scientific data

OPEN

DATA DESCRIPTOR

A high-quality chromosome-level genome assembly of a feather star *Glyptometra* sp. from a deep seamount

Jing Wang[1,2,3], Shao'e Sun[4,5,6], Zijie Mei[4,5,6], Zhongli Sha[4,5,6] & Nansheng Chen[1,2,3] ✉

Feather stars are an important group for studying the evolution and environmental adaptability. However, the absence of a high-quality feather star genome from deep sea has hindered a thorough understanding of their adaptive evolution. In this study, we reported the first chromosome-level genome assembly of *Glyptometra* sp. from a deep seamount, Zhenbei seamount, in South China Sea. The *Glyptometra* sp. CNS01629 genome assembly has a total length of 1.14 Gb with a contig N50 length of 19.0 Mb and a scaffold N50 length of 84.75 Mb, encompassing 13 chromosomes, highlighting high genome continuity. Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment yielded a completeness estimate of 98.1% for this assembly. We predicted a total of 20,814 protein coding genes (PCGs), with 20,309 functionally annotated. The availability of this high-quality chromosome-level genome of *Glyptometra* sp. CNS01629 contributes to our understanding of the origin and evolution of crinoids.

Background & Summary

Crinoidea is the most basal group in the phylum Echinodermata. Crinoids were one of the most significant Paleozoic and Mesozoic benthic marine animal groups[1,2]. After the Permian-Triassic extinction event, almost all branches of crinoids became extinct, leaving behind over 6000 species in the form of fossil records[1,3]. According to WoRMS, the existing crinoids are divided into 4 orders, 33 families, and 193 genera, with over 650 species[4]. They are usually divided into the free-living feather stars (or comatulids) and the sessile sea lilies, a distinction that, based on recent molecular analysis, corresponds to two different ways of life rather than a true taxonomic relationship[5,6]. Among them, the comatulids (Comatulida) appeared in the Late Triassic, and are the most diverse crinoids in recent marine ecosystems, with over 550 extant species[7]. Comatulids are the only extant crinoid group that is globally distributed in both shallow and deep waters[8]. The ancestors of crinoids lived mostly in shallow waters during the Paleozoic and Mesozoic eras, with most being sea lilies with stems[9]. With the intensification of the competition for shallow sea predation due to the outbreak of the Mesozoic oceanic revolution[10], shallow crinoids gradually migrated to the semi deep sea and deep sea to avoid predation pressure. Therefore, feather stars are an important group for studying the evolution and environmental adaptability.

There has been an increasing interest in feather stars within the scientific community. Most studies focused on morphological classification, molecular phylogenetic analysis, embryogenesis and metamorphosis, nervous system development and behavioral science. Recent research on molecular phylogenetic analysis showed inconsistency between morphology-based and molecular sequence-based phylogenetic analysis, revealing complexity of comatulid evolution and confusion in classification system. For example, phylogenetic analysis of 4~5 genes

[1]Laboratory of Marine Ecology and Environmental Sciences, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, 266071, China. [2]Laboratory for Marine Ecology and Environmental Science, Qingdao Marine Science and Technology Center, Qingdao, 266237, China. [3]Center for Ocean Mega-Science, Chinese Academy of Sciences, Qingdao, 266071, China. [4]Laboratory of Marine Organism Taxonomy and Phylogeny, Shandong Province Key Laboratory of Marine Biodiversity and Bio-resource Sustainable Utilization, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, 266071, China. [5]Laboratory for Marine Biology and Biotechnology, Laoshan Laboratory, Qingdao, 266237, China. [6]College of Marine Science, University of Chinese Academy of Sciences, Beijing, 100039, China. ✉e-mail: chenn@qdio.ac.cn

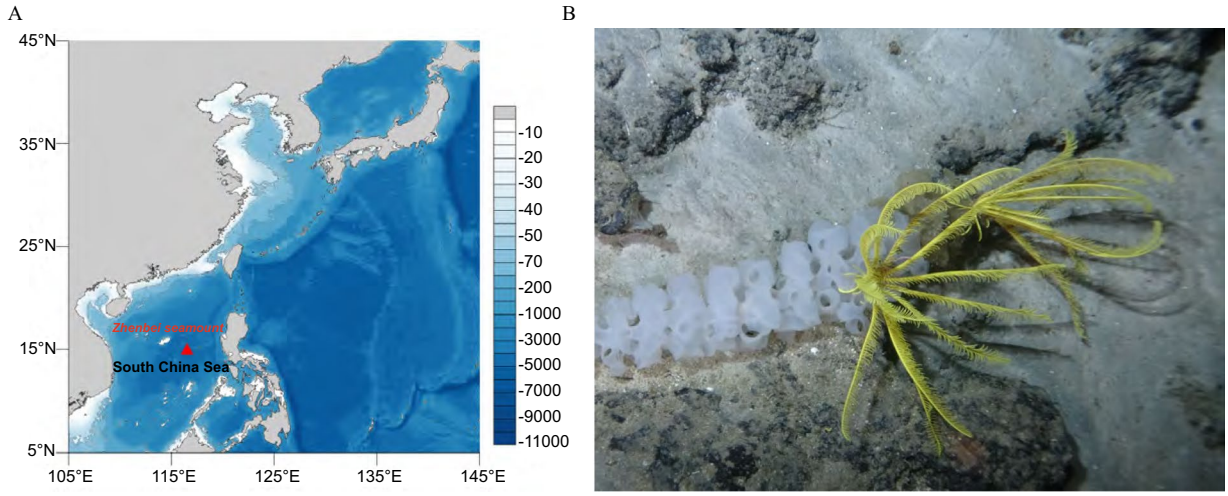

Fig. 1 (**A**) Illustration of the sampling site of Zhenbei seamount in South China Sea; (**B**) *in situ* morphology of *Glyptometra* sp. CNS01629 (left) and its same-species companion *Glyptometra* sp. CNS01628 (right) inhibiting a sponge.

