scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Coding and non-coding RNA sequencing during *Thalassiosira gravida* resting cell formation

Rosa Maria Sepe[1,2], Ida Orefice[1]✉, Marco Di Marsico[3], Marina Montresor[4], Giovanna Romano[1], Valeria Di Dato[1]✉ & Silvestro Greco[5]

Diatoms are abundant unicellular microalgae that play a fundamental role in global aquatic ecosystems, producing bioactive secondary metabolites with different functions, from signalling to defence. They exhibit complex life cycles comprising distinct phases: active growth, quiescence state in some species (with the formation of spores or resting cells), and sexual reproduction. The transition to the quiescent state permits long-term persistence, allowing species to survive under unfavourable environmental conditions. Nevertheless, the molecular mechanisms governing the switch between growth and quiescent state remain largely unknown. Here, we present the first multi-timepoint sequencing of coding and non-coding RNAs during the formation of resting cells, induced by nitrogen depletion, in the marine diatom *Thalassiosira gravida* (valid name of *Thalassiosira rotula*). This dataset provides new omics data associated with the initiation and maintenance of the quiescent state, supporting further investigation of its underlying molecular pathways.

Background & Summary

Diatoms, a major group of eukaryotic microalgae, are known for their unique silica cell wall and their significant role in aquatic ecosystems[1]. One fascinating aspect of their life cycle is the formation of quiescent states, which are crucial for their survival and ecological success. These states are dormant forms that diatoms adopt in response to unfavourable environmental conditions, such as nutrient depletion, low light availability, or extreme temperatures[2]. Quiescent states can take the form of spores, which are morphologically distinct from vegetative cells, or resting cells (RCs), which can be very similar to their vegetative counterparts. Both forms are characterized by reduced metabolic activity and physiological adaptations allowing them to survive in sediments for extended periods, up to several decades[3,4]. They act as a reservoir of genetic diversity and ensure long-term persistence of diatom populations. When environmental conditions improve, spores or RCs can germinate and repopulate the water column, contributing to seasonal blooms and maintaining the stability of aquatic ecosystems[5,6]. Investigating the induction of the transition from the vegetative to the quiescent state is not only ecologically significant, but also biologically meaningful, as the molecular mechanisms that trigger this shift remain largely unclear.

The genus *Thalassiosira* comprises numerous species capable of forming RCs and/or spores[7,8], like *Thalassiosira nordenskioeldii*[9], *Thalassiosira scotia*[10] and *Thalassiosira pseudonana*[11]. Recent studies have demonstrated that the RCs of *T. pseudonana* modifies metabolic pathways involving the inhibition of anabolic processes and the activation of catabolic ones for energy production[11]. Similar energy-related metabolic modifications have also been observed during the formation of diatoms' spores and dinoflagellates' cysts[12,13]. For instance, the diatom *Chaetoceros socialis* upregulates catabolic pathways, such as the tricarboxylic acid cycle, the glyoxylate cycle, and the fatty acid beta oxidation to utilize lipids as an energy source[12].

[1]Ecosustainable Marine Biotechnology Department, Stazione Zoologica Anton Dohrn, Via Ammiraglio Ferdinando Acton 55, Giardini Molosiglio, 80133, Naples, Italy. [2]Ecosustainable Marine Biotechnology Department, Calabria Marine Centre, CRIMAC (Centro Ricerche ed Infrastrutture Marine Avanzate in Calabria), Stazione Zoologica Anton Dohrn, Contrada Torre Spaccata, 87071, Amendolara, CS, Italy. [3]Sequentia Biotech, Carrer Dr. Trueta 179, 3° 5ª, 08005, Barcelona, Spain. [4]Integrative Marine Ecology Department, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy. [5]Integrative Marine Ecology Department, Calabria Marine Centre, CRIMAC (Centro Ricerche ed Infrastrutture Marine Avanzate in Calabria), Stazione Zoologica Anton Dohrn, Contrada Torre Spaccata, 87071, Amendolara, CS, Italy. ✉e-mail: ida.orefice@szn.it; valeria.didato@szn.it

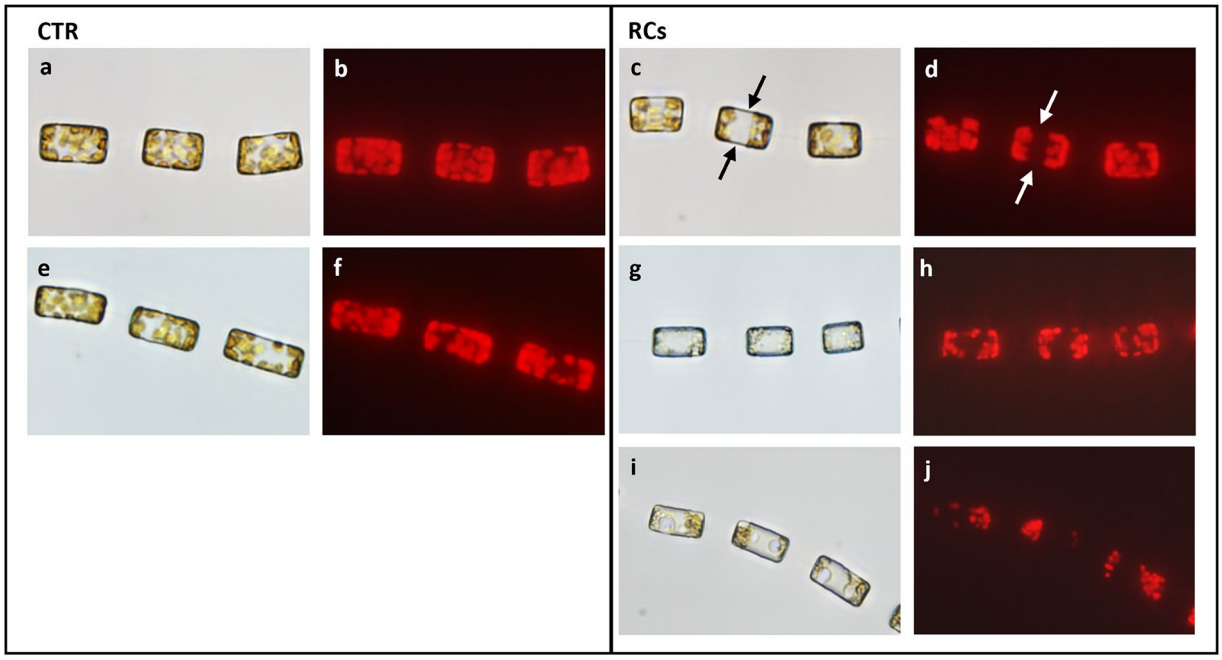

Fig. 1 Vegetative and resting cells (RCs) of *T. gravida* under different culture conditions. (**a**–**d**): Control cultures grown in complete f/2 medium (**a, b**) and cultures grown in nitrogen-depleted medium (**c, d**) for 3 days (T3). Black and white arrows: chloroplast displacement from the central region toward the cell periphery; (**e**–**h**): control cultures grown in complete f/2 medium (**e, f**) and cultures grown in nitrogen-depleted medium (**g, h**) for 7 days (T7), showing resting cells with pronounced morphological alterations, including a glassy appearance and peripheral compression of chloroplasts; (**i, j**) RCs after 30 days of dark incubation at 4 °C (T1M). (**a, c, e, g, i**) bright field; (**b, d, f, h, j**) chlorophyll a autofluorescence. All micrographs were acquired using a Zeiss Axio Observer 7 inverted microscope (Jena, Germany) at 40 × magnification.

