Supporting Information

Article title

Resolving the microalgal gene landscape at the strain level: A novel hybrid transcriptome of *Emiliania huxleyi* CCMP3266

Access DOI:

https://doi.org/10.1128/AEM.01418-21

Published in:

Applied and Environmental Microbiology

Authors & Affiliations

Martin Sperfeld,[a]# Dayana Yahalomi,[b] Einat Segev[a]#

[a]Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot, Israel

[b]Nancy and Stephen Grand Israel National Center for Personalized Medicine, Weizmann Institute of Science, Rehovot, Israel

#Address correspondence to Martin Sperfeld, martin.sperfeld@weizmann.ac.il and Einat Segev, Einat.Segev@weizmann.ac.il

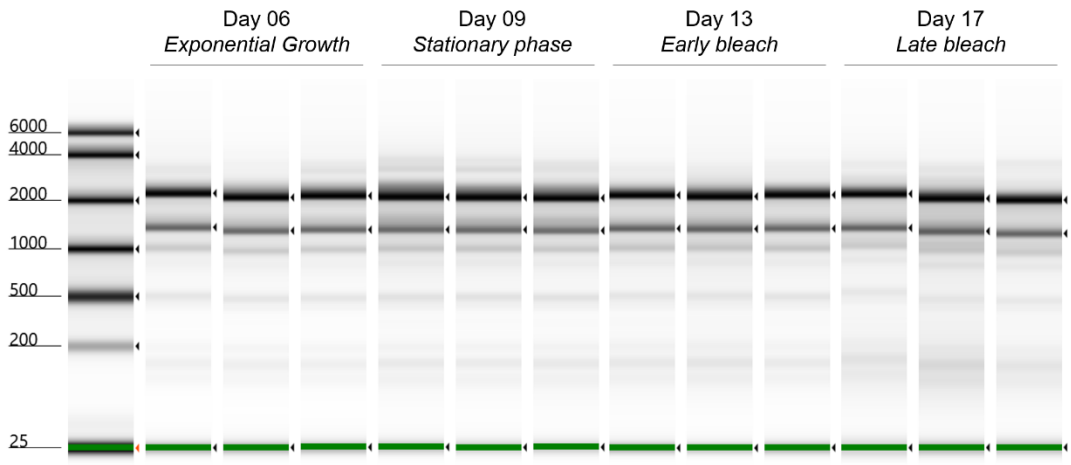
The following Supporting Information is available for this article:

Figures

Tables

Data

(a)



(b)

TapeStation RIN	8.5	8.4	8.7	7.9	8.0	8.1	8.1	7.9	8.1	8.1	7.3	7.9
TapeStation [ng/ul]	134	136	148	217	235	242	156	235	156	192	188	163
Qubit rep. 1 [ng/ul]	187	162	192	280	276	336	212	296	204	279	289	264
Qubit rep. 2 [ng/ul]	183	158	190	276	268	324	208	284	200	276	284	252
Nanodrop [ng/ul]	204	187	199	307	309	342	210	291	210	298	297	249
Nanodrop 260/280nm	2.06	2.08	2.07	2.05	2.07	2.06	2.05	2.07	2.08	2.03	2.04	2.04
Nanodrop 260/230nm	2.10	2.30	2.08	2.34	2.35	2.28	2.42	2.10	2.24	2.36	2.19	2.22

Fig. S1 Quality control of 12 total RNA extracts used for PacBio and Illumina library preparation. (a) Electronic gel image of RNA extracts generated with the TapeStation 4150 instrument. Intensities were scaled to samples. The sharp upper band (28S rRNA) is roughly two times more intense than the sharp lower band (18S rRNA), thus confirming the integrity of the RNA extracts. Additional faint bands are possibly derived from ribosomal RNA of *E. huxleyi* organelles. (b) RNA quality metrics confirm the integrity and purity of the RNA samples. The average RIN value (RNA integrity number; calculated from rRNA band intensity) was 8.1 ± 0.4. RNA concentrations measured with Qubit were 180 (± 16) ng/µl, 297 (± 34) ng/µl, 237 (± 51) ng/µl and 277 (± 13) ng/µl for extracts from Day 06, Day 09, Day 13 and Day 17, respectively. The average $260_{nm}/280_{nm}$ and $260_{nm}/230_{nm}$ absorption ratios were 2.06 (± 0.02) and 2.25 (± 0.11), respectively.

