scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the tropical sea cucumber *Stichopus monotuberculatus*

Ting Chen[1,7], Yun Yang[2,3,7], Xuan Wang[1,4], Zhou Qin[1,2,4], Zhenyu Xie[3], Dingding Fan[1], Chunhua Ren[1], Hongyan Sun[2], Peng Luo[1], Xiao Jiang[1], Hao Long[3], Chang Chen[1], Wenjie Pan[1,4], Zixuan E.[1,4], Jiasheng Huang[1,4], Qianying Huang[5], Jianfeng Xu[1,2,4], Zepeng Zhang[2], Chuhang Cheng[6], Suzhong Yu[1,4], Yanhong Wang[1], Fajun Jiang[6], Aifen Yan[5] ✉ & Chaoqun Hu[1] ✉

In this study, a chromosome-level genome of the tropical sea cucumber *Stichopus monotuberculatus* was generated by a combination of Nanopore long-read, Illumina short-read, and Hi-C sequencing technologies. The final assembly was 810.54 Mb in length, with contig N50 and scaffold N50 values of 10.15 Mb and 35.36 Mb, respectively. This assembly comprised 23 pseudo-chromosomes, covering 99.82% of the genome. Completeness analysis using BUSCO indicated that 97.8% of the metazoan conserved genes were presented in their entirety. A total of 29,596 protein-coding genes were predicted, with functional annotations available for 94.43% of these genes. The high-quality genome assembly produced in this study may provide an essential foundation for future researches on resource conservation and genetic breeding of *S. monotuberculatus*.

Background & Summary

Sea cucumbers, belong to the phylum Echinodermata and the class Holothuroidea, are one of the largest groups of marine invertebrates found in diverse habitats on the sea floor worldwide[1]. The number of known sea cucumber species exceeds 1,800 in global[2], of which over 80 species hold significant economic value, encompassing both dietary and medicinal applications[3]. In addition, sea cucumbers play crucial roles in the marine ecosystem through bioturbation, organic matter processing, nutrient recycling, seawater chemistry balancing, biodiversity supporting and energy transferring in the food chain[1]. It has been further reported that sea cucumbers safeguard coral reefs by mitigating disease[4]. However, the enormous commercial demand has led to overfishing, resulting in a significant decrease in many sea cucumbers' populations[5]. Therefore, artificial breeding and release of seedlings were used to restore the wild sea cucumber resources[6,7].

Stichopus monotuberculatus is a tropical sea cucumber that is commonly distributed in the coral reefs of the Indo-Pacific Ocean, ranging from the Red Sea and Madagascar to Easter Island, and from Japan to Australia (Fig. 1a)[3,8]. Characterized by the unique morphology, the adult *S. monotuberculatus* typically grows to ~20 cm in length, displaying a brownish-yellow body wall and a quadrangular shape (Fig. 1a)[8,9]. *S. monotuberculatus* is similar to *Stichopus horrens* and *Stichopus naso* in appearance, whereas genetic barcoding indicates that *S. monotuberculatus* and *S. horrens* fall within the same clade[10,11]. The abundant high-quality protein and collagen fibers in the body wall of *S. monotuberculatus* make it highly valuable for consumption[12], earning it a premium edible sea cucumber in Asia[3]. In order to meet the commercial demand and restore the damage of wild resources

[1]Key Laboratory of Breeding Biotechnology and Sustainable Aquaculture, CAS Key Laboratory of Tropical Marine Bio-resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China. [2]Guangdong Laboratory for Lingnan Modern Agriculture, College of Marine Sciences, South China Agricultural University, Guangzhou, 510642, China. [3]Laboratory of Marine Resource Utilization in the South China Sea, Hainan University, Haikou, 570228, China. [4]University of Chinese Academy of Sciences, Beijing, 100049, China. [5]School of Medicine, Foshan University, Foshan, 528225, China. [6]Guangxi Key Laboratory of Marine Environmental Science, Guangxi Academy of Marine Sciences, Guangxi Academy of Sciences, Nanning, 530007, China. [7]These authors contributed equally: Ting Chen, Yun Yang. ✉e-mail: yanaifen@fosu.edu.cn; hucq@scsio.ac.cn

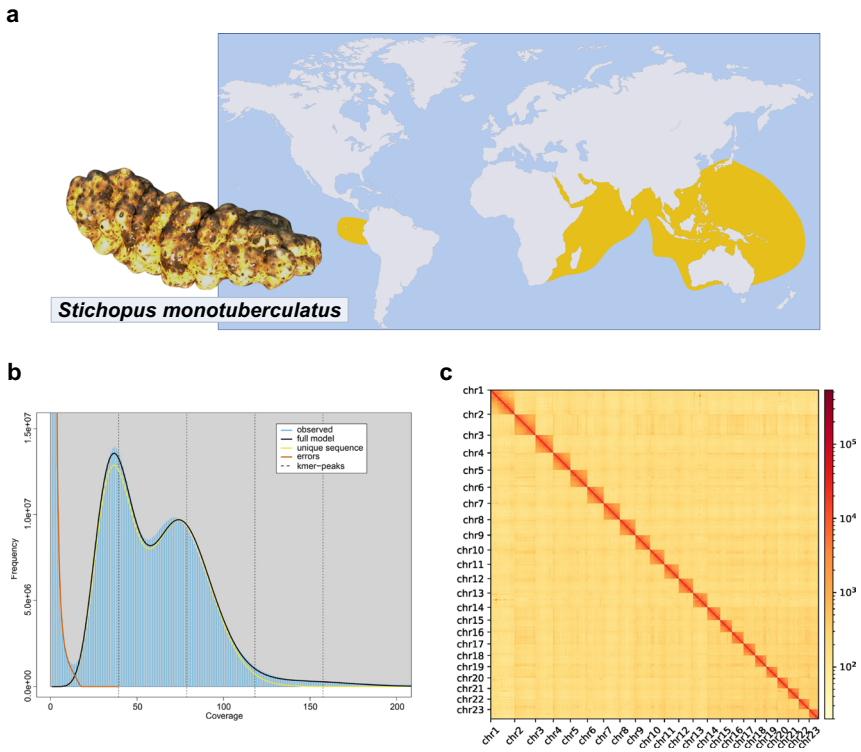

Fig. 1 *S. monotuberculatus* and its genomics feature. (**a**) Appearance and global distribution of *S. monotuberculatus*; (**b**) 21-mer seq frequency distribution in the *S. monotuberculatus* genome; (**c**) Hi-C contact heatmap of the *S. monotuberculatus* genome.

due to overfishing, the artificial spawning of *S. monotuberculatus* has recently been developed[9], and various related aquaculture techniques were also continually improved[8,13,14]. Moreover, research were further performed on taxonomic clarification[10,15], sex determination[16], genetic diversity[17], food resource[18], gut microbiota[19] and immune mechanism[20–25] in *S. monotuberculatus*. Despite the mitochondria DNA[26], transcriptomic sequencing[27], and preliminary unassembled genome draft[28] of *S. monotuberculatus* having been conducted, there remains a significant gap in integrated genomic research, restricting a more comprehensive understanding of the evolutionary and genetic traits this species.

