Transcriptomes from the diatoms *Thalassiosira* and *Minidiscus* from the English Channel and Antarctica.

Mariela Guajardo[1], Valeria Jimenez[2], Daniel Vaulot[2,3], and Nicole Trefault[1,*]

[1]Centro GEMA- Genómica, Ecología y Medio Ambiente, Facultad de Ciencias, Universidad Mayor, Santiago, 8580745, Chile.
[2]Sorbonne Université, CNRS, UMR7144, Ecology of Marine Plankton team, Station Biologique de Roscoff, 29680 Roscoff, France
[3]Asian School of the Environment, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798
[*]corresponding author(s): Nicole Trefault (nicole.trefault@umayor.cl)

ABSTRACT

Diatoms are key members of the oceanic phytoplankton and major contributors to global marine primary production and biogeochemical cycles. The family Thalassiosiraceae is a successful and diverse diatom group, which dominates in particular in coastal water. Despite recent 'omic' efforts dedicated to diatoms, more reference sequence data are needed to elucidate the molecular mechanisms that enable their success in a wide range of environments. We present ten transcriptomes from species belonging to the genera *Thalassiosira* and *Minidiscus* isolated from the English Channel and the Western Antarctic Peninsula. We describe the assembly process, quality evaluation, annotation procedure, and gene expression analysis. The data generated are of high quality, with good assembly and annotation metrics. Similarity analysis shows a clear separation according to environment and species. These data will be of high interest for phytoplankton genomic researchers since it includes the first transcriptomes from Thalassiosiraceae strains isolated from Antarctic waters and the first ones for species of the genus *Minidiscus*.

ORCID Numbers

- Mariela Guajardo: 0000-0001-7865-9751
- Daniel Vaulot: 0000-0002-0717-5685
- Nicole Trefault: 0000-0002-4388-6791

Background & Summary

Diatoms are unicellular photosynthetic eukaryotes and key members of phytoplankton in all oceans and aquatic systems. They contribute to 20% of the global annual marine primary production[1] and take part in major marine biogeochemical cycles including those of carbon[2], silicate[3], nitrogen[4,5] and phosphorus[6]. Thalassiosiraceae are one of the most studied and diverse diatom family. They are recognized by their distinctive morphological features: valves with radial symmetry, absence of a raphe system and elaborated frustule ornamentation[7]. Thalassiosiraceae inhabit brackish, nearshore and open-ocean environments. Within this family, the genera *Thalassiosira* and *Minidiscus* make important contribution to the carbon export in various coastal and offshore oceanic regions[8]. They are major components of summer blooms in Antarctic coastal and oceanic waters, where they can be responsible for up to 90% of primary production when blooms occur[9,10].

'Omics' approaches has been widely applied in diatom research, providing a better understanding of diatom evolution and ecology[11–18]. However, for many environments and many diatom species, including *Thalassiosira* and *Minidiscus*, genomic and transcriptomic data remains scarce. Currently, complete genomes are only available for *Thalasiossira pseudonana*[11] and *Thalasiossira oceanica*[15], both isolated from the North Atlantic ocean. In addition, 65 transcriptome projects corresponding to 15 species of Thalassiosiraceae are available in the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)[16].

In this study, we present transcriptome data for ten diatoms strains, belonging to the genera *Thalassiosira* and *Minidiscus* isolated from two contrasting coastal environments, the temperate English Channel and the cold Antarctic West Peninsula. These data will enrich the sequence information from polar diatoms, as well as the functional roles of diatoms from the Thalassiosiraceae family.

Methods

Culture conditions and RNA extraction

Minidiscus spp. (RCC4590, RCC4582 and RCC4584), *Thalassiosira* spp. (RCC4219 and RCC4606) and *Thalassiosira minima* (RCC4583 and RCC4593) were isolated in January 2015 from Fildes Bay, King George Island, Western Antarctic Peninsula. *Minidiscus spinulatus* (RCC4659), *Minidiscus variabilis* (RCC4665) and *Minidiscus comicus* (RCC4660) were isolated from the English Channel (Northern Atlantic) (Table 1). All strains were obtained from the Roscoff Culture Collection (RCC, www.roscoff-culture-collection.org).

