www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

High-quality genome assembly and annotation of *Thalassiosira rotula* (synonym of *Thalassiosira gravida*)

F. Di Costanzo[1,10], M. Di Marsico[2], I. Orefice[3,4], J. B. Kristoffersen[5], P. Kasapidis[5], T. Chaumier[6], L. Ambrosino[7], M. Miralto[7], R. Aiese Cigliano[2], F. Verret[5], L. Tirichine[6,8], M. Trindade[9], L. Van Zyl[9], V. Di Dato[3,11] ✉ & G. Romano[3,4,11] ✉

Diatoms are unicellular eukaryotic microorganisms thriving in most aquatic environments thanks to the expression of biosynthetic pathways for secondary metabolites involved in defence and adaptation to environmental changes. The sequencing of the transcriptome of the cosmopolitan diatom *Thalassiosira rotula* Meunier 1910 (synonym of *Thalassiosira gravida* Cleve 1896) and of the metagenome of its associated microbiome revealed the presence of biosynthetic pathways synthesising molecules and compounds useful for the algae survival and with potential biotechnological applications. Here we present the genome of a Neapolitan *T. rotula* strain, which is 672 Mbp in size due to a high proportion of repetitive elements (63.59%) and segmental duplications (14%), while the number of predicted genes resulted to be comparable to that of smaller diatom genomes. DNA methylation was predominantly located in transposable elements.

Background & Summary

Diatoms are unicellular, predominantly photosynthetic eukaryotes widespread in all aquatic environments[1] and play a key role in marine ecosystems functioning, being responsible for 40% of the organic carbon produced yearly in the sea and contributing to the biogeochemical cycling of several nutrients[1–3]. Diatoms' ecological success can be traced back to their evolutionary history, characterized by a secondary endosymbiotic event, that confers them a unique combination of metabolic features[4,5]. In addition, horizontal-gene transfer from bacteria further contributed to the metabolic variability and might have been a major driving force during diatom evolution, together with genomic rearrangements due to gene families' expansions/contractions[1,5,6]. Other elements contributing to genomic rearrangement in diatoms, and thus genome diversification, are transposable elements (TEs), a group of repetitive sequences able to change their position and expand within the genome through retro-transcription of a RNA intermediate (Class I), "cut and paste" of DNA elements or replicative transposition of a DNA intermediate (Class II)[7,8]. Interestingly, diatoms seem to possess peculiar long terminal repeat (LTR) retrotransposons (Class I), named Ty1-Copia-like elements (CoDi), detected in *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* genomes[7,9]. In *P. tricornutum*, two CoDi1 are highly expressed in nitrate starvation and in response to the toxic aldehyde decadienal, suggesting that these Class I retrotransposons could contribute to the adaptation and response to environmental stress[7,9].

[1]Ecosustainable Marine Biotechnology Department, Stazione Zoologica Anton Dohrn, Calabria Marine Centre, C.da Torre Spaccata, Amendolara, Italy. [2]Sequentia Biotech, Carrer Dr. Trueta 179, 3° 5ª, 08005, Barcelona, Spain. [3]Ecosustainable Marine Biotechnology Department, Stazione Zoologica Anton Dohrn Napoli, Via Ammiraglio Ferdinando Acton 55, 80135, Naples, Italy. [4]National Future Biodiversity Center (NFBC), Palermo, Italy. [5]Institute of Marine Biology, Biotechnology and Aquaculture, Hellenic Centre for Marine Research (HCMR), Gournes Pediados, 71003, Heraklion, Greece. [6]Nantes Université, CNRS, US2B, UMR 6286, Nantes, F-44000, France. [7]Research Infrastructures for Marine Biological Resources Department, Stazione Zoologica Anton Dohrn Napoli, Via Ammiraglio Ferdinando Acton 55, 80135, Naples, Italy. [8]Institute for Marine and Antarctic Studies (IMAS), Ecology and Biodiversity Centre, University of Tasmania, Hobart, TAS, 7004, Australia. [9]Institute for Microbial Biotechnology and Metagenomics (IMBM), Department of Biotechnology, University of the Western Cape, Cape Town, 7535, South Africa. [10]Present address: Integrative Marine Ecology Department, Stazione Zoologica Anton Dohrn Napoli, Villa Comunale, 80121, Naples, Italy. [11]These authors contributed equally: V. Di Dato, G. Romano. ✉e-mail: valeria.didato@szn.it; giovanna.romano@szn.it

Assembly	Number
Contigs (>= 0 bp)	2941
Contigs (>= 1000 bp)	2912
Contigs (>= 5000 bp)	2659
Contigs (>= 10000 bp)	2479
Contigs (>= 25000 bp)	2216
Contigs (>= 50000 bp)	1902
Total length (>= 0 bp)	672666689
Total length (>= 1000 bp)	672648741
Total length (>= 5000 bp)	671907326
Total length (>= 10000 bp)	670572246
Total length (>= 25000 bp)	666075883
Total length (>= 50000 bp)	654559745
Contigs	2941
Largest contig (bp)	4182965
Total length (bp)	672666689
GC (%)	42.87
N50 (bp)	531435
N75 (bp)	275336
L50	370
L75	805

Table 1. Summary of the *T. rotula* genome assembly in terms of number and dimension of defined contigs.

To date, 120 genomes from 80 diatom species have been sequenced and released in public repositories[10], having sizes ranging from 0.01022 Mbp of the smallest *F. kerguelensis* genome to 558.9 of the largest *Paralia guyana* genome. Fifteen diatom genomes have been deeply studied and several paper describing their features have been published. These include 9 pennates, i.e. *P. tricornutum*[6], *Pseudo-nitszchia multistriata*[11], *Pseudo-nitzschia multiseries*[12], *Fragilariopsis cylindrus*[13], *Seminavis robusta*[14], *Fistulifera solaris*[15], *Nitzschia inconspicua*[16], the non-photosynthetic raphid species *Nitzschia* sp. Nitz4[17] and the araphid *Synedra acus*[18], and 6 centrics, i.e. *Thalassiosira oceanica*[19], *T. pseudonana*[1], *Cyclotella cryptica*[20], *Skeletonema marinoi*[21,22], *Skeletonema costatum*[23,24] and *Skeletonema tropicum*[25]. The size of these genomes ranges from 24.9 Mbp of *F. solaris* to 218.7 Mbp of *P. multiseries*[12,15] with the gene numbers ranging from ~9,000 in *Nitzschia* sp.[17] to ~22,000 in *S. marinoi*[21], with *T. oceanica* and *S. robusta* behaving as outliers, having 34,642 and 36,254 genes, respectively[14,19,26]. The percentage of TEs over the total genome size has been reported for several species except for *T. oceanica*, *S. marinoi*, *Nitszchia* sp. and *S. acus*[17–19,21]. Among those species for which TE percentages have been estimated, the values vary widely, ranging from 1.79% in *T. pseudonana* to 73% in *P. multiseries* genomes[7,13,27].

The availability of the different diatom genomes opened the way to a deeper understanding of the evolution of this group of eukaryotic microalgae including highlighting their adaptation strategies[11]. Indeed, each of the sequenced species contributed to the characterization of different aspects of diatom metabolisms. For example, the genome sequencing in *P. multistriata* clarified the mechanisms of sexual reproduction and the evolutionary rate of sex-related genes in this species, while the genome sequencing of *P. tricornutum*, *T. pseudonana*, *Nitzschia* sp. Nitz4, *F. cylindrus* and *F. solaris* shed light on nutrient assimilation, signalling, photosynthesis, cold-adaptation, and lipids metabolism, respectively[1,6,11,13,15,17].

Here we describe the genome features of a strain of the centric diatom *Thalassiosira rotula* Meunier, 1910 (synonym of *Thalassiosira gravida* Cleve 1896) isolated from the Gulf of Naples, Italy, a planktonic cosmopolitan species that can dominate phytoplankton assemblages[28]. From now on, *T. gravida* have been referred with its synonym *T. rotula*.

