Yu et al. *Zool. Res.* 2022, 43(6): 931−934
https://doi.org/10.24272/j.issn.2095-8137.2022.189

ZR | Zoological Research

Letter to the editor Open Access

High-quality chromosome-level genome assembly of the melon-headed whale (*Peponocephala electra*)

DEAR EDITOR,

The melon-headed whale (*Peponocephala electra*), a small toothed whale in the Delphinidae family, inhabits tropical and subtropical oceans. It is an attractive model species for studying secondary aquatic adaptation and evolution. Here, we successfully assembled a high-quality chromosome-level genome of *P. electra* using PacBio and Hi-C sequencing technologies. A 2.29 Gb 120 contig assembly with a contig N50 of 82.36 Mb and scaffold N50 of 102.72 Mb was obtained, with more than 96.93% (2.22 Gb) anchored to 22 pseudochromosomes. The genome assembly showed high completeness, with a BUSCO score of 96.6%. In total, 21 492 protein-coding genes were predicted in the newly assembled genome, 87.61% of which were functionally annotated. Repetitive elements accounted for 43.12% of the genome, dominated by long interspersed nuclear element (LINE) retrotransposons (25.10%). Based on phylogenetic analysis, *P. electra* and the long-finned pilot whale (*Globicephala melas*) formed a clade, which diverged 3.54–6.10 million years ago (Ma). We also identified gene family dynamics and positively selected genes specific to *P. electra*, which may be related to head and nervous system development. The high-quality reference genome of *P. electra* provides a valuable genomic resource for future studies of cetacean biology and evolution.

The melon-headed whale (*P. electra*) (also known as the “little killer whale”, “many-toothed blackfish”, and “electra dolphin”) is found in tropical to subtropical waters worldwide. The species belongs to the blackfish (Globicephalinae) group of smaller toothed whales in the family Delphinidae and is taxonomically closer to dolphins than to other whales. It is one of the smallest blackfish (adults up to 2.7 m long, weighing approximately 160 kg) and is characterized by a rounded melon-shaped head and lack of a clearly defined beak (Perrin et al., 2009). Tooth count in *P. electra* decreases with age, with the loss of teeth in the upper jaw preceding that in the lower jaw, which may be a suction feeding adaptation (Kurihara et al., 2016). It is a deep-dwelling whale that can dive up to 471.5 m and remain underwater for 12 min (West et al., 2018). The species possesses a complex social structure and usually travels in large groups of up to 1 000 individuals (Jefferson et al., 2008), a behavior likely related to its large brain size (Fox et al., 2017). Compared to other toothed whales, *P. electra* is particularly long-lived, with a maximum lifespan (47 years) more than twice that of similarly sized pygmy killer whales (*Feresa attenuata*; 21 years) (Tacutu et al., 2018). Overall, *P. electra* can serve as a useful model species for understanding evolution, longevity, and mammalian intelligence. However, its underlying genetic information has not yet been reported. Here, we constructed a chromosome-level genome assembly of *P. electra* using PacBio and Hi-C data.

An adult male *P. electra* individual collected in July 2021 at Tumen Port, Linhai, Taizhou City, Zhejiang Province, China (N28°50′51″, E121°39′14″) was used for genome sequencing. All sample collection and use protocols were approved by the Committee of Animal Care and Use in Nanjing Normal University (approval No: IACUC-20200501). Muscle tissue was collected for nucleic acid extraction. DNA was extracted using the phenol/chloroform DNA extraction method. Following DNA extraction, a sequencing library was prepared according to standard PacBio protocols and sequenced on the PacBio Sequel II system. We obtained 99.49 Gb (49×) of PacBio HiFi circular consensus sequencing (CCS) reads (14.5 kb average length and 14.7 kb N50) with a predicted accuracy of at least Q20 (99%) using *pbccs* (Supplementary Table S1). *K*-mer analysis (*k*=21) using GenomeScope v2.0 combined with Jellyfish v2.3.0 estimated a 2.03 Gb genome with 0.29%

Received: 30 June 2022; Accepted: 22 September 2022; Online: 22 September 2022

Foundation items: This work was supported by the National Natural Science Foundation of China (NSFC, 32070409 to S.X.X.), Key Project of the NSFC (32030011 to G.Y.), Priority Academic Program Development of Jiangsu Higher Education Institutions to G.Y. and S.X.X., and Qinglan Project of Jiangsu Province to S.X.X.

heterozygosity (Figure 1A). The PacBio HiFi long reads were used for *de novo* genome assembly with Hifiasm v0.16.1-r375. We obtained a 2.30 Gb genome assembly with 120 contigs and a contig N50 of 82.36 Mb (Supplementary Table S2).