(18S rDNA, 28S rDNA, 16S rDNA, COI and Cytb) suggested that two stalked crinoid groups, Bourgueticrinina and Guillecrinina, were reduced to families within Comatulida[5,6]. With the development of the next-generation sequencing technology, increasingly more mitochondrial genomes of comatulids were obtained and applied for phylogenetic analysis, *Anneisia* genus and Antedonidae family were found to be polyphyletic and feather stars have a complex evolutionary relationship based on phylogenetic analysis of 13 mitochondrial PCGs and 2 rRNAs from 25 crinoid taxa[11].

Regarding embryogenesis, it was found that crinoids have swimming ciliated lecithotrophic larvae, with sea lilies exhibiting an early dipleurula-like stage followed by a barrel-shaped doliolaria stage and feather stars featuring only a doliolaria stage[12,13]. Mercurio *et al*. defined a standardized staging system for the feather star *Antedon mediterranea*, characterized the morphological and molecular development of the main structures of the feather star body plan, and showed that the larval apical organ includes serotonergic, GABAergic and glutamatergic neurons, which developed within a conserved anterior molecular signature[14]. In terms of phototactic behavior, Nonclercq *et al*. revealed the mechanisms of crinoid opsin-based photoreception by focusing on the phototactic behavior of a model feather star species *Antedon bifida*, annotated and described the complete opsin gene repertoire (*Abif-opsins 4.1*, *4.2* and *4.3*), and assessed the precise anatomical location of several opsin-based photoreceptors[15].

Overall, the studies of comatulids have been mainly focused on the comatulids in shallow waters, with comatulids in deep sea are mostly neglected. In this study, we constructed a chromosomal-level genome of a feather star (*Glyptometra* sp. CNS01629) in a deep seamount in the South China Sea, using combined short and long reads sequencing, and Hi-C technology-assisted genome assembly. Comparative genomic analysis with 12 species in Echinodermata helped us to identify species-specific genes. This high-quality chromosome-scale genome provides a robust foundation for future research into comatulid genetics, functional gene discovery, and the evolutionary genomics of echinoderms and deuterostome.

Methods

Sample collection. The feather star sample CNS01629 (an internal identifier), used in this study was collected at 775.8 m from the Zhenbei seamount in the South China Sea (station FX-Dive 315, 116°34′E, 15°04′N) by the Remotely Operated Vehicle (ROV) Faxian during the cruise of the R/V Kexue from July 15 to August 13 in 2022 (Fig. 1). The sample was photographed on board, small tissues of fresh specimens were cut off and stored at −80 °C for DNA extraction in further genomic studies, and the remaining main part was fixed by 95% ethanol for morphological classification. The sample CNS01629 is identified as *Glyptometra* sp. based on some key characteristics: The genital pinnules taper evenly from the usually more or less broadened earlier segments to a delicate tip, the portion beyond the gonad being much longer than the gonad itself; the division series and arms are rounded dorsally, with or without a fine median abruptly raised line and a tubercular ornamentation on the division series and arm bases; and the oral pinnules are little or not at all longer than the genital and distal pinnules[16]. The material (catalogue number MBM288789) is deposited in the Marine Biological Museum of the Chinese Academy of Sciences (MBMCAS) at Qingdao, China.

Nucleic acid extraction, library preparation, and sequencing. An arm tissue was used for genomic DNA extraction. High-quality genomic DNA was extracted using a modified CTAB method[17]. The quality and quantity of the extracted DNA were examined using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and electrophoresis on a 1% agarose gel, respectively. To construct DNA short-read library, we used Covaris S220 to shear extracted DNA (1 ug) to 200 bp as the input material for the library construction. The sequencing library

Technology	Total raw data (Gb)	Reads number	Sequence Depth (×)
PacBio HiFi	56	1,399	56.9
DNBSEQ-T7	135	449,606,898	136.2
Hi-C	192	642,197,477	195.1

Table 1. Summary of sequencing data for *Glyptometra* sp. CNS01629 assembly.

was generated using the VAHTS Universal DNA Library Prep Kit for MGI (Vazyme, Nanjing, China) following the manufacturer's instruction. The library was quantified using Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and sequenced on DNBSEQ-T7 platform (BGI, China)[18]. To construct the SMRTbell library, we used the SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences). Briefly, 15 μg of the genomic DNA was carried into the first enzymatic reaction to remove single-stranded overhangs followed by treatment with repair enzymes to repair any damage that may be present on the DNA backbone. After DNA damage repair, ends of the double-stranded fragments were polished and subsequently tailed with an A-overhang. Ligation with T-overhang SMRTbell adapters was performed at 20 °C for 60 mins. Following ligation, the SMRTbell library was purified with 1X AMPure PB beads. The size distribution and concentration of the library were assessed using the FEMTO Pulse automated pulsed-field capillary electrophoresis instrument (Agilent Technologies, Wilmington, DE) and the Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). Following library characterization, 3 μg was subjected to a size selection step using the BluePippin system (Sage Science, Beverly, MA) to remove SMRTbells ≤ 15 kb. After size selection, the library was purified with 1X AMPure PB beads. Library size and quantity were assessed using the FEMTO Pulse and the Qubit dsDNA HS reagents Assay kit. Sequencing primer and Sequel II DNA Polymerase were annealed and bound, respectively, to the final SMRTbell library. The library was loaded at an on-plate concentration of 120 pM using diffusion loading. SMRT sequencing was performed using a single 8 M SMRT Cell on the Sequel II System with Sequel II Sequencing Kit.