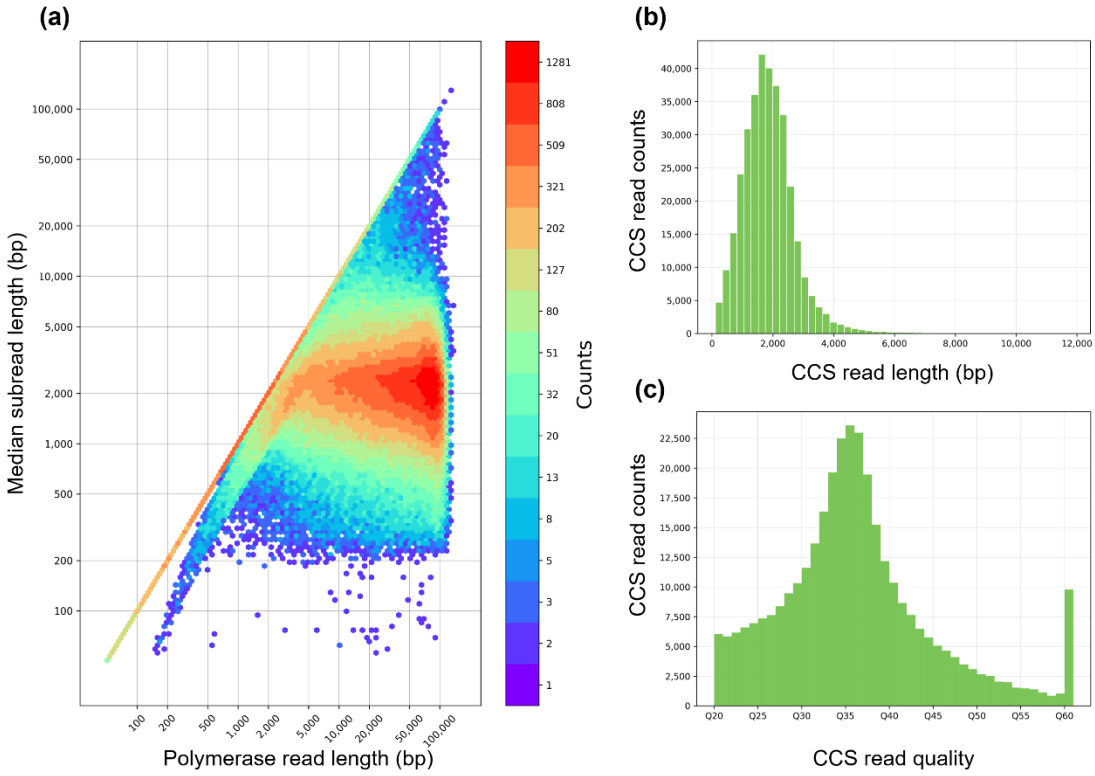


Fig. S2 Diagnostic plots of PacBio polymerase and CCS reads reflecting library quality. The PacBio Sequel I system utilizes a polymerase that replicates circular cDNA templates for several passes, generating long, contiguous molecules (= polymerase read) with multiple subreads of the insert of interest. (a) Relation of polymerase read length and median subread length is depicted as density plot. A horizontally distributed high-density area with a shift to the right side of the x-axis shows multiple replications of the insert of interest per polymerase read. A median subread length of around 2,500 bp is in accordance with a library peak at 2,565 bp (Fig. S3) (b) Polymerase reads were further processed by removing adapters and collapsing subreads into circular consensus sequencing reads (CCS, ≥ Q20). CCS read length reflects the length of the investigated cDNA molecule. The majority of CCS reads were evenly distributed at around 2,000 bp. (c) Distribution of CCS read quality.

