scientific data

OPEN

DATA DESCRIPTOR

The draft genome sequences of the cosmopolitan centric diatom, the genus *Skeletonema*

Satoshi Nagai[1]✉, Ryuhei Minei[2], Ikuyo Takemura[2], Hiromasa Tabata[2], Mats Töpel[3,4], Matthew Pinder[3], Olga Kourtchenko[3], Kang Lee-Kuo[5], Nguyen Van Nguyen[6], Sirje Sildever[7], Atsushi Ogura[2] & Takashi Gojobori[8]

The genus *Skeletonema* stands out as one of the most prevalent and globally distributed diatoms within coastal marine ecosystems. Additionally, it holds significance as a pivotal organism across diverse research domains, encompassing biochemistry, ecophysiology, molecular biology, ecology, oceanography, and aquaculture. This study presents the draft genome assemblies and annotations of 11 strains in 9 *Skeletonema* species. We generated approximately 13.2–45.4 Gbp (DNA) and 6.0–15.4 Gbp (RNA) of sequencing data and produced draft haploid genomes with total lengths of 40.3–69.3 Mbp and N50 scaffold sizes of 0.35–1.09 Mbp containing 90–98% of orthologs conserved in Stramenopiles. We also identified 5.1–28.4 Mbp (11.0–41.1% of the draft genomes) of repetitive genome sequences and a total number of 15,275–21,376 protein-coding genes. The draft genomes of *Skeletonema* will facilitate further study, such as population dynamics, phylogeography, the process of speciation, and the evolution of gene function, and help to provide deep insight into the cosmopolitanism of this most successfully prosperous genus.

Background & Summary

Diatoms, unicellular, primarily diploid during their vegetative phase, photosynthetic, and eukaryotic microalgae, are widely distributed across marine and freshwater systems[1], exhibiting a vast array of shapes and sizes within their tens of thousands of species[2]. Despite their explosive diversification, diatoms have only emerged since the early Mesozoic period[3]. Notably, diatom cells feature a unique characteristic—a silica-based cell wall called a frustule. Their vegetative cells undergo asexual reproduction, but the mechanism of frustule formation reduces cell size. The linear decrease in valve diameter with each cell division is termed the "McDonald and Pfitzer's rule"[4]. Substantial reduction in cell size prevents further division, resulting in cell death[5–7]. Consequently, diatoms employ auxospore formation (considered a true sexual process) and vegetative cell enlargement (pseudo-auxospore formation, i.e., an asexual process) to restore their cell size[1,7]. Diatoms, constituting the most dominant microalgal group in coastal waters, sporadically form dense blooms[8,9]. The global net primary production (NPP) of all terrestrial and marine autotrophs is around 105 petagrams (Pg) of carbon annually[10]. Marine diatoms, a diverse group of phytoplankton found worldwide, are estimated to contribute up to 25% (26 Pg C yr^{-1}) of this total, surpassing the annual primary production of any terrestrial biome[10–12]. Diatoms account for up to 50% of marine primary production and are considered essential components of the biological carbon pump[13].

Since the release of the *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* genomes[14,15], some additional diatom genomes representing both centric and pennate species have been published[16,17]. Genome information has been accumulated, and advanced technology, such as Hi-C, optimal genome mapping, and

[1]Japan Fisheries Research and Education Agency, Yokohama Branch, Yokohama, Kanagawa, 236-8648, Japan. [2]Nagahama Institute of Bioscience and Technology, 1266 Tamura, Nagahama, Shiga, 526-0829, Japan. [3]University of Gothenburg, Department of marine sciences, Medicinaregatan 7B, Gothenburg, 41390, Sweden. [4]IVL - Swedish Environmental Research Institute, Valhallavägen 81, 114 28, Stockholm, Sweden. [5]Bachelor Degree Program in Marine Biotechnology / Taiwan Ocean Genome Center, National Taiwan Ocean University, Keelung, 20224, Taiwan, ROC. [6]Research Institute for Marine Fisheries, 224 Le Lai, Hai Phong, Vietnam. [7]Department of Marine Systems, Tallinn University of Technology, Akadeemia tee 15A, 12618, Tallinn, Estonia. [8]Marine Open Innovation Institute, 9-25 Hinode-cho, Shimizu-ku, Shizuoka, 424-0922, Japan. ✉e-mail: snagai0101@gmail.com

species	strain	source	Coordinates	Estimated genome size (Mbp)*	Illumina short-reads			Nanopore Long reads			RNA (Gbp)		
					Raw reads (Gbp)	Read depth	Accession number	Raw reads (Gbp)	Read depth	Accession number	Raw reads (Gbp)	Read depth	Accession number
S. ardens	NIES-2837	Dokai Bay, Japan	33.88 N, 130.78E	46.6	30.0	(x 644)	DRR539406	15.4	(x 330)	DRR539417	7.3	(x 157)	DRR539428
S. costatum	SC_Ari12	Ariake Sea, Japan	33.11 N, 130.37E	69.3	9.1	(x 131)	DRR539407	7.8	(x 113)	DRR539418	15.4	(x 222)	DRR539429
S. costatum	SCHP35	Hai Phong, Vietnum	20.72 N, 107.04E	62.7	13.9	(x 222)	DRR539408	11.0	(x 175)	DRR539419	—	—	
S. dohrnii	SD_Ari07	Ariake Sea, Japan	33.11 N, 130.37E	51.6	23.1	(x 448)	DRR539409	8.0	(x 155)	DRR539420	7.5	(x 145)	DRR539430
S. dohrnii	SD_MZ01	Nangan island, Taiwan	26.15 N, 119.95E	51.9	10.3	(x 198)	DRR539410	16.0	(x 308)	DRR539421	6.3	(x 121)	DRR539431
S. grevillei	NIES-2840	Okinawa, Japan	26.64 N, 127.88E	47.6	13.5	(x 284)	DRR539411	21.2	(x 445)	DRR539422	7.4	(x 155)	DRR539432
S. japonicum	SJ_T01	Tokyo Bay, Japan	35.35 N, 139.66E	61.4	10.3	(x 168)	DRR539412	4.3	(x 70)	DRR539423	6.2	(x 101)	DRR539433
S. menzelii	NIES-2842	Dokai Bay, Japan	33.88 N, 130.78E	46.6	16.9	(x 363)	DRR539413	19.0	(x 408)	DRR539424	7.5	(x 161)	DRR539434
S. pseudocostatum	NIES-2843	Dokai Bay, Japan	33.88 N, 130.78E	40.3	16.4	(x 407)	DRR539414	22.1	(x 549)	DRR539425	7.5	(x 186)	DRR539435
S. subsalsum	SS_0375	Mäntyniemi, Finland	60.40 N, 26.63E	60.2	12.0	(x 199)	DRR539415	1.2	(x 20)	DRR539426	7.2	(x 120)	DRR539436
S. tropicum	ST_Kao	Kaohsung Harbor, Taiwan	22.62 N, 120.26E	62.3	10.2	(x 164)	DRR539416	7.8	(x 125)	DRR539427	6.0	(x 96)	DRR539437