This species was among the first studied to unveil the diatom-grazer interaction, proving its ability to impair the reproductive success of copepods through the wound-activated production of polyunsaturated aldehydes (PUAs)[29,30]. Its transcriptome sequencing revealed the expression of biosynthetic pathways responsible for the synthesis of molecules with high pharmaceutical value that have elevated chemical synthetic costs, such as prostaglandins (Pgs), secologanin, a precursor of several alkaloids, and polyketides[31,32]. Moreover, the metagenome sequencing of the associated microbiome, revealed interesting biosynthetic clusters potentially synthesising bioactive molecules such as polyketides, antibiotics, beta-lactamases/cephalosporinase and osmolytes[33]. The novelty of the occurrence of these pathways in diatoms, as identified in the *T. rotula* transcriptome and in its associated microbiome, stimulated a deeper exploration of its genome. The availability of *T. rotula* genome will provide a complimentary resource toward gaining a better understanding of these biosynthetic pathways, their main roles, and their possible regulation in response to biotic and abiotic factors.

With regards to genome assembly, a combination of Illumina and PacBio sequencing outputs suggested an estimated genome size of 677,656,337 bp. The reduced and gap-closed assembly showed a final genome size of 672 Mbp composed of 2,941 contigs, with the largest measuring 4.1 Mbp, (N50–530 Kbp; L50–370 Kbp) (Table 1). Variant calling analysis, considering all the variant types, resulted in 2,501,923 heterozygous and 1,797 homozygous loci. When increasing quality filters and considering only single nucleotides polymorphisms (SNP)

Annotation	Number
CDS	80,384
Exons	80,639
Five_prime_UTR	3,191
Total genes	35,230
Not functionally annotated genes	16,321
Three_prime_UTR	625
Segmental duplications	13,336
Transposons	744,830

Table 2. Summary of the genome structural annotation.

Category of RNA	Type of RNA	N
rRNA (N = 54)	5_8S_rRNA (5.8S ribosomal RNA)	8
	5S_rRNA (5S ribosomal RNA)	14
	LSU_rRNA_archaea (Eukaryotic large subunit ribosomal RNA)	1
	LSU_rRNA_bacteria (Eukaryotic large subunit ribosomal RNA)	5
	LSU_rRNA_eukarya (Eukaryotic large subunit ribosomal RNA)	7
	SSU_rRNA_archaea (small subunit ribosomal RNA)	1
	SSU_rRNA_bacteria (small subunit ribosomal RNA)	5
	SSU_rRNA_eukarya (small subunit ribosomal RNA)	12
	SSU_rRNA_microsporidia (small subunit ribosomal RNA)	1
tRNA (N = 870)	tRNA	850
	Selenocysteine transfer RNA (tRNA-Sec)	20
sRNA (N = 3)	Cis8_sRNA (Ruegeria cis8 sRNA)	1
	5_ureB_sRNA (5′ ureB small RNA)	1
	6S (6S/SsrS RNA)	1
snRNA (spliceosomal RNA) (N = 14)	U2 (U2 spliceosomal RNA)	4
	U4 (U4 spliceosomal RNA)	2
	U5 (U5 spliceosomal RNA)	4
	U6 (U6 spliceosomal RNA)	4
snoRNA (N = 7)	SNORD24 (Small nucleolar RNA SNORD24)	4
	snoZ157 (Small nucleolar RNA Z157/R69/R10)	1
	U3 (Small nucleolar RNA U3)	2
Bacteria_small_SRP (Bacterial small signal recognition particle RNA) (N = 1)	Bacteria_small_SRP (Bacterial small signal recognition particle RNA)	1
Histone 3′ UTR stem-loop (N = 34)	Histone 3′ UTR stem-loop	34
Catalytic Intron (group I-II) (N = 25)	Intron gpI (catalytic intron group I)	1
	Intron gpI (catalytic intron group II)	24
Antisense RNA (N = 2)	IsrR (Antisense RNA which regulates isiA expression)	2
Rnase (N = 2)	RNase_MRP (Rnase MRP)	1
	RNaseP_bact_a (Bacterial Rnase P class A)	1
mmgR RNA (Makes More Granules Regulator RNA) (N = 1)	suhB - Makes More Granules Regulator RNA (mmgR)	1
Riboswitch (N = 5)	Cobalamin riboswitch	1
	Glycine riboswitch	1
	TPP riboswitch	3
tmRNA (transfer-messenger RNA) (N = 1)	tmRNA (transfer-messenger RNA)	1

Table 3. Types of predicted non-coding RNAs and their occurrences. Abbreviation: N = number of each RNA.

as variant types, these numbers reduced to 1,528,119 and 1,297, respectively, suggesting that *T. rotula* might be diploid, with a heterozygosity rate of 0.23%.

The genome annotation enabled the prediction of 80,384 coding sequences (CDS), 80,639 exons and 35,230 genes (Table 2), among which 16,321 genes (46% of the total) remain unannotated. Almost half of the encoded functions are unknown and therefore, many functions have yet to be elucidated and characterized.

Table 3 report the predicted non-coding RNA families (i.e., rRNAs, tRNAs, snoRNAs, etc.) annotated in the *T. rotula* genome grouped by category of RNA. Unfortunately, this methodology was not able to predict miRNAs sequences.

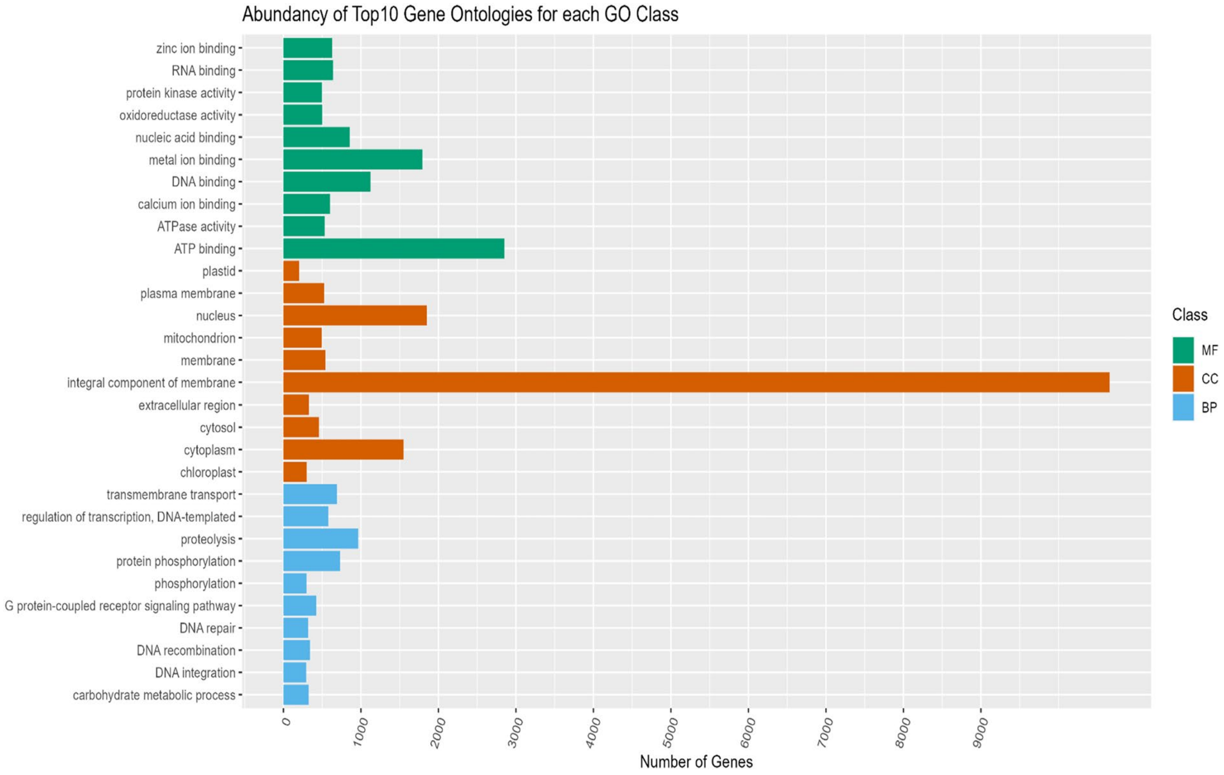

Fig. 1 Abundance of Top 10 Gene Ontologies (GO) for each Class. X axis: number of genes associated with each GO term; Y axis: GO terms for each class. MF = molecular function; CC = cellular component; BP = biological process.