In this study, we utilized a combination of Nanopore long-read sequencing, Illumina short-read sequencing, and Hi-C technology to produce a high-quality chromosome-level genome of *S. monotuberculatus*. The final assembly is estimated to be 810.54 Mb, with a contig N50 of 10.15 Mb and a scaffold N50 of 35.36 Mb, consisting of 23 pseudochromosomes and achieving 99.82% assembly coverage. The Benchmarking Universal Single-Copy Orthologs (BUSCO) integrity assessment indicates that 97.8% of the conserved metazoan genes are complete. The *S. monotuberculatus* genome contains 29,596 protein-coding genes, with 94.43% (27,948 genes) annotated with functional information. This high-quality genome assembly will facilitate a deeper understanding of the genomic structure and genetic traits of *S. monotuberculatus*, laying a solid foundation for future wild resource conservation, genetic breeding and aquaculture. Furthermore, this study can also give new insights to evolution and ecological adaptation mechanisms of echinoderms.

Methods

Sample collection and nucleic acid extraction. All samples used in this study were obtained from adult female *S. monotuberculatus*, collected from the natural habitat at Tanmen Port, Qionghai City, Hainan Province (19.33° N, 110.49° E). The longitudinal muscles of a female specimen were excised, washed three times with phosphate-buffered saline (PBS), quickly frozen in liquid nitrogen, and subsequently stored at −80 °C. High-quality DNA was extracted from the longitudinal muscles using the QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany) to facilitate both long-read and short-read whole-genome sequencing.

Transcriptomic sequencing was carried out on 9 tissues, including the body wall, coelomocytes, intestine, muscle, oral tentacles, Polian vesicles, respiratory tree, rete mirabile and skin. The coelomocytes were harvested from coelomic fluids that were filtered by 100-μm sterile nylon mesh and centrifuged immediately at 4 °C and 1,000 × g for 10 min. Other tissues were carefully separated using scissors and forceps, then processed similarly to the muscle specimen. Total RNA was extracted using the RNAprep Pure Plant Plus Kit (Tiangen Biotech Co. Ltd., Beijing, China) for transcriptomic sequencing.

Library preparation and sequencing. For Illumina short-read sequencing, high-quality DNA was randomly fragmented using the Covaris ultrasonic disruptor (Woburn, MA, USA). The sequencing pair-end

Libraries	Insert sizes (bp)	Clean reads number (Mb)	Clean bases (Gb)	Read Length (bp)	Sequencing coverage (×)
Illumina	350	1,035.33	148.04	150	182.64
Nanopore	20,000	1,142.86	126.07	18,500	155.54
Hi-C	300–700	1,313.54	194.07	150	239.43
RNA Seq	—	1,270.42	174.53	150	—

Table 1. Statistics of the sequencing data used for genome assembly.

Rank	Total base (bp)	Total reads (bp)	Max len (bp)	Avg len (bp)	N50	L50	N90	L90	Fold (×)
>0	126,072,140,153	11,428,566	490,594	11,031.32	36,673	1,112,377	3,939	4,740,334	155.54
>10000	102,656,238,593	2,973,737	490,594	34,520.95	43,649	819,973	17,884	2,218,104	126.65
>50000	41,263,390,171	604,461	490,594	68,264.77	66,904	244,863	52,832	524,168	50.91
>100000	4,160,948,064	35,083	490,594	118,602.97	114,383	15,487	102,156	30,966	5.13

Table 2. Statistics of the sequencing data for Nanopore.

Feature	Value
Raw paired-end reads	656,659,852
Unique Mapped Paired-end Pairs	395,045,680
Unique Mapped Ratio	60.2%
Q20(%)	98.422
Q30(%)	95.987
GC Content(%)	37.6
Chromosome/total	23

Table 3. Statistics of sequencing data from Hi-C.

K-mer length	K-mer number	K-mer depth	Genome size	Heterozygous ratio (%)	Repeat (%)
21	120,544,041,989	159	735,232,327	1.86	29.67

Table 4. K-mer frequency and genome size evaluation of the *S. monotuberculatus* genome.

libraries, with a 350 bp insert size, were prepared using the Nextera DNA Flex Library Prep Kit (Illumina, San Diego, CA, USA). Sequencing was performed on the Illumina NovaSeq6000 platform. Low-quality reads were removed from raw reads through the SOAPnuke (v2.1.4)[29] tool and clean data were used for subsequent analyses. A total of 148.04 Gb of short-read data (equivalent to a coverage of 182.64×) were obtained (Table 1).

For long-read Nanopore sequencing, libraries were prepared using the SQK-LSK110 ligation kit (Oxford Nanopore Technologies, Oxford, UK). The libraries were loaded onto primed R9.4 Spot-On Flow Cells and sequenced on a PromethION sequencer (Oxford Nanopore Technologies) with a 48-h run. Base-calling analysis of raw data was performed using the Oxford Nanopore GUPPY (v0.3.0)[30]. A total of 126.07 Gb of Nanopore continuous long-read data (equivalent to a coverage of 155.54×) were obtained (Table 1, Table 2).

For Hi-C sequencing with freshly harvested muscle samples, a formaldehyde cross-linking step was performed, followed by digestion of DpnII enzyme (NEB, Ipswich, MA, USA). *In situ* Hi-C chromosome conformation capture was performed following the DNase-based method[31]. The libraries were sequenced in 150 bp paired-end mode on an Illumina NovaSeq, resulting in 194.07 Gb of clean data (Table 1; Table 3).

For transcriptomic sequencing, 2 μg total RNA was used in a sample. Sequencing libraries were generated using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB), and index code was added for each sample. The sequencing was performed for 9 tissues, employing a strand-switching approach using Illumina HiSeq X Ten platform. A total of 174.53 Gb of clean data were generated in final.