Experimental design and analysis strategy are presented in Figure 1. Cells were grown in T175 cell culture flasks using L1 growth medium[19]. Strains from the English Channel were grown at 20ºC and Antarctic strains at 4ºC. Cultures were exposed to a 12:12h light:dark photoperiod with a mean light intensity of 100 µmol photons m^{-2} s^{-1}. The growth of cultures was monitored daily using an Accuri C6 flow cytometer (Beckton Dickinson). To maximize the diversity of transcripts obtained, samples for RNA extraction were taken at four different times during the cell growth: mid-exponential at day and night and early stationary at day and night. These four samples were combined for sequencing. . Two days before extraction, cultures were treated with Penicillin, Neomycin and Streptomycin (PNS) mixture of antibiotics at 1X (Thermo Fisher Scientific) in order to decrease bacterial growth. A total of 1x10^{8} cells were filtered onto 0.8 µm polycarbonate filters (Sigma Aldrich), flash frozen in liquid nitrogen and stored at -80 ºC. The frozen filters were processed using a TRIzol - PureLinkRNA Mini Kit (Invitrogen) hybrid extraction protocol[20]. RNA samples were treated with Turbo DNA-free kit (Thermo Fisher Scientific), according to the manufacturer rigorous DNase treatment protocol for removal of genomic DNA from RNA samples. Two aliquots of 3µl were separated for quantification and quality control, and the remaining material was flash frozen and stored at -80ºC. RNA concentration was measured using the Qubit RNA BR Assay Kit in a Qubit 2.0 Fluorometer (Life Technologies). RNA integrity was evaluated using the RNA Nano 6000 Assay Kit on the Agilent Bioanalyzer 2100 system (Agilent Technologies). For check for genomic DNA contamination, we amplified the 18S rRNA gene, using the eukaryotic primers Euk 63F and 1818R[21] with the following PCR conditions: denaturation at 98° for 5 min; 25 cycles of 98° for 20 s, 52° for 30 s, 72° for 90 s; and 72° for 5 min. Two µl of each PCR product were loaded on a 1% agarose gel with 2 µL of SYBR Safe dye (Molecular Probes, Eugene) and ran at 100 V for 30 min. RNA samples with high quality, no contamination detected and a RIN value > 6, were selected for sequencing. For each diatom species, RNA material obtained from the four different times during the cell growth as described above (mid-exponential day, mid-exponential night, early stationary day and early stationary night) was merged at equimolarity. Merged samples were sent for sequencing at the INRA sequencing platform GeT-PlaGe in Toulouse, France (http://get.genotoul.fr/la-plateforme/get-plage/). Merged samples were quantified and their quality checked using the NanoDrop spectrophotometer (Thermo Scientific) and the 5400 Fragment Analyzer system (Agilent Thechnologies). Poly A selection was carried out for the selection of mRNA prior to library construction (Figure 1), using the Illumina TruSeq Stranded mRNA kit. Sequencing was performed on an Illumina HiSeq 3000 instrument to generate about 30,000,000 pair-end reads per sample (Table 1).

Quality control, digital normalization and assembly of transcriptomes

Read quality was analyzed with FastQC (v0.11.5)[22] before and after trimming. A conservative trimming approach[23] was used with Trimmomatic (version 0.33)[24] to remove residual Illumina adapters and nucleotides off the start and end of reads if they were below a given threshold Phred quality score (Q < 25).

To decrease the memory requirements for each assembly, we applied a digital normalization using the Eel Pond mRNAseq Protocol (https://github.com/dib-lab/eel-pond) with the khmer[25] software package (v2.0) prior to assembly. First, reads were interleaved, normalized to a k-mer coverage of 20 and a memory size of 4^9. Low-abundance k-mers from reads with a coverage above 18 were trimmed. Digital normalization approach with khmer uses the same algorithm implemented in Trinity, but requires less memory and accelerates the assembly stage (Figure 1).

Transcriptomes were assembled using two strategies. We tested Trinity 2.2.0[26] and rnaSPAdes 3.14.1[27] using default parameters (Table 2). We chose the optimal strategy based on the complementarity of the following quality parameters: contig N50 size, completeness based on the percentage of single-copy orthologs, and score given by Transrate (Figure 1).

de novo assembly evaluation and annotation

The coding capacity of the assemblies was evaluated first using BUSCO (Benchmarking Universal Single-Copy Orthologs) V4.1[28]. BUSCO scores are calculated based on the presence of 303 Eukarya specific genes. To further confirm the quality of produced assemblies, scores were calculated by re-mapping the input reads against assembly using Transrate V 1.0 (Figure 1). Transrate provides reference-free quality assessment for *de novo* transcriptomes with a score value based on the evaluation of chimeras, structural errors, incomplete assembly, and base errors. After the selection of the best assembly strategy, open reading frames (ORFs) were identified with default parameters using Transdecoder v5.5.0 (https://github.com/TransDecoder).