Thalassiosira gravida Cleve, 1896 is the taxonomically valid name of *Thalassiosira rotula* Meunier, 1910[14]. It is a widespread diatom species that blooms in many areas of the world's oceans[15,16]. The scientific interest in this species started when it was demonstrated its ability to impair the reproductive success of copepods through production of polyunsaturated aldehydes (PUAs)[17,18]. Its transcriptome sequencing revealed the expression of biosynthetic pathways responsible for the synthesis of key secondary metabolites, such as prostaglandins, secologanin and polyketides[19,20]. Moreover, *T. gravida* harbours associated bacteria capable to synthesize bioactive molecules of interest[21]. In addition, germination experiments carried out using serial dilution of surface sediments sampled in the Gulf of Naples (Mediterranean Sea, Italy), have provided evidence that this species produces RCs[6,22]. *In vitro* studies have demonstrated that this species can withstand three weeks of nitrate exhaustion by forming physiological RCs, characterized by shrunken, chlorotic protoplasts, reduction in photosynthetic capacity and accumulation of particulate organic carbon[23]. However, there is a general lack of information regarding the molecular mechanism underlying the formation of RCs in this diatom species.

To address this knowledge gap, quiescence state in *T. gravida* was induced by nitrogen depletion (-N). Cells were collected at four defined time points (T) for molecular analyses, corresponding to distinct stages of the morphological transition toward the quiescent state: start of the cells inoculation (**T0**); beginning of the RCs formation in -N cultures (**T3**, 3 day of culture growth), when cells start showing morphological changes and reduced cell division (Fig. 1c,d); complete formation of RCs (**T7**, 7 day of culture growth), when all -N culture cells exhibited a glassy appearance and chloroplasts compressed along the cell wall (Fig. 1g,h); 30 days after having maintained the full RC cultures in darkness at 4 °C (**T1M**, 30 days after **T7**) to verify long-term retaining of the quiescent state (glassy appearance and chloroplast compression, Fig. 1i,j).

Sequencing of both coding and non-coding RNAs was performed at all time-points, which allowed to identify differentially expressed genes involved in the formation and maintenance of RCs. Transcriptome assembly and annotation were guided by previously sequenced transcriptomic[19] and genomic[24] data.

The present work provides the first dataset for a transcriptomic time-course analysis of *T. gravida* growth and its transition from a vegetative to a quiescent state, contributing with valuable insights in the occurring molecular changes. These data are expected to provide a framework for future genome-enabled studies and pave the way for molecular investigations of the quiescence state of diatoms.

Methods

Culture growth conditions and antibiotic treatment for axenic cultures production. The experiments were carried out using a clonal strain of *T. gravida*, FE80, isolated in 2011 from the Gulf of Naples (40°48.5′ N, 14°15′ E), Mediterranean Sea (Italy)[19]. For experimental purposes, cultures were made axenic and grown in sterile, filtered oligotrophic seawater at a salinity of ~37‰, enriched with f/2 nutrients (f/2

Replicate #	Name	Condition	Time Point	Cultures days	Treatment
1	T0CTR1	CTR	T0	0 – onset of the experiment	Complete f/2
2	T0CTR2				
3	T0CTR3				
1	T0-N1	-N	T0	0 – onset of the experiment	Nitrogen depleted f/2
2	T0-N2				
3	T0-N3				
2	T3CTR2	CTR	T3	3	Complete f/2
3	T3CTR3				
4	T3CTR4				
1	T3-N1	-N	T3	3	Nitrogen depleted f/2
3	T3-N3				
4	T3-N4				
2	T7CTR2	CTR	T7	7	Complete f/2
3	T7CTR3				
4	T7CTR4				
2	T7-N2	-N	T7	7	Nitrogen depleted f/2
3	T7-N3				
4	T7-N4				
1	T1M-N1	-N	T1M	7 + 30 in the dark at 4 °C	Nitrogen depleted f/2
2	T1M-N2				
3	T1M-N3				

Table 1. List of samples collected for sequencing. Summary of *T. gravida* samples collected for RNA-seq and small RNA-seq analyses, with indication of treatment and sampling time points. Samples have been numbered from 1 to 4 but only the best 3 RNA (based on quality analysis) have been chosen for the sequencing.

Comparison #	Code	Test sample		Reference sample	
		Time Point	Treatment	Treatment	Time Point
1	**T0_N vs T0_CTR**	T0	-N	CTR	T0
2	**T3_N vs T3_CTR**	T3	-N	CTR	T3
3	**T7_N vs T7_CTR**	T7	-N	CTR	T7
4	**T7_N vs T3_CTR**	T7	-N	CTR	T3
5	**T3_CTR vs T0_CTR**	T3	CTR	CTR	T0
6	**T7_CTR vs T0_CTR**	T7	CTR	CTR	T0
7	**T7_CTR vs T3_CTR**	T7	CTR	CTR	T3
8	**T3_N vs T0_N**	T3	-N	-N	T0
9	**T7_N vs T0_N**	T7	-N	-N	T0
10	**T1M_N vs T0_N**	T1M	-N	-N	T0
11	**T7_N vs T3_N**	T7	-N	-N	T3
12	**T1M_N vs T3_N**	T1M	-N	-N	T3
13	**T1M_N vs T7_N**	T1M	-N	-N	T7

Table 2. List of pairwise comparisons performed for differential expression analyses. Summary of all pairwise comparisons used in differential expression (DE) analyses for both RNA-seq and small RNA-seq datasets of *T. gravida*. Comparisons were designed to evaluate transcriptional differences between treatments and across time points during the transition from vegetative to resting cell stages.

medium)[25]. Cultures were incubated at 18 ± 2 °C, under a 12 h:12 h light:dark cycle, ~100 μmol photons $m^{-2} s^{-1}$. Axenicity was achieved as described in Di Dato *et al.*[19]. Briefly, a 200 mL inoculum of *T. gravida* was initially cultured in f/2 medium supplemented with a mix of antibiotics at the following final concentrations: 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 Kanamycin (PanReac Applichem, A1493) and 0.02 mg/mL Cefotaxime (Sigma-Aldrich, C7039). After 48 h, the inoculum was diluted 1:10 to a final volume of 1 L f/2 medium supplemented with the same antibiotics mix and concentration and allowed to grow under standard conditions