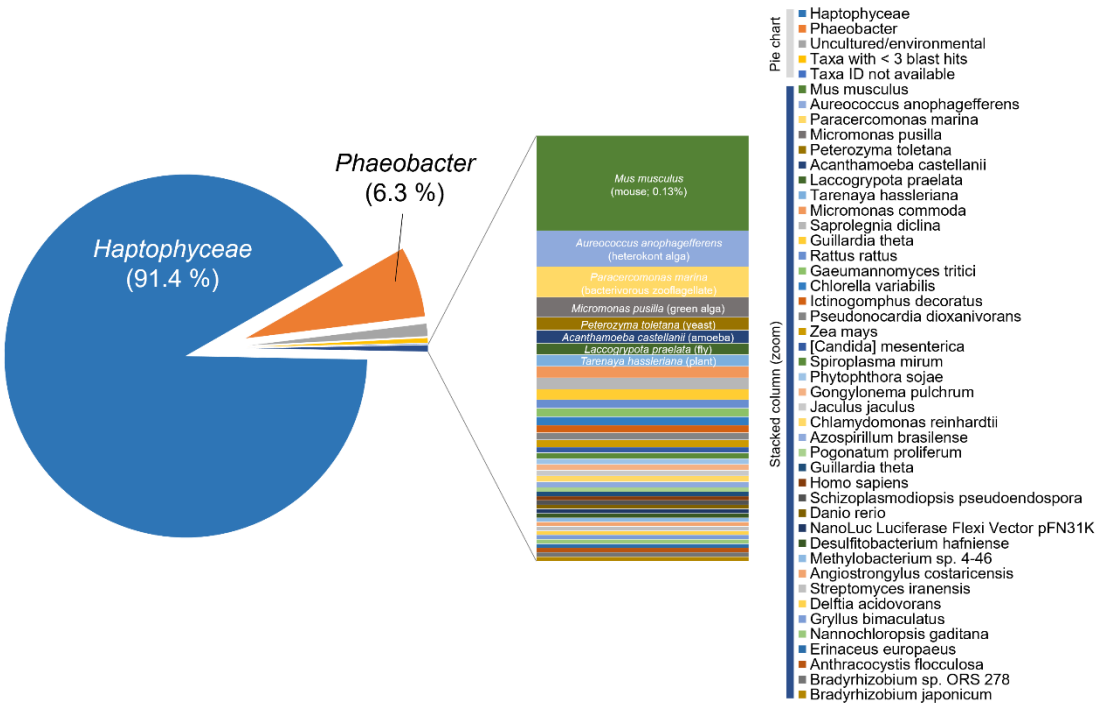


Figure S3 Screen for contaminations in the hybrid transcriptome assembly. The contigs of the unfiltered rnaSPAdes assembly output were subjected to a blastn search (nr/nt database; -evalue 1e-5). The majority of the contigs had either similarity to *Haptophyceae* (*E. huxleyi* family name; 91.4%) or to members of the bacterial genus *Phaeobacter* (6.3%). Other contigs could not be assigned to specific taxa (grey: uncultured/environmental; yellow: ≤ 3 blast hits with single taxon; dark blue: Tax ID not available), or were assigned to taxa that can be ruled out as contaminants (e.g. 0.13% of transcripts had similarity with mouse). Noteworthy, among a total of 69,353 contigs, 15,322 had no significant blastn hit. Contigs with no significant blastn hit mapped largely to non-coding regions of the CCMP1516 genome assembly (68%; data not shown). Therefore, we refrained from removing unassigned contigs from the hybrid transcriptome assembly, but filtered out contigs that mapped to the genome of *P. inhibens* DSM17395.