Table 1. Basic information on the strains of Skeletonema and the sequencing data of Illumina short-reads and Nanopore long-reads. *scaffold sizes by HaploMerger2

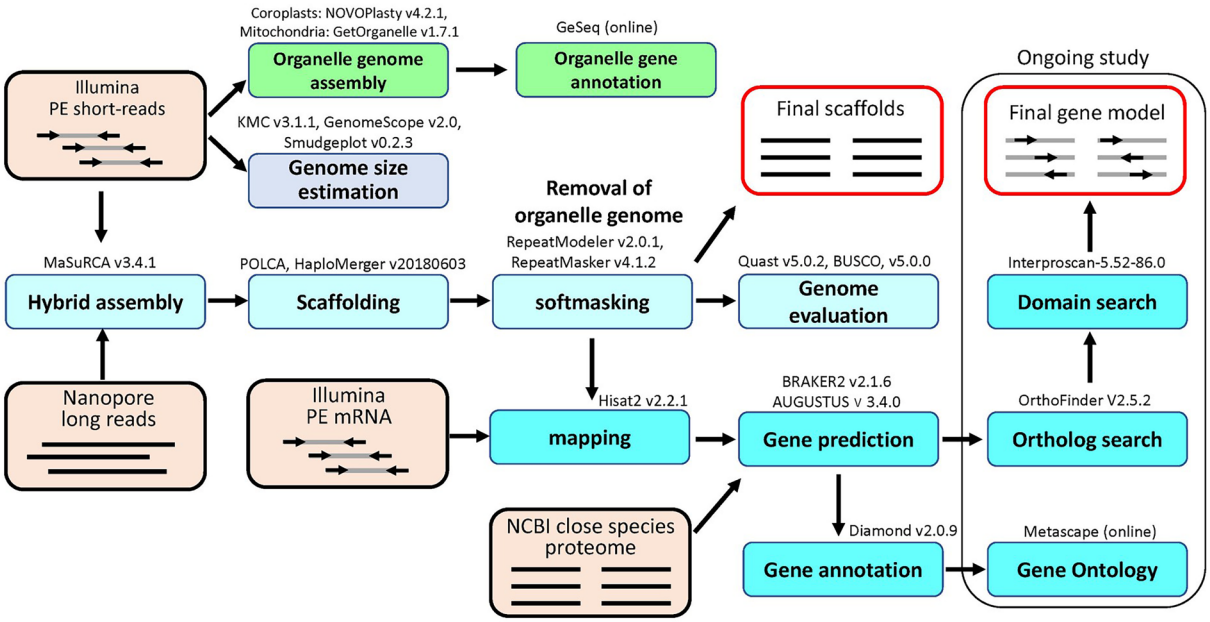

Fig. 1 The analytical flow of the whole genome in *Skeletonema*.

long-reads sequencing has been developed[16–18]. Thanks to technological innovations, genome models are now available with higher accuracy than ever before.

Investigations employing electron microscopy and rDNA sequences from marine strains have unveiled that *S. costatum* sensu lato (s.l.) consists of a series of genetically and morphologically distinct species[18–21]. Since then, 27 *Skeletonema* species have been identified and taxonomically accepted[22]. This study conducted a comparative genome analysis to obtain the basic genome information in *Skeletonema*, showing the wide distribution range, rapid growth, and forming of extensive blooms. We sequenced the genome of nine species (11 strains) in the genus *Skeletonema*, using a hybrid assembly pipeline in combination with Illumina short-reads and Nanopore long-reads (Table 1). The analytical flow of the whole genome in *Skeletonema* is shown in Fig. 1.

		SA	SC_ Ari12	SCHP35	SD_ Ari07	SD_ MZ01	SG	SJ	SMen	SP	SS	ST_Kao
GenomeScope	Total length (Mbp)	64.6	85.3	70.1	47.6	60.4	60.7	71.8	45.5	43.3	72.2	64.8
HaploMerger2	Scaffolds	107	348	160	132	152	173	288	103	94	333	135
	Largest contig (Mbp)	4.2	2.2	3.5	2.4	3.5	2.1	2.5	4.6	3.5	3.0	3.0
	Total length (Mbp)	46.6	69.3	62.7	51.6	51.9	47.6	61.4	46.6	40.3	60.2	62.3
	GC (%)	44.8	45.4	45.4	45.2	45.3	44.7	44.8	47.3	44.8	45.6	44.9
	N50 (Mbp)	0.9	0.5	0.9	0.7	0.8	0.5	0.5	0.9	0.7	0.3	1.1
	BUSCO (complete)	97	90	93	90	94	98	93	96	95	98	95
	BUSCO (single)	94	89	91	89	93	95	91	90	94	96	94
	BUSCO (duplicated)	3	1	2	1	1	3	2	6	1	2	1
Gene prediction	gene number	17,458	17,105	18,107	16,532	16,787	17,293	15,333	21,376	15,275	16,243	16,539
	average gene length	1652.8	1635.6	1606	1689.1	1705.3	1628.7	1678.7	1515.1	1691.1	1613.1	1656.7
	average exon number/gene	1.6	1.7	1.7	1.7	1.7	1.6	1.7	1.5	1.6	1.6	1.7
	average exon length	996.4	944.2	946	946.7	958.1	984.3	974.8	1004.5	989.7	948.8	943.5
	average intron length	124	127.9	127	149.1	147.6	127.3	129.9	123.8	145.7	126.6	150.1
	BUSCO (complete)	97	90	92	91	94	98	92	94	94	97	93
	BUSCO (single)	81	89	89	89	93	96	90	89	93	96	91
	BUSCO (duplicated)	16	1	3	2	1	2	2	5	1	1	2

Table 2. Statistics of the genome assembly and gene models in *Skeletonema*.