To obtain a chromosome-level assembly, we generated a Hi-C library from muscle tissue, with DpnII as the chromatin-digesting restriction enzyme. Hi-C reads then were mapped to the contig assembly using bwa v0.7.17 and scaffolded using SALSA v2.2. The SALSA output files were loaded into Juicebox v1.11.08 to visualize the Hi-C contact map. A total of 269.85 Gb of clean Hi-C reads were generated, and 2.22 Gb (96.93%) of the long-read genome assembly were anchored to 22 pseudochromosomes (Figure 1B), consistent with the general Delphinidae karyotype (2*n*=44) (Brookwell et al., 2021). The final *P. electra* chromosome-level genome had a scaffold N50 of 102.72 Mb. This is the most contiguous Delphinidae genome assembly to date, with a contig N50 at least one order of magnitude larger than other published toothed whale genomes (e.g., 82.36 vs. 9.53 Mb for *Tursiops truncatus*) (Figure 1C). Further evaluation using Merqury v1.3 revealed that base-call accuracy of the assembly was QV68.82 (0.14 errors per 1 Mb), indicating very high accuracy (Supplementary Table S2). Using the 9 226 mammalian genes in BUSCO v5.2.2, we achieved 96.6% completeness (Figure 1D). Thus, the new *P. electra* genome assembly is a well-assembled and nearly complete reference quality genome.

Repetitive sequences were identified by homology- based and *de novo* strategies. RepeatMasker v4.1.2 and RepeatProteinMasker v4.1.2 were used to identify repetitive elements at the DNA and protein levels, respectively. RepeatModeler v2.0.3 was performed for *de novo* prediction, and Tandem Repeats Finder v4.09.1 was applied to annotate tandem repeats. In total, 43.12% (1 011.15 Mb) of the *P. electra* genome contained repetitive sequences (Supplementary Table S2), mainly long interspersed nuclear elements (LINEs, 25.10%), followed by long terminal repeats (LTRs, 5.59%), DNA transposons (3.33%), short interspersed nuclear elements (SINEs, 2.54%), and simple repeats (2.40%) (Supplementary Figure S1 and Table S2). Approximately 4.16% of the genome was annotated as unknown repetitive sequences.

Protein-coding genes were annotated using homology-based, *de novo*, and RNA sequencing (RNA-seq)-assisted prediction methods. *Ab initio* gene prediction and gene model testing were performed using BRAKER v2.1.5 to train Augustus v3.4.0 and GeneMark-ES/ET/EP 4.69_lic, integrating evidence from the OrthoDB database and transcriptome. Gene structure annotations of the *P. electra* genome were then generated using BRAKER. We extracted total RNA from muscle using TRIzol reagent. The RNA-seq library was generated using a TruSeq RNA v2 kit and sequenced. RNA-seq alignments were produced using HISAT2 v2.2.1 and assembled into transcripts with the genome-guided assembler Stringtie v2.2.1.

Human (*Homo sapiens*), mouse (*Mus musculus*), cattle (*Bos taurus*), dog (*Canis lupus familiaris*), goat (*Capra hircus*), common bottlenose dolphin (*T. truncatus*), Pacific white-sided dolphin (*Lagenorhynchus obliquidens*), killer whale (*Orcinus orca*), long-finned pilot whale (*Globicephala melas*), blue whale (*Balaena mysticetus*), and minke whale (*Balaenoptera acutorostrata*) protein sequences were retrieved from the NCBI database and used as homology evidence. Homology-based gene functions were assigned using DIAMOND v2.0.14 against the UniProtKB database with the sensitive mode "--more-sensitive -e 1e-5". We annotated protein domains and GO and KEGG pathways using eggNOGmapper v2.1.7. BUSCO v5.2.2 was used to assess the completeness of the final gene set. We predicted 21 492 protein-coding genes in the *P. electra* genome, with average coding sequence, exon, and intron lengths of 1 432, 179, and 3 405 bp, respectively (Supplementary Figure S2 and Table S2). According to BUSCO assessment, 91.40% of the complete mammalian BUSCO genes were found, indicating high annotation quality. Furthermore, 18 828 (87.61%) and 18 499 (86.07%) of the predicted genes in the *P. electra* genome were annotated in the UniProt and eggNOG databases, respectively.