Genome assessment and assembly. The K-mer analysis was conducted to estimate the genome size, proportion of repetitive sequences and heterozygosity using the Jellyfish (v2.3.0)[30] and GenomeScope (v2.0.0)[32]. Jellyfish was used to count K-mer from the raw sequencing reads, configuring the K-mer and hash sizes at 21 M and 100 M, respectively (Fig. 1b; Table 4). Employing 10 threads, the analysis was accounted for both DNA strands. The K-mer histogram generated was further analyzed via GenomeScope, setting the K-mer size at 21 and assuming a diploidy level of 2. After discarding K-mers reflecting abnormal depth, the analysis yielded 120,544,041,989 K-mers, with a peak depth at 159 (Table 4). Based on these results, the *S. monotuberculatus* genome size was estimated to be 735.23 Mb, comprising about 1.86% heterozygosity and 29.67% repetitive sequences.

The initial long-read assembly of genome was performed using SMARTdenovo[33], followed by polishing with nanopore sequencing data by Racon (v1.4.11)[34]. The subsequent error correction was accomplished by

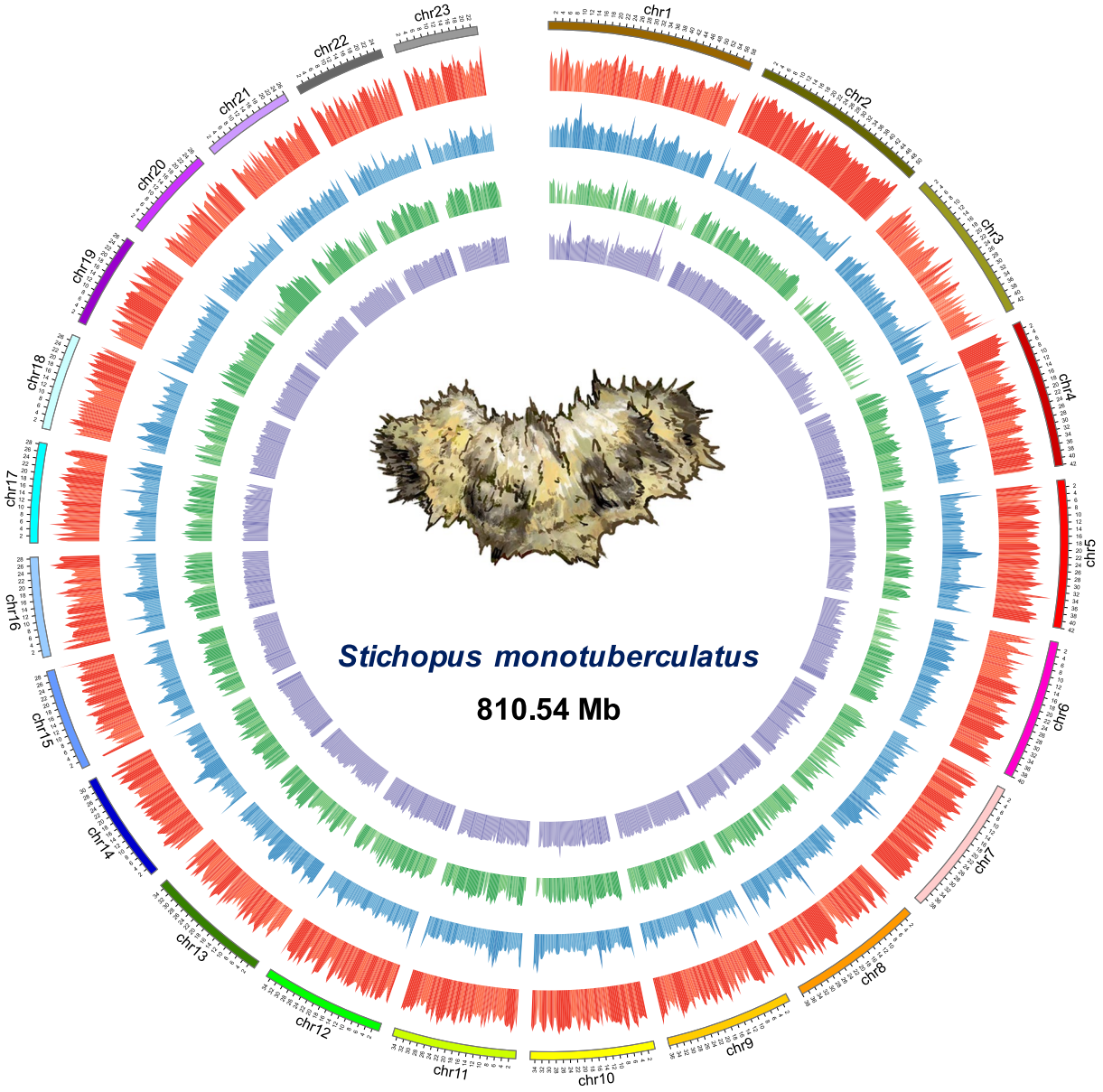

Fig. 2 Circos plot depicting the features of the *S. monotuberculatus* genome. From inner to outer circles: chromosomes, gene densities, repeat sequences, SNP percentage, and NGS sequencing depth. Average values for these features are 0.600, 0.200, 0.019 and 173, respectively. The windows size for all circles was 500k bp.

integrating 148.04 Gb of Illumina sequencing data via the Pilon (v1.23)[35] software. Modifications to address heterozygosity were conducted using the Purge_haplotigs (v1.0.4)[36] pipeline. Hi-C data was processed through the ALLHiC (v1.1)[36] pipeline to organize scaffolds into chromosome-length sequences, supporting accurate scaffold placements into 23 pseudo-chromosomes (Fig. 1c; Table 3). The chromosomal architecture and diverse genomic features were visualized using Circos software (v0.69)[37] (Fig. 2; Table 5). The assembly spans 810.55 Mb across 42 scaffolds, achieving a scaffold N50 of 35.36 Mb (Table 6).

Repetitive element annotation. RepeatMasker[38] (v 4.09 with RepBase 20181026) and EDTA[39] (a whole-genome *de-novo* TE annotation pipeline) were used to create High-quality TE annotations. The EDTA pipeline included LTRharvest, the parallel version of LTR_FINDER, LTR_retriever, GRF, TIR-Learner, HelitronScanner, and RepeatModeler along with customized filtering scripts. The gene-like sequences and redundancy results were removed using nucleotide coding sequences (CDS). Overall, sequences constituting 25.02% of the assembled genome were identified as repeats, of which the most abundant repetitive element was DNA elements (20.3%), followed by long interspersed nuclear elements (LINEs, 3.24%), short interspersed nuclear elements (SINEs, 0.16%), SINEs (0.16%), terminal repeats (LTRs, 0.06%) (Fig. 3a; Table 7).