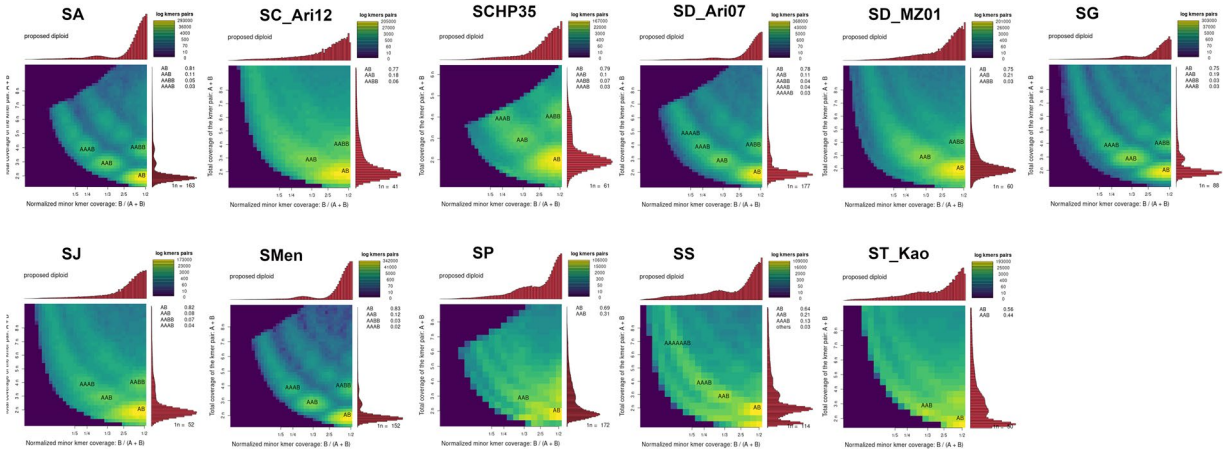

Fig. 2 The result of the analysis by Smudgeplot in *Skeletonema*.

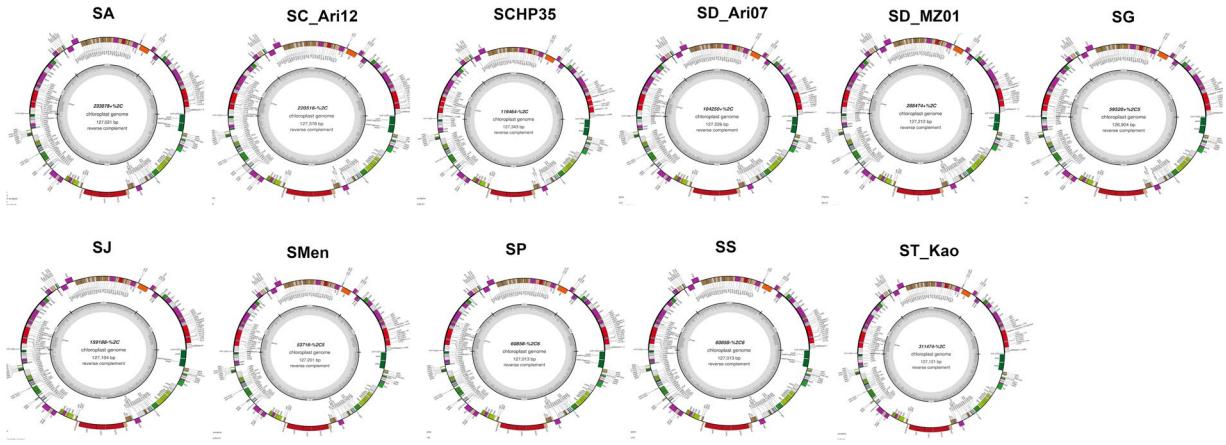

Fig. 3 The result of Chloroplast genome annotation by GeSeq in *Skeletonema*.

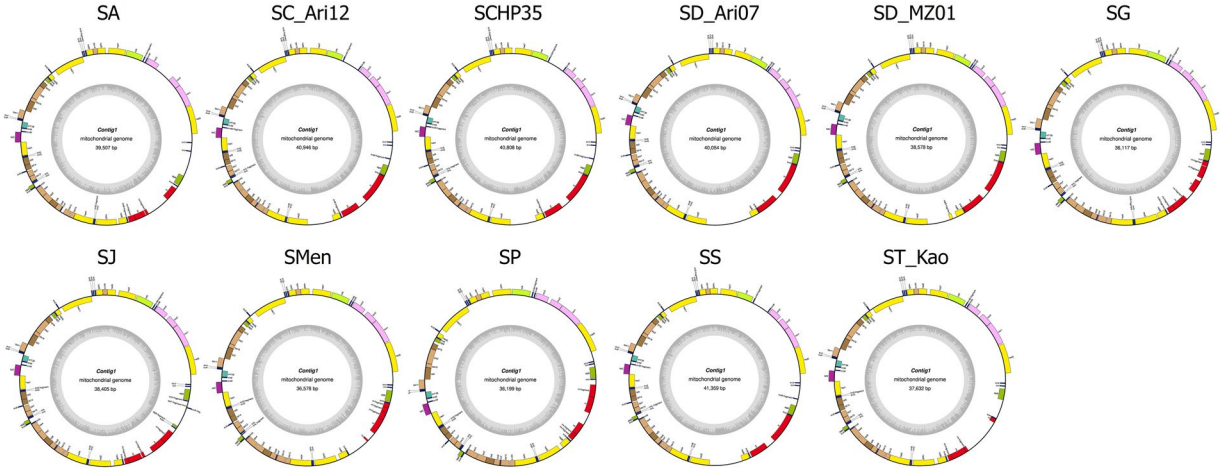

Fig. 4 The result of mitochondrial genome annotation by GeSeq in *Skeletonema*.