We used OrthoFinder v2.5.4 to identify 19 677 orthologous groups in seven cetaceans and five terrestrial mammals (Supplementary Table S3). Among the 21 492 protein-coding genes annotated in the genome, 18 448 (86.02%) were clustered into 16 262 gene families, including 61 with 163 species-specific genes (Supplementary Figure S3). Single-copy orthologous genes (9 947) were aligned using MAFFT v7.490 with the L-INS-i iterative refinement method. We used trimAl v1.4 to remove alignment gaps and unreliable regions. The aligned sequences were then concatenated using *catsequences* (https://github.com/ChrisCreevey/catsequences). IQ-TREE v2.1.2 was used to construct a maximum-likelihood phylogenetic tree. All branches had 100/100 bootstrap support, showing consistent phylogeny with previous studies (McGowen et al., 2020) and supporting a sister relationship between *P. electra* and the long-finned pilot whale (*G. melas*) (Supplementary Figure S4). Species divergence time was estimated using MCMCTREE in PAML v4.9. Three divergence time points from TimeTree were used to calibrate the divergence times between (a) humans and mice (88–90 Ma), (b) humans and dogs (94 Ma), and (c) cattle and whales (56–60 Ma). Our analysis indicated that *P. electra* and long-finned pilot whales diverged 3.54–6.1 Ma (Supplementary Figure S3).

Syntenic blocks between cattle, bottlenose dolphin, and *P. electra* were defined using MCscanX, based on the core-orthologous gene sets identified by DIAMOND v2.0.14 with an *E*-value threshold of 1×10^{-10} (at least 10 syntenic genes and a maximum of six intervening genes allowed), and visualized using TBtools v1.098722. Comparison of the high-quality chromosome-level genome of bottlenose dolphin with that of *P. electra* showed a highly similar synteny, consistent with the high similarities of their genomes, thereby validating the accuracy of the *P. electra* genome assembly at the chromosome level (Figure 1E). A somewhat surprising result was the degree of synteny between cattle and *P. electra*, species that shared a common ancestor ~55 Ma (Supplementary Figure S3). Comparing *P. electra* and cattle, we identified several fission and fusion events. For example, Chr14 in the *P. electra* genome is a fusion of Chr13 and Chr25 in cattle (Figure 1E). In addition, the number of syntenic blocks