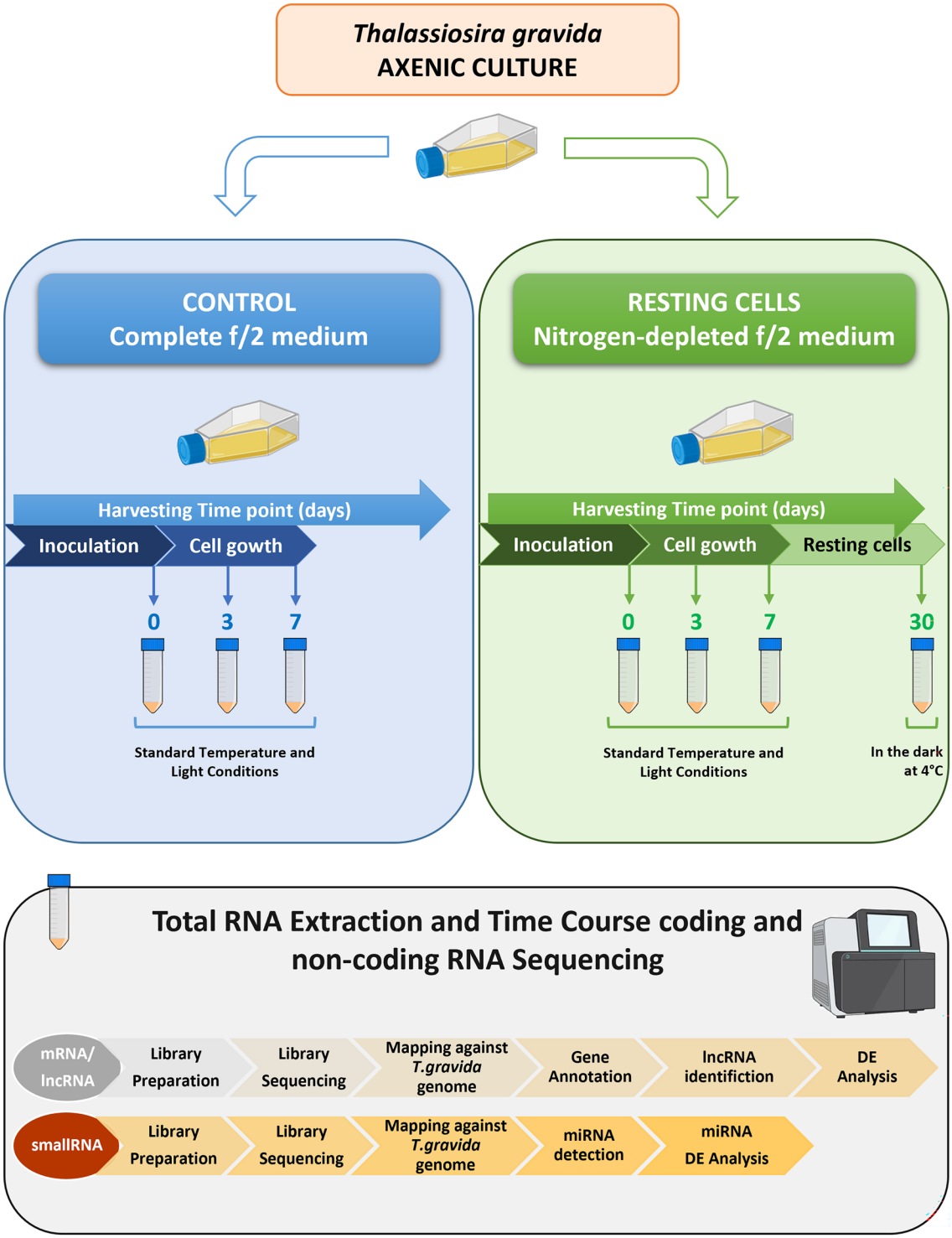

Fig. 2 Experimental design and bioinformatic workflow. Overview of the experimental setup and sequencing workflow used to characterize coding and non-coding RNAs during RCs formation in *T. gravida*. Downstream analyses included mapping on the *T. gravida* genome, gene annotation, identification of lncRNAs and miRNAs, and differential expression analysis.

described above. Peptone tests conducted at each sampling point confirmed the absence of bacterial contamination throughout the entire experiment (data not shown)[19].

Induction of RCs and assessment of awakening capacity. Experimental conditions for inducing RCs in *T. gravida* were established as follows: nitrogen-depleted f/2 medium, a chamber volume ratio of occupied to free space of 1:3, incubation for 7 days at light and temperature conditions as described above. An axenic culture

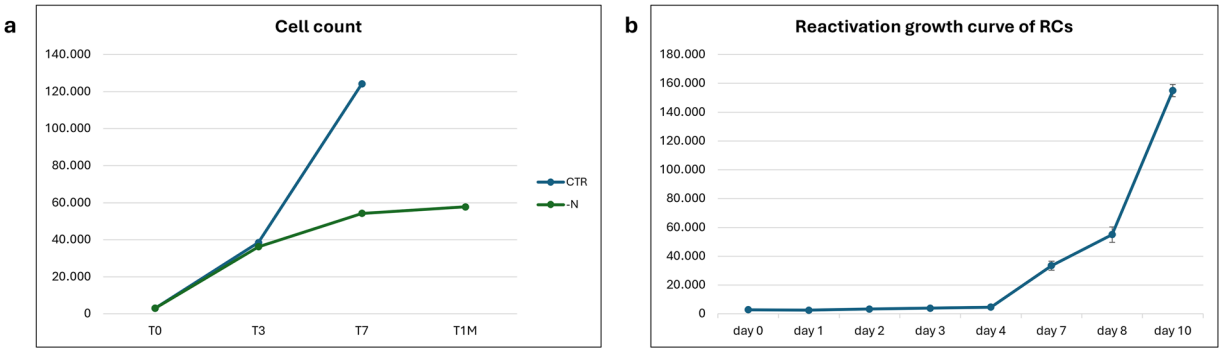

Fig. 3 Cell counts of *T. gravida* cultured in complete and nitrogen-depleted medium at different time points. (**a**) Cell density of *T. gravida* measured at different time points in complete f/2 medium (CTR, blue) and nitrogen-depleted medium (−N, green). T0: Cells Inoculation; T3: exponential phase, 3 days after T0; T7: stationary phase, 7 days after T0; T1M: 1 Month after the incubation of the T7 at 4 °C in the dark (only for the -N condition). Each enumerated time point corresponded to the experimental harvesting Ts. (**b**) Awakening of the Resting Cells (RCs). Day 0 correspond to cells maintained at 4 °C for one Month (T1M, RCs) immediately diluted in complete f/2 and incubated in standard laboratory growth conditions (See Materials and Methods). Awakened RCs growth has been evaluated by daily cell's enumeration for 10 days.

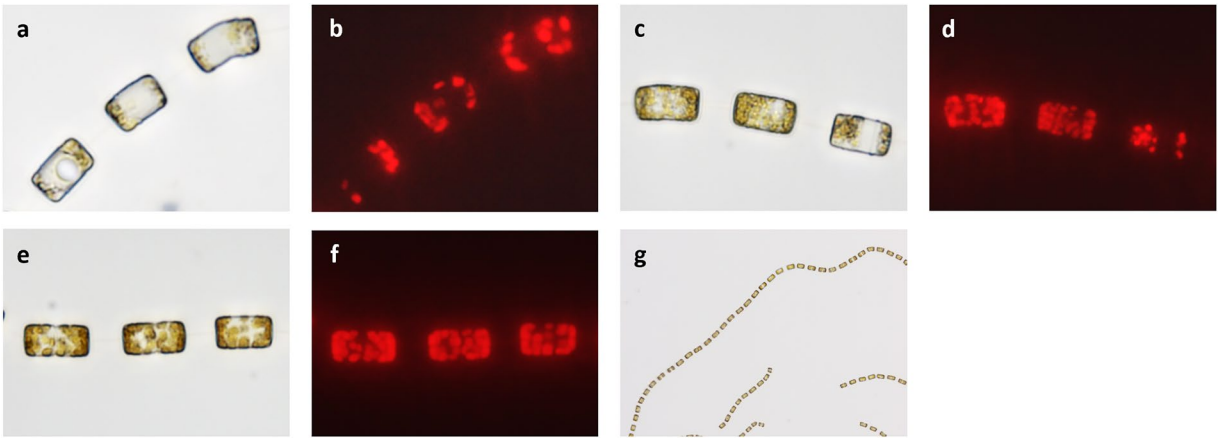

Fig. 4 Awakening of resting cells (RCs). RCs (T1M) transferred into 20 mL of complete f/2 medium and incubated under standard laboratory growth conditions progressively regained their vegetative morphology (a-f), with chloroplasts redistributed along the entire cell. (**a, b.**) RCs 3 days after incubation in complete f/2 medium. (**c, d.**) RCs 7 days after incubation in complete f/2 medium. (**e, f.**) RCs 8 days after incubation in complete f/2 medium. (**g**) RCs 10 days after incubation in complete f/2 medium forming long cell chains. (**a, c, e, g**) bright field; (**b, d, f**) chlorophyll a autofluorescence. All micrographs were acquired using a Zeiss Axio Observer 7 inverted microscope (Jena, Germany) at 40 × magnification (**a–f**) and 10 × magnification (**g**).

in late exponential phase (5 days growing culture) was inoculated at 3×10^3 cells/mL in four replicates, in a final volume of 300 mL (T300 flasks TPP 90301) of complete f/2 medium for the control (CTR) or of nitrogen depleted f/2 medium (-N) for RCs induction. RCs persistence was tested for two and half months, under axenic conditions, at 4 °C in the dark. To perform a time course transcriptomic analysis of RCs formation and persistence, both CTR and -N cultures were collected at three time points (T): T0, cultures inoculation day; T3, 3 days after the start of the cultures; T7, 7 days after the start of the cultures. One more T was included only for the -N cultures, T1M: T7 leaved for 30 days at 4 °C in the dark.