A Hi-C library was constructed according to previous studies[19]. Briefly, the arm tissue was cross-linked with 4% formaldehyde for 30 min at room temperature. The crosslinking reaction was quenched with 0.125 M final concentration glycine for 5 min. Then the reaction system was put on ice for 15 mins, and centrifuged at 2500 r/min at 4°C for 10 mins. The pellet was resuspended with 20 μl of lysis buffer (1 M Tris-HCl, pH 8, 1 M NaCl, 10% CA-630, and 13 units protease inhibitor), the supernatant was then centrifuged at 5000 r/min at room temperature for 10 mins. The pellet was washed twice in 100 μl ice cold 1x NEB buffer and then centrifuged for 5 mins at 5000 r/min. The nuclei were re-suspended by 100 μl NEB buffer and solubilized with dilute SDS followed by incubation at 65 °C for 10 mins. After quenching the SDS by Triton X-100, an overnight digestion was applied to the samples with a 4-cutter restriction enzyme MboI (NEB, 400 units) at 37 °C on a rocking platform. The following steps involved marking the DNA ends with biotin-14-dCTP (Invitrogen) and blunt-end ligation of the cross-linked fragments. The proximal chromatin DNA was re-ligated by ligation enzyme. The nuclear complexes were reverse cross-linked by incubation with the proteinase K at 65°C. DNA was purified by the phenol-chloroform extraction. Biotin was removed from non-ligated fragment ends using T4 DNA polymerase. Ends of sheared fragments by sonication (300–500 base pairs) were repaired by the mixture of T4 DNA polymerase, T4 polynucleotide kinase and Klenow DNA polymerase. Biotin-labeled Hi-C samples were specifically enriched using streptavidin C1 magnetic beads. After adding A-tails to the fragment ends and following ligation by the adapters, Hi-C sequencing libraries were amplified by PCR (12–14 cycles) and sequenced on DNBSEQ-T7 platform (BGI, China). The 192 Gb raw sequence data were produced. After removing the reads with adapters, and with the proportion of N (N indicates undetermined base) greater than 10%, and then removing the paired reads when the number of low-quality (below 5) bases in a single-end sequencing read exceeded 20% of its length, 101 Gb clean data were obtained, with Q20 = 96.06% and Q30 = 90.72%, which were used for chromosomal level genome assembly (Table 1).

For RNA-seq, middle segments of arms and distal arms were originally collected. Total RNA was extracted using Trizol reagent (Invitrogen, CA, USA). RNA concentration, purity and integrity was monitored by NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and a Bioanalyzer 2100 system (Agilent Technologies, CA, USA). RNA contamination was assessed by 1.5% agarose gel. To construct the library for RNA-seq, we used Covaris S220 to shear extracted RNA (1 ug per sample) to 200 bp. The mRNA was purified from the total RNA using poly(T) oligo attached magnetic beads. Sequencing libraries were generated using the VAHTS Universal V6 RNA-seq Library Kit for MGI (Vazyme, Nanjing, China) following the manufacturer's recommendations and index codes were added to attribute sequences to each sample. The library quantification was measured using Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). Subsequently, sequencing was performed on a DNBSEQ-T7 platform. To construct the library for Iso-seq, full-length cDNA of mRNA was synthesized using a SMARTer™ PCR cDNA Synthesis Kit (Clonetech, CA, USA). Eukaryotic mRNA 3′ terminal has a poly-A tail structure, Primers with Oligo dT were used to pair A-T bases with poly-A as primers for reverse synthesis of cDNA, and primers were added to the end of full-length cDNA of reverse synthesis. The full-length cDNA was amplified by PCR, and the product was purified by PB magnetic beads to remove some small fragments of cDNA less than 1 kb. Repair the end and connect the SMRT dumbbell connector. The unconnected fragments were digested by exonuclease and purified by PB magnetic beads to obtain the sequencing library. The library quantification and size were measured using Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Subsequently, SMRT sequencing was performed on a PacBio Sequel II platform.

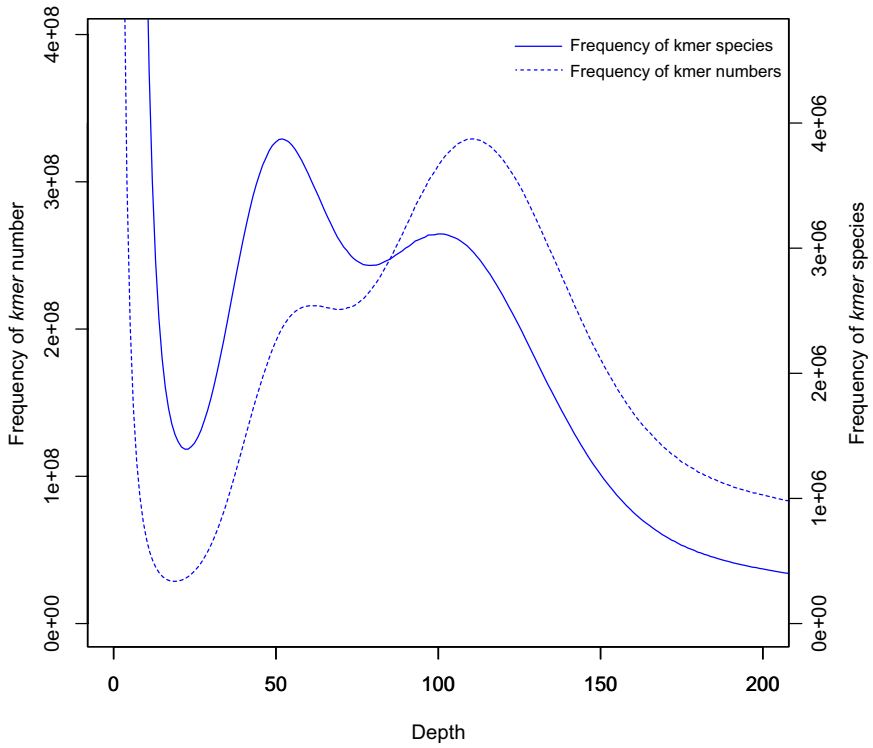

Fig. 2 A *k*-mer (17-mer) distribution curve for estimation of the genome size, heterozygosity rate, and repeat content.