Class	*T. rotula*	
	Bp covered	**Percentage (%)**
LTR_unknown	140,462,761	20.88
Copia_LTR_retrotransposon	154,006,397	22.89
Gypsy_LTR_retrotransposon	57,322,463	8.52
CACTA_TIR_transposon	15,452,925	2.3
hAT_TIR_transposon	3,653,289	0.54
Mutator_TIR_transposon	35,701,697	5.31
PIF_Harbinger_TIR_transposon	3,745,578	0.56
Tc1_Mariner_TIR_transposon	12,011,579	1.79
helitron	5,384,199	0.8
TOTAL TEs content	427,740,888	63.59

Table 4. Classification of *T. rotula* TEs. Bp covered: total bp number spanning the genome; Percentage (%): percentage of their genome coverage. Abbreviations: Bp = base pair.

The most represented Gene Ontology (GO) terms in the whole *T. rotula* genome were "ATP binding", "metal ion binding" and "DNA binding" in the Molecular Function (MF) class; "integral component of membrane", "nucleus" and "cytoplasm" in the Cellular Component (CC); "proteolysis", "protein phosphorylation" and "transmembrane transport" in the Biological Process (BP) (Fig. 1).

A large portion of the of *T. rotula* genome was composed of TEs (744,830 elements; 63.59%), of which 52.29% were classified as LTR retrotransposons (class I, types: generic, Copia and Gypsy), 10.5% were classified as Terminal Inverted Repeats (TIR) transposons (class II, types: Mutator, CACTA, Mariner, Harbinger, hAT) and 0.8% as helitrons (non TIR elements of class II) (Table 4).

Roughly 94 Mb, (14% of the total genome size), was composed of Segmental Duplications (SD = 13,336) with an average size of 7 Kb (Fig. 2a). A total of 1,319 duplicated gene families and 2,039 genes were found interspersed among the SDs. The most relevant enriched functions (GO terms) associated with this subset of genes were: "telomere maintenance", "mRNA polyadenylation" and "G protein-coupled receptor signalling pathway" for the BP class; "G protein-GABA receptor activity", "DNA helicase activity" and "magnesium chelatase activity" for MF; "septin ring", "histone acetyltransferase complex", and "Nrd1 complex" for CC (Fig. 2b).

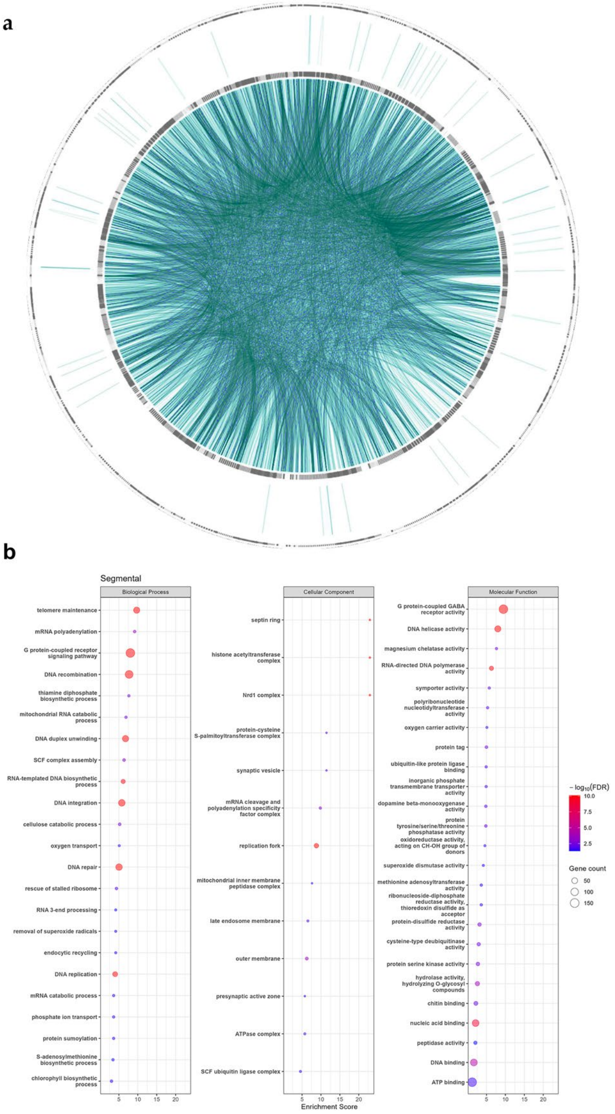

Fig. 2 Segmental duplications analysis. (**a**) Asgart Plot showing only Segmental Duplications (SD) with a duplication rate higher than 90%. A total of 13,336 SDs were identified intersected by 2,039 genes. (**b**) Bubble Plot of the Gene Ontology (GO) Enrichment Analysis on the genes present in the SD. X-axis: Enrichment Score; Y-axis: GO terms. Gradient of colors is associated with the -$\log_{10}$ (FDR, false discovery rate) of each enriched GO, while the dimension of the bubbles is proportional to the number of genes associated to that GO.

Methylation analysis showed that 69.2% of CpG dinucleotides are methylated. The genomic compartment "gene" showed a statistically significant different level of methylation (24.9%) compared to that of "TEs" (84%). Of the TEs, "Class I" TEs showed higher average methylation levels (85.8%) compared to "Class II" TEs (70.9%) (Fig. 3a).

DNA methylation entropy in the sequencing reads, defined as the combination of methylation status of contiguous CpG dinucleotides, indicated median variability of methylation patterns (Fig. 3b). Again, the "genes" compartment showed statistically significant lower methylation entropy when compared with the TEs compartment. Similarly, there were differences between Class I and II TEs with the Class I TEs having higher average methylation entropy than Class II TEs.

Methods

Cell cultures. *Thalassiosira rotula*, strain FE80, was isolated in 2011 in the Gulf of Naples (40°48.5′ N,14°15′ E), Mediterranean Sea (Italy)[34] and is currently available in the Roscoff Culture Collection with the code RCC7813. Clonal cultures were established by isolating single cells or short chains from phytoplankton net samples collected from the surface layer of the water column. Cultures were grown in sterile filtered oligotrophic seawater at 36 ppt salinity amended with f/2 nutrients, vitamins and metals[35], and maintained at 18 °C, 12:12 h light:dark photoperiod under 100 µmol photons m^2 s^{-1} irradiance.