Chromosome ID	Number of contigs	Length (bp)
Chr1	17	58,848,218
Chr2	4	51,320,803
Chr3	15	43,552,950
Chr4	6	42,278,664
Chr5	6	42,005,325
Chr6	5	40,092,979
Chr7	11	39,912,285
Chr8	5	38,003,003
Chr9	6	36,207,149
Chr10	5	35,364,055
Chr11	10	35,311,802
Chr12	6	35,276,996
Chr13	6	35,202,105
Chr14	2	31,569,361
Chr15	2	29,065,116
Chr16	3	28,473,956
Chr17	3	28,466,419
Chr18	2	27,667,838
Chr19	2	27,238,875
Chr20	5	26,204,974
Chr21	6	26,067,128
Chr22	4	25,729,855
Chr23	7	23,849,629
unanchored	19	2,843,734
Total	157	807,709,485

Table 5. Chromosome and corresponding statistical results after Hi-C assisted assembly.

	Scaffold		Contig	
	Length(bp)	Number	Length(bp)	Number
Max length	58,848,218		30,409,708	
N10	51,320,803	2	22,713,389	4
N20	42,278,664	4	17,610,103	8
N30	40,092,979	6	14,393,173	13
N40	38,003,003	8	11,786,873	19
N50	**35,364,055**	**10**	**10,153,360**	**27**
N60	35,276,996	12	8,522,540	35
N70	29,065,116	15	6,324,438	46
N80	28,466,419	17	4,649,266	61
N90	26,204,974	20	3,142,654	82
Total length	810,553,219		810,541,719	
Number >=100 bp	42		157	
Number >=2000 bp	42		157	

Table 6. Assembly statistics of the *S. monotuberculatus* genome.

Noncoding RNA annotation. Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) were predicted by Barrnap (v0.9) and tRNAscan-SE (v2.0.11)[40] using default parameters, respectively. For other non-coding RNAs, such as small nuclear RNAs (snRNAs) and microRNAs (miRNAs), an alignment with the Rfam database (v14.8)[41] was conducted, followed by annotation using Infernal (v1.1.4)[42]. A total of 19 miRNAs, 2,023 tRNAs, 444 rRNAs, and 202 snRNAs were identified in the *S. monotuberculatus* genome (Fig. 3b; Table 8).

Gene prediction and functional annotation. To obtain a high-quality gene set derived from the genome, three methods, namely RNAseq, homology, and *de novo*, were employed, and the results were integrated using MAKER3 (v3.01.03)[43]. For RNAseq, a total of 28 samples from 9 different tissues were sequenced, resulting in the generation of 186.82 Gb of data (Table 9). The reads were aligned to the genome using HISAT2 (v2.2.1)[44] and possible gene structures were assembled using StringTie (v2.1.7)[45]. For Homology-based predictions, protein sets selected from 6 representative genomes, including *Homo sapiens* (GCA_000001405.29), *Drosophila*

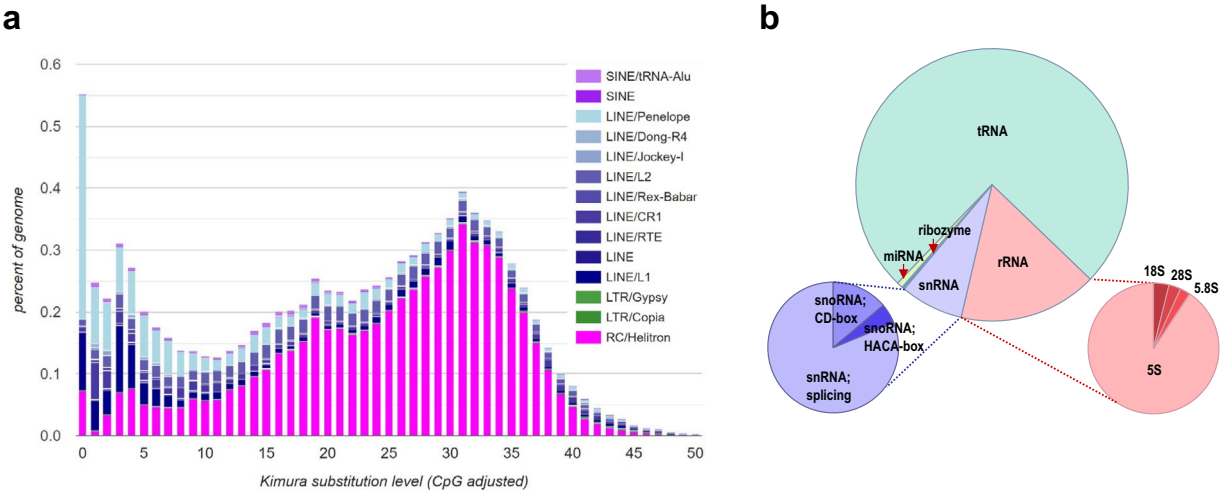

Fig. 3 Prediction and annotation of repetitive element (**a**) and Noncoding RNA (**b**).

	Denovo + Repeat Masker		Denovo		Repeat Masker	
TE_subtype	length(bp)	length/genome	length(bp)	length/genome	length(bp)	length/genome
DNA	164,514,308	20.3	164,450,591	20.29	107,395	0.01
LINE	26,286,578	3.24	25,727,563	3.17	1,636,208	0.20
LTR	508,257	0.06	345,535	0.04	209,365	0.03
MITE	6,703,076	0.83	6,703,076	0.83	0	0.00
SINE	1,331,371	0.16	1,240,192	0.15	94,483	0.01
Satellite	109,509	0.01	0	0.00	109,509	0.01
Unknown	12,013,719	1.48	12,009,655	1.48	4,064	0.00

Table 7. Statistics of repetitive sequence of the *S. monotuberculatus* genome.

	Type	Copy number	Total length (bp)	Percentage (%)
	miRNA	19	1,465	0.0002%
	tRNA	2,023	155,199	0.0191%
rRNA	rRNA	444	125,197	0.0154%
	18S	17	24,404	0.0030%
	28S	13	53,424	0.0066%
	5.8S	11	1,734	0.0002%
	5S	402	45,520	0.0056%
snRNA	snRNA	202	29,931	0.0037%
	snoRNA; CD-box	28	3,583	0.0004%
	snoRNA; HACA-box	10	1,783	0.0002%
	snRNA; splicing	164	24,565	0.0030%
ribozyme		9	1,399	0.0002%

Table 8. Statistics of non-coding RNA annotation.

melanogaster (GCA_000001215.4), *Strongylocentrotus purpuratus* (GCF_000002235.5)[46], *Acanthaster planci* (GCF_001949145.1)[47], *A. japonicus* (GCA_037975245.1)[48], and *Holothuria leucospilota* (GCA_029531755.1)[49], were aligned to the *S. monotuberculatus* genome with BLASTx[50]. For *de novo* method, genes were predicted using Augustus (v3.3.2)[51] and Fgenesh, and integrated within MAKER3. Genes supported only by the *de novo* method were retained if they aligning to UniProt, as they may represent fast-evolving or novel genes in the species. A total of 29,596 genes were identified, with 79.4% supported by at least two methods, and 3,468 genes (12.5%) supported only by the *de novo* method. The comparison of the number and features of predicted genes in the *S. monotuberculatus* genome with those of 4 other echinoderm species including *S. purpuratus*[46], *A. planci*[47], *A. japonicus*[48], and *H. leucospilota*[49] were shown in Fig. 4a & Table 10.