Genome size estimation. The short-reads from DNBSEQ-T7 platform were quality filtered by fastp v0.23.4[20]. The quality filtered reads were used for genome size estimation. We generated the 17-mer occurrence distribution of sequencing reads from the DNA short-read library using the *k*-mer method. The occurrences of *k*-mers were counted by jellyfish 2.2.7[21]. Based on the *k*-mer analysis, the main peak was observed at depth = 109 (Fig. 2). The genome size calculated using the formula Kmer-number/depth was 996.5 Mb, and the adjusted genome size was 984.3 Mb. The heterozygosity rate and repeat content of *Glyptometra* sp. CNS01629 were 1.54%, and 64.82%, respectively.

Genome assembly. For preliminary assembly, SMRTbell libraries were sequenced on a PacBio Sequel II system and consensus reads (HiFi reads) were generated using ccs software (https://github.com/pacificbiosciences/unanimity) with the parameter '-minPasses 3'. To further validate and improve the assemblies, we generated 56 Gb (56.9×) PacBio HiFi reads (Table 1). These long (~15 kb) and highly accurate (>99%) HiFi reads were assembled using hifiasm v0.16.1 with the parameter '-l 3'[22]. For anchored contigs, the clean reads generated from the Hi-C library and were mapped to the *Glyptometra* sp. CNS01629 preliminary assembly using ALLHiC (version 0.9.8). The Hi-C heatmap was visualized using Juicebox 1.11.08[23]. The genome landscape illustrating the length, repeat element density, gene density and GC content was created by circos-0.69-9[24]. The assembled genome of *Glyptometra* sp. CNS01629 consisted of 13 chromosomes (Fig. 3; Table 2), with a contig anchoring rate of 93.19%. The assembled *Glyptometra* sp. CNS01629 genome has a total length of 1.14 Gb with a contig N50 length of 19.0 Mb and a scaffold N50 length of 84.75 Mb (Table 3).

Genome annotation. *Repeats annotation.* Repeat sequences of the *Glyptometra* sp. CNS01629 genome were identified by homology alignment and de novo prediction. For homology-based analysis, known transposable elements (TEs) within the *Glyptometra* sp. CNS01629 genome were identified according to the Repbase TE library[25,26] employing RepeatMasker (V 4.1.2-p1)[27] and RepeatProteinMask (v4.1.2). For de novo prediction, we constructed a de novo repetitive elements library of the *Glyptometra* sp. CNS01629 genome using RepeatModeler (v2.0.3). Then all repeat sequences with >100 bp lengths and with gap 'N' content of less than 5% constituted the raw TE library. A custom library (a combination of Repbase and the de novo TE library which was processed by vsearch (v1.11.2) to yield a non-redundant library) was supplied to RepeatMasker to identify the repeat contents[27]. We also identified tandem repeats using the Tandem Repeat Finder (TRF, v4.09.1) package[28]. The majority (66.12%) of the *Glyptometra* sp. CNS01629 genome was estimated to be repetitive elements (Table 4), which is similar to the predicted content of repetitive elements using *k*-mer analysis (64.82%). The top 3 most frequent categories of repetitive elements in the *Glyptometra* sp. CNS01629 genome were DNA transposons (34.74%), long terminal repeats (LTRs, 15.85%), and long interspersed nuclear elements (LINEs, 11.81%). The distribution of repetitive elements was highly uneven in the *Glyptometra* sp. CNS01629 genome, with peaks occurring at regions with low gene density (Fig. 4).