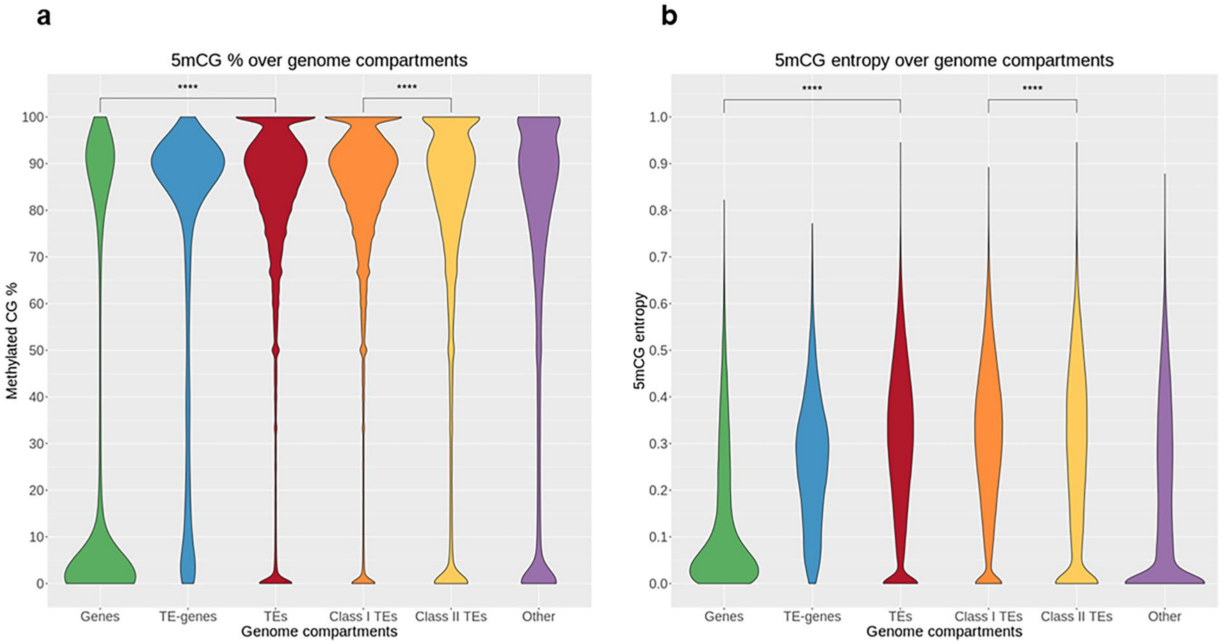

Fig. 3 Violin plots of the DNA methylation distribution and entropy levels over diverse genome compartments. (**a**) Proportion of CpG methylation and their distributions on genome compartments. Compartments percentages: Gene, 24.9%; TE-genes, 61.4%; TEs, 84%; Class I-TEs, 85.8%; Class II-TEs, 70.9%; Other, 50.7%. The term "Other" refers to loci that were neither gene, TE-genes, nor TEs. Means were compared between genes/TEs and Class I/Class II TEs with an unpaired Wilcoxon test (p-values $\leq 10^{-4}$). (**b**) DNA methylation entropy in the sequencing reads. Values close to 0 indicate low variability of methylation patterns between reads while values close to 1 show high heterogeneity among the reads. Entropy distribution means were compared between genes/TEs and Class I/Class II TEs with an unpaired Wilcoxon test, yielding a significant difference in both cases (p-values $\leq 10^{-4}$).

Axenic FE80 cultures were prepared by diluting 1:10 an exponentially growing culture in 250 mL of fresh f/2 medium supplemented with 0.1 mg/mL streptomycin (PanReac AppliChem, A1852), 0.06 mg/mL penicillin (PanReac AppliChem, A1837), 1 mg/mL ampicillin (PanReac AppliChem, A0839), 0.1 mg/mL kanamicin (PanReac AppliChem, A1493) and 0.02 mg/mL cefotaxime (Sigma-Aldrich, C7039) as previously done[31]. After three days of growth under the growth conditions reported above, a second antibiotic treatment was done by diluting again the cultures 1:10 to a final volume of 1 L with fresh f/2 medium containing the same concentration of antibiotics as above. After three more days, the cultures were refreshed with f/2 medium without antibiotics and then maintained in culture with weekly refreshes until reaching the necessary biomass to obtain 40 µg of high quality high molecular weight genomic DNA (gDNA). At each dilution step, the culture purity was assessed by checking the turbidity level of 1 mg/mL peptone solution added with 1 mL of algae culture, maintained in the dark at room temperature for 1 week. Absence of turbidity in the test tubes was used as a sign of absence of contamination. Bacteria contamination was also monitored by Dapi staining[31]. Algae biomass was collected at each refresh by centrifugation at 3500 rpm for 10 min at 4 °C (Eppendorf Centrifuge 5810 R), after which the supernatant was discarded, and the cell pellets immediately frozen in liquid nitrogen and stored at −80 °C.

DNA extraction and quality check. gDNA was obtained using a Phenol-Chloroform extraction method[36]. The extracted gDNA was dissolved in TE buffer (10 mM TrisHCl pH 7.6 and 1 mM EDTA pH 8.0) and stored at −20 °C until sequencing. High molecular weight integrity was verified by running 200 ng of gDNA on 0.8% agarose gel under 80 V continuous electric field for 40 min.

gDNA purity and concentration were verified by spectrophotometric and fluorometric reading respectively on a nanodrop reader (Thermo Fisher Scientific) and a Qubit fluorometer (Qubit 4 Fluorometer, Invitrogen by Thermo Fisher scientific) using the Qiagen dsDNA HS Assay kit. gDNA was used for Illumina, PacBio and MinION sequencing.

Library preparation and sequencing. To generate the genomic sequences, three Pacbio and one Illumina run were performed. Library preparations followed manufacturer instructions. For the Pacbio sequencing, appropriately sized double-stranded DNA fragments were generated by random shearing of DNA. Then, the template single molecule real-time sequencing (SMRTbell) library was produced by ligating universal hairpin adapters onto double-stranded DNA fragments. At the end of the protocol, the hairpin dimers formed during this process were removed using a magnetic bead purification step with size-selective conditions and adapter dimers were efficiently removed using PacBio's MagBead kit. Finally, failed ligation products were removed using exonucleases.

For the sequencing, after the exonuclease and AMPure PB purification steps, sequencing primers were annealed to the SMRTbell templates, followed by binding of the sequence polymerase to the annealed templates[37].

For the Illumina run, a TruSeq DNA PCR-free library (https://emea.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-dna-pcr-free.html) was generated using manufacturer instructions, generating a library with a 350 bp insert. Sequencing was performed on an Illumina NovaSeq6000 producing 15 Gbp of paired-end 150 bp reads.

For methylation profiling, Nanopore sequencing was performed on an Oxford Nanopore MinION sequencer. Two runs were done; the first following the manufacturer's protocol for the SQK-LSK109 sequencing kit and used part of one R9.4.1 flow cell. The second run followed the manufacturer's protocol for the SQK-LSK114 sequencing kit and used one R10.4.1 flow cell. 5mCG methylation was called using Guppy v6.5.7 with the 'sup' model. 2.6 Gbp were produced in the first run; 8.8 Gbp in the second run.

Quality control and error correction. Quality of Illumina raw sequence data was assessed using FastQC software, version 0.11.5[38]. After reviewing the FastQC report, sequences were trimmed using the software Trimmomatic[39] version 0.39 in accordance with the software manual. Trimmed reads were re-analysed with FastQC to verify the improved data quality.

***De novo* assembly.** The Illumina sequencing produced 108,255,798 paired-end reads with a length range of 35–150 bp, after trimming. The PacBio sequencing produced 6,860,415 reads with an average length of 8.3 Kbp and an N50 of 9.2 Kbp. To estimate the genome size, a Genomescope (http://genomescope.org/genomescope2.0/) analysis was performed on Illumina reads. PacBio reads were used to produce two draft *de novo* genome assemblies using both Canu[40] version 2.1.1 and Flye[41] version 2.5. Canu draft assembly was subjected to a Redundans[42] (version 0.14a) run to remove redundant contigs, perform the scaffolding and close the gaps in the genome draft assembly. Illumina reads were used to perform the scaffolding and the gap closing. Pacbio reads were also used to produce a second draft assembly using Flye. Also, this draft was subjected to a Redundans run. Flye was used with the flag–subassemblies to join the two draft genome assemblies produced. This intermediate assembly was subjected again to Redundans run using both Illumina and PacBio reads. The reduced assembly was finally corrected through the Pilon[43] algorithm using the Illumina reads. In order to exclude possible contaminations in the obtained assembly, the software Kraken2 v2.1.1 (https://ccb.jhu.edu/software/kraken2/) was used to analyse the assembled sequences against a database containing archaea, bacteria, viruses, plasmids, human sequences (UniVec_Core) and against another containing fungi, protozoa and plants (PlusPFP-16). Matches were filtered to consider only those with a confidence value of at least 0.05. Only one contig gave a match with a bacterial sequence, however a manual inspection with MegaBLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) against the "nt" database indicated a partial hit (less than 50% of alignment length), that did not provide a clear evidence for a contamination, as a consequence no contigs were filtered out.

Quality and accuracy of the final assembly were evaluated using Quast[44] version 5.0.2 and BUSCO[45] version 5.2.0, respectively. The reduced and gap closed assembly resulted in a dimension of 672 Mbp and 2,941 contigs, the largest of which measured 4.1 Mbp, an N50 equal to 530 Kbp, and an L50 equal to 370. BUSCO analysis was performed on both "Eukaryota" and "Stramenopiles" lineages.

Ploidy evaluation. To evaluate the ploidy of the *T. rotula* genome, a variant calling analysis was performed with Freebayes[46] version 1.3.4 using *T. rotula* as reference genome and the short Illumina reads as input with the following options:–min-mapping-quality 30,–min-base-quality 30,–min-coverage 4. Filtering on the produced VCF file was performed by minimum depth (10 reads) and quality score of 10 with VCFtools[47] version 0.1.17.