Cells were harvested by centrifugation at 2,465 × *g* (Eppendorf Centrifuge 5810 R) for 15 min at 4 °C in 50 mL tubes, followed by a second centrifugation at 10,286 × *g* (Beckman AVANTI 30 Centrifuge) for 15 min at 4 °C in 2 mL tubes. Supernatants were discarded and the cell pellets were immediately frozen in liquid nitrogen and stored at −80 °C until RNA extraction.

To evaluate the awakening capacity and recovery of the vegetative status, T1M RCs were collected by centrifugation as described above, re-suspended in 20 mL of fresh f/2 medium and incubated under standard growth conditions. Cultures were monitored daily, for 10 days, to assess their return to the vegetative state by microscopy observation using a Zeiss Axio Observe epifluorescence microscope (ZEISS, Germany). To estimate cell concentration, 2 mL of each culture were collected at each experimental T, fixed with Lugol's solution (Sigma-Aldrich), and counted using a Sedgewick-Rafter chamber under a Zeiss Axiovert 200 microscope (ZEISS, Germany) at 20 × magnification.

	mRNA/Lnc Seq analysis			smallRNA Seq analysis		
Name	reads	trimmed_reads	Kept (%)	reads	trimmed_reads	Kept (%)
T0-CTR1	48317379	48252356	99.865	25764640	25479997	98.895
T0-CTR2	54443559	54364771	99.855	22980023	22650235	98.565
T0-CTR3	46932372	46857371	99.84	25334654	25148551	99.265
T0-N1	45244545	45173670	99.843	46372389	45727301	98.609
T0-N2	45395281	45337150	99.872	46260404	45933141	99.293
T0-N3	50759344	50691041	99.865	50338900	50126641	99.578
T3-CTR2	43199595	43148451	99.882	42494429	42301003	99.545
T3-CTR3	43997127	43940500	99.871	35391834	35153863	99.328
T3-CTR4	49132118	49066885	99.867	41467554	41301942	99.601
T3-N1	49571344	49482040	99.82	52924183	52751716	99.674
T3-N3	48426244	48354508	99.852	40275573	40084614	99.526
T3-N4	50279755	50157170	99.756	43720276	43560207	99.634
T7-CTR2	46370091	46210685	99.656	41394645	41160793	99.435
T7-CTR3	46696954	46522171	99.626	37927735	37753398	99.54
T7-CTR4	50160097	50001346	99.684	42911587	42734811	99.588
T7-N2	47055038	46965331	99.809	62836514	62505375	99.473
T7-N3	49291902	49205149	99.824	90593205	90005260	99.351
T7-N4	48614338	48540624	99.848	23018624	22958049	99.737
T1M-N1	48271673	48190493	99.832	38054274	37938196	99.695
T1M-N2	47973974	47867884	99.779	25714803	25526332	99.267
T1M-N3	46783646	46703442	99.829	31336181	31104018	99.259

Table 3. Summary of sequencing reads retained after trimming. Number of reads obtained before and after adapter and quality trimming for each mRNA/lncRNA and small RNA library of *T. gravida*. "Reads" indicates the total number of raw reads; "Trimmed reads" indicates the number of reads retained after trimming; "Kept (%)" represents the percentage of reads remaining after quality filtering.

RNA extraction and quality check. Total RNA was extracted using a Total RNA Purification Kit (product # 17270 Norgen Biotek Corporation) according to the manufacturer's instructions, including a DNase digestion step (Product # 25710, RNase-Free DNase I Kit Norgen Biotek Corporation).

RNA integrity was assessed using a Plant RNA nano chip run on an Agilent 2100 Bioanalyzer platform (Agilent Technologies 5301 Stevens Creek Blvd. Santa Clara, California 95051 USA). Samples purity (260/230 and 260/280 ratios) was determined using a NanoDrop One Spectrophotometer (Thermo Fisher Scientific Inc., Waltham, Massachusetts, USA) and concentrations were measured using a Qubit Fluorometer (Thermo Fisher Scientific Inc./Qubit™ RNA Broad Range (BR) Assay Kit).

Table 1 lists the Ts and the corresponding samples collected. Only three replicates, among the four, have been chosen for the downstream analysis based on the value of the RNA integrity number.

Library preparation and sequencing. Next-generation sequencing experiments, including a post-shipment quality control of the samples, were performed by BMKGENE - Multi-omics Sequencing Services (Biomarker Technologies, Technologiepark Münster, Johann-Krane-Weg 42, 48149 Münster, Germany). Indexed libraries were prepared from 2 µg of purified RNA per sample. Hieff NGS Ultima Dual-mode mRNA Library Prep Kit for Illumina (Yeasen), model:13533ES96, was used to generate library for mRNA sequencing. The library preparations were sequenced on the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA) and reads were generated. PE150 sequencing reagent: NovaSeq 6000 S4 Reagent Kit (San Diego). VAHTS Small RNA Library Prep Kit for Illumina v2, NR811-02 was used to prepare library for small RNA sequencing. For sequencing of the constructed library, the Illumina NovaSeq X sequencing platform was used, with NovaSeqTM X Series25B Rgt Kit.

Processing of raw data. The high quality reads were aligned against the *T. gravida* genome with the STAR[26] aligner using the -twopassMode as Basics.

Bulk RNASeq mapping and annotation. In total, 21 RNA-seq libraries were analysed. Libraries were trimmed using Trimmomatic, allowing a minimum read length of 35 bp and a minimum sequencing quality of 20. Mapping was performed using STAR (v 2.7.9a)[26] with default parameters and the *T. gravida* genome[24]. LncPipe scripts have been used to perform a reference-guided annotation of each sample, with several customization on the input and output files since no reference information or known lncRNAs were available for *T. gravida*.

Identification of lncRNAs. lncRNAs have been identified using the workflow suggested by the LncPipe[27] with some customization. Since this pipeline has been developed to work with well annotated organisms such as *Homo sapiens* or *Mus musculus*, our workflow was adapted to be effective with species with no official annotation available. To identify the novel LncRNAs, BAM file from each sample was assembled and annotated with the software stringtie v2.0.5[28–31] by setting the annotation from *T. gravida* genome[24] as a guide, and specifying the strandedness of the libraries as fr-firststrand. With stringtie all the informations reported in the reference annotations