The annotation was carried out using 4 different annotation software tools: Diamond blastp option against UnipProt/SwissProt[29], hmmscan V3.3.2 (http://hmmer.org/) against Pfam-A release 33.1[30], TmHMM V2.0c[31] and SignalP V5.0[32]. Results obtained were loaded and merged using Trinotate v3.2.1 (https://github.com/Trinotate). In order to extend the annotation of predicted ORFs we incorporated annotations of orthologous genes using the eggNOG-mapper tool[33] and enriched functional annotations by incorporation of the results from the Mercator web-based annotation pipeline[34] and the dammit annotation tool (`https://github.com/dib-lab/dammit`). For each ORF, all generated annotations were compared and better annotation terms were added to the existing ones (Figure 1). For the cases with more than one hit, one gene name per contig was selected according to the lowest e-value match ($< 1e^{-05}$).

Gene expression analysis

Orthologous genes were analysed with DESeq2 R package[35]. The matrix of gene expression counts was 'regularized log' transformed and used for hierarchical clustering based on pairwise sample distances (Figure 3). The similarity of the different transcriptomes was evaluated by Principal Component Analysis (PCoA) performed with the Scikit-learn (version 0.24.0) Python library. Data visualization was carried out using the matplotlib (version 3.3.4) and seaborn (version 0.11.1) Python libraries.

Data Records

Raw sequencing reads are available at NCBI under SRA accessions SRR13846805 - SRR13846814 (BioProject PRJNA706094). In addition, assemblies, peptide translation and annotation data are available from Zenodo[36–38].

Technical Validation

RNA integrity

RNA integrity was first assessed by automated electrophoresis to evaluate OD ratios. Degradation was minimum with observed 260/280 OD ratios larger than 1.9 in all samples. Amplification of the 18S rRNA gene and visualization in agarose gel electrophoresis confirmed that no contamination due to genomic DNA was present in the RNA samples. RNA integrity number (RIN) ranged from 5.2 to 6.6 (Table 1).

Quality validation and assembly

More than 25 millions of 150 bp paired-end Illumina reads were obtained from each of the 10 cDNA libraries (Table 1). Sequences with low quality(< 25) and containing adapters were removed. Between 95,6% to 97.1% of sequences passed the evaluation and were considered of high quality for further analysis (Figure 2). The per base quality scores were high, and most sequence quality scores were > 20 (Figure 2a and 2b). The GC content is normally distributed (Figure 2c) with a mean of 45.4%. The smooth distribution observed is in general indicative of the absence of specific contaminant as adapter dimers or other bias produced during library construction.

Assemblies using Trinity and Spades contained more than 40,000 contigs each. We selected the method implemented in Trinity as the best assembly strategy applied to filtered and normalized raw reads (Table 2). Trinity produced the longest and more contiguous transcriptome assembly with the highest Transrate score. Trinity assemblies contained a high number of complete BUSCO genes. Post assembly analyses showed that between 16 and 19% of the total BUSCO genes were missing.

Distance heatmap (Figure 3) and PCA (Figure 3b) of gene profiles from the 10 Thalassiosiraceae transcriptomes revealed that they clustered primarily according to the genus (*Thalassiosira* vs. *Minidiscus*) and secondarily according to the environment from which the strains had been isolated and. Samples from Antarctic waters clustered into two separate groups: *Minidiscus* spp. from Antarctica vs. English Channel.

Annotation

Annotation of predicted ORFs was performed with 4 different methodologies (Table 3). dammit was the annotation tool with the highest percentage of annotated ORFs which results from the fact that dammit uses several reference databases: Pfam-A (version 28.0), Rfam (version 12.1), OrthoDB (version 8), and BUSCO (version 4). The percentage of annotated genes was lower for strains from the Antarctic environment, which could be due to the lack of genomic and transcriptomic data from Antarctica in the public databases compared to temperate waters.

Code availability

The commands, tools and versions used to analyse the transcriptomic data are available at: `https://github.com/MariIGM/Thalassisiraceae-transcriptomes-project-ThTSP`

References

1. Bowler, C., Vardi, A. & Allen, A. E. Oceanographic and biogeochemical insights from diatom genomes. *Annu. review marine science* **2**, 333–365, https://doi.org/10.1146/annurev-marine-120308-081051 (2010).

2. Smetacek, V. Diatoms and the ocean carbon cycle. *Protist* **150**, 25–32, https://doi.org/10.1016/S1434-4610(99)70006-4 (1999).