No.	species	strain	Choloplast genome		Mitochondrial genome		rRNA genes	
			length (Kbp)	Accession number	length (Kbp)	Accession number	length (bp)	Accession number
1	*S. ardens*	NIES-2837	127,021	LC814759	39,657	LC814760	8,323	LC814761
2	*S. costatum*	SC_Ari12	127,378	LC814762	40,946	LC814763	8,551	LC814764
3	*S. costatum*	SCHP35	127,343	LC814765	40,808	LC814766	8,570	LC814767
4	*S. dohrnii*	SD_Ari07	127,226	LC814768	40,054	LC814769	8,218	LC814770
5	*S. dohrnii*	SD_MZ01	127,212	LC814771	38,578	LC814772	8,427	LC814773
6	*S. grevillei*	NIES-2840	126,904	LC814774	36,117	LC814775	5,876	LC814776
7	*S. japonicum*	SJ_T01	127,104	LC814777	38,405	LC814778	7,393	LC814779
8	*S. menzelii*	NIES-2842	127,201	LC814780	36,578	LC814781	4,139	LC814782
9	*S. pseudocostatum*	NIES-2843	127,013	LC814783	36,199	LC814784	8,272	LC814785
10	*S. subsalsum*	SS_0375	127,348	LC814786	41,359	LC814787	8,625	LC814788
11	*S. tropicum*	ST_Kao	127,121	LC814789	37,632	LC814790	9,650	LC814791

Table 3. Genome sizes and accession numbers for sequencing and complete organelle genomes and nuclear rRNA genes in *Skeletonema*.

KMC v3.2.0[23] and GenomeScope v2.0[24] estimated the haploid genome sizes as 43.3–85.5 Mbp (Table 2). Smudgeplot v0.2.4[24] revealed that the vegetative cells in *Skeletonema* is diploid (Fig. 2). After scaffolding, the number of contigs was 94–348, and the estimated haploid genome sizes were 40.3–69.3 Mbp. Lengths of N50 were 0.35–1.09 Mbp, and the longest scaffold of length 2.1–4.6 Mbp, as calculated by QUAST[25]. GC contents were 44.7–45.6% (45.3 ± 0.7, average $\pm$ SD), and a one-sample t-test indicated no significant differences between the strains. BUSCO assessment showed that 90–98% (92.4 ± 2.3, average $\pm$ SD) of orthologs conserved in Stramenopiles were present in this genome assembly (sum of the percentages of single-copy and duplicate), suggesting that our draft genome possessed a sufficient gene repertoire from Stramenopiles (Table 2).

The complete genome sequences in chloroplast and mitochondria were determined and the sizes were 126.9–127.4 Kbp and 36.1–41.4 Kbp, respectively (Figs. 3, 4, Table 3). Gene organization of the plastid and mitochondrial genomes was similar and conservative among species, but non-coding sequences in the mitochondria genome were variable. Repetitive regions of *Skeletonema* were analysed. Interestingly, the unclassified element was the most abundant and accounted for 56–78% of the total repeat elements (Fig. 5). The estimated repeat regions of total length 5.1–28.4 (16.0 ± 7.9, mean $\pm$ SD) Mbp accounted for 11.0–41.1 (27.9 ± 10.3) % of the genomes (Fig. 6). Repetitive elements were found to contribute 11.3 Mbp (19.9%) and 2.55 Mbp (7.6%) to de novo *Phaeodactylum tricornutum* and *Tharassiosira pseudonana* genome assemblies, respectively[14], and larger in the seven strains of five *Skeletonema* species than *T. pseudonana* in this study. The core genome size, namely the genome size without repeat elements in *Skeletonema*, was 32.1–41.1 (38.4 ± 2.4, average $\pm$ SD) Mbp, and it was more consistent than the repeat element sizes (Fig. 6) among strains. Interestingly, a significant correlation was detected between core genome size and protein-coding gene number in the Spearman test, showing that the gene numbers affect the core genome size in *Skeletonema* (Fig. 7). However, whole genome size is largely affected by the repeat element sizes (Fig. 6).

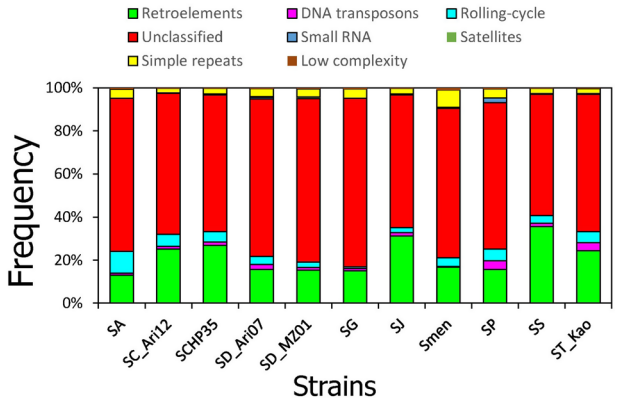

Fig. 5 The frequency of each repeat element identified by RepeatMasker in *Skeletonema*.

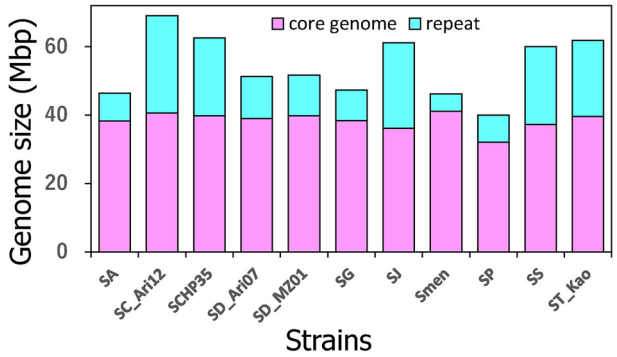

Fig. 6 The comparison of repeat size frequency in the whole genome size in *Skeletonema*.

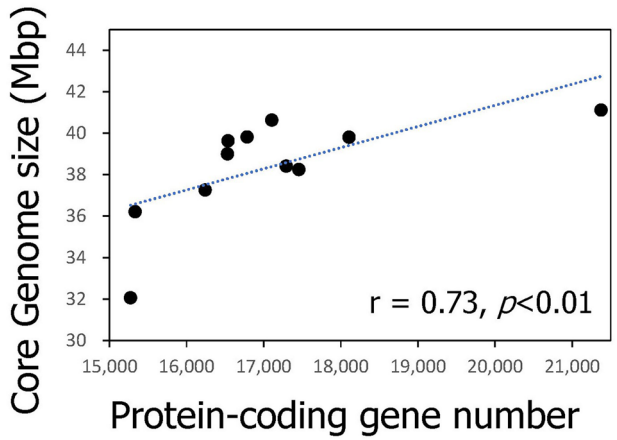

Fig. 7 The relationship between the core genome size and protein-coding gene number in *Skeletonema*.