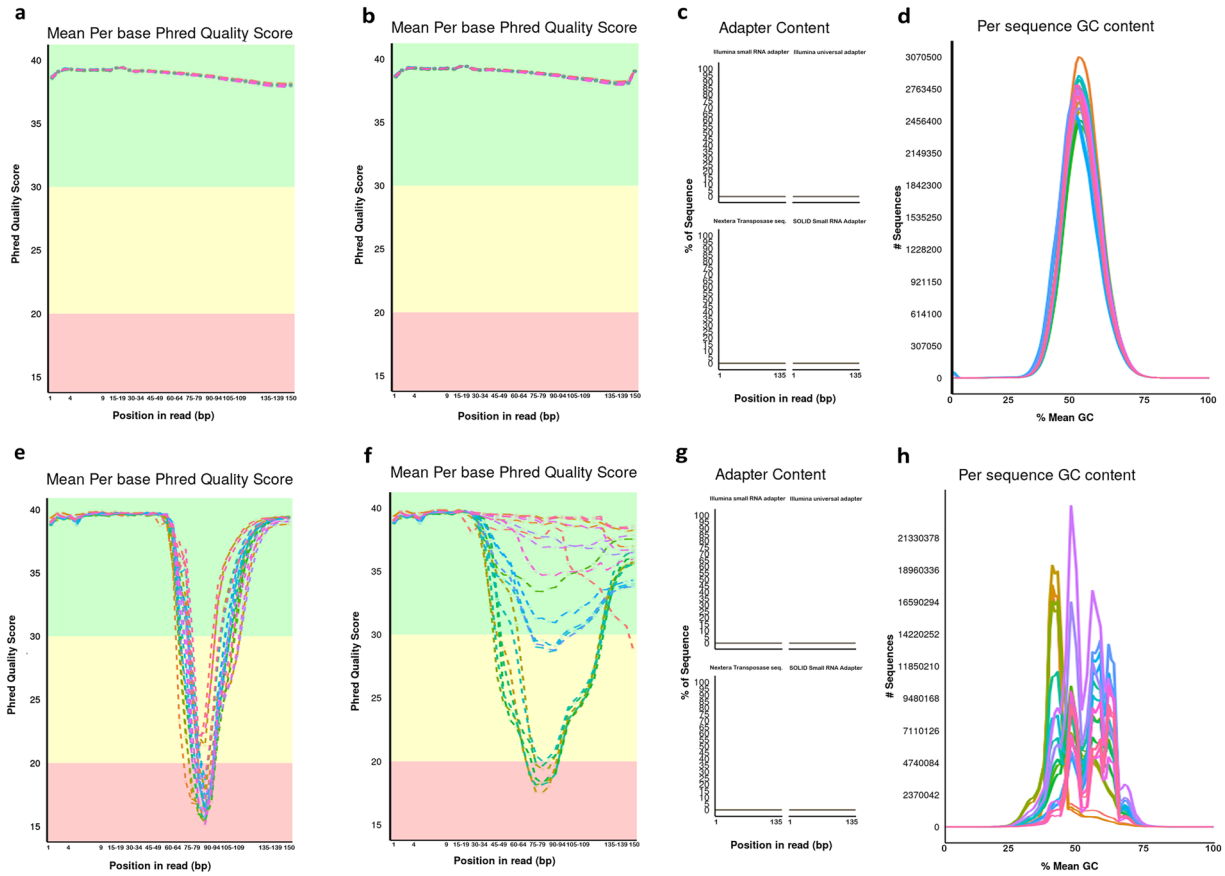

Fig. 5 Quality assessment of raw and trimmed sequencing data. Quality metrics of mRNA/lncRNA and small RNA libraries generated from *T. gravida*. (**a–d**) mRNA/lncRNA sequencing data; (**e–h**) small RNA sequencing data. (**a, e**) Mean per-base Phred quality scores before trimming; (**b, f**) mean per-base Phred quality scores after trimming; (**c, g**) adapter content per sequence; (**d, h**) per-sequence GC content. Quality assessment was performed using FastQC.

was kept as well as the novel isoforms identified by the software using the RNAseq samples. Once this step was accomplished, stringtie merge was used to pull these informations all together and reduce the redundancy by using as guide the reference annotation from *T. gravida* genome[24] and as input the resulting annotation file from each stringtie transcript assembly from the previous step. A comparison of the original annotation with the newest was performed with gffcompare v0.10.1[32] determining the production of a tmap (contains transcript-level comparisons between assembled transcripts and reference annotations) and a refmap file. Only those transcripts classified as "unknown" (u) and "Exonic overlap with a known gene" (x) were extracted from the tmap file. To exclude all the transcripts with a length smaller than 200 nt, the resulting annotation was filtered resulting in a GTF file and a FASTA file of transcripts from the GTF one. The transcript file was used as input of PLEK (v1.2)[33] and CPC2 (v1.0.1)[34] to calculate their coding potential. All the transcripts showing a coding potential lower than 0.364 were considered lncRNA. The final annotation was parsed and modified to also integrate the lncRNA in the file. Finally, the annotation containing only the coding genes with the one referred to the lncRNAs were merged. On this annotation file it was performed a re-mapping step with STAR[26] and for each sample using featureCounts was generated a count table setting minimum quality to 30 and reversely stranded counting.

Identification of small RNA. Libraries were trimmed using Trimmomatic[35], allowing a minimum read length of 15 bp, kmer length of 8 and a minimum sequencing quality of 20. To identify miRNA it was used mirdeep2[36]. All the FASTQ file were converted and merged in a unique FASTA using fastq_to_fasta (https://manpages.ubuntu.com/manpages/focal/man1/fastq_to_fasta.1.html). Reads were then mapped against the *T. gravida* reference genome[24] using the mapper script from mirdeep2, setting fasta as input file, converting RNA to DNA alphabet, removing all the entries showing sequences containing letters different from a, c, g, t, u, n, A, C, G, T, U, N, and discarding all the reads shorter than 18 base pair. microRNA detection was performed with mirdeep2 on the results of the mapping with no related species from miRBase. Quantification of the mirdeep results was performed with the quantifier script from the mirdeep suite of tools. A total of 84 miRNAs were identified.

Principal component analysis and differential expression analysis. To stabilize the variance across different expression levels, raw counts were transformed through the Variance Stabilizing Transformation (VST)

Name	total_reads	Uniquely Mapped %	Multimapped %	Multimapped Toomany %	Unmapped tooshort %	Unmapped Other %
T0CTR1	48252356	81.69	11.67	1.94	4.49	0.21
T0CTR2	54364771	81.87	11.51	1.99	4.38	0.25
T0CTR3	46857371	81.61	11.44	2	4.69	0.27
T0-N1	45173670	82.38	11.53	1.69	4.17	0.24
T0-N2	45337150	82.27	11.61	1.71	4.15	0.26
T0-N3	50691041	81.89	11.52	1.74	4.62	0.24
T3CTR2	43148451	84.37	10.11	1.12	4.14	0.26
T3CTR3	43940500	84.41	10.01	1.17	4.16	0.26
T3CTR4	49066885	84.06	9.97	1.24	4.51	0.22
T3-N1	49482040	82.81	10.24	2.54	4.12	0.29
T3-N3	48354508	82.28	10.21	2.99	4.25	0.28
T3-N4	50157170	82.4	10.41	2.81	4.12	0.25
T7CTR2	46210685	83.31	9	3.18	4.34	0.17
T7CTR3	46522171	82.22	9.55	3.52	4.53	0.18
T7CTR4	50001346	83.03	8.91	3.44	4.43	0.19
T7-N2	46965331	85.13	8.74	1.31	4.57	0.24
T7-N3	49205149	85.76	9.2	1.22	3.61	0.21
T7-N4	48540624	85.53	8.89	1.19	4.08	0.3
T1M-N1	48190493	81.55	9.77	4.39	4.06	0.23
T1M-N2	47867884	80.86	9.77	4.69	4.48	0.2
T1M-N3	46703442	81.59	9.34	4.79	4.05	0.23

Table 4. Overview of read alignment statistics for mRNA/lncRNA and small RNA libraries of *T. gravida*. "Uniquely mapped %" indicates the proportion of reads that aligned to a single genomic location; "Multimapped (%)" refers to reads that mapped to multiple positions on the genome; "Multimapped toomany (%)" includes reads with an excessive number of mapping locations; "Unmapped tooshort (%)" represents reads that were too short to align; "Unmapped other (%)" includes reads not mapped due to causes other than short length or excessive mismatches.

offered by the DESeq2 package[37]. The transformed data was then used for the Principal Component Analysis (PCA), where the top 500 most variable genes were selected based on row-wise variance. PCA was performed using the prcomp() function on the transposed data, and the proportion of variance explained by each principal component was calculated.