3. Treguer, P. *et al.* The silica balance in the world ocean: a reestimate. *Science* **268**, 375–379, https://doi.org/10.1126/science.268.5209.375 (1995).

4. Allen, A. E., Vardi, A. & Bowler, C. An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr. opinion plant biology* **9**, 264–273, https://doi.org/10.1016/j.pbi.2006.03.013 (2006).

5. Allen, A. E. *et al.* Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203–207, https://doi.org/10.1038/nature10074 (2011).

6. Brembu, T., Mühlroth, A., Alipanah, L. & Bones, A. M. The effects of phosphorus limitation on carbon metabolism in diatoms. *Philos. Transactions Royal Soc. B: Biol. Sci.* **372**, 20160406, https://doi.org/10.1098/rstb.2016.0406 (2017).

7. Medlin, L., Kooistra, W., Gersonde, R. & Wellbrock, U. Evolution of the diatoms (bacillariophyta). ii. nuclear-encoded small-subunit rrna sequence comparisons confirm a paraphyletic origin for the centric diatoms. *Mol. Biol. Evol.* **13**, 67–75, https://doi.org/10.1093/oxfordjournals.molbev.a025571 (1996).

8. Leblanc, K. *et al.* A global diatom database–abundance, biovolume and biomass in the world ocean. *Earth Syst. Sci. Data* **4**, 149–165, https://doi.org/10.5194/essd-4-149-2012 (2012).

9. Wilson, D. L., Smith Jr, W. O. & Nelson, D. M. Phytoplankton bloom dynamics of the western ross sea ice edge—i. primary productivity and species-specific production. *Deep. Sea Res. Part A. Oceanogr. Res. Pap.* **33**, 1375–1387, https://doi.org/10.1016/0198-0149(86)90041-5 (1986).

10. Tsuda, A. *et al.* A mesoscale iron enrichment in the western subarctic pacific induces a large centric diatom bloom. *Science* **300**, 958–961, https://doi.org/10.1126/science.1082000 (2003).

11. Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86, https://doi.org/10.1126/science.1101156 (2004).

12. Montsant, A. *et al.* Identification and comparative genomic analysis of signaling and regulatory components in the diatom thalassiosira pseudonana 1. *J. Phycol.* **43**, 585–604, https://doi.org/10.1111/j.1529-8817.2007.00342.x (2007).

13. Bowler, C. *et al.* The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239–244, https://doi.org/10.1038/nature07410 (2008).

14. Oudot-Le Secq, M.-P. & Green, B. R. Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Gene* **476**, 20–26, https://doi.org/10.1016/j.gene.2011.02.001 (2011).

15. Lommer, M. *et al.* Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome biology* **13**, 1–21, https://doi.org/10.1186/gb-2012-13-7-r66 (2012).

16. Keeling, P. J. *et al.* The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12**, e1001889, https://doi.org/10.1371/journal.pbio.1001889 (2014).

17. Mock, T. *et al.* Evolutionary genomics of the cold-adapted diatom fragilariopsis cylindrus. *Nature* **541**, 536–540, https://doi.org/10.1038/nature20803 (2017).

18. Tirichine, L., Rastogi, A. & Bowler, C. Recent progress in diatom genomics and epigenomics. *Curr. opinion plant biology* **36**, 46–55, https://doi.org/10.1016/j.pbi.2017.02.001 (2017).

19. Guillard, R. & Hargraves, P. *Stichochrysis immobilis* is a diatom, not a chrysophyte. *Phycologia* **32**, 234–236, https://doi.org/10.2216/i0031-8884-32-3-234.1 (1993).

20. Poong, S.-W., Lim, P.-E., Lai, J. W.-S. & Phang, S.-M. Optimization of high quality total rna isolation from the microalga, chlorella sp.(trebouxiophyceae, chlorophyta) for next-generation sequencing. *Phycol. Res.* **65**, 146–150, https://doi.org/10.1111/pre.12165 (2017).

21. Lepere, C. *et al.* Whole-genome amplification (wga) of marine photosynthetic eukaryote populations. *FEMS Microbiol. Ecol.* **76**, 513–523, https://doi.org/10.1111/j.1574-6941.2011.01072.x (2011).

22. Andrews, S. F. A quality control tool for high throughput sequence data.(2016) http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc (2016).

23. MacManes, M. D. On the optimal trimming of high-throughput mrna sequence data. *Front. genetics* **5**, 13, https://doi.org/10.3389/fgene.2014.00013 (2014).

24. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).

25. Crusoe, M. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis [version 1; peer review: 2 approved, 1 approved with reservations]. *F1000Research* **4**, https://doi.org/10.12688/f1000research.6924.1 (2015).

26. Haas, B. J. *et al.* De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nat. protocols* **8**, 1494–1512, 10.1038/nprot.2013.084 (2013).

27. Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A. D. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *GigaScience* **8**, giz100, https://doi.org/10.1093/gigascience/giz100 (2019).

28. 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, https://doi.org/10.1093/bioinformatics/btv351 (2015).

29. Consortium, U. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research* **47**, D506–D515, https://doi.org/10.1093/nar/gky1049 (2019).

30. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic acids research* **47**, D427–D432, https://doi.org/10.1093/nar/gky995 (2019).

31. Möller, S., Croning, M. D. & Apweiler, R. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**, 646–653, https://doi.org/10.1093/bioinformatics/17.7.646 (2001).

32. Armenteros, J. J. A. *et al.* Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat. biotechnology* **37**, 420–423, https://doi.org/10.1038/s41587-019-0036-z. (2019).

33. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggnog-mapper. *Mol. biology evolution* **34**, 2115–2122, https://doi.org/10.1093/molbev/msx148 (2017).

34. Schwacke, R. *et al.* Mapman4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. plant* **12**, 879–892, https://doi.org/10.1016/j.molp.2019.01.003 (2019).

35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21, https://doi.org/10.1186/s13059-014-0550-8 (2014).

36. Guajardo, M., Jimenez, V., Vaulot, D. & Trefault, N. (Assemblies) Transcriptomes from Thalassiosira and Minidiscus diatoms from English Channel and Antarctic coastal waters, https://doi.org/10.5281/zenodo.4591037 (2021). Type: dataset.

37. Guajardo, M., Jimenez, V., Vaulot, D. & Trefault, N. (Peptide translation) Transcriptomes from Thalassiosira and Minidiscus diatoms from English Channel and Antarctic coastal waters, https://doi.org/10.5281/zenodo.4596789 (2021). Type: dataset.

38. Guajardo, M., Jimenez, V., Vaulot, D. & Trefault, N. (Annotations) Transcriptomes from Thalassiosira and Minidiscus diatoms from English Channel and Antarctic coastal waters, https://doi.org/10.5281/zenodo.4609198 (2021). Type: dataset.

Acknowledgements

This work was funded by INACH DG_10-17 and FONDECYT No. 1190879 grants. MG was funded through the Agency for Research and Development (ANID)/ DOCTORADO BECAS CHILE/2019 - 21191821 fellowship. Collaboration between Chile and France was funded through ECOS-CONICYT No. C16B02 and CNRS International Research Network "Diversity, Evolution and Biotechnology of Marine Algae" (GDRI No. 0803). DV and VJ were supported by ANR contract PhytoPol (ANR-15-CE02-0007). We thank the Roscoff Collection for providing the strains used in this work and Sarah Romac for her help with RNA extraction.

Author contributions statement

MG, DV and NT designed the experiments. MG and VJ conducted the experiments. MG performed the bioinformatic analyses. MG and NT analysed the results and wrote the first draft of the paper. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables

Table 1. Description of samples for the obtention of Thalassiosiraceae transcriptomes.

Sample ID	Species	Origin	Strain	RIN	RNA concentration (ng/µl)	Raw reads * 10^6
ThTSP-01	*Minidiscus spinulatus*	English Channel	RCC4659	6.6	145.5	31,6
ThTSP-02	*Minidiscus variabilis*	English channel	RCC4665	6.8	61.69	26,1
ThTSP-03	*Minidiscus comicus*	English Channel	RCC4660	5.5	114.3	31,1
ThTSP-04	*Thalasiosira sp.*	Western Antarctic Peninsula	RCC4219	5.5	197.0	24,4
ThTSP-05	*Thalassiosira minima*	Western Antarctic Peninsula	RCC4593	5.2	143.2	43,2
ThTSP-06	*Minidiscus sp.*	Western Antarctic Peninsula	RCC4590	6.5	224.3	40,3
ThTSP-07	*Minidiscus sp.*	Western Antarctic Peninsula	RCC4582	6.0	123.4	28,6
ThTSP-08	*Thalasiosira sp.*	Western Antarctic Peninsula	RCC4606	5.4	366.8	27,3
ThTSP-09	*Thalassiosira minima*	Western Antarctic Peninsula	RCC4583	5.3	166.6	30,3
ThTSP-10	*Minidiscus sp.*	Western Antarctic Peninsula	RCC4584	6.2	93.24	36,4

Table 2. Summary of assembly statistics. C* Complete BUSCO eukarya genes F** Fragmented BUSCO eukarya genes M*** Missing BUSCO eukarya genes.