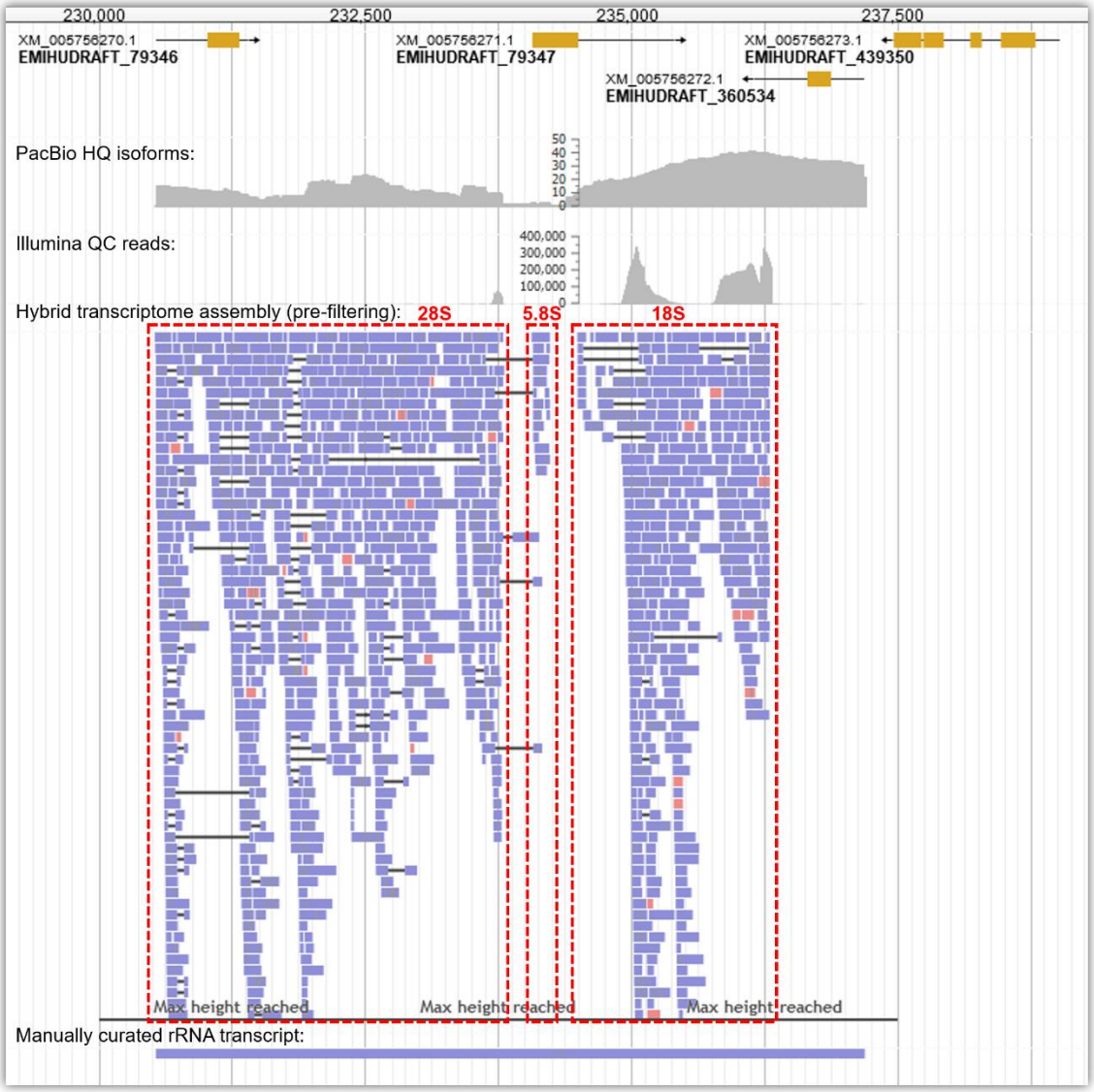


Fig. S4 Curation of rRNA fragments from the hybrid transcriptome of *E. huxleyi* CCMP3266. The upper track (yellow bars) shows a selected rRNA locus within the *E. huxleyi* CCMP1516 reference genome (scaffold NW_005194754.1). *E. huxleyi* CCMP3266 PacBio HQ isoforms, Illumina QC reads, and contigs of the hybrid transcriptome assembly (pre-filtering) were mapped to the reference genome (gray shading and purple bars). Visualization of the mapping reveals erroneously assembled rRNA fragments in the hybrid transcriptome assembly (red, dashed boxes). Erroneously assembled rRNA fragments were replaced by three manually curated, nucleus-, plastid and mitochondrion-encoded rRNA consensus sequences (Table S3).