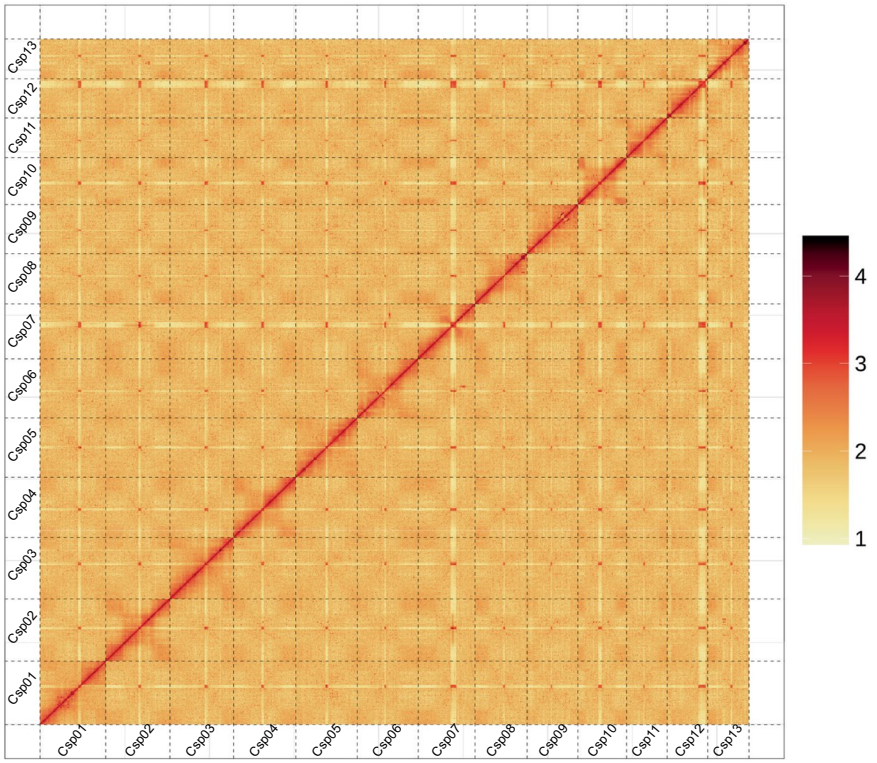

Fig. 3 Hi-C analysis of *Glyptometra* sp. CNS01629 genome contigs. Chromosomes are arranged in the size order from left to right and from bottom to top. The color bar illuminates the logarithm of the contact density from red (4) to white (0) in the plot.

Chromosome	Length (bp)	Contig number
Csp01	97,022,918	16
Csp02	96,961,588	8
Csp03	94,157,808	7
Csp04	93,400,800	9
Csp05	92,311,335	5
Csp06	90,437,529	5
Csp07	84,750,652	9
Csp08	78,475,324	9
Csp09	76,220,805	6
Csp10	72,370,411	7
Csp11	61,647,183	3
Csp12	61,019,220	20
Csp13	60,966,786	5

Table 2. The length and contig number of chromosomes in *Glyptometra* sp. CNS01629 genome.

Protein coding gene annotation. Protein coding gene (PCG) structures were predicted using three basic strategies: de novo, homology-based, and transcriptome sequencing-based prediction. Prior to gene prediction, the assembled *Glyptometra* sp. CNS01629 genome was masked using RepeatMasker. For homology-based gene prediction, gene sets from three echinoderms *Anneissia japonica*, *Asterias amurensis*[29], and *Apostichopus japonicus*[30] were used. Blastall (v2013.11.29), Solar (v0.9.6) and Genewise (v2.4.1) were used to align the protein sequences of *Glyptometra* sp. CNS01629 to the three other species for homology-based gene prediction. In terms of transcriptome sequencing-based prediction, we applied both RNA-seq for splicing support and expression evidence and Iso-seq for defining full-length transcript structures in this study. For transcriptome sequencing-based prediction, we first mapped short RNA reads by RNA-seq to the reference genome using HISAT2 (v2.2.1), then aligned the Circular Consensus Sequencing (CCS) reads from Iso-seq with the reference genome using minimap2, and both of them were used for gene prediction using StringTie (v2.2.1) with the default parameters. In addition, the transcripts from the short RNA reads assembled by Trinity (v2.8.5) and CCS were clustered to eliminate redundancy by Cd-hit (V4.8.1)[31], and the final non-redundancy transcript was predicted by PASA

Sample ID	Contig length	Scaffold length	Contig number	Scaffold number
Total length (bp)	1,137,167,871	1,137,177,471	451	355
Max length (bp)	49,813,272	97,022,918	—	—
Min length (bp)	2,313	2,313	—	—
Number >= 2000	—	—	451	355
N50	19,016,247	84,750,652	21	7
N60	17,013,094	78,475,324	27	8
N70	13,979,695	76,220,805	35	9
N80	10,913,414	61,647,183	44	11
N90	3,768,898	60,966,786	62	13
GC content (%)	37.22	37.22	—	—
Placed scaffolds		1,059,742,359		13
Unplaced		77,435,112		342
Anchored rate (%)		93.19%		

Table 3. Assembly statistics of *Glyptometra* sp. CNS01629 genome assembly.

Type	Number	Length(bp)	Percentage (%)
SINE	64,797	10,859,168	0.95
LINE	634,224	134,310,287	11.81
LTR	827,893	180,244,318	15.85
DNA	1,786,491	395,003,836	34.74
Unknown	132,318	21,099,795	1.86
Total		751,951,940	66.12

Table 4. Summary statistics of repetitive sequences in *Glyptometra* sp. CNS01629 genome.

(2.5.2). Based on the genome sequence, we used Augustus (v3.5)[32–34] and SNAP (v2013.11.29) to perform ab initio gene prediction. Models used for gene predictors were trained from a set of high-quality proteins generated from the RNA-seq dataset and Iso-seq dataset. The the final non-redundancy transcript from CCS and Trinity transcripts were used to predict Open Reading Frames (ORFs) using PASA (2.5.2), and full-length cDNA was screened as a training set[35]. Finally, EvidenceModeler (EVM, v1.1.1) was applied for combining predicted results from three strategies above. A total number of 20,814 PCGs were annotated in the *Glyptometra* sp. CNS01629 genome (Table 5). The average gene length, average coding sequence (CDS) length, average exon length, and average intron length were 21,413 bp, 1,654 bp, 201 bp, and 2,728 bp, respectively.

Gene functions were assigned according to the best match by aligning the protein sequences to the Swiss-Prot[36] using Blastp (v2.2.26, with a threshold of E-value $\leq 1e^{-5}$). The motifs and domains were annotated using InterProScan (v5.59–91.0)[37] by searching against publicly available databases Pfam[38]. The Gene Ontology (GO) IDs for each gene were assigned according to the corresponding InterPro entry. We predicted the proteins function by transferring annotations from the closest BLAST hit (E-value $< 10^{-5}$) in the Swissprot[36] database and DIAMOND (v0.8.22) hit (E-value $< 10^{-5}$)/ BLAST hit (E-value $< 10^{-5}$) in the National Center for Biotechnology Information (NCBI) Non-Redundant (NR). We also mapped gene set to a KEGG pathway and identified the best match for each gene. Of these PCGs, 20,309 (97.6%) had functional annotations assigned using comparisons to public databases, including NCBIs nr, Swiss-Prot, KEGG, InterPro, Pfam, and Gene Ontology (Table 6), supporting the high quality of the genome.