Sample ID	Assembler	Length Mbp	Number of transcript	N50	Transrate score	Transrate optimal score	C* BUSCO genes	F** BUSCO genes	M*** BUSCO genes
ThTSP-01	rnaSPAdes	63.81	44,010	2,354	0.26	0.4	156	44	55
	Trinity	114.88	65,297	2,979	0.18	0.43	192	18	45
ThTSP-02	rnaSPAdes	72.2	48,173	2,132	0.27	0.37	163	34	58
	Trinity	113.41	60715	3,048	0.6	0.2	196	14	45
ThTSP-03	rnaSPAdes	47.9	35,253	2,114	0.32	0.50	176	32	47
	Trinity	73.4	40,455	3,002	0.36	0.51	197	16	42
ThTSP-04	rnaSPAdes	50.8	31,288	2,477	0.37	0.60	179	26	50
	Trinity	90.9	44779	3,007	0.38	0.53	193	19	43
ThTSP-05	rnaSPAdes	57.5	31,161	2,717	0.35	0.47	169	31	55
	Trinity	113.4	47,415	3,403	0.02	0.11	197	16	42
ThTSP-06	rnaSPAdes	65.7	41,560	2,664	0.29	0.50	168	28	59
	Trinity	130.0	65,166	3,572	0.01	0.10	195	17	43
ThTSP-07	rnaSPAdes	63.6	38,120	2,812	0.29	0.51	169	27	59
	Trinity	119.5	59,494	3,626	0.24	0.52	193	16	46
ThTSP-08	rnaSPAdes	54.9	33,381	2,584	0.32	0.51	169	31	55
	Trinity	100.4	46,134	3,251	0.35	0.54	195	14	46
ThTSP-09	rnaSPAdes	58.1	39,246	2,513	0.31	0.55	171	27	57
	Trinity	107.6	51,331	3,265	0.40	0.58	197	17	41
ThTSP-10	rnaSPAdes	59.0	31,376	2,812	0.16	0.28	175	24	56
	Trinity	114.3	55,285	3,516	0.02	0.11	191	16	48

Table 3. Comparison of the different annotation strategies used. Values represent the percentage of predicted ORFs annotated using four different annotation tools: Trinotate, eggNOG-mapper, Mercator and dammit.

	Number of predicted ORFs	Trinotate %	eggNOG-mapper %	Mercator %	dammit %	Total number of annotated ORFs
ThTSP-01	72,214	65.8	52.4	65.8	88.5	**63,915**
ThTSP-02	68,119	70.1	54,9	70.1	89.1	**60,715**
ThTSP-03	46,249	70.7	53.4	69.2	87.4	**40,455**
ThTSP-04	53,423	68.9	55.04	68.9	83.8	**44,779**
ThTSP-05	64,811	62.6	51,4	62.6	73.1	**47,415**
ThTSP-06	78,207	60.7	48.1	60.7	83.3	**65,166**
ThTSP-07	70,763	60.7	47.7	60.8	84.1	**59,494**
ThTSP-08	57,128	66.1	54.11	66.1	80.8	**46,134**
ThTSP-09	61,472	65.9	53.4	65.9	83.5	**51,331**
ThTSP-10	67,387	61.8	48.3	61.9	82.0	**55,285**