Prediction of RNA families. Infernal v1.1.4[48] was used to perform sequence similarity searches of each sequence of the new assembled version of *T. rotula* genome versus the RFAM database (RNA families database, Dec2021)[49]. The output from Infernal was filtered by removing all the hits with an E-value threshold $E > 0.01$. A second filtering step was performed to remove all partial/fragmented matches with incomplete hits from the reference collection.

Transposable elements. Transposable elements (TEs) were annotated on the final genome assembly using the EDTA[50] software version 2.0.0 with the CDS sequences of *Thalassiosira pseudonana* as relatively close species. The EDTA annotation ended with a softmasked version of the genome and a non-redundant TEs library.

Segmental duplication. Asgart software[51] version 2.3 was run to identify genome segmental duplications using the parameters: "Gap size" = 3000 and the "minimum length of sequences" = 5000. The chord plot was created by setting the minimal identity rate of duplication to 0.9. The overlap between the segmental duplication and the coding genes was evaluated using BEDTools (version 2.30.0) intersect[52] setting the minimum overlap to 10%.

Genome guided transcriptome assembly. *T. rotula* mRNAseq data published in Di Dato *et al.* 2019[31] were downloaded from the supplementary materials associated to the corresponding main article to generate a *de novo* genome guided transcriptome assembly. Those data were mapped on the newly produced assembly using the software STAR[53] version 2.7.9a in double pass mode. Newly produced BAM files were merged and sorted and then used as input for Trinity[54] version 2.11.0, including super-transcripts (constructed sequences generated by collapsing unique and common sequence regions among splicing isoforms into a single linear sequence), enabling jaccard_clip and setting to 100,000 the max intron length (–genome_guided_max_intron). Clustering of transcriptome sequences was performed using cd-hit-est setting the parameters −g = 1 and −c = 0.95[55]. ORFs at least 20 amino acids long were identified running TransDecoder.LongOrfs[56] (version 5.5.0) on the cd-hit-est

Accession ID	Genomes	Accession ID	Genomes
GCF_000142945.1[78]	*Phytophthora infestans*	GCF_000240725.1[79]	*Nannochloropsis gaditana*
GCA_002980425.1[80]	*Paralagenidium karlingii*	GCA_004519485.1[81]	*Nannochloropsis oceanica*
GCA_905220665.1[82]	*Albugo candida*	GCA_008828725.1[83]	*Saccharina japonica*
GCA_900088475.1[84]	*Hyphochytrium catenoides*	GCA_000310025.1[85]	*Ectocarpus siliculosus*
GCF_000186865.1[86]	*Aureococcus anophagefferens*	Nemde1[87]	*Nemacystus decipiens*
GCA_012489335.1[88]	*Triparma laevis*	GCA_012862495.1[89]	*Thraustochytrium aureum*
GCA_900642245.1[90]	*Fragilaria radians*	GCA_004332575.1[91]	*Aurantiochytrium acetophilum*
GCA_002256025.1[92]	*Asterionella formosa*	GCA_014084085.1[93]	*Hondaea fermentalgiana*
Semro1[94]	*Seminavis robusta*	GCA_018398765.1[95]	*Chromera velia*
GCF_000150955.2[96]	*Phaeodactylum tricornutum*	GCA_001179505.1[97]	*Vitrella brassicaformis*
GCA_900660405.1[98]	*Pseudo-nitzschia multistriata*	GCF_000520115.1[99]	*Aphanomyces invadans*
GCA_900095095.1[100]	*Fragilariopsis cylindrus*	GCA_002081575.1[101]	*Thraustotheca clavata*
GCA_014885115.2[102]	*Asterionellopsis glacialis*	GCA_002081595.1[103]	*Achlya hypogyna*
GCA_008632985.1[104]	*Psammoneis japonica*	GCF_000151545.1[105]	*Saprolegnia parasitica*
GCA_019693575.1[106]	*Thalassiosira oceanica*	GCA_002286825.1[107]	*Lagenidium giganteum*
GCA_018806925.1[108]	*Skeletonema costatum*	GCA_001029375.1[109]	*Pythium insidiosum*
GCF_000149405.2[110]	*Thalassiosira pseudonana*	GCA_001887855.2[111]	*Sclerospora graminicola*
GCA_013187285.1[112]	*Cyclotella cryptica*	GCA_002099245.1[113]	*Peronospora tabacina*
GCA_015148565.1[114]	*Spumella vulgaris*	GCA_004359215.2[115]	*Bremia lactucae*
GCA_015143345.1[116]	*Mallomonas annulata*	GCF_900000015.1[117]	*Plasmopara halstedii*
GCA_900617105.1[118]	*Hydrurus foetidus*	GCA_000173235.2[119]	*Hyaloperonospora arabidopsidis*
GCA_015146655.1[120]	*Ochromonas danica*	GCA_000143045.1[121]	*Globisporangium ultimum*
GCA_015146095.1[122]	*Dinobryon divergens*	GCA_001600495.1[123]	*Pilasporangium apinafurcum*

Table 5. Genomes used to perform the phylogenetics analysis with the related accession ID.

resulting sequences. To identify ORFs with homology to known proteins, the longest ORFs were analysed with DIAMOND BLAST[57] version 0.9.29 using the Stramenopiles protein database downloaded from NCBI. Finally, with TransDecoder.Predict[56] (version 5.5.0), the likely coding regions were predicted using the result of the BLAST analysis as supporting data keeping only a single best ORF per transcript. Finally, proteins obtained from this analysis and Stramenopiles proteins were merged and used as input for the gene annotation.

Genome annotation. Braker2[58] version 2.1.5 was run using the softmasked final assembly, the merged BAM file produced with the publicly available RNAseq data, and the protein sequences generated with TransDecoder[56]. Proteins were aligned to genome using the software Exonerate[59] version 2.4.0. UTR training examples were generated, and GeneMark-EP+[60] was run using "fungus" as branch point model. Annotation produced by Braker2 and the AUGUSTUS[61] (version 3.3) training parameters (–species) were used as input for the gene annotation performed by MAKER2[62]. To evaluate the accuracy of the transcriptome the resulted gene annotation was analysed using BUSCO[63] on both "Eukaryota" and "Stramenopiles" lineages.

Functional annotation. The functional annotation of *T. rotula* protein sequences was obtained by loading the sequences into the webserver PANNZER2[64] with the following options: minimum query coverage = 0.4 or minimum subject coverage = 0.4, and minimum alignment length = 50. The output included, for each gene, the description, and the associated Gene Ontology (GO) terms, when available. Kyoto Encyclopedia of Genes and Genomes (KEGG)[65] annotation was obtained by submitting the protein sequences to the KEGG Automatic Annotation Server (KAAS)[66] using the following gene dataset references: *Chlamydomonas reinhardtii*, *Breviolum minutum*, *Phaeodactylum tricornutum*, *Fragilariopsis cylindrus*, *T. pseudonana*, *Arabidopsis thaliana*, *Ostreococcus lucimarinus*, *Micromonas pusilla*.

Nanopore reads mapping and constitution of genomic compartments. Nanopore reads that were basecalled for 5mCG modifications were mapped to the assembly with Minimap2[67] (version 2.28) using the "map-ont" preset, yielding a mapping rate of 99.49%. Genomic compartments were built as follow: genes with an overlap of at least half the length of a Transposable Element (TE) on the same strand were considered TE-genes and removed from the list of genes. Repeats that were classified as "LTR" were extracted from the complete repeat annotations done by EDTA to build the Class I TEs compartment; similarly, repeats that were "DNA" or "MITE" including the Class II TEs genome fraction. Finally, all regions from the assembly that have not been classified as either genes, TE-genes or repeats fell into the "Other" category.