Functional annotation of the predicted protein-coding genes was performed by BLASTp (v2.11.0, e-value 1e-5)[52] against entries in the UniProt[53] and UniProtKB/Swiss-Prot databases. Domain and GO information were identified using InterProScan (v5.52-86.0)[54], resulting in successful annotation of 96.05% of the genes to

Sample	Tissue	Clean reads (Mbp)	Clean bases (Gbp)	Q20 bases	Q30 bases
SM-T-BW-1	body wall	48.24	7.23	98.32%	95.07%
SM-T-CC-3	coelomocytes	44.18	6.57	98.12%	94.62%
SM-T-IN-1	intestine	48.28	6.67	98.84%	97.00%
SM-T-IN-2	intestine	44.38	6.15	98.82%	96.92%
SM-T-IN-3	intestine	82.14	11.47	98.72%	96.69%
SM-T-IN-4	intestine	43.36	5.96	98.70%	96.69%
SM-T-IN-5	intestine	42.14	5.90	98.77%	96.85%
SM-T-MS-1	muscle	42.12	5.91	98.74%	96.77%
SM-T-MS-2	muscle	46.8	6.37	98.80%	96.93%
SM-T-MS–3	muscle	44.64	6.24	98.80%	96.85%
SM-T-MS-4	muscle	44.28	6.27	98.70%	96.64%
SM-T-MS-5	muscle	40.43	5.68	98.69%	96.66%
SM-T-OT-3	oral tentacles	39.34	5.45	98.81%	96.97%
SM-T-OT4	oral tentacles	48.6	6.89	98.77%	96.77%
SM-T-OT-5	oral tentacles	40.72	5.67	98.84%	97.03%
SM-T-PV-2	Polian vesicles	51.15	7.67	98.31%	95.00%
SM-T-PV-4	Polian vesicles	52.01	7.80	98.27%	94.92%
SM-T-RT-1	respiratory tree	47.79	6.62	98.63%	96.52%
SM-T-RT-2	respiratory tree	43.83	6.15	98.57%	96.34%
SM-T-RT-3	respiratory tree	45.99	6.44	98.66%	96.54%
SM-T-RT-4	respiratory tree	47.14	6.56	98.83%	96.98%
SM-T-RT-5	respiratory tree	44.32	6.12	98.77%	96.85%
SM-T-RM-1	rete mirabile	47.90	7.18	98.42%	95.30%
SM-T-RM-4	rete mirabile	47.58	6.56	98.85%	97.02%
SM-T-RM-5	rete mirabile	41.35	5.86	98.69%	96.61%
SM-T-SK2-1	skin	46.03	6.90	98.30%	95.02%
SM-T-SK2-2	skin	54.78	8.22	98.32%	95.09%
SM-T-SK2-3	skin	42.23	6.31	98.22%	94.91%
Total		**1311.75**	**186.82**	—	—

Table 9. Statistics of tissue transcriptomic sequencing data.

the existing databases (Table 11). A common set of 8,844 genes is shared among all 4 species belonging to the class Holothuroidea, namely, *A. japonicus*[48], *S. monotuberculatus*, *H. leucospilota*[49], and *H. scabra*. The number of genes specific to each species are 893, 727, 1,380, and 688 in *A. japonicus*, *S. monotuberculatus*, *H. leucospilota*, and *H. scabra*, respectively. A Venn diagram is presented to visually illustrate both the shared characteristics and differences in the gene profiles of these 4 sea cucumber species (Fig. 4b).

Gene family analysis. Gene family clustering was conducted with 9 species belonging to the phylum Echinodermata, including *A. planci* (GCF_001949145.1), *Lytechinus variegatus* (GCA_018143015.1), *S. purpuratus* (GCF_000002235.5), *Chiridota heheva* (GCA_020152595.1), *S. monotuberculatus* (PRJNA1123322), *A. japonicus* (GCA_037975245.1), *Holothuria glaberrima* (GCA_009936505.2), *Holothuria scabra* (PRJNA1074116) and *H. leucospilota* (GCA_029531755.1), using OrthoFinder[55] with MMseqs. 2[56], followed by phylogenetic tree construction based on single-copy genes. The analysis of gene families' expansion and contraction was carried out across using CAFE (v2.1)[57] (Fig. 4c). The number of gene families in the Most Recent Common Ancestor (MRCA) was determined to be 8915. Within the class Stichopodidae, there were 351 expanded gene families and 269 contracted gene families. Among them, 29 expansions and 22 contractions were found to be statistically significant ($p < 0.01$). In the case of *S. monotuberculatus*, there were 687 expanded gene families and 956 contracted gene families, with 68 expansions and 155 contractions exhibiting statistical significance ($p < 0.01$), possibly indicating adaptations to varying environmental conditions or ecological niches.

Data Records

All the sequencing reads and the chromosome-level genome assembly sequences related to this project have been deposited in NCBI BioProject PRJNA1123322[58]. In this study, whole genome sequencing reads are available under the accession numbers SRR29673227 to SRR29654456, while RNA-seq sequencing reads can be found under the accession numbers SRR29673218 to SRR29673229. The whole genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number JBEVTQ000000000[59]. Further related data, including assembly, annotation, functional annotation, and gene families, have been submitted to the Figshare database[60].