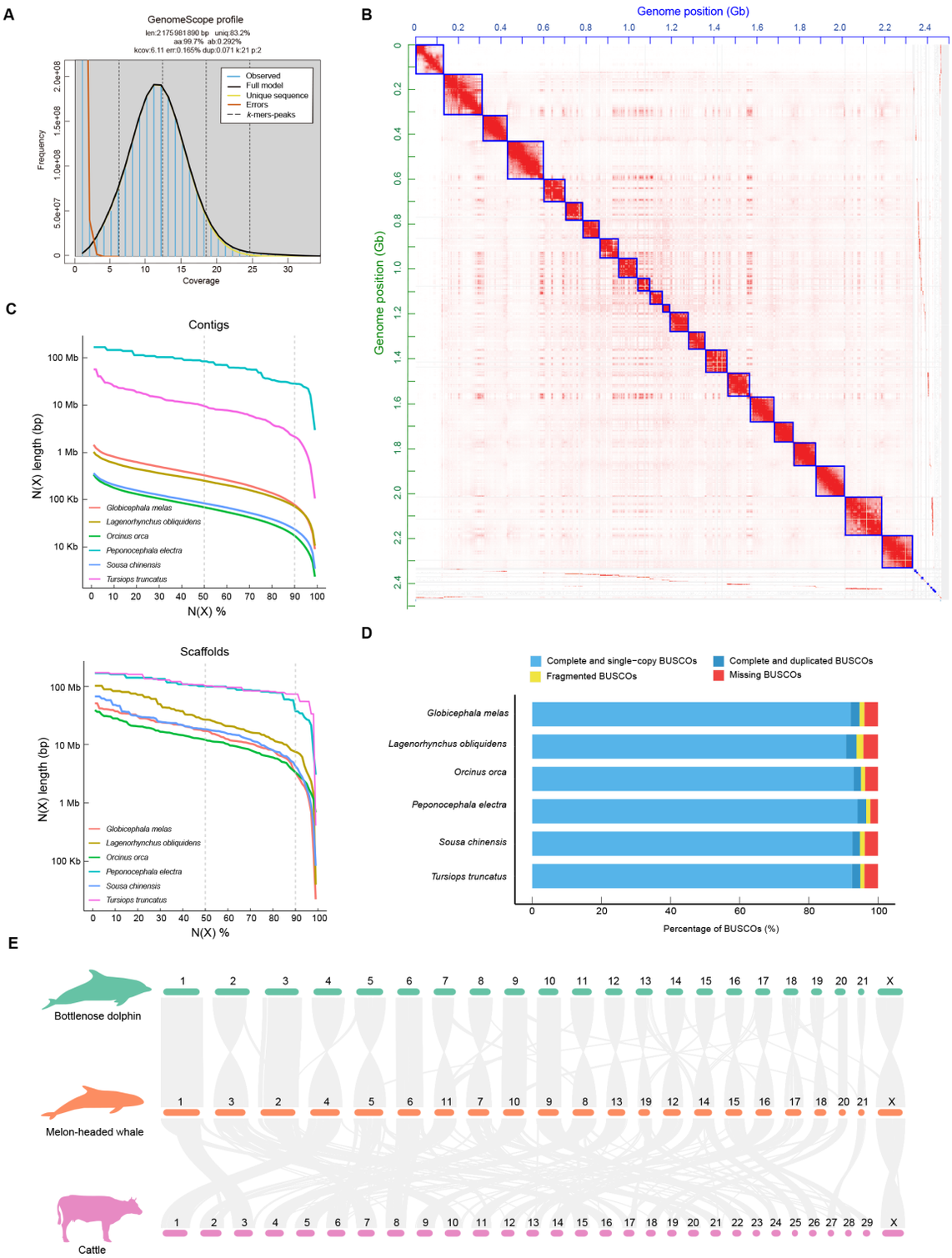

Figure 1 Overview of *P. electra* genome assembly

A: Estimation of genome size and heterozygosity rate for *P. electra* based on distribution of *k*-mers (*k*=21) of PacBio HiFi reads. len: Genome haploid length, aa: Homozygous, ab: Heterozygous, kcov: Mean *k*-mer coverage for heterozygous bases, err: Read error rate, dup: Average rate of read duplications, k: *k*-mer size, p: Ploidy. B: Genome-wide heatmap of Hi-C interactions among 22 chromosomes of *P. electra*. C: Plots comparing contig and scaffold statistics of six species in Delphinidae family. N(X)% value implies that X% of the assembly consists of contigs (top) or scaffolds (bottom) of at least this size. Y-axis represents genome sequence length. D: Comparison of genome completeness using mammalian conserved genes in BUSCO. E: Chromosome synteny among cattle, bottlenose dolphin, and *P. electra*.

showed that the X chromosome is conserved between cetacean and cattle (Figure 1E).

Gene family evolution was determined using CAFÉ v4. The overall *P*-value of each branch and node was calculated using the Viterbi method in CAFÉ. A total of 829 expanded and 1 600 contracted gene families (P<0.05) were detected in the *P. electra* genome compared to the last common ancestor shared with *G. melas* (Supplementary Figure S3). These numbers are similar to gene family dynamics of the Chinese white dolphin (*Sousa chinensis*) based on a similar dataset (Jia et al., 2019). Gene family analysis is inherently sensitive to genome assembly and gene annotation errors. Such analysis requires gene sets from "high-quality" genome chromosome-level assemblies. Using BLAST, CAFE analysis can result in a gene family contraction call in the event of sequence gaps. Nevertheless, we observed an expansion for genes enriched in several GO terms, including "wnt signaling pathway", "head development", and "neuron projection morphogenesis" (P<0.01) using Metascape (Supplementary Figure S5). We speculate that these genes may play a role in the development of the pointed, melon-shaped head in *P. electra* (Marchini et al., 2021), but await the availability of additional chromosome-level cetacean genomes.