Non coding RNA gene annotation. We used tRNAscan-SE (v1.4) algorithms[39] with default parameters to identify the genes associated with tRNA. For rRNA identification, we chose rRNA sequences of relative species as references, predict rRNA sequences using Blast (v2.2.26). MiRNAs and snRNAs were identified by cmsearch (v1.1.4) against the Rfam (v14.1) database[38] with default parameters. All the ncRNA annotation were summarized in Table 7.

Genome quality assessment. To examine the assembly integrity, the assembled genome and predicted PCGs were subjected to BUSCO (v5.8.0)[40] using the metazoan_db10 database to evaluate the completeness of the genome, respectively. We found that 98.1% of the core genes used by BUSCO were identified as full length in the *Glyptometra* sp. CNS01629 genome (Table 8). We further tested the completeness of the predicted PCGs using BUSCO, which identified 96.6% of the core genes as full length.

Data Records

All raw sequencing data of the short-read sequencing data, PacBio sequencing data, Hi-C sequencing data, RNA-seq data and Iso-seq data for *Glyptometra* sp. CNS01629 were deposited at NCBI under the BioProject accession number PRJNA1310249. The raw short-read sequencing data, PacBio sequencing data, and Hi-C sequencing data are available in the Sequence Reads Archive (SRA) with the accession numbers SRR35528329[41],

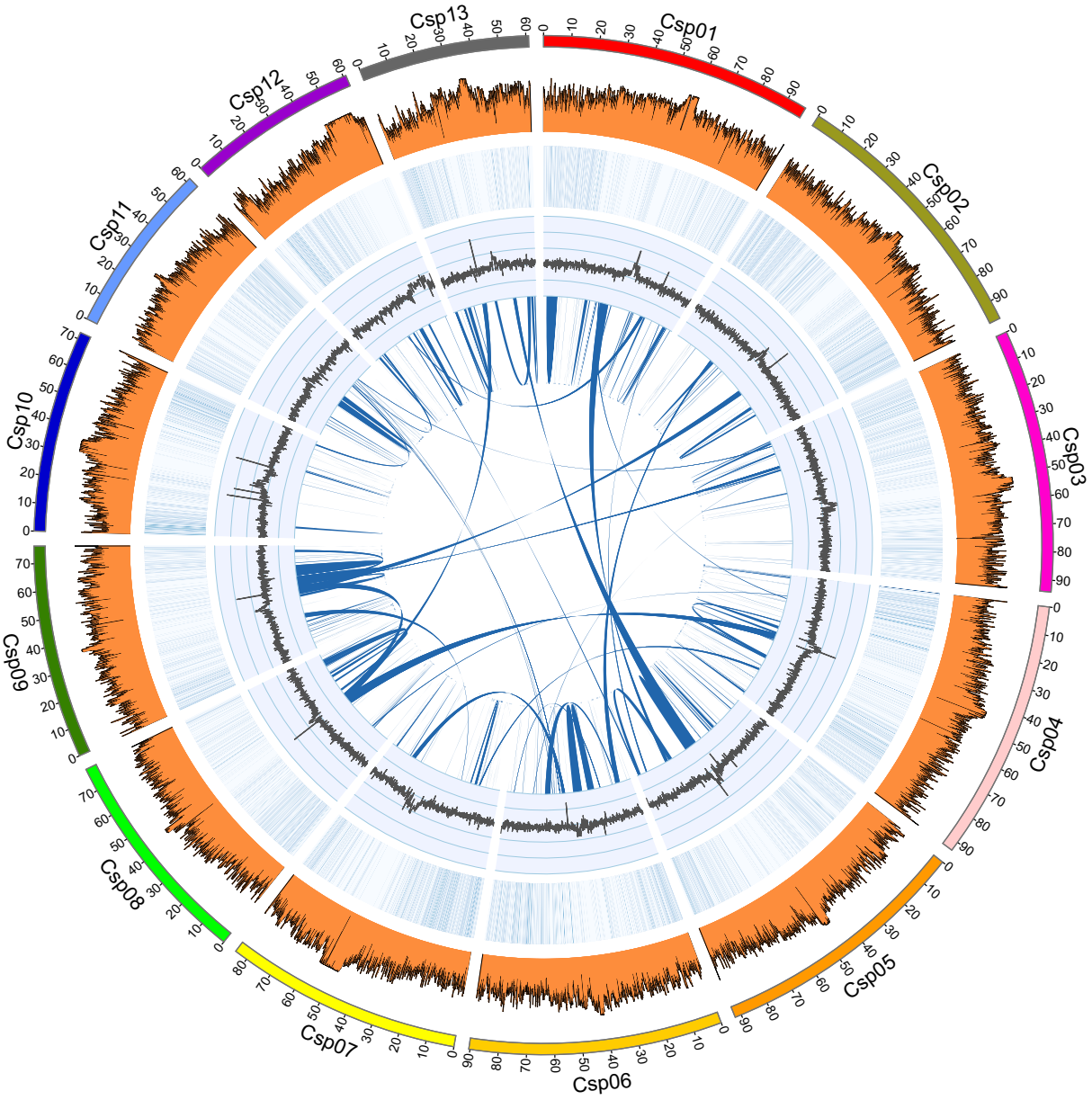

Fig. 4 The genomic landscape of *Glyptometra* sp. CNS01629 (from outer to inner circles): the 13 chromosomes, repetitive element density, gene density, GC density, and a colinear relationship across the genome, drawn in 100 kb non-overlapping windows.