Selection of methylation probability filter thresholds and calculation of average levels. The base modification probability distribution in the sequencing reads was calculated on the whole genome using modkit sample-probs (https://github.com/nanoporetech/modkit version 0.3.0) with the "–no-sampling–only-mapped" flags activated, to exclude soft-clipped and inserted bases from the distribution. The methylation calls having a probability score below the 10th percentile of this distribution were discarded. The same

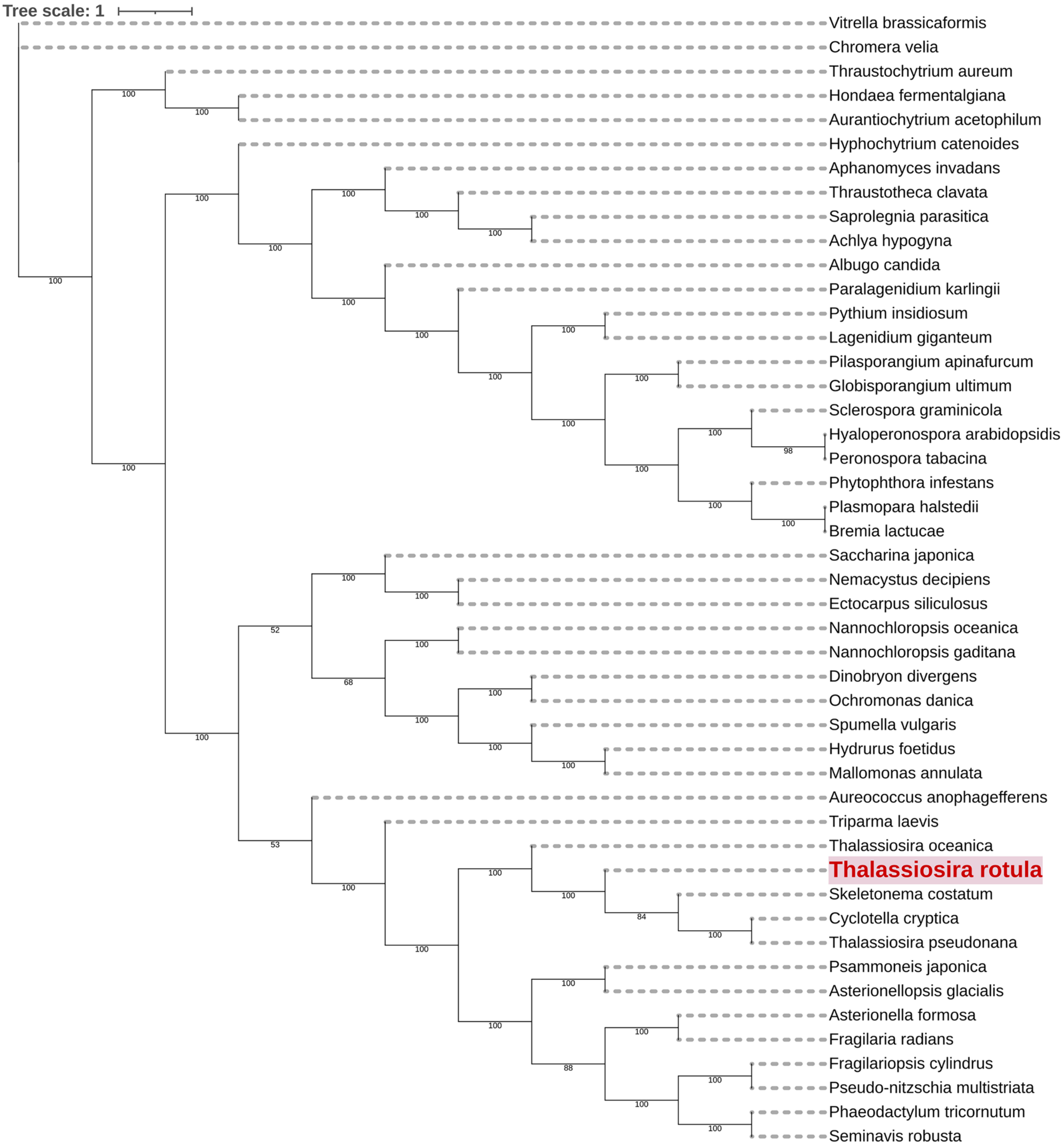

Fig. 4 Phylogenomic tree reporting all the Stramenopiles with annotated genomes. Bootstrap have been calculated using default parameters in RAxML-NG which uses maximum-likelihood (ML) optimality criterion using as substitution model LG + G4.

procedure was then performed on the six genomic compartments with the help of the "–include-bed" flag, resulting in filters thresholds that were specific to each of them. The global proportion of passing CpG methylation calls over the total number of CpG dinucleotides in the reads was then computed with modkit summary with options "–no-sampling–only-mapped–tsv–include-bed", using each compartment respective thresholds ("–filter-threshold" option). Results were plotted with R (version 4.3.3).

Computation and plotting of methylation levels per feature. Modkit call-mods was run on the BAM file from the mapping step using the whole-genome methylation probability threshold computed earlier ("–filter-threshold" option). This resulted in a BAM file with updated 5mCG probabilities, where the "ML" tags of passing CpG methylation calls were set to the maximum value (255) and the discarded calls probabilities set to 0. This updated BAM was then inputted to mbtools region-frequency (https://github.com/jts/mbtools version 0.1) with the "–cpg" flag on and each compartment feature coordinates successively ("–region-bed" option). Finally, the results of the ratios "passing CpG methylation calls/total number of CpG dinucleotides" for each genomic

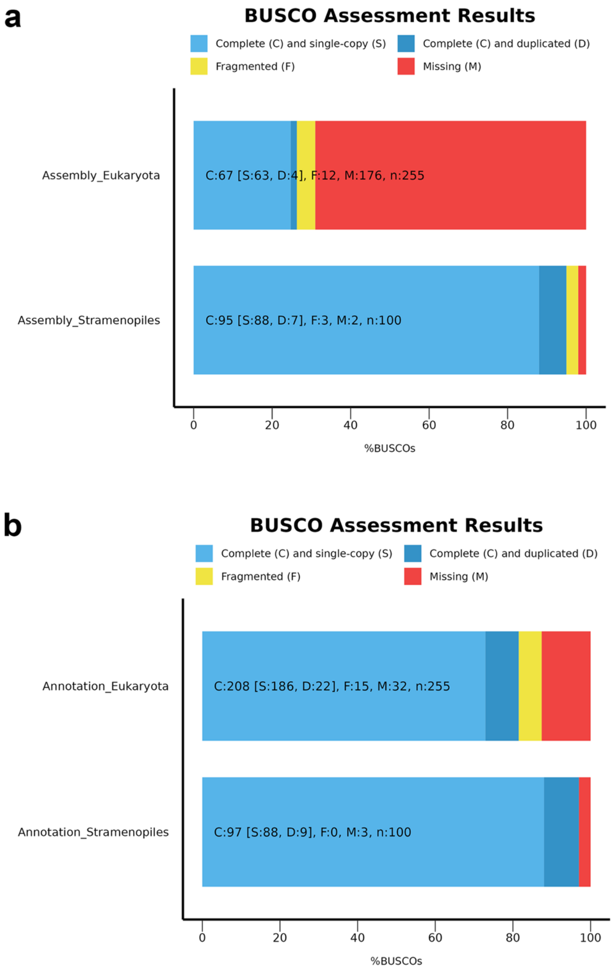

Fig. 5 BUSCO analysis performed using as reference lineages Eukaryota and Stramenopiles. (**a**) Genome analysis. At a lower taxonomic level, a lower percentage of single and complete BUSCOs were identified, while, using the higher taxonomic level, completeness of the assembly reached 95%. (**b**) Proteome analysis. At lower taxonomic level, a lower percentage of single and complete BUSCOs were identified, while, using the higher taxonomic level, completeness of the transcriptome was higher than 95%.

feature was aggregated by compartment and their distributions were plotted with R packages ggpubr 0.6.0 (https://rpkgs.datanovia.com/ggpubr/) and ggplot2 3.4.4[68].

Computation and plotting of methylation entropy per feature. The methylation entropy of each genomic feature was computed by modkit entropy for each genome compartment using their individual thresholds ("–regions" and "–filter-threshold" options, respectively). Similarly to their methylation levels, the features' entropies were aggregated by compartment and a violin plot was generated with R using the same packages and versions as above.

Phylogenomics. A GO Enrichment analysis was performed using the genes identified in the three different CAFE5 significant results. RAxML-NG[69], a method which uses maximum-likelihood (ML) optimality criterion using as substitution model LG + G4 and default parameters for bootstrap, as a tree inference program, has been used to infer the phylogenomic tree. All the genomes used to produce the phylogenomics tree with their accession numbers are reported in Table 5.