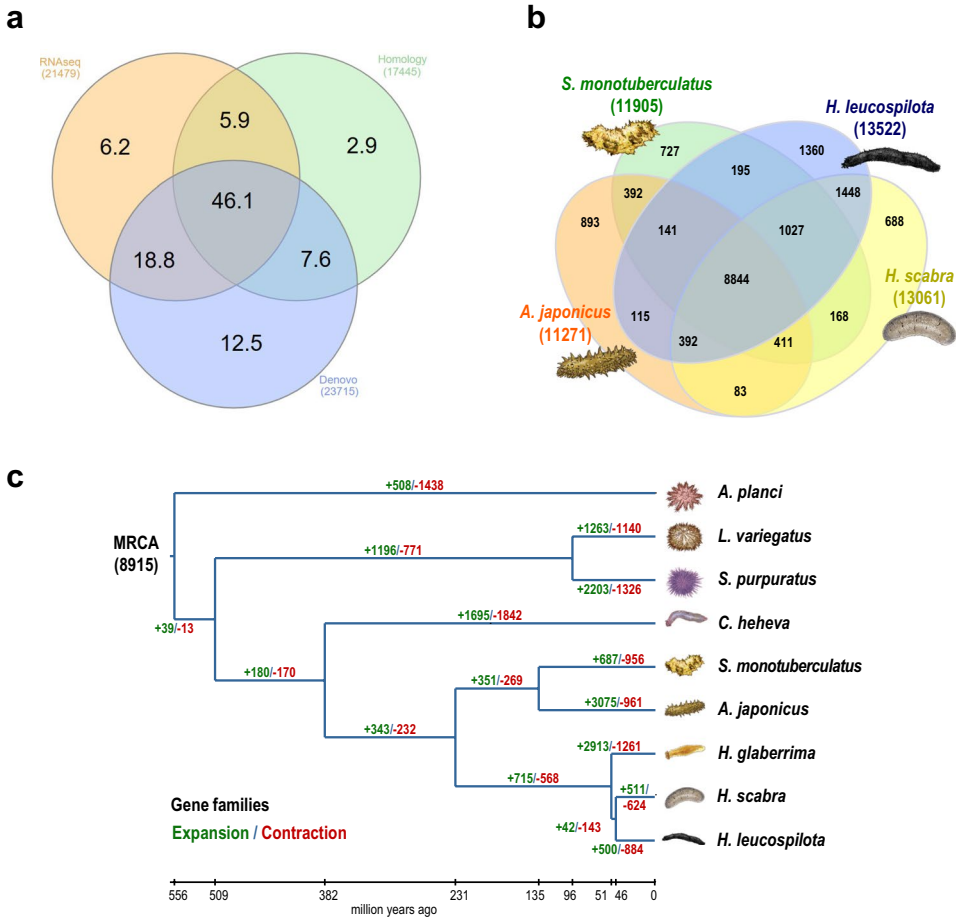

Fig. 4 Prediction and annotation of gene, and gene family analysis. (**a**) Prediction and annotation of gene; (**b**) Venn diagram of common and unique gene family in 4 sea cucumbers; (**c**) Phylogenetic and gene family evolution analysis among 9 echinoderm species. The scale below represents the divergence time. The number of expanded (+green) and contracted (−red) gene families were shown alongside the species.

Species	Number	Single exon gene	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
Acanthaster planci	16,463	8.39%	13,480	1,851	10	192	1,348
Strongylocentrotus purpuratus	28,821	15.71%	10,885	1,422	6	219	1,725
Holothuria leucospilota	37,209	25.02%	18,434	1,350	6	240	2,960
Apostichopus japonicus	19,828	11.60%	14,590	1,573	8	195	1,845
Stichopus monotuberculatus	29,596	11.67%	16,701	1,384	7	201	2,251

Table 10. Comparison of the number and features of predicted genes in the *S. monotuberculatus* genome with those of four other echinoderm species.

Feature	Numbers	Percent of all genes (%)
UniProt	27,705	93.61%
UniProtKB/Swiss-Prot	17,815	60.19%
InterProscan	21,004	70.97%
GeneOntology	16,738	56.55%
Total	28,427	96.05%

Table 11. Function annotation of predicted protein-coding genes.

Sequence depth	Bases	Coverage of genome
>=1	807,827,688	99.67%
>=5	806,508,393	99.50%
>=10	805,610,584	99.39%
>=20	804,089,644	99.20%
>=30	802,490,062	99.01%
>=50	797,353,293	98.37%
>=100	723,020,128	89.20%
>=200	242,610,016	29.93%
>=400	10,966,122	1.35%

Table 12. Reads mapping of NGS data.

Feature	*S. monotuberculatus*		*A.japonicus*	
	Number	Percent (%)	Number	Percent (%)
Complete BUSCOs (C)	899	94.3	898	94.2
Complete and single-copy BUSCOs (S)	892	93.5	891	93.4
Complete and duplicated BUSCOs (D)	7	0.8	7	0.8
Fragmented BUSCOs (F)	33	3.5	29	3.1
Missing BUSCOs (M)	21	2.2	25	2.7
Total BUSCO groups searched	954	100.0	954	100.0

Table 13. Statistical result of BUSCO evaluation results of genome assembly.

Technical Validation

Nucleic acid quality. The DNA quality and concentration were assessed using 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). DNA samples showing slight degradation were deemed suitable for sequencing library preparation. RNA purity was evaluated with the kaiaoK5500® Spectrophotometer (Kaiao, Beijing, China). RNA integrity and concentration were determined using the RNA Nano 6000 Assay Kit on the Bioanalyzer 2100 system (Agilent Technologies, Santa Clara, CA, USA). Samples with an RNA integrity number (RIN) exceeding 9.50 were considered appropriate for library construction.

Genome assembly and annotation quality. The QV pipeline of Merqury[61] was used to estimate the assembly QV based on k-mer analysis. The script "best_k.sh" in Merqury was employed to determine the optimal k-mer length, which was found to be 19. Meryl was then utilized to calculate the number of k-mers in the Illumina WGS reads with default settings. The QV evaluation was performed in Merqury using the output from Meryl and the assembly. The findings demonstrated a k-mer completeness of 79.10% and a k-mer-based QV of 43.36.

To assess the integrity and precision of the genome, alignment of sequenced reads was performed against the genome. High alignment integrity indicates robust genome integrity. Utilizing Illumina short-read sequencing, the genome was sequenced at the coverage of 182.64×, resulting in the generation of 148 Gb of high-quality reads. This approach achieved a mapping rate of 99.57%, with 90.43% of reads properly paired. The genome exhibited a coverage of 99.67%, with regions exceeding 30× coverage representing 99.01% of the genome (Table 12).

Genome completeness was further evaluated through BUSCO analysis using the metazoa_odb10 dataset, which contains 954 conserved genes across 65 genomes (Table 13). The analysis revealed an impressive overall completeness of 97.8%, with 94.3% of BUSCOs being complete, and minor proportions being fragmented (3.5%) or missing (2.2%). These results highlight the strength of our assembly and demonstrate the excellent quality of the genome. These metrics confirm the completeness and accuracy of our genome assembly, providing valuable data for further wild resource conservation, genetic breeding and aquaculture of *S. monotuberculatus*.