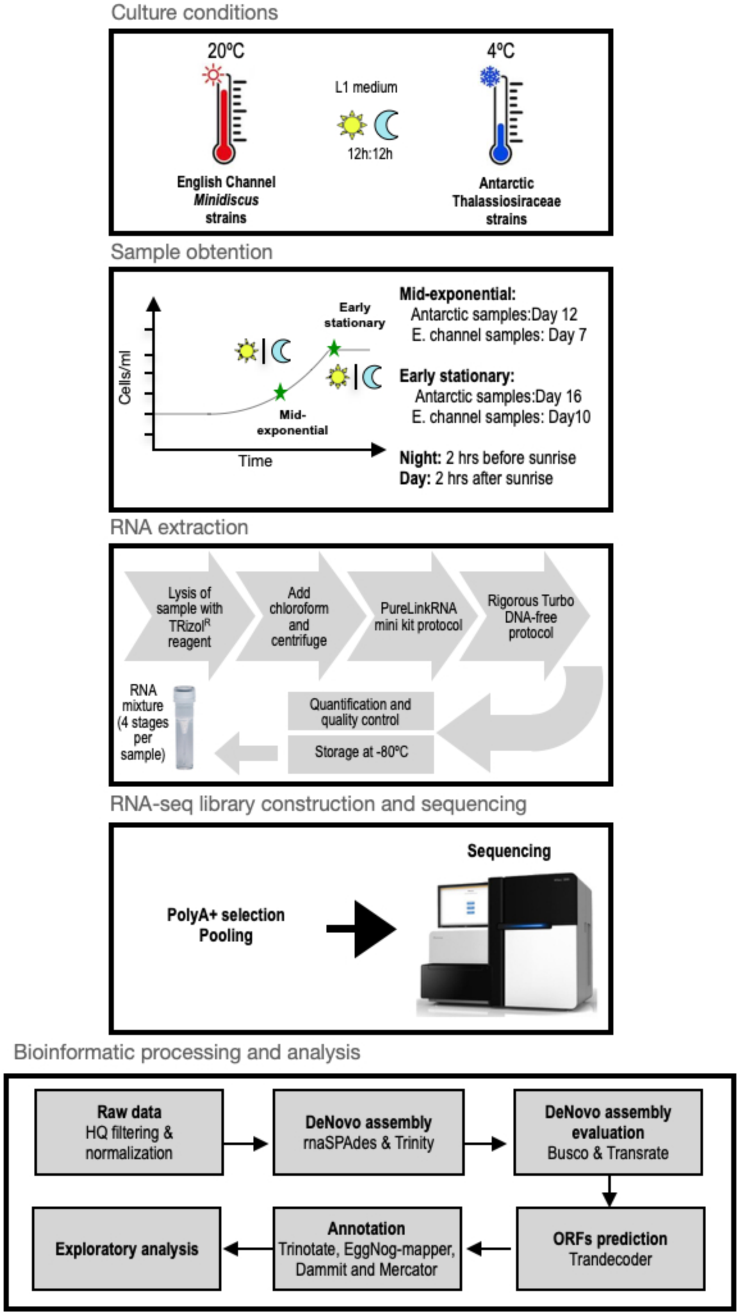
Culture conditions
20ºC
L1 medium
12h:12h
4ºC
English Channel
Minidiscus
strains
Antarctic
Thalassiosiraceae
strains
Sample obtention
Cells/ml
Time
Early
stationary
Mid-
exponential
Mid-exponential:
Antarctic samples:Day 12
E. channel samples: Day 7
Early stationary:
Antarctic samples:Day 16
E. channel samples: Day10
Night: 2 hrs before sunrise
Day: 2 hrs after sunrise
RNA extraction
Lysis of sample with TRizol^R reagent
Add chloroform and centrifuge
PureLinkRNA mini kit protocol
Rigorous Turbo DNA-free protocol
Quantification and quality control
Storage at -80ºC
RNA mixture (4 stages per sample)
RNA-seq library construction and sequencing
PolyA+ selection
Pooling
Sequencing
Bioinformatic processing and analysis
Raw data HQ filtering & normalization
DeNovo assembly rnaSPAdes & Trinity
DeNovo assembly evaluation Busco & Transrate
ORFs prediction Trandecoder
Annotation Trinotate, EggNog-mapper, Dammit and Mercator
Exploratory analysis

Figure 1. Flowchart of the experimental design and data processing pipeline for the construction and analyses of transcriptomes from Thalassiosiraceae strains. Cultures were exposed to normal growth conditions. To increase the gene repertoire obtained, RNA was extracted at four different times of the growth curve. Two distinct assembly methods were tested and the better assembly according to quality parameters was selected.

