu, D. (2005) *Study on Comparative Morphology of the Olfactory Organ and Phylogeny of the Snake-eel Fishes from China*, Shanghai Ocean University, Shanghai, Master's thesis
45 Ashley, M.V. and Dow, B.D. (1994) The Use of Microsatellite Analysis in Population Biology: Background, Methods and Potential Applications. In *Molecular Ecology And Evolution: Approaches And Applications. Experientia Supplementum* (Schierwater, B., Streit, B., Wagner, G.P. and Desalle, R., eds), Birkhäuser, Basel, https://doi.org/10.1007/978-3-0348-7527-1˙10
46 Chen, B.J., Sun, Z.C., Lou, F.R., Gao, T.X. and Song, N. (2020) Genomic characteristics and profile of microsatellite primers for *Acanthogobius ommaturus* by genome survey sequencing. *Biosci. Rep.* **40**, 1–8, https://doi.org/10.1042/BSR20201295
47 Qiu, B.X., Fang, S.B., Ikhwanuddin, M., Wong, L.L. and Ma, H.Y. (2020) Genome survey and development of polymorphic microsatellite loci for *Sillago sihama* based on Illumina sequencing technology. *Mol. Biol. Rep.* **47**, 3011–3017, https://doi.org/10.1007/s11033-020-05348-z

Bioscience Reports (2022) **42** BSR20220460
https://doi.org/10.1042/BSR20220460

48 Yang, T.Y., Huang, X.X., Ning, Z.J. and Gao, T.X. (2021) Genome-Wide Survey Reveals the Microsatellite Characteristics and Phylogenetic Relationships of *Harpadon nehereus*. *Curr. Issues Mol. Biol.* **43**, 1282–1292, https://doi.org/10.3390/cimb43030091

49 Zhao, R.R., Lu, Z.C., Cai, S.S., Gao, T.X. and Xu, S.Y. (2021) Whole genome survey and genetic markers development of crocodile flathead *Cociella crocodilus*. *Anim. Genet.* **52**, 891–895, https://doi.org/10.1111/age.13136

50 Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J. and Deka, R. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *P. Natl. Acad. Sci. U. S. A.* **94**, 1041–1046, https://doi.org/10.1073/pnas.94.3.1041

51 Katti, M.V., Ranjekar, P.K. and Gupta, V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18**, 1161–1167, https://doi.org/10.1093/oxfordjournals.molbev.a003903