The predicted protein-coding gene numbers were 15,275–21,376. Bacterial genome contamination was <1% in the eight strains, showing successful cultivation under axenic conditions. More than 90% of the annotated genes originated from eukaryotes, mainly from Bacillariophyceae. The result indicates that the gene model is consistent with the systematic position of *Skeletonema*. In contrast, 3–30% of bacterial genome contamination was unfortunately confirmed in the remaining three strains (Fig. 8). However, the BUSCO analysis with Stramenopiles conserved gene databases found 90–98% (93.8 ± 2.5, average ± SD) completeness in our annotation dataset (Table 2), which was the same as those estimated in the genome assembly, suggesting that our draft genome possessed a sufficient gene repertoire from Stramenopiles. We believe that these *Skeletonema* genomes will be crucial references and help deepen our understanding of diatom genome structure and function.

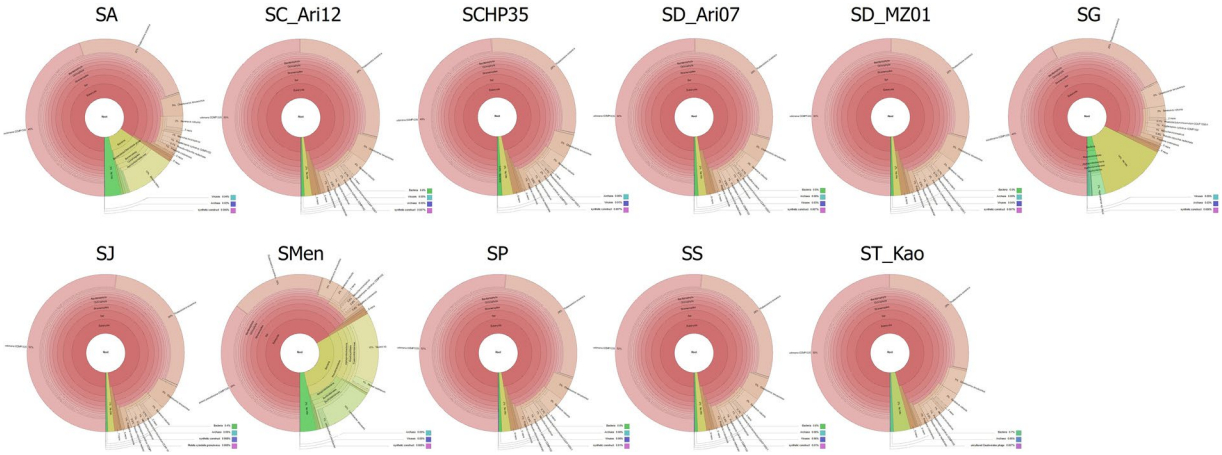

Fig. 8 Krona charts representing taxonomic composition based on the result of Diamond search of the predicted genes in *Skeletonema*. Bacterial genome contamination was <1% in the eight strains, showing successful cultivation under axenic conditions.

No.	species	strain	DNA (short_reads)				RNA (short_reads)			
			company	sequencers	Library kit	insert size	company	sequencers	Library kit	insert size
1	*Skeletonema ardens*	NIES-2837	Macrogen	Novaseq. 6000	TruSeq DNA PCR Free	350	BGI	DNBSEQ-G400	BGI kit	200-400
2	*S. costatum*	SC_Ari12	Macrogen	Novaseq. 6000	TruSeq DNA PCR Free	350	BGI	DNBSEQ-G400	BGI kit	200-400
3	*S. costatum*	SCHP35	BGI	DNBSEQ-G400	BGI kit	300-400	—	—		—
4	*S. dohrnii*	SD_Ari07	Macrogen	Novaseq. 6000	TruSeq DNA PCR Free	350	BGI	DNBSEQ-G400	BGI kit	200-400
5	*S. dohrnii*	SD_MZ01	BGI	DNBSEQ-G400	BGI kit	300-400	BGI	DNBSEQ-G400	BGI kit	200-400
6	*S. grevillei*	NIES-2840	Macrogen	Novaseq. 6000	TruSeq DNA PCR Free	350	BGI	DNBSEQ-G400	BGI kit	200-400
7	*S. japonicum*	SJ_T01	BGI	DNBSEQ-G400	BGI kit	300-400	BGI	DNBSEQ-G400	BGI kit	200-400
8	*S. menzelii*	NIES-2842	Macrogen	Novaseq. 6000	TruSeq DNA PCR Free	350	BGI	DNBSEQ-G400	BGI kit	200-400
9	*S. pseudocostatum*	NIES-2843	Macrogen	Novaseq. 6000	TruSeq DNA PCR Free	350	BGI	DNBSEQ-G400	BGI kit	200-400
10	*S. subsalsum*	SS_0375	BGI	DNBSEQ-G400	BGI kit	300-400	BGI	DNBSEQ-G400	BGI kit	200-400
11	*S. tropicum*	ST_Kao	BGI	DNBSEQ-G400	BGI kit	300-400	BGI	DNBSEQ-G400	BGI kit	200-400

Table 4. The information on the kit for library preparation kit and insert size for genome and transcriptome sequencing by short-read sequencers.

Methods

Sample collection and DNA and RNA extractions. Eleven clonal strains of 9 *Skeletonema* species were established from a bloom in various regions in Japan, Vietnam, Taiwan, and Sweden by micropipetting single chains (Table 1). f/2 medium was modified by adding 10 μM of selenious acid (H_2SeO_3) and without copper sulfate hydrate in the stock solution of the metal mixture[26]. The clonal strain was maintained in 25 ml of f/2 medium based on enrichment of natural seawater collected from Tokyo Bay (salinity adjusted to 30 PSU) in a 75 mL capacity plastic tube at a temperature of 25 °C under an irradiance of 100 μmol m^{-2} s^{-1} provided by cool-white fluorescent lamps with a 12:12 h L:D cycle. For the whole genome analysis, the strain was incubated in 3–4 × 400 mL of the modified f/2 medium with 500 mL plastic flasks under the same conditions for the maintenance culture for 5–7 days until the cultures reached exponential or stationary growth phases. The vegetative cells were harvested by filtrating through 1-μm-pore-size polycarbonate filters (Nucleopore membrane, GE Healthcare, Tokyo, Japan). Genomic DNA was extracted from the harvested cells on the filter with a modified SDS-Proteinase K method (TE buffer: 10% SDS: 20 mg mL^{-1} Proteinase K (Qiagen) = 16:15:1)[27]. Purified precipitates were dissolved in TE buffer (pH 8.0) and stored at −30 °C until further processing.