For bulk/Lnc RNAs Differential Expression Analysis (DEA), a count matrix was generated as input for the differential expression analysis. The count matrix was build using featureCounts[38] only counting reads with a minimum alignment quality of 30. The count matrix was filtered using the HTSeqfilter package[39] (https://bioconductor.org/packages/release/bioc/html/HTSFilter.html), and differentially expressed (DE) Genes were defined using a p.adj cutoff of 0.05 and log2FC > |1|. An Euclidean clusterization analysis was also performed using the pheatmap[40] R packages.

For miRNA, the count matrix was filtered using the HTSeqfilter package, and differentially expressed (DE) genes were defined using a p.adj cutoff of 0.05 and log2FC > |1|.

Table 2 lists the comparisons done.

Data Records

The objective of this study was to perform a molecular characterization of *T. gravida* resting cells (RCs) induced by nitrogen depletion, using a time-resolved transcriptomic analysis coupled with the identification of novel long non-coding RNAs (lncRNAs) and small RNAs. An overview of the experimental and bioinformatic workflow is shown in Fig. 2, illustrating the main steps from library preparation to transcript annotation and miRNA identification.

Metadata describing all samples analysed in this study, including growth conditions, nitrogen status, time points, biological replicates, sequencing platform, and corresponding SRA accession numbers for both RNA-seq and small RNA-seq datasets, together with all predicted miRNA sequences and their secondary structures, have been deposited in Mendeley[41].

All raw sequencing reads generated from Illumina platforms have been deposited in Sequence Read Archive (SRA) at NCBI under the following SRP accession number "SRP572204"[42] with BioProject ID "PRJNA1240261"[43].

Technical Validation

Validation of the quiescent state and assessment of cell reactivation. Culture cells counting confirmed the entering of the culture in a quiescent state. At T3 growth rates diverged between the two conditions: CTR cultures continued to proliferate up to T7, while -N ones showed reduced proliferation rate becoming RCs cultures that remained stable until T1M (Fig. 3a). Their capacity to resume growth was assessed by transferring

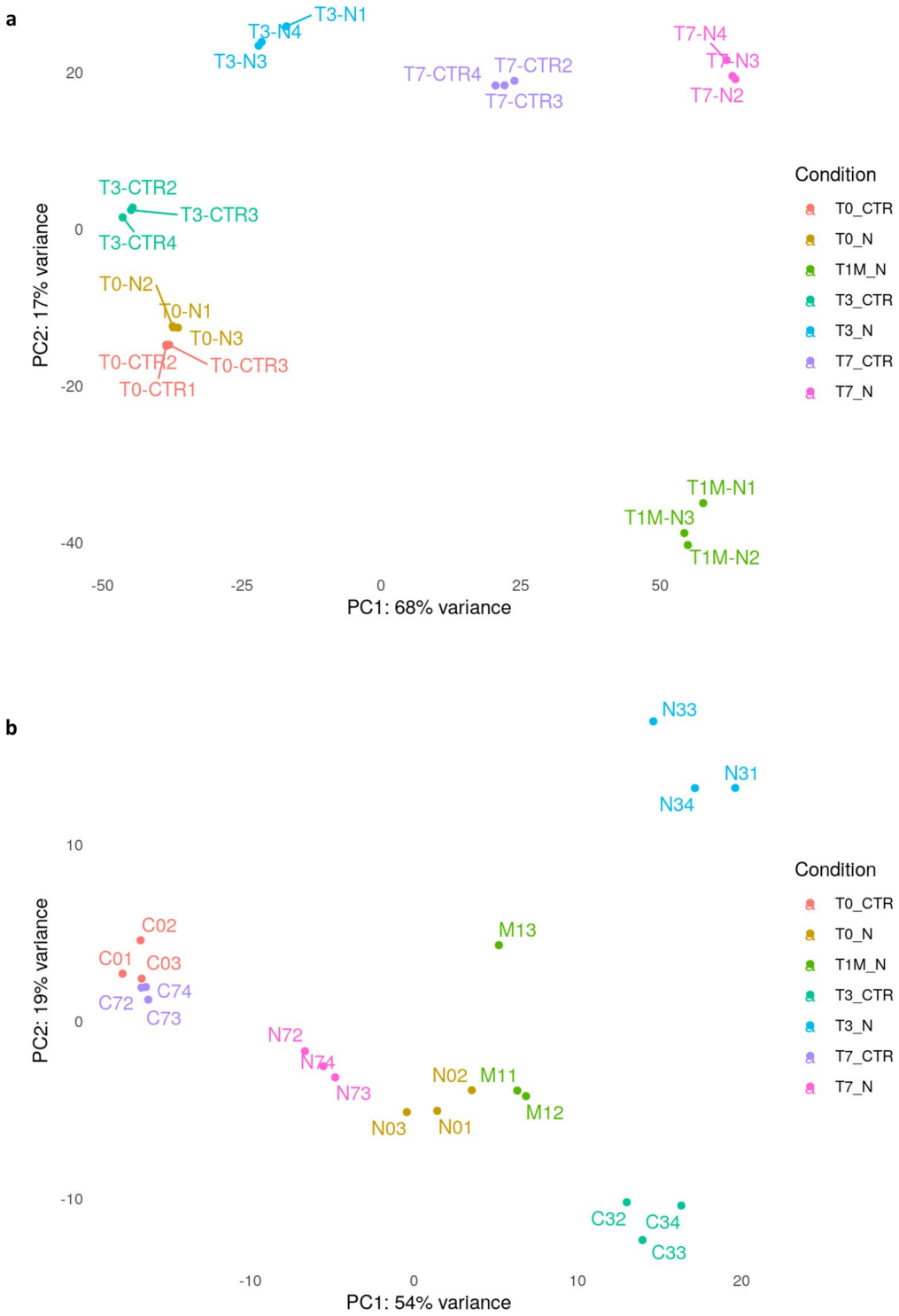

Fig. 6 Principal component analysis of transcriptomic profiles. (**a**) Principal component analysis (PCA) of mRNA/lncRNA; (**b**) PCA of small RNA libraries from *Thalassiosira gravida* cultures grown under control (CTR, complete f/2 medium) and nitrogen-depleted (−N) conditions.

the cells back into nitrogen-replete conditions (complete f/2 medium). Following an initial adaptation period, the cultures re-start their growth entering a classical exponential growth phase (Fig. 3b). Within few days, the cells progressively regained their vegetative morphology, losing the vitreous aspect with chloroplasts redistributed along the entire cell (Fig. 4). These observations demonstrated that the quiescent state was reversible and that the cells maintained full viability and proliferative potential during dormancy.