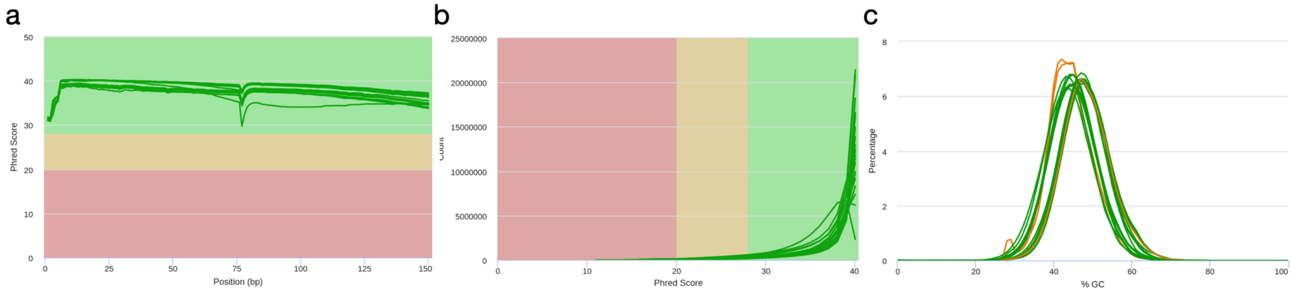

Figure 2. Quality parameters after the RNA sequencing procedure. (a) Mean quality scores for all samples. (b) Per sequence quality scores. (c) GC content.

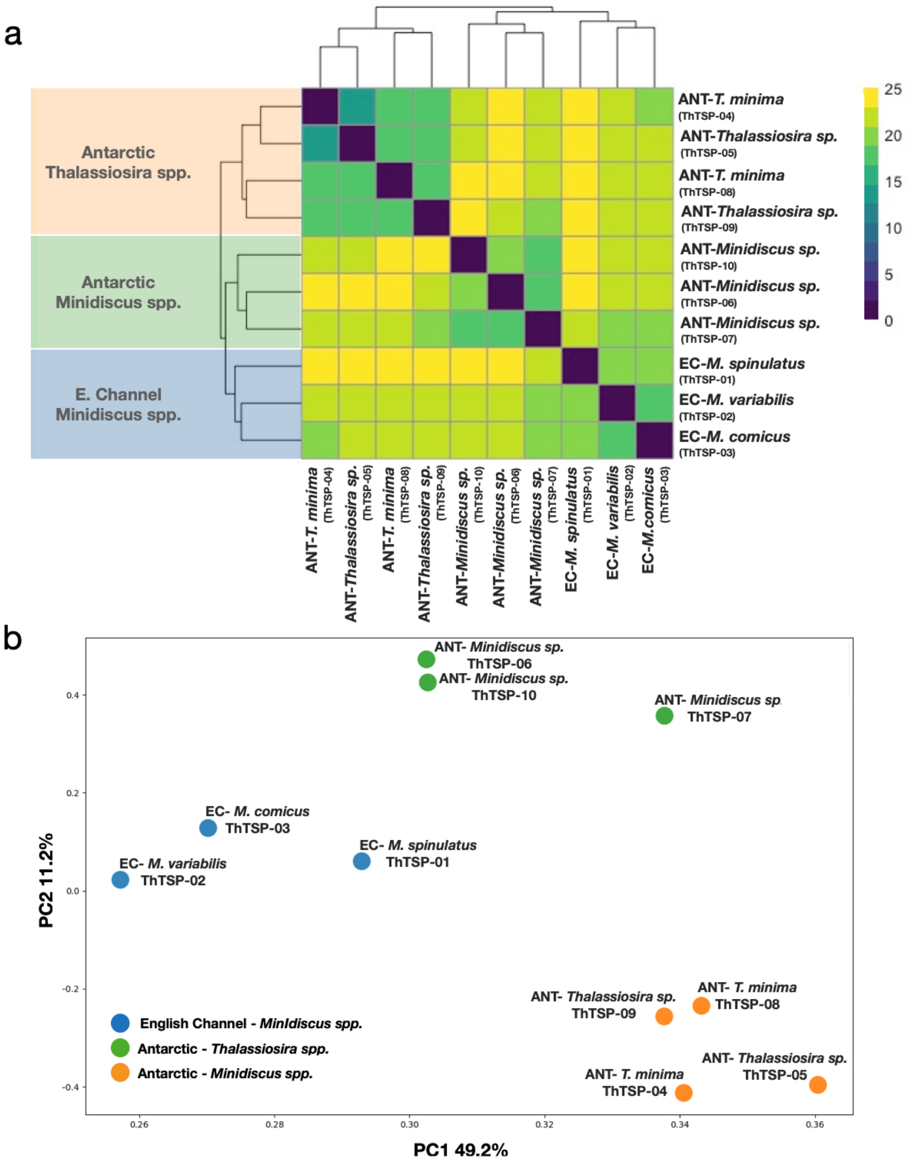

Figure 3. Analysis of transcriptome similarity. a) Hierarchical clustering analysis and heatmap based on euclidean distances between normalized gene count. b) Principal component analysis (PCA) between transcriptomes using normalized gene count. EC, English Channel and ANT, Antarctica.