The strains were incubated under the same incubation conditions with DNA sequencing for RNA extraction, but only one of 400 mL of the modified f/2 medium in a 500 mL plastic culture flask. At the end of the incubation, the cultures were harvested by filtering mentioned at the DNA sequencing, and the total RNA was immediately extracted by PureLink RNA mini Kit with TRIzol (Thermo Fisher Scientific, MA, USA) and stored at −80 °C until further processing.

No.	species	strain	accession number
1	*Skeletonema ardens*	NIES-2837	BAAHPM010000001-BAAHPM010000104
2	*S. costatum*	SC_Ari12	BAAHPN010000001-BAAHPN010000346
3	*S. costatum*	SCHP35	BAAHPO010000001-BAAHPO010000158
4	*S. dohrnii*	SD_Ari07	BAAHPP010000001-BAAHPP010000130
5	*S. dohrnii*	SD_MZ01	BAAHPQ010000001-BAAHPQ010000150
6	*S. grevillei*	NIES-2840	BAAHPR010000001-BAAHPR010000169
7	*S. japonicum*	SJ_T01	BAAHPS010000001-BAAHPS010000286
8	*S. menzelii*	NIES-2842	BAAHPT010000001-BAAHPT010000096
9	*S. pseudocostatum*	NIES-2843	BAAHPU010000001-BAAHPU010000093
10	*S. subsalsum*	SS_0375	BAAHPV010000001-BAAHPV010000330
11	*S. tropicum*	ST_Kao	BAAHPW010000001-BAAHPW010000121

Table 5. The accession number information of the assembly genome on the strains of *Skeletonema*.

Library preparation and sequencing. Before the library preparation, a quality check (quantity of DNA, fragment size, contamination of DNA/RNA, protein or salt ions) was done by gel electrophoresis, Qubit fluorometer (Life technology, Carlsbad, CA, USA), and a general spectrophotometer. Gel electrophoresis for genomic DNA and RNA was performed using an agarose gel system and Agilent 2100 Bioanalyzer (Agilent, Tokyo, Japan), respectively. The judgment for qualified and disqualified in genomic DNA and RNA samples was mainly done by sequencing companies based on the quantity and concentration of DNA/RNA and the presence or absence of degradation. Only qualified samples were used for the library preparation.

Sequencing of DNA and RNA was performed using the Novaseq. 6000 (Illumina, San Diego, CA, USA) or DNBSEQ-G400 (MGI Tech, Shenzhen, Guangdong, China), the next-generation sequencers. For genome sequencing, library construction of Pair End libraries (150PE) were performed by the default protocol and these libraries were used for sequencing (Table 1). The kit for library preparation and insert size for genome and transcriptome sequencing by short-read sequencers were shown in Table 4. A total of 9.1–30.0 Gbp of sequences were obtained (Table 1), which were approximately 131–644 x coverage of *Skeletonema* genomes (40.3–69.3 Mbp, see below). For long-read sequencing using MinION (Oxford Nanopore Technology, Oxford, UK), the extracted genomic DNA was fragmented to ~20 kbp using Covaris g-TUBE (Covaris, Woburn, MA, USA). After purification using AMPure XP beads (Beckman Coulter, Brea, CA, USA), library preparation was performed using the SQK-LSK109 Ligation Sequencing kit (Oxford Nanopore Technologies) based on the manufacturer's protocol. The libraries were prepared and loaded onto R9.4.1 chemistry flow cell (FLO-MIN106) and sequenced using MinKNOW v 19.06.7. After sequencing, Guppy v3.2.2 (Nanopore) was used for base calling. A total of 1.1–22.1 Gbp of long-read data were obtained (Table 1), which were 20–445 x coverage of *Skeletonema* genomes. The raw reads were checked using Seqkit v2.1.0[28] and quality filtered using Seqtk v1.3-r117-dirty[29]. In RNA, a total of 6.0–15.4 Gbp of sequences were obtained, which were approximately 96–222 x coverages of *Skeletonema* genomes (Table 1). Quality check and trimming of the raw reads were done by fastp v0.22.0[30] with default setting.

Genome assembly. We estimated the overall characteristics of the *Skeletonema* genomes, including its genome size, heterozygosity, ploidy, and repeat content calculated from Illumina short–reads. The analytical flow of the whole genome in *Skeletonema* is shown in Fig. 1. KMC v3.2.0[28] and GenomeScope v2.0[24] estimated the haploid genome sizes as 43.3–85.5 Mbp (Table 2). Smudgeplot v0.2.4[24] was employed to estimate the ploidy of the vegetative cells in *Skeletonema*). We applied a hybrid de novo assembly approach based on Illumina short-reads and Nanopore long-reads. Short– and long–reads were assembled to contigs using MaSuRCA v4.0.8[31]. For gap-closing, assembled contigs were scaffolded into the draft genome using HaploMerger2 v20180603[32]. The resultant draft haploid genomes had total lengths, scaffold numbers, N50, and the longest scaffold of length, as calculated by QUAST v5.1.0rc1[32] (Table 2). We evaluated the gene completeness of our draft genome using BUSCO v5.3.0[33,34]. GetOrganelle v1.7.5.0[35] and NOVOPlasty4.3.1[36] determined the complete genome sequences of chloroplast and mitochondria. The organelle gene annotation was done with GeSeq[37]. The rRNA gene sequences, including the intergenic spacer (IGS) regions, were also obtained through the analyses of GetOrganelle, and the Nucleotide BLAST (Standard database with Nucleotide collection (nr/nt) option) in NCBI identified the species in each strain with high identity showing >99.0% (Table 3).

Repeat analysis. Repetitive regions of *Skeletonema* were identified using a combination of de novo and homology-based approaches. For homology-based prediction, known repetitive elements were identified using RepeatMasker v4.1.2[38] to search against published RepBase sequences. For de novo prediction, RepeatModeler v2.0.3[39] was executed on the *Skeletonema* assemblies to build a de novo repeat library in each species. Then, RepeatMasker was used to annotate repetitive elements using the libraries.