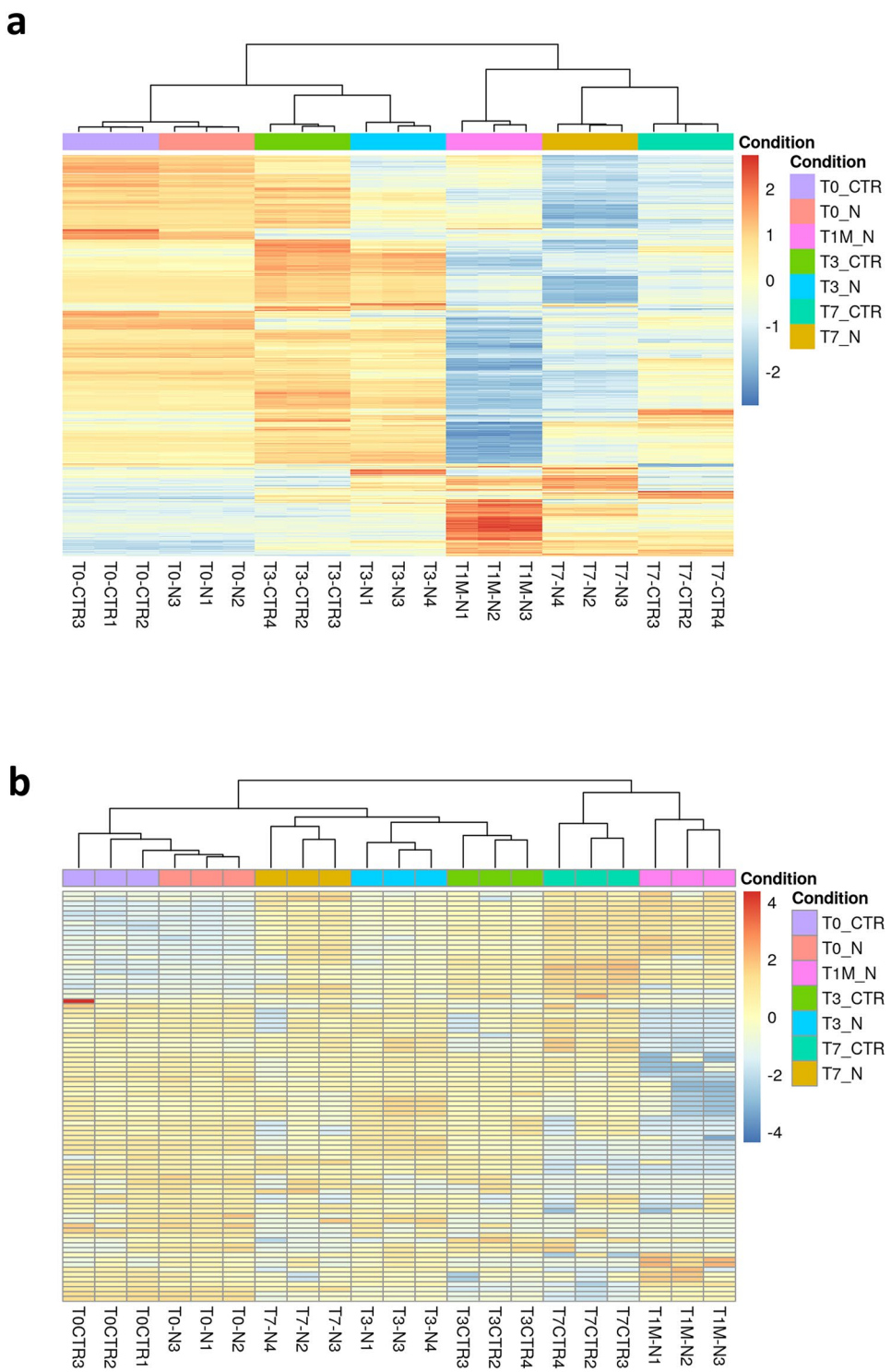

Fig. 7 Heatmaps of variance-stabilized expression data. (**a**) heatmap of the 1,000 most variable genes after variance stabilizing transformation (VST) of mRNA/lncRNA. Centering was done for each gene's values across samples. (**b**) VST of miRNA. Each miRNA value across samples was centered and plotted. The colour gradient represents gene expression relative to the mean across samples, where red indicates upregulation and blue indicates downregulation.

Processing of raw data. Table 3 shows the number of fragments present and those that passed the trimming step for both mRNA/lncRNA and smallRNA sequencing.

For each sample and type of sequencing, almost all reads were kept with a percentage over 99% (Fig. 5).

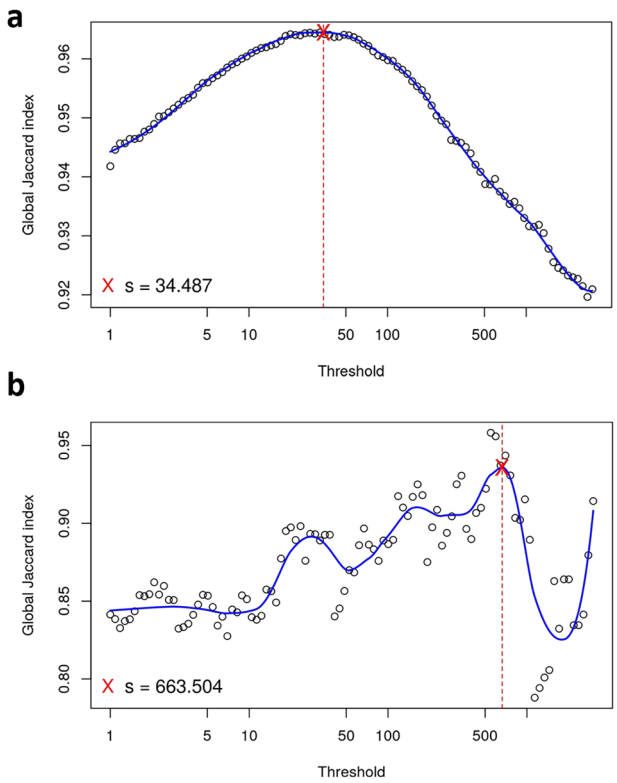

Fig. 8 Low-expression filtering using HTSFilter. Global Jaccard similarity index (s, Y-axis) calculated by the HTSFilter algorithm as a function of the minimum normalized read count threshold (X-axis). (**a**) mRNA/lncRNA dataset; (**b**) miRNA dataset.

Table 4 reports the mapping statistics for mRNA/LncRNA showing over 80% of the reads mapping uniquely, indicating a good level of sequencing accuracy. Among 8.74 and 11.67% of reads mapped many times.

Quality control. The principal component analysis (PCA) revealed a clear separation between control (CTR) and nitrogen depleted (-N) samples, as well as distinct clustering among time points (Fig. 6). Each group of biological replicates clustered tightly, confirming the goodness of the experimental sample replications and the right performance of the different Ts. The separation observed among Ts, reflects the transcriptional changes occurring throughout growth and during the induction of the resting cells.

After variance stabilizing transformation (VST) of the bulk mRNA/lncRNA dataset, the 1,000 most variable genes were selected and visualized in a heatmap (Fig. 7a). Hierarchical clustering of the samples revealed that replicates from the same experimental condition grouped together, highlighting consistent transcriptional profiles within conditions and distinct expression patterns between control and nitrogen-depleted samples.

Similarly, smallRNA hierarchical clustering (Fig. 7b) shows that those belonging to the same experimental condition tend to cluster together.

Global jaccard similarity. A global Jaccard similarity index was calculated to evaluate the degree of overlap and reproducibility among sample sets. The highest similarity among replicates was observed for the mRNA/lncRNA-seq data (s = 34.487) and for the miRNA-seq data (s = 663.504) (Fig. 8). The optimal cutoff (s) identifies the point of maximum similarity among samples. Genes or transcripts with expression values below this threshold are filtered out, while those above are retained for differential expression analysis (DEA).

Usage Notes

This dataset provides a comprehensive, time-resolved transcriptomic resource describing both coding RNA and non-coding RNA, lncRNA and small RNA (putative miRNA) dynamics during resting cell formation induced by nitrogen depletion in the marine diatom *T. gravida*. It includes sequencing data at multiple time points, enabling marine ecologists, algal physiologists, and molecular biologists to reuse these data and to explore molecular mechanisms of quiescence. Furthermore, the temporal resolution of this dataset enables comparative analyses of quiescence-related pathways among species or environmental conditions, fostering systems-level studies of dormancy, survival strategies, and metabolic reprogramming in phytoplankton.