Data Records

All raw reads derived from PacBio, Illumina and MinIon sequencing have been deposited in Sequence Read Archive at NCBI under the following SRP accession number: SRP529004[70]. Specifically, raw reads are available under the accession number SRR30420826[71] for PacBio, SRR30420827[72] for Illumina and SRR30689427[73], SRR30689426[74] and SRR30689425[75] for the three MinIon runs performed. Genome assembly data have been deposited in GenBank under the accession number JBKRFO000000000[76]. Genome assembly and annotation have been also deposited in the Mendeley Data repository[77].

Technical Validation

Phylogenomic analysis. Phylogenomic analysis was performed using as input all the annotated Stramenopiles genomes with the newly assembled *T. rotula* one. The phylogenomic tree confirmed the placement of this study's *T. rotula* strain in the order *Thalassiosirales*, by grouping its genome in the same clade, that is a sister to a clade grouping the pennate diatoms, together with *T. pseudonana*, *T. oceanica* and other centric diatoms (Fig. 4).

Evaluation of genome assembly and annotation quality. To identify possible errors in the assembly and evaluate its accuracy, both Illumina and Pacbio reads were mapped on the final assembly, showing a percentage of alignment of 99.23% and 96.65%, respectively. Finally, the quality of the assembled genome and the predicted proteins were validated with BUSCO using two lineages: "Eukaryota" and "Stramenopiles" (Fig. 5). Although the Eukaryota result showed a low percentage of genome completeness and single-copy BUSCOs, the Stramenopiles lineage showed 95% of Complete BUSCOs (Fig. 5a). The BUSCO analysis performed on the predicted proteins using the Stramenopiles gene-set assessed the completeness of the proteome at 97% (Fig. 5b).

Code availability

All software and pipelines were executed according to the manual and protocols of the published bioinformatics tools. The version and parameters of software have been described in Methods.

Received: 31 July 2024; Accepted: 12 February 2025;
Published online: 20 February 2025

References

1. Armbrust, E. V. *et al*. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86 (2004).
2. Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles* **9**, 359–372 (1995).
3. Field, null, Behrenfeld, null, Randerson, null & Falkowski, null. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
4. Allen, A. E. *et al*. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203–207 (2011).
5. Benoiston, A.-S. *et al*. The evolution of diatoms and their biogeochemical functions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160397 (2017).
6. Bowler, C. *et al*. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244 (2008).
7. Maumus, F. *et al*. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC genomics* **10**, 624 (2009).
8. Bourque, G. *et al*. Ten things you should know about transposable elements. *Genome Biology* **19**, 199 (2018).
9. Hermann, D. *et al*. Introduction to the vast world of transposable elements - What about the diatoms? *Diatom Research* **29**, 91–104 (2014).
10. Bacillariophyta Genomes, NCBI Datasets. *NCBI* https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=2836.
11. Basu, S. *et al*. Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytol* **215**, 140–156 (2017).
12. Armbrust, E. V. *Whole Genome Sequencing of the Pennate Diatom* Pseudo-Nitzschia Multiseries. https://www.osti.gov/biblio/1487853 https://doi.org/10.25585/1487853 (2012).
13. Mock, T. *et al*. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* **541**, 536–540 (2017).
14. Osuna-Cruz, C. M. *et al*. The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat Commun* **11**, 3320 (2020).
15. Tanaka, T. *et al*. Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell* **27**, 162–176 (2015).
16. Oliver, A. *et al*. Diploid genomic architecture of *Nitzschia inconspicua*, an elite biomass production diatom. *Sci Rep* **11**, 15592 (2021).
17. Onyshchenko, A., Roberts, W. R., Ruck, E. C., Lewis, J. A. & Alverson, A. J. The genome of a nonphotosynthetic diatom provides insights into the metabolic shift to heterotrophy and constraints on the loss of photosynthesis. *New Phytol* **232**, 1750–1764 (2021).
18. Galachyants, Y. *et al*. Sequencing of the complete genome of an araphid pennate diatom *Synedra acus* subsp. radians from Lake Baikal. *Doklady. Biochemistry and biophysics* **461**, 84–8 (2015).
19. Lommer, M. *et al*. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol* **13**, R66 (2012).
20. Traller, J. *et al*. Genome and methylome of the oleaginous diatom *Cyclotella cryptica* reveal genetic flexibility toward a high lipid phenotype. *Biotechnology for Biofuels* **9**, (2016).
21. Johansson, O. N. *et al*. *Skeletonema marinoi* as a new genetic model for marine chain-forming diatoms. *Sci Rep* **9**, 5391 (2019).
22. Liu, S., Xu, Q. & Chen, N. Expansion of photoreception-related gene families may drive ecological adaptation of the dominant diatom species *Skeletonema marinoi*. *Science of The Total Environment* **897**, 165384 (2023).
23. Ogura, A. *et al*. Comparative genome and transcriptome analysis of diatom, *Skeletonema costatum*, reveals evolution of genes for harmful algal bloom. *BMC Genomics* **19**, 765 (2018).
24. Sorokina, M. *et al*. Draft genome assembly and sequencing dataset of the marine diatom *Skeletonema cf. costatum* RCC75. *Data in Brief* **41**, 107931 (2022).
25. Liu, S. & Chen, N. Chromosome-level genome assembly of marine diatom *Skeletonema tropicum*. *Sci Data* **11**, 403 (2024).
26. Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B. & Mock, T. Diatom molecular research comes of age: model species for studying phytoplankton biology and diversity. *The Plant Cell* **32**, 547–572 (2020).
27. Filloramo, G. V., Curtis, B. A., Blanche, E. & Archibald, J. M. Re-examination of two diatom reference genomes using long-read sequencing. *BMC Genomics* **22**, 379 (2021).
28. Whittaker, K. A., Rignanese, D. R., Olson, R. J. & Rynearson, T. A. Molecular subdivision of the marine diatom Thalassiosira rotula in relation to geographic distribution, genome size, and physiology. *BMC Evol Biol* **12**, 209 (2012).
29. Ianora, A. & Poulet, S. A. Egg viability in the copepod Temora stylifera. *Limnology and Oceanography* **38**, 1615–1626 (1993).
30. Miralto, A. *et al*. The insidious effect of diatoms on copepod reproduction. *Nature* **402**, 173–176 (1999).
31. Di Dato, V. *et al*. Unveiling the presence of biosynthetic pathways for bioactive compounds in the *Thalassiosira rotula* transcriptome. *Sci Rep* **9**, 9893 (2019).