Type	Gene Set	Number	Average transcript length(bp)	Average CDS length(bp)	Average exons per gene	Average exon length(bp)	Average intron length(bp)
De novo	Augustus	38,226	11,271.60	1,237.90	5.27	234.96	2,350.62
	SNAP	38,228	35,609.07	626.29	4.29	146.05	10,639.25
Homolog	*Anneissia japonica*	21,555	14,478.89	1,414.16	6.74	209.90	2,277.13
	Apostichopus japonicus	17,007	8,765.14	975.33	4.61	211.57	2,157.82
	Asterias amurensis	15,012	11,281.08	1,147.60	5.64	203.37	2,182.55
RNAseq	transcripts	35,228	29,302.58	3,449.23	8.34	413.71	3,523.58
	PASA	32,151	18,196.88	1,330.63	6.20	214.51	3,241.65
EVM	EVM	35,360	13,578.06	1,261.65	5.62	224.61	2,667.56
Pasa-update*	pasaupdate	35,016	14,480.83	1,297.52	5.79	224.28	2,754.93
Final set*	Final	20,814	21,412.68	1,654.11	8.22	201.35	2,738.49

Table 5. Statistics of PCG structure prediction in *Glyptometra* sp. CNS01629 genome.

Type	Number	Percentage (%)
Total	20,814	100.00
NR	19,963	96.00
Swissport	15,459	74.30
KEGG	16,212	77.90
InterPro	19,299	92.80
Pfam	15,337	73.70
GO	12,336	59.30
Annotated	20,309	97.60
Unannotated	505	2.43

Table 6. Statistics of gene function annotation.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA	miRNA	3,644	113.973	415,319	0.037
tRNA	tRNA	11,013	73.002	803,972	0.071
rRNA	rRNA	5,725	221.495	1,268,059	0.112
	18S	3,386	308.035	1,043,008	0.092
	28S	674	107.350	72,354	0.006
	5.8S	0	0.000	0	0.000
	5S	1,665	91.710	152,697	0.013
snRNA	snRNA	463	160.955	74,522	0.007
	CD-box	78	119.410	9,314	0.001
	HACA-box	51	209.588	10,689	0.001
	splicing	322	161.807	52,102	0.005
	scaRNA	5	164.000	820	0.000
	Unknown	7	228.143	1,597	0.000

Table 7. Statistics of ncRNA annotation in *Glyptometra* sp. CNS01629 genome.

	Genome assembly		Gene set	
Type	**Number of genes**	**Percentage (%)**	**Number of genes**	**Percentage (%)**
Complete BUSCOs (C)	936	98.1%	922	96.6%
Complete and single-copy BUSCOs (S)	890	93.3%	883	92.6%
Complete and duplicated BUSCOs (D)	46	4.8%	39	4.1%
Fragmented BUSCOs (F)	3	0.3%	15	1.6%
Missing BUSCOs (M)	15	1.6%	17	1.8%
Total BUSCO groups	954		954	

Table 8. Completeness assessment of *Glyptometra* sp. CNS01629 by BUSCO.

SRR35528324[42], SRR35528328[43], respectively. The raw RNA-seq data are available in the SRA with the accession numbers SRR35528325, SRR35528326, and SRR35528327[44–46], and the raw Iso-seq data are available in the SRA with the accession number SRR35528323[47]. The final genome assembly data and annotation files of repeats annotation and PCG annotation are available in NCBI[48] and figshare[49], respectively.

Technical Validation

To evaluate the assembly completeness and integrity, we implemented two complementary validation approaches. First, the completeness of *Glyptometra* sp. CNS01629 genome assembly was evaluated using the BUSCO[40] (in the metazoa_odb10 database), and the completeness was 98.1% (93.3% single-copied genes and 4.8% duplicated genes) (Table 8). The Hi-C heatmap revealed a well-structured interaction pattern in and around the chromosome inversion regions (Fig. 3). The short paired reads were realigned to the assembled genome using BWA[50]. There were 98.38% short-reads mapped to the genome assembly of *Glyptometra* sp. CNS01629 and the coverage reached to 99.85%. The package Merqury[51] was employed for assessing the accuracy of genome assemblies using a reference-free and *k*-mer-based approach, which resulted in a high assembly quality value (QV) of 40.71. All available evidence robustly supports the completeness and accuracy of *Glyptometra* sp. CNS01629 genome assembly.

Data availability

The data that support the findings of this study are openly available in public repositories NCBI and figshare. All raw sequencing data of the short-read sequencing data, PacBio sequencing data, Hi-C sequencing data, RNA-seq data and Iso-seq data for *Glyptometra* sp. CNS01629 were deposited at NCBI under the BioProject accession number PRJNA1310249. The raw short-read, PacBio, and Hi-C sequencing data are available in the Sequence Reads Archive (SRA) with accession numbers SRR35528329, SRR35528324, and SRR35528328, respectively. The raw RNA-seq data are available in the SRA with accession numbers SRR35528325, SRR35528326, and SRR35528327, and the raw Iso-seq data are available in the SRA with accession number SRR35528323. The final genome assembly data and annotation files of repeats annotation and PCG annotation are available in NCBI (accession number JBUEDP000000000) and figshare at https://doi.org/10.6084/m9.figshare.30145666 (version2), respectively.