Code availability

All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatic software. The settings and parameters of software were listed below:
SOAPnuke (v2.1.4): Used for filtering low-quality reads from Illumina short-read sequencing data with default parameters.
GUPPY (v0.3.0): Used for base-calling analysis of Nanopore sequencing data with default parameters.
Jellyfish (v2.3.0): Used for K-mer counting in genome size estimation, K-mer size is 19.
GenomeScope (v2.0.0): Applied for analyzing K-mer histograms and estimating genome characteristics such as size, heterozygosity, and repetitive content with default parameters.
SMARTdenovo: Used for initial assembly of long-read Nanopore data with default parameters.
Racon (v1.4.11): Used for polishing the assembly with Nanopore data with default parameters.
Pilon (v1.23): Applied for further error correction using Illumina data with default parameters.

Purge_haplotigs (v1.0.4): Utilized for addressing heterozygosity in the genome assembly with -j 80 -s 80.
ALLHiC (v1.1): Used for processing Hi-C data and organizing scaffolds into chromosome-length sequences.
Merqury (v1.3): Used for genome assembly quality with best K-mer 19.
BUSCO (v5.7.1): Applied for genome completeness assessment with metazoa_odb10 dataset.
Circos (v0.69): Used for visualizing chromosomal architecture and various genomic features.
RepeatMasker (v4.09): Applied for repetitive element annotation with default parameters.
EDTA: Used for *de novo* transposable element (TE) annotation with default parameters.
Barrnap (v0.9): Used for predicting ribosomal RNAs (rRNAs) with default parameters.
tRNAscan-SE (v2.0.11): Used for transfer RNA (tRNA) prediction with default parameters.
Infernal (v1.1.4): Used for annotating small nuclear RNAs (snRNAs) and microRNAs (miRNAs) with default parameters.
MAKER3 (v3.01.03): Used for integrating gene prediction results with default parameters.
HISAT2 (v2.2.1): Used for aligning RNA-seq reads to the genome with default parameters.
StringTie (v2.1.7): Used for assembling possible gene structures from RNA-seq data with default parameters.
Augustus (v3.3.2): Used for *de novo* gene prediction with general sets.
OrthoFinder: Used for gene family clustering with mmseq2 aligner.
MCMCTree from **PAML** (v4.10.7): Used for estimating divergence times.
Evolview2: Used for visualizing phylogenetic trees.
CAFE (v2.1): Used for gene family expansion and contraction analysis.

Received: 16 July 2024; Accepted: 7 October 2024;
Published online: 18 November 2024

References

1. Purcell, S. W., Conand, C., Uthicke, S. & Byrne, M. Ecological roles of exploited sea cucumbers. *Oceanogr Mar Biol* **54**, 367–386 (2016).
2. Ahyong, S. *et al*. World register of marine species (WoRMS). WoRMS Editorial Board (2024).
3. Purcell, S. W. *et al*. *Commercially important sea cucumbers of the world – Second edition*. Food and Agriculture Organization of the United Nations FAO (2023).
4. Clements, C. S., Pratte, Z. A., Stewart, F. J. & Hay, M. E. Removal of detritivore sea cucumbers from reefs increases coral disease. *Nat Commun* **15**, 1338 (2024).
5. Anderson, S. C., Flemming, J. M., Watson, R. & Lotze, H. K. Serial exploitation of global sea cucumber fisheries. *Fish Fish* **12**, 317–339 (2011).
6. Z, E. *et al*. Applications of Environmental DNA (eDNA) in Monitoring the endangered status and evaluating the stock enhancement effect of tropical sea cucumber *Holothuria Scabra*. *Mar Biotechnol* **25**, 778–789 (2023).
7. Yang, Y. *et al*. Pipeline for identification of genome-wide microsatellite markers and its application in assessing the genetic diversity and structure of the tropical sea cucumber *Holothuria leucospilota*. *Aquacult Rep* **37**, 102207 (2024).
8. Cheng, C. *et al*. Aquaculture of the tropical sea cucumber, *Stichopus monotuberculatus*: Induced spawning, detailed records of gonadal and embryonic development, and improvements in larval breeding by digestive enzyme supply in diet. *Aquaculture* **540**, 736690 (2021).
9. Hu, C. *et al*. Larval development and juvenile growth of the sea cucumber *Stichopus* sp. (Curry fish). *Aquaculture* **300**, 73–79 (2010).
10. Byrne, M., Rowe, F. & Uthicke, S. Molecular taxonomy, phylogeny and evolution in the family Stichopodidae (Aspidochirotida: Holothuroidea) based on COI and 16S mitochondrial DNA. *Mol Phylogenet Evol* **56**, 1068–1081 (2010).
11. Uthicke, S., Byrne, M. & Conand, C. Genetic barcoding of commercial Beche-de-mer species (Echinodermata: Holothuroidea). *Mol Biol Evol* **10**, 634–646 (2010).
12. Zhong, M., Chen, T., Hu, C. & Ren, C. Isolation and characterization of collagen from the body wall of sea cucumber *Stichopus monotuberculatus*. *J Food Sci* **80**, C671–C679 (2015).
13. Gray, B. C. T., Byrne, M., Clements, M., Foo, S. A. & Purcell, S. W. Length-weight relationship for the dragonfish, *Stichopus* cf. *monotuberculatus* (Holothuroidea). *Fish Res* **268**, 106851 (2023).
14. Huang, L. *et al*. Effects of acute salinity stress on physiology and immunoenzymatic activity in juvenile sea cucumber, *Stichopus monotuberculatus*. *Aquaculture* **578**, 740094 (2024).
15. Wen, J. *et al*. The application of PCR–RFLP and FINS for species identification used in sea cucumbers (Aspidochirotida: Stichopodidae) products from the market. *Food Control* **21**, 403–407 (2010).
16. Wu, F. *et al*. Identification of sex-specific molecular markers and development of PCR-based sex detection techniques in tropical sea cucumber (*Stichopus monotuberculatus*). *Aquaculture* **547**, 737458 (2022).
17. Yuan, L., Hu, C., Zhang, L. & Xia, J. Population genetics of a tropical sea cucumber species (*Stichopus monotuberculatus*) in China. *Conserv Genet* **14**, 1279–1284 (2013).
18. Jia, C. *et al*. Comparative analysis of in situ eukaryotic food sources in three tropical sea cucumber species by metabarcoding. *Animals-Basel* **12**, 2303 (2022).
19. Jia, C. *et al*. Comparative Analysis of Gut Bacterial community composition in two tropical economic sea cucumbers under different seasons of artificial environment. *Int J Mol Sci* **25**, 4573 (2024).
20. Yan, A. *et al*. Identification and functional characterization of a novel antistasin/WAP-like serine protease inhibitor from the tropical sea cucumber, *Stichopus monotuberculatus*. *Fish Shellfish Immun* **59**, 203–212 (2016).
21. Ren, C. *et al*. The first echinoderm poly-U-binding factor 60 kDa (PUF60) from sea cucumber (*Stichopus monotuberculatus*): Molecular characterization, inducible expression and involvement of apoptosis. *Fish Shellfish Immun* **47**, 196–204 (2015).
22. Ren, C., Chen, T., Jiang, X., Wang, Y. & Hu, C. The first characterization of gene structure and biological function for echinoderm translationally controlled tumor protein (TCTP). *Fish Shellfish Immun* **41**, 137–146 (2014).
23. Ren, C. *et al*. The first echinoderm gamma-interferon-inducible lysosomal thiol reductase (GILT) identified from sea cucumber (*Stichopus monotuberculatus*). *Fish Shellfish Immun* **42**, 41–49 (2015).
24. Chen, T. *et al*. Calmodulin of the tropical sea cucumber: Gene structure, inducible expression and contribution to nitric oxide production and pathogen clearance during immune response. *Fish Shellfish Immun* **45**, 231–238 (2015).
25. Ren, C. *et al*. Two proprotein convertase subtilisin/kexin type 9 (PCSK9) paralogs from the tropical sea cucumber (*Stichopus monotuberculatus*): Molecular characterization and inducible expression with immune challenge. *Fish Shellfish Immun* **56**, 255–262 (2016).
26. Zhong, S., Liu, Y., Zhao, Y. & Huang, G. The complete mitochondrial genome of sea cucumber *Stichopus monotuberculatus* (aspidochirotida: Stichopodidae). *Mitochondrial Dna B* **4**, 3305–3306 (2019).
27. Ma, B. *et al*. Integrative Application of transcriptomics and metabolomics provides insights into unsynchronized growth in sea cucumber (*Stichopus monotuberculatus*). *Int J Mol Sci* **23**, 15478 (2022).