The Codeml program in PAML v4.9 was employed to detect positively selected genes (PSGs) in the *P. electra* genome using the branch site model. A total of 190 PSGs were identified in *P. electra* (Supplementary Table S4), which were functionally enriched in "metanephric nephron development", "neuron development", and "forebrain development" (P<0.01; Supplementary Figure S6). Of these PSGs, 10 may be involved in the adaptive evolution of *P. electra* for complex social behavior and brain expansion, including cut-like homeobox 1 (*CUX1*), FAT atypical cadherin 4 (*FAT4*), neurotrophic receptor tyrosine kinase 3 (*NTRK3*), and diencephalon/mesencephalon homeobox 1 (*DMBX1*). However, further studies are needed to determine the putative roles and function of PSGs in *P. electra*.

In the present study, we assembled a high-quality chromosome-level genome of *P. electra*, the first reference genome for cetaceans in the subfamily Globicephalinae. This study provides a useful genomic resource for further studies on aquatic adaptations in cetacean species and the molecular mechanisms underlying their biological traits.

DATA AVAILABILITY

The raw genome sequencing data for *P. electra* were deposited in the NCBI Sequence Read Archive (SRA) database under BioProjectID PRJNA800182 and were deposited in the Genome Sequence Archive (GSA) database under Accession No. CRA004842. The genome assembly of *P. electra* was submitted to the Science Data Bank database (http://cstr.cn/31253.11.sciencedb.j00139.00033). The genome assembly, genome annotation, coding sequences, protein sequences, and functional annotation files were deposited in Figshare: https://doi.org/10.6084/m9.figshare.19407029.v1.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

S.X.X. designed the study. Z.P.Y., X.L., and B.Z. were responsible for genome assembly and genome annotation. L.S. collected the data. Z.P.Y. prepared the original manuscript. S.X.X., I.S., and G.Y. revised the manuscript. All authors read and approved the final version of the manuscript.

Zhen-Peng Yu[1], Xing Liu[1], Biao Zhang[1], Lei Shan[1], Inge Seim[1], Guang Yang[1], Shi-Xia Xu[1,*]

[1] *Jiangsu Key Laboratory for Biodiversity and Biotechnology*, *College of Life Sciences*, *Nanjing Normal University*, *Nanjing*, *Jiangsu* 210023, *China*

*Corresponding author, E-mail: xushixia@njnu.edu.cn

REFERENCES

Brookwell R, Finlayson K, van de Merwe JP. 2021. A comparative analysis of the karyotypes of three dolphins–*Tursiops truncatus* Montagu, 1821, *Tursiops australis* Charlton-Robb et al., 2011, and *Grampus griseus* Cuvier, 1812. *Comparative Cytogenetics*, **15**(1): 53−63.

Fox KCR, Muthukrishna M, Shultz S. 2017. The social and cultural roots of whale and dolphin brains. *Nature ecology & Evolution*, **1**(11): 1699−1705.

Jefferson TA, Webber MA, Pitman RL. 2008. Marine Mammals of the World: A Comprehensive Guide to Their Identification. London: Academic Press.

Jia KT, Bian C, Yi YH, Li YP, Jia P, Gui D, et al. 2019. Whole genome sequencing of Chinese white dolphin (*Sousa chinensis*) for high-throughput screening of antihypertensive peptides. *Marine Drugs*, **17**(9): 504.

Kurihara N, Amano M, Yamada TK. 2016. Decrease in tooth count in melon-headed whales. *Journal of Zoology*, **300**(1): 8−17.

Marchini M, Hu DN, Lo Vercio L, Young NM, Forkert ND, Hallgrímsson B, et al. 2021. *Wnt* signaling drives correlated changes in facial morphology and brain shape. *Frontiers in Cell and Developmental Biology*, **9**: 644099.

McGowen MR, Tsagkogeorga G, Álvarez-Carretero S, Dos Reis M, Struebig M, Deaville R, et al. 2020. Phylogenomic resolution of the cetacean tree of life using target sequence capture. *Systematic Biology*, **69**(3): 479−501.

Perrin WF, Würsig B, Thewissen JGM. 2009. Encyclopedia of Marine Mammals. 2nd ed. London: Academic Press.

Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. 2018. Human ageing genomic resources: new and updated databases. *Nucleic Acids Research*, **46**(D1): D1083−D1090.

West KL, Walker WA, Baird RW, Webster DL, Schorr GS. 2018. Stomach contents and diel diving behavior of melon-headed whales (*Peponocephala electra*) in Hawaiian waters. *Marine Mammal Science*, **34**(4): 1082−1096.