Code availability

All software and pipelines were executed in accordance with the manuals and programs of the published bioinformatic tools. All programs used in this study are publicly available, with version and parameters explicitly described in the Methods section.

Received: 30 September 2025; Accepted: 25 February 2026;
Published online: 05 March 2026

References

1. Hess, H. Permian. In: Hess, H., Ausich, W. I., Brett, C. E. & Simms, M. J. (eds.) *Fossil crinoids*. Cambridge University Press, Cambridge, pp. 160–163 (1999).
2. Baumiller, T. K. *et al*. Post-Paleozoic crinoid radiation in response to benthic predation preceded the Mesozoic marine revolution. *Proc. Natl. Acad. Sci. USA* **107**, 5893–5896 (2010).
3. Baumiller, T. K. & Hagdorn, H. Taphonomy as a guide to hnctional morphology of Holocrinus, the first post-Paleozoic crinoid. *Lethaia* **28**, 221–228 (1995).
4. Messing, C. *et al*. World List of Crinoidea. Crinoidea. Accessed through: World Register of Marine Species at: https://www.marinespecies.org/aphia.php?p=taxdetails&id=123081 (2025).
5. Hemery, L., Michel, R., Améziane, N. & Eléaume, M. High-resolution crinoid phyletic inter-relationships derived from molecular data. *Cah. Biol. Mar.* **54**, 511–523 (2013).
6. Rouse, G. W. *et al*. Fixed, free, and fixed: The fickle phylogeny of extant Crinoidea (Echinodermata) and their Permian–Triassic origin. *Mol. Phylogenet. Evol.* **66**, 161–181 (2013).
7. Hess, H. & Messing, C. G. *Treatise on Invertebrate Paleontology, part T, Echinodermata 2, Crinoidea* Vol. 3 (The University of Kansas, 2011).
8. Meyer, D. L. Evolutionary implications of predation on recent comatulid crinoids from the Great Barrier Reef. *Paleobiology* **11**, 154–164 (1985).
9. Ameziane, N. & Roux, M. Biodiversity and historical biogeography of stalked crinoids (Echinodermata) in the deep sea. *Biodiversity Conserv.* **6**, 1557–1570 (1997).
10. Oji, T. Is predation intensity reduced with increasing depth? Evidence from the West Atlantic stalked crinoid *Endoxocrinus parrae* (Gervais) and implications for the Mesozoic marine revolution. *Paleobiology* **22**, 339–351 (1996).
11. Xu, Q. *et al*. Structural features and phylogenetic implications of crinoid echinoderms based on thirteen novel mitochondrial genomes. *J. Mar. Sci. Eng.* **12**, 361 (2024).
12. Barbaglio, A. *et al*. Larval development in the feather star *Antedon mediterranea*. *Invertebr. Reprod. Dev.* **56**, 124–137 (2012).
13. Nakano, H., Hibino, T., Oji, T., Hara, Y. & Amemiya, S. Larval stages of a living sea lily (stalked crinoid echinoderm). *Nature* **421**, 158–160 (2003).
14. Mercurio, S. *et al*. A feather star is born: embryonic development and nervous system organization in the crinoid *Antedon mediterranea*. *Open Biol.* **14**, 240115 (2024).
15. Nonclercq, Y. *et al*. Opsin-based photoreception in crinoids. *bioRxiv*, 2024.2008.2014.607903 (2024).
16. Clark, A. H. A monograph of the existing crinoids, vol. 1. The comatulids, pt. 4c: Superfamily Tropiometrida (the families Thalassometridae and Charitometridae) (Smithsonian Institution, 1950).
17. Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* **15**, 8–15 (1997).
18. Zhu, X. *et al*. Genome sequencing and analysis of Thraustochytriidae sp. SZU445 provides novel insights into the polyunsaturated fatty acid biosynthesis pathway. *Mar. Drugs* **18**, 118 (2020).
19. Rao, S. S. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
20. Chen, S. F. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2** (2023).
21. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
22. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
23. Robinson, J. T. *et al*. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258 (2018).
24. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
25. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
26. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
27. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc.* **25**, 4.10.11–14.10.14 (2009).
28. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
29. Wang, Y. *et al*. Chromosome-level genome assembly of the northern Pacific seastar *Asterias amurensis*. *Sci. Data* **10**, 767 (2023).
30. Sun, L., Jiang, C., Su, F., Cui, W. & Yang, H. Chromosome-level genome assembly of the sea cucumber Apostichopus japonicus. *Sci. Data* **10**, 454 (2023).
31. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
32. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439 (2006).
33. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).

34. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
35. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
36. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
37. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
38. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–230 (2014).
39. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
40. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528329 (2025).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528324 (2025).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528328 (2025).
44. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528325 (2025).
45. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528326 (2025).
46. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528327 (2025).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35528323 (2025).
48. *NCBI GenBank*, https://identifiers.org/ncbi/insdc:JBUEDP000000000.1 (2026).
49. Wang, J. & Chen, N. Chromosome-level genome assembly of *Glyptometra* sp. CNS01629. https://doi.org/10.6084/m9.figshare.30145666 (2025).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

Acknowledgements

We acknowledge the support of the Research Vessel KEXUE of the National Major Science and Technology Infrastructure from the Chinese Academy of Sciences (CAS). We are grateful to the pilots and crew of FAXIAN ROV. This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB42000000) and the Taishan Scholar Project Special Fund (to Nansheng Chen), and National Natural Science Foundation of China (42476100, to Shaoʾe Sun).

Author contributions

N.C. conceived and supervised the research project. J.W. collected and processed the sample. S.S., Z.M. and Z.S. identified the sample and carried out taxonomical identification, J.W. performed analysis and wrote the manuscript. N.C. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.