28. Zhong, S. P. *et al*. The draft genome of the tropical sea cucumber *Stichopus monotuberculatus* (Echinodermata, Stichopodidae) reveals critical genes in fucosylated chondroitin sulfates biosynthetic pathway. *Front Genet* **14**, 1182002 (2023).
29. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1–6 (2017).
30. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129 (2019).
31. Ramani, V. *et al*. Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods* **170**, 61–68 (2020).
32. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
33. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a *de novo* assembler using long noisy reads. *GigaByte* **2021**, 15–15 (2021).
34. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* **27**, 737–746 (2017).
35. Walker, B. J. *et al*. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One* **9**, e112963 (2014).
36. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
37. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
38. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *P Natl Acad Sci USA* **117**, 9451–9457 (2020).
39. Ou, S. J. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
40. 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).
41. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–D124 (2005).
42. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
43. Cantarel, B. L. *et al*. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188–196 (2008).
44. Kim, D., Landmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).
45. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
46. Elsik, C. G. *et al*. The genome of the sea urchin. *Science* **315**, 766–766 (2007).
47. Hall, M. R. *et al*. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature* **544**, 231–234 (2017).
48. Sun, L. A., Jiang, C. X., Su, F., Cui, W. & Yang, H. S. Chromosome-level genome assembly of the sea cucumber *Apostichopus japonicus*. *Sci Data* **10**, 454 (2023).
49. Chen, T. *et al*. The *Holothuria leucospilota* genome elucidates sacrificial organ expulsion and bioadhesive trap enriched with amyloid-patterned proteins. *P Natl Acad Sci USA* **120**, e2213512120 (2023).
50. Camacho, C. *et al*. BLAST plus: architecture and applications. *Bmc Bioinformatics* **10**, 421 (2009).
51. Gao, Z. *et al*. TSTD:A Cross-modal Two Stages Network with New Trans-decoder for Point Cloud Semantic Segmentation. In: *Pattern Recognition and Computer Vision, PRCV 2023, Part VIII* (2024).
52. Li, Y. C. & Lu, Y. C. BLASTP-ACC: Parallel Architecture and Hardware Accelerator Design for BLAST-Based Protein Sequence Alignment. *IEEE T Biomed Circ S* **13**, 1771–1782 (2019).
53. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489.
54. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
55. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
56. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).
57. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
58. Chen, T. *Stichopus monotuberculatus* isolate:south_sea_xycz | cultivar:nanhai. *NCBI BioProject* https://identifiers.org/ncbi/insdc.sra:SRP517039 (2024).
59. Chen, T. *Stichopus monotuberculatus* isolate:south_sea_xycz | cultivar:nanhai Genome sequencing and assembly. *GenBank* https://identifiers.org/ncbi/insdc:JBEVTQ000000000 (2024).
60. Chen, T. Chromosome-level genome assembly and annotation of the tropical sea cucumber *Stichopus monotuberculatus*. *Figshare* https://doi.org/10.6084/m9.figshare.26124061 (2024).
61. 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

This study was graciously supported by grants from the National Natural Science Foundation of China (42176132), the Guangxi Natural Science Foundation (2021GXNSFAA196074), the Science and Technology Program of Nansha District (NSJL202103), the Research on breeding technology of candidate species for Guangdong modern marine ranching (2024-MRB-00-001), the National Key R & D Program of China (2022YFD2401301), the Guangdong Province Project (2024A1515011418, 2024A1515010899), and the Innovation Team Project of High Level Local Universities from Shanghai Education Committee (HJWK-2021-21).

Author contributions

Ting Chen, Chunhua Ren, Aifen Yan and Chaoqun Hu conceived and designed this study. Ting Chen, Yun Yang, Zhenyu Xie, Hao Long, Wenjie Pan, Zixuan E, Jiasheng Huang, Qianying Huang, Jianfeng Xu, Zepeng Zhang, Chuhang Cheng and Suzhong Yu collected the samples. Zhou Qin and Dingding Fan assembled and annotated the genome. Ting Chen, Yun Yang, Xuan Wang, Zhou Qin and Dingding Fan performed gene family and genome evolutionary analyses. Ting Chen, Yun Yang, Xuan Wang, Zhou Qin and Dingding Fan performed bioinformatic analyses. Ting Chen, Zhenyu Xie, Chunhua Ren, Hongyan Sun, Peng Luo, Xiao Jiang, Chang Chen, Yanhong Wang, Fajun Jiang, Aifen Yan, and Chaoqun Hu contributed reagents/analytic tools. Ting Chen, Yun Yang and Dingding Fan wrote the manuscript. Chunhua Ren, Aifen Yan and Chaoqun Hu revised it. All authors reviewed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.Y. or C.H.

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.