Gene prediction and annotation. The organelle sequences were excluded from the assembly data, and repeat regions were masked to use the assembly data for the gene prediction. RNA–seqs reads were mapped to the assembled genome sequences using HISAT2 v2.2.1[40] with default settings, and gene prediction was performed using the BRAKER2 v2.1.6 pipeline[41], which integrates RNA-Seq data through GeneMark-ET (GENEMARK v.4.68)[42] and refines gene models using AUGUSTUS v3.4.0[43], trained with the protein sequence data of *Thalassoisira*

pseudonana, the closest species to *Skeletonema*. This process resulted in the annotation of 15,275–21,376 protein-coding genes in the *Skeletonema* genomes (Table 2). BRAKER was run with default parameters, incorporating transcript evidence to enhance the precision of the predicted gene structures. The closest protein homolog of each entry in the gene models of *Skeletonema* using Diamond v2.0.13[44], and visualized results by Krona[45].

Data Records

All DNA and RNA raw reads have been deposited in DDBJ with the accession numbers of DRR539406–DRR539437 (See Table 1)[46]. The complete genome sequences of chloroplast, mitochondria and nuclear rRNA genes were deposited with the accession numbers of LC814759–LC814791 (See Table 3)[47].

The assembly genome data have also been deposited with the accession numbers of BAAHPM010000001-BAAHPM010000104, BAAHPN010000001-BAAHPN010000346, BAAHPO010000001-BAAHPO010000158, BAAHPP010000001-BAAHPP010000130, BAAHPQ010000001-BAAHPQ010000150, BAAHPR010000001-BAAHPR010000169, BAAHPS010000001-BAAHPS010000286, BAAHPT010000001-BAAHPT010000096, BAAHPU010000001-BAAHPU010000093, BAAHPV010000001-BAAHPV010000330, and BAAHPW010000001-BAAHPW010000121 (See Table 5)[48]

Technical Validation

Technical validation quality assessment of the genome assembly. The total assembly lengths are 40.3–69.3 Mbp, and the scaffold N50s are 0.3–1.1 Mbp (Table 2). BUSCO analysis was performed with Stramenopiles conserved genes databases to assess the completeness of the genome assembly, resulting in values of 90–98%.

Gene prediction and annotation validation. Gene models within the assembly were forecasted using Augustus, trained with the BUSCO assessment results. The ultimate gene set encompassed a range of 15,275 to 21,376 genes, as detailed in Table 2. The BUSCO value, ranging from 90% to 98%, closely paralleled those observed in the genome assembly. This congruence suggests the robust reliability of the generated genome models.

Code availability

All analyses were conducted on Linux systems. The leading software tools' version, code, and parameters are described below.

(1) Seqkit, version 2.1.0, parameters used: default.
(2) Seqtk, version 1.3-r117-dirty, parameters used: default.
(3) fastp, version 0.22.0, parameters used: default.
(4) KMC, version 3.2.0, parameters used: k21, ci1, cs10000.
(5) GenomeScope, version 2.0, parameters used: ploidy 2, kmer_length 21.
(6) MaSuRCA, version 4.0.8, parameters used: LIMIT_JUMP_COVERAGE = 300, CA_PARAMETERS = cgwErrorRate = 0.15, FLYE_ASSEMBLY = 0.
(7) HaploMerger2, version 20180603, parameters used: default; hm.batchA and hm.batchB.
(8) QUAST, version 5.1.0rc1, parameters used: default.
(9) BUSCO, version 5.2.2, parameters used: lineage_dataset Stramenopiles.
(10) RepeatMasker, version 4.1.2, parameters used: engine ncbi, xsmall, Database: Dfam with RBRM.
(11) RepeatModeler, version 2.0.3, parameters used: default, Database: The scaffolds assembled with MaSuRCA and HaploMerger2.
(12) Augustus, version 3.4.0, parameters used: species = Database trained with BUSCO, alternatives-from-evidence = true, hintsfle = Output of RepeatMasker.
(13) Diamond, version 2.0.14, parameters used: more-sensitive, max-target-seqs. 1, evalue 1e-5.

Received: 7 June 2024; Accepted: 24 June 2025;
Published online: 05 August 2025

References

1. Round, F.E., Crawford, R.M. & Mann, D.G. The diatoms. *Biology and morphology of the genera*. Cambridge: Cambridge University Press. p. 747 (1990).
2. Gordon, R. & Drum, R. W. The chemical basis for diatom morphogenesis. *Int Rev Cytol.* **150**, 243–372 (1994).
3. Medlin, L. K., Kooistra, W., Gersonde, R., Sims, P. & Wellbrock, U. Is the origin of diatoms related to the end-Permian mass extinction? *Nova Hedwigia* **65**, 1–11 (1997).
4. Drebes, G. Sexuality. in *The Biology of Diatoms*, ed. By D. Werner. Oxford: Blackwell Scientific Publ., 250–283 (1997).
5. Lewis, W. M. The diatom sex clock and its evolutionary significance. *Am Nat.* **123**, 73–80 (1984).
6. Geitler, L. Der formwechsel der pennaten diatomeen (kieselalgen). *Arch Protistenkd.* **78**, 1–226 (1932).
7. Nagai, S., Hori, Y., Manabe, T. & Imai, I. Restoration of cell size by vegetative cell enlargement in *Coscinodiscus wailesii* (Bacillariophyceae). *Phycologia* **34**, 533–535 (1995).
8. Karentz, D. & Smayda, T. J. Temperature and seasonal occurrence patterns of 30 dominant phytoplankton species in Narragansett Bay over a 22-year period (1959-1980). *Mar Ecol Prog Ser.* **18**, 277–293 (1984).
9. Nishikawa, T. *et al.* Nutrient and phytoplankton dynamics in Harima-Nada, eastern Seto Inland Sea, Japan during a 35-year period from 1973 to 2007. *Estuar Coasts.* **33**, 417–27 (2010).
10. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science.* **281**, 237–40 (1998).
11. Nielsen, L. T., Hallegraeff, G. M., Wright, S. W. & Hansen, P. J. Effects of experimental seawater acidification on an estuarine plankton community. *Aquat. Microb. Ecol.* **65**, 271–285, https://doi.org/10.3354/ame01554 (2012).