Data availability

All raw sequencing reads (Illumina NovaSeq 6000, Illumina NovaSeqX) are available under the NCBI Sequence Read Archive SRP572204 (2025) and NCBI BioProject PRJNA1240261. The metadata reporting information on the RNAseq and the small RNAseq data including plot representing structure of the predicted miRNAs and each predicted miRNA mature and precursor sequence, precursor coordinate are available at the Mendeley Data repository under the associated https://doi.org/10.17632/7h65fhbtzp.1.

Code availability

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

Received: 26 March 2025; Accepted: 27 January 2026;
Published online: 05 February 2026

References

1. Benoiston, A.-S. *et al*. The evolution of diatoms and their biogeochemical functions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372**, 20160397 (2017).
2. Ellegaard, M. & Ribeiro, S. The long-term persistence of phytoplankton resting stages in aquatic 'seed banks'. *Biol. Rev.* **93**, 166–183 (2018).
3. Härnström, K., Ellegaard, M., Andersen, T. J. & Godhe, A. Hundred years of genetic structure in a sediment revived diatom population. *Proc. Natl. Acad. Sci.* **108**, 4252–4257 (2011).
4. Andersson, B. *et al*. Cross-contamination risks in sediment-based resurrection studies of phytoplankton. *Limnol. Oceanogr. Lett*. **8** (2022).
5. Ellegaard, M. *et al*. Dead or alive: sediment DNA archives as tools for tracking aquatic evolution and adaptation. *Commun. Biol.* **3**, 169 (2020).
6. Montresor, M., Prisco, C., Sarno, D., Margiotta, F. & Zingone, A. Diversity and germination patterns of diatom resting stages at a coastal Mediterranean site. *Mar. Ecol. Prog. Ser.* **484**, 79–95 (2013).
7. McQuoid, M. R. Influence of salinity on seasonal germination of resting stages and composition of microplankton on the Swedish west coast. *Mar. Ecol. Prog. Ser.* **289**, 151–163 (2005).
8. McQuoid, M. R. & Hobson, L. A. Diatom Resting Stages. *J. Phycol.* **32**, 889–902 (1996).
9. Sugie, K. & Kuma, K. Resting spore formation in the marine diatom *Thalassiosira nordenskioeldii* under iron- and nitrogen-limited conditions. *J. Plankton Res.* **30**, 1245–1255 (2008).
10. Johansen, J. R., Doucette, G. J. & Fryxell, G. A. The Genus *Thalassiosira* (Bacillariophyceae): Morphology of Heterovalvate Resting Spores of *T. scotia*. *Am. J. Bot.* **72**, 1861–1870 (1985).
11. Wang, G. *et al*. Resting cell formation in the marine diatom *Thalassiosira pseudonana*. *New Phytol.* **243**, 1347–1360 (2024).
12. Pelusi, A. *et al*. Gene expression during the formation of resting spores induced by nitrogen starvation in the marine diatom *Chaetoceros socialis*. *BMC Genomics* **24**, 106 (2023).
13. Guo, X., Wang, Z., Liu, L. & Li, Y. Transcriptome and metabolome analyses of cold and darkness-induced pellicle cysts of *Scrippsiella trochoidea*. *BMC Genomics* **22**, 526 (2021).
14. Sar, E. A., Sunesen, I., Lavigne, A. S. & Lofeudo, S. *Thalassiosira rotula*, a heterotypic synonym of *Thalassiosira gravida*: morphological evidence. *Diatom Res.* **26**, 109–119 (2011).
15. 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).
16. Brzezinski, M. A. & Nelson, D. M. Differential cell sinking as a factor influencing diatom species competition for limiting nutrients. *J. Exp. Mar. Biol. Ecol.* **119**, 179–200 (1988).
17. Ianora, A. & Poulet, S. A. Egg viability in the copepod *Temora stylifera*. *Limnol. Oceanogr.* **38**, 1615–1626 (1993).
18. Miralto, A. *et al*. The insidious effect of diatoms on copepod reproduction. *Nature* **402**, 173–176 (1999).
19. Di Dato, V. *et al*. Unveiling the presence of biosynthetic pathways for bioactive compounds in the *Thalassiosira rotula* transcriptome. *Sci. Rep.* **9**, 9893 (2019).
20. Di Dato, V. *et al*. Variation in prostaglandin metabolism during growth of the diatom *Thalassiosira rotula*. *Sci. Rep.* **10**, 5374 (2020).
21. Di Costanzo, F. *et al*. Three Novel Bacteria Associated with Two Centric Diatom Species from the Mediterranean Sea, *Thalassiosira rotula* and *Skeletonema marinoi*. *Int. J. Mol. Sci.* **22**, 13199 (2021).
22. Diatom resting stages in surface sediments. a pilot study comparing Next Generation Sequencing and Serial Dilution Cultures. *Cryptogam. Algol.* **38**, 31–46 (2017).
23. Peters, E. & Thomas, D. N. Prolonged nitrate exhaustion and diatom mortality: a comparison of polar and temperate *Thalassiosira* species. *J. Plankton Res.* **18**, 953–968 (1996).
24. Di Costanzo, F. *et al*. High-quality genome assembly and annotation of *Thalassiosira rotula* (synonym of *Thalassiosira gravida*). *Sci. Data* **12**, 1–14 (2025).
25. 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).
26. Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr. Protoc. Bioinforma.* **51**, 11.14.1–11.14.19 (2015).
27. Zhao, Q. *et al*. LncPipe: A Nextflow-based pipeline for identification and analysis of long non-coding RNAs from RNA-Seq data. *J. Genet. Genomics Yi Chuan Xue Bao* **45**, 399–401 (2018).
28. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLOS Comput. Biol.* **18**, e1009730 (2022).
29. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
30. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
31. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
32. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Research* **9**, ISCB Comm J-304 (2020).
33. Li, A., Zhang, J. & Zhou, Z.-Y. PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* **15**, 311 (2014).
34. Kang, Y.-J. *et al*. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).
35. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).

36. Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W. & Rajewsky, N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* **40**, 37–52 (2012).
37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
38. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).
39. Rau, A., Gallopin, M., Celeux, G. & Jaffrézic, F. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinforma. Oxf. Engl.* **29**, 2146–2152 (2013).
40. pheatmap function - RDocumentation. https://www.rdocumentation.org/packages/COMPASS/versions/1.10.2/topics/pheatmap.
41. Sepe, R. M. *et al*. Non-coding and coding RNA expression profile during Thalassiosira gravida resting cell formation. *Mendeley* https://doi.org/10.17632/7h65fhbtzp.1 (2025).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP572204.
43. https://identifiers.org/ncbi/bioproject:PRJNA1240261.

Acknowledgements

The sequencing costs, research activities, lab materials, reagents and RMS were funded by CRIMAC - Centro Ricerche ed Infrastrutture Marine Avanzate in Calabria (CRIMAC) - Fondo FSC (Fondo per lo Sviluppo e la Coesione) 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. (Project code: PIR02_00002 - CUP: C64I20000320001). Sequentia Biotech SL (Barcelona, Spain) performed the sequencing, designed the scripts and performed data analysis. We are grateful to Elvira Mauriello, technician of the SZN RIMAR (Research Infrastructures for marine biological resources) department, for her support on RNA integrity analysis.

Author contributions

S.G., funded the project, M.M., G.R., V.D.D., I.O. conceived and designed the study and coordinated the experiments. I.O., V.D.D., R.M.S. performed the experiments. M.M. and G.R. supervised the resting cell induction conditions. V.D.D. and I.O. supervised the analyses. M.D.M. performed the bioinformatic analysis, R.M.S. and V.D.D. drafted the manuscript and all authors contributed to data interpretation, edited, and approved the final manuscript.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.O. or V..D.