32. Di Dato, V. *et al*. Variation in prostaglandin metabolism during growth of the diatom *Thalassiosira rotula*. *Sci Rep* **10**, 5374 (2020).
33. Di Costanzo, F. *et al*. Three novel bacteria associated with two centric diatom species from the Mediterranean Sea, *Thalassiosira rotula* and *Skeletonema marinoi*. *IJMS* **22**, 13199 (2021).
34. Lauritano, C. *et al*. Bioactivity Screening of Microalgae for Antioxidant, Anti-Inflammatory, Anticancer, Anti-Diabetes, and Antibacterial Activities. *Frontiers in Marine Science* **3**, 68 (2016).
35. Guillard, R. R. L. Culture of phytoplankton for feeding marine invertebrates. in *Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on Culture of Marine Invertebrate Animals Greenport* (eds. Smith, W. L. & Chanley, M. H.) 29–60. https://doi.org/10.1007/978-1-4615-8714-9_3 (Springer US, Boston, MA, 1975).
36. Vanstechelman, I., Sabbe, K., Vyverman, W., Vanormelingen, P. & Vuylsteke, M. Linkage mapping identifies the sex determining region as a single locus in the pennate diatom *Seminavis robusta*. *PLOS ONE* **8**, e60132 (2013).
37. Pacific Biosciences of California (2010–2014). Template preparation and sequencing guide, © Copyright 2010–2014, Inc. https://www.pacb.com/wp-content/uploads/2015/09/Guide-Pacific-Biosciences-Template-Preparation-and-Sequencing.pdf.
38. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics* https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
40. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).
41. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546 (2019).
42. Pryszcz, L. P. & Gabaldón, T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res* **44**, e113 (2016).
43. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9**, e112963 (2014).
44. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
45. 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).
46. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* **1207**, (2012).
47. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
48. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
49. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192–D200 (2020).
50. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology* **20**, 275 (2019).
51. Delehelle, F., Cussat-Blanc, S., Alliot, J.-M., Luga, H. & Balaresque, P. ASGART: fast and parallel genome scale segmental duplications mapping. *Bioinformatics* **34**, 2708–2714 (2018).
52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
53. Dobin, A. & Gingeras, T. R. Mapping RNA-seq reads with STAR. *Curr Protoc Bioinformatics* **51**, 11.14.1–11.14.19 (2015).
54. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* **29**, 644–652 (2011).
55. 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).
56. Haas, B. J. TransDecoder. TransDecoder (2024).
57. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
58. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Methods Mol Biol* **1962**, 65–95 (2019).
59. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
60. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* **2**, lqaa026 (2020).
61. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–439 (2006).
62. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
63. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr Protoc* **1**, e323 (2021).
64. Törönen, P. & Holm, L. PANNZER-A practical tool for protein function prediction. *Protein Sci* **31**, 118–128 (2022).
65. Kanehisa, M. The KEGG database. *Novartis Found Symp* **247**, 91–101; discussion 101–103, 119–128, 244–252 (2002).
66. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**, W182–185 (2007).
67. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
68. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, 2016).
69. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP529004 (2024).
71. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30420826 (2024).
72. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30420827 (2024).
73. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30689427 (2024).
74. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30689426 (2024).
75. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30689425 (2024).
76. Di Costanzo, F. *et al*. Thalassiosira rotula isolate FE80/RCC7813, whole genome shotgun sequencing project. *Genbank* https://identifiers.org/ncbi/insdc:JBKRFO000000000.1 (2025).
77. Di Costanzo, F. *et al*. *Thalassiosira rotula*, Mendeley Data, V2, https://doi.org/10.17632/c24hb3w4y2.2 2 (2024).
78. *Phytophthora infestans* T30-4 genome assembly ASM14294v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000142945.1/.
79. *Nannochloropsis gaditana* CCMP526 genome assembly ASM24072v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000240725.1/.

80. *Paralagenidium karlingii* genome assembly ASM298042v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_002980425.1/.
81. *Nannochloropsis oceanica* genome assembly ASM451948v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_004519485.1/.
82. *Albugo candida* genome assembly Assembly of Albugo candida isolate Ac2vPB. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_905220665.1/.
83. *Saccharina japonica* genome assembly ASM882872v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_008828725.1/.
84. *Hyphochytrium catenoides* genome assembly hypho_2016. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_900088475.1/.
85. *Ectocarpus siliculosus* genome assembly ASM31002v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_000310025.1/.
86. *Aureococcus anophagefferens* genome assembly v 1.0. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000186865.1/.
87. Nishitsuji, K. *et al*. Draft genome of the brown alga, Nemacystus decipiens, Onna-1 strain: Fusion of genes involved in the sulfated fucan biosynthesis pathway. *Sci Rep* **9**, 4607 (2019).
88. *Triparma laevis f. inornata* genome assembly Parmale_b_g1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_012489335.1/.
89. *Thraustochytrium aureum* genome assembly Tau_assembly01. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_012862495.1/.
90. *Fragilaria radians* genome assembly sac1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_900642245.1/.
91. *Aurantiochytrium acetophilum* genome assembly ASM433257v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_004332575.1/.
92. *Asterionella formosa* genome assembly ASM225602v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_002256025.1/.
93. *Hondaea fermentalgiana* genome assembly ASM1408408v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_014084085.1/.
94. *Seminavis robusta* genome assembly. https://bioinformatics.psb.ugent.be/gdb/seminavis/.
95. *Chromera velia* CCMP2878 genome assembly ASM1839876v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_018398765.1/.
96. *Phaeodactylum tricornutum* CCAP 1055/1 genome assembly ASM15095v2. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000150955.2/.
97. *Vitrella brassicaformis* CCMP3155 genome assembly Vbrassicaformis. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_001179505.1/.
98. *Pseudo-nitzschia multistriata* genome assembly ASM90066040v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_900660405.1/.
99. *Aphanomyces invadans* genome assembly Apha_inva_NJM9701_V1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000520115.1/.
100. *Fragilariopsis cylindrus* CCMP1102 genome assembly. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_900095095.1/.
101. *Thraustotheca clavata* genome assembly Thrcla_v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_002081575.1/.
102. *Asterionellopsis glacialis* genome assembly ASM1488511v2. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_014885115.2/.
103. *Achlya hypogyna* genome assembly ACH_lane_v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_002081595.1/.
104. *Psammoneis japonica* genome assembly ASM863298v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_008632985.1/.
105. *Saprolegnia parasitica* CBS 223.65 genome assembly ASM15154v2. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000151545.1/.
106. *Thalassiosira oceanica* genome assembly ASM1969357v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_019693575.1/.
107. *Lagenidium giganteum* genome assembly Lag_gig_ARSEF373_v1.0. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_002286825.1/.
108. *Skeletonema costatum* genome assembly FSU_Scostatum_1.0. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_018806925.1/.
109. *Pythium insidiosum* genome assembly Pythium_insidiosum_1.0. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_001029375.1/.
110. *Thalassiosira pseudonana* CCMP1335 genome assembly ASM14940v2. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_000149405.2/.
111. *Sclerospora graminicola* genome assembly SGv3. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_001887855.2/.
112. *Cyclotella cryptica* genome assembly ASM1318728v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_013187285.1/.
113. *Peronospora tabacina* genome assembly Ptab_968-S26. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_002099245.1/.
114. *Spumella vulgaris* genome assembly ASM1514856v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_015148565.1/
115. *Bremia lactucae* genome assembly BlacSF5v2. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_004359215.1/.
116. *Mallomonas annulata* genome assembly ASM1514334v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_015143345.1/.
117. *Plasmopara halstedii* genome assembly Plasmopara halstedii genome. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_900000015.1/.
118. *Hydrurus foetidus* genome assembly Hfo_v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_900617105.1/.
119. *Hyaloperonospora arabidopsidis* Emoy2 genome assembly HyaAraEmoy2_2.0. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_000173235.2/.
120. *Ochromonas danica* genome assembly ASM1514665v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_015146655.1/.
121. *Globisporangium ultimum* DAOM BR144 genome assembly. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_000143045.1/.
122. *Dinobryon divergens* genome assembly ASM1514609v1. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_015146095.1/.
123. *Pilasporangium apinafurcum* genome assembly JCM_30514_assembly_v001. *NCBI* https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCA_001600495.1/.

Acknowledgements

Part of the research leading to these results received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984, ASSEMBLE Plus project. The sequencing project have been supported by: Ministero dell'Università e Ricerca, Italia-SZN Institutional Funding; "Ocean Medicine"- MSCA-RISE-2015 - Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE), Grant ID 690944; Progetto Grande Rilevanza South Africa-Italy, "Genomics for a Blue Economy" (PGR05972), Ministero degli Affari Esteri e della Cooperazione Internazionale Italia; National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP C63C22000520001, Project title "National Biodiversity Future Center - NBFC". K.J.B., and F.V. were supported by the project "Centre for the study and sustainable exploitation of Marine Biological Resources (CMBR)" (MIS 5002670), which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund). F.V. was also supported by the project RADIO (RNA Silencing in Diatoms), which has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 483; CRIMAC - Centro Ricerche ed Infrastrutture Marine Avanzate in Calabria (CRIMAC) - Fondo FSC 2014–2020 - Piano Stralcio «Ricerca e Innovazione 2015–2017» – Programma Nazionale Infrastrutture di Ricerca (PNIR), linea d'azione 1. Cofinanziamento Infrastrutture di Ricerca (IR)», Ministero dell'Università e Ricerca, Italia.

Author contributions

G.R. and V.D.D. designed the project and contributed equally to the work as corresponding authors and last authors; V.D.D., F.D.C., I.O., F.V., J.B.K., L.V.Z. performed experimental procedures; J.B.K., L.V.Z., M.D.M., R.A.C., L.A., M.M., T.C. designed the scripts and performed data analysis; L.T., R.A.C, F.V., M.T. supervised the analysis; P.K. funded J.B.K.; F.D.C., V.D.D., I.O., G.R., M.D.M., L.A., T.C. drafted the manuscript and all Authors contributed to data interpretation, edited, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to V.D.D. or G.R.

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.