12. Bach, L. T. & Taucher, J. CO_2 effects on diatoms: a synthesis of more than a decade of ocean acidification experiments with natural communities. *Ocean Sci.* **15**, 1159–1175, https://doi.org/10.5194/os-15-1159-2019 (2019).
13. Tréguer, P. J. & De La Rocha, C. L. The World Ocean Silica Cycle. *Annu. Rev. Mar. Sci.* **5**, 477–501, https://doi.org/10.1146/annurev-marine-121211-172346 (2013).
14. Armbrust, E. V. *et al*. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86 (2004).
15. Bowler, C. *et al*. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
16. Ogura, A. *et al*. Comparative genome and transcriptome analysis of diatom, *Skeletonema costatum*, reveals evolution of genes for harmful algal bloom. *BMC Genomics* **19**(1), 765 (2018).
17. Filloramo, G. V., Bruce, A., Curtis, B. A., Emma Blanche, E. & Archibald, J. M. Re-examination of two diatom reference genomes using long-read sequencing. *BMC Genomics* **22**, 379 (2021).
18. Lieberman-Aiden, E. *et al*. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
19. Zingone, A., Percopo, I., Sims, P. A. & Sarno, D. Diversity in the genus Skeletonema (Bacillariophyceae) I. A reexamination of the type of S. Grevillei sp. *nov. J. Phycol.* **41**, 140–50 (2005).
20. Sarno, D., Kooistra, W. C. H. F., Medlin, L. K., Percopo, I. & Zingone, A. Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy S. costatum-like species, with the description of four new species. *J. Phycol.* **41**, 151–76 (2005).
21. Sarno, D., Kooistra, W. C. H. F., Balzano, S., Hargraves, P. E. & Zingone, A. Diversity in the genus Skeletonema (Bacillariophyceae). III. Phylogenetic position and morphological variability of Skeletonema costatum and *Skeletonema grevellei*, with the description of *Skeletonema ardens* sp. *nov. J. Phycol.* **43**, 156–70 (2007).
22. Guiry, M. D. in Guiry MDG, G.M. AlgaeBase. In: *National University of Ireland*. Galway: World-wide electronic publication. (2021).
23. Kokot, M., Dlugosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* **33**, 2759–2761 (2017).
24. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
25. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
26. Nagai, S., Matsuyama, Y., Oh, S.-J. & Itakura, S. Effect of nutrients and temperature on encystment of the toxic dinoflagellate *Alexandrium tamarense* (Dinophyceae) isolated from Hiroshima Bay, Japan. *Plankton Biol. Ecol.* **51**, 103–109 (2004).
27. Kamikawa, R., Hosoi-Tanabe, S., Nagai, S., Itakura, S. & Sako, Y. Development of a quantification assay for the cysts of the toxic dinoflagellate *Alexandrium tamarense* using real-time polymerase chain reaction. *Fisheries Science* **71**, 985–989 (2005).
28. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* https://doi.org/10.1371/journal.pone.0163962 (2016).
29. seqtk, Toolkit for processing sequences in FASTA/Q formats. Available from: https://github.com/lh3/seqtk.
30. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
31. Zimin, A. V. *et al*. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
32. Huang, S., Kang, M. & Xu, A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* **33**, 2577–2579 (2017).
33. 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).
34. Waterhouse, R. M. *et al*. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
35. Jin, J. J. *et al*. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology* **21**, 241 (2020).
36. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* https://doi.org/10.1093/nar/gkw955 (2016).
37. Tillich, M. *et al*. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* **45**, W6–W11 (2017).
38. RepeatMasker v4.1.2, Toolkit for screening DNA sequences for interspersed repeats and low complexity DNA sequences. Available from: https://www.repeatmasker.org/
39. RepeatModeler v2.0.3, Toolkit for a de novo transposable element (TE) family identification and modeling package. Available from: GitHub - Dfam-consortium/RepeatModeler: De-Novo Repeat Discovery Tool.
40. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high through put sequencing data. *Bioinformatics* **31**, 166–169, https://doi.org/10.1093/bioinformatics/btu638 (2015).
41. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* **42**(15), e119, https://doi.org/10.1093/nar/gku557 (2014).
42. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic gene prediction from RNA-Seq data and protein sequences. *Bioinformatics* **37**(23), 4202–4204, https://doi.org/10.1093/bioinformatics/btab409 (2021).
43. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
44. Buchfnk, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
45. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a web browser. *BMC Bioinform.* **12**, 384 (2011).
46. *DNA Data Bank of Japan*. https://ddbj.nig.ac.jp/search/entry/sra-study/DRP011640 (2024).
47. *DNA Data Bank of Japan*. https://ddbj.nig.ac.jp/search/entry/bioproject/PRJDB17309 (2024).
48. *DNA Data Bank of Japan*. https://getentry.ddbj.nig.ac.jp/top-j.html/BAAHPM010000001-BAAHPW010000121 (2025).

Acknowledgements

This study was supported by a research grant from the Japan Fisheries Research and Education Agency (FRA) in 2019-2022, a Grant-in-Aid for Scientific Research (Kiban-B) by the Japan Society for the Promotion of Science (18KK0812) [SN], (21H02273) [SN], (21H02274) [SN]. This work was also funded by JST/JICA, Science and Technology Research Partnership for Sustainable Development (JPMJSA1705) [SN], Vietnam MARD's Environmental grant (HAB Guides for Bivalves) [NN], the National Science and Technology Council ROC grant (NSTC 113-2611-M-019-019) [KLK], the Kajima Fundation, International Joint Research Grants (2025ICR-N05) [SN], the European Regional Development Fund and the programme Mobilitas Pluss (MOBTP160), and by the Estonian Research Council grant (PSG735) [SS].

Author contributions

S.N. and A.O. conceived the concept and experimental design. S.N. and N.N. collected samples. K.L.K. and O.K. provided culture strains. S.N., H.T., A.O. and T.G. performed wet-laboratory experiments. A.O., I.T., R.M. and S.N. performed bioinformatics analysis including visualization. S.N. wrote the draft version. A.O., M.T., M.P., S.N., S.S. and T.G. checked contributed writing-review and editing. All authors have read and commented on the submitted version of the article.

Competing interests

The authors declare no competing interests.

Additional information

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

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.