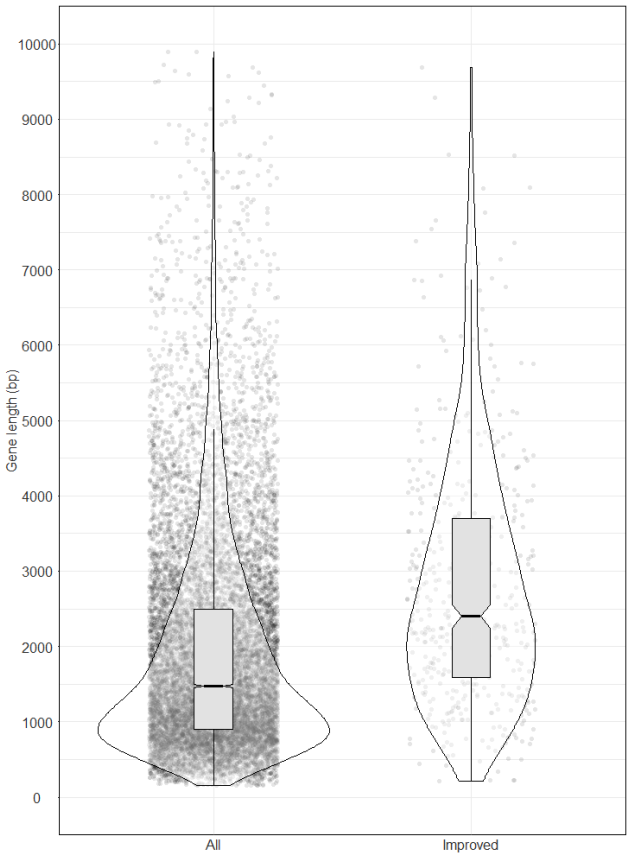


Fig. S5 The hybrid transcriptome approach improved the reconstruction of transcripts from long *E. huxleyi* reference genes (> 5% higher coverage). Single dots represent the length of all 13,168 *E. huxleyi* reference genes (left), compared to the length of 480 *E. huxleyi* genes with improved transcript length (right). Box plots show median gene lengths (centered notch), including lower and upper quartiles (hinges). Violin plots show gene lengths distribution. The median length of all *E. huxleyi* reference genes is 1,488 bp, while genes with improved transcript length have a median length of 2,424 bp.

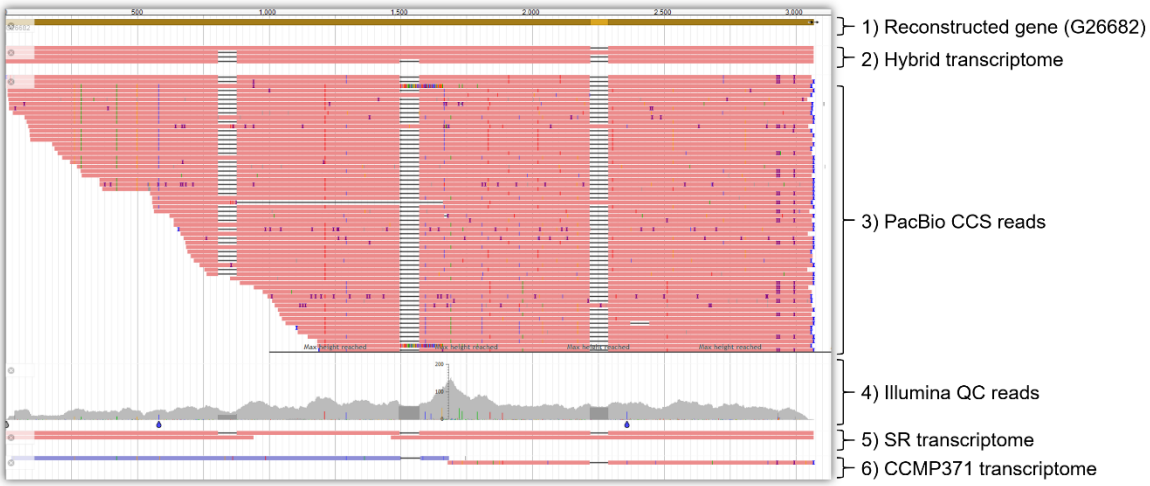


Fig. S6 Example for *E. huxleyi* CCMP3266 gene reconstruction. Visualized is a CCMP3266 gene sequence (track 1; gene locus G26682) that was reconstructed by COGENT from four full-length transcript isoforms of the hybrid transcriptome (track 2). The reconstructed CCMP3266 gene is supported by PacBio CCS reads (track 3) and Illumina QC reads (track 4), and has three introns, as indicated by thin black lines (PacBio CCS reads) and grey dark areas (Illumina QC reads). The control SR transcriptome (track 5), which was entirely assembled from Illumina QC reads, contains only a single full-length transcript. No full-length transcripts are present in the control CCMP371 transcriptome (track 6). Track 2, 3, 5 and 6 show individual sequences (light red: +strand; light blue: -strand). Track 4 is depicted as coverage track (light gray) with numbers on the scale indicating how often a read mapped to a specific position. The visualization was done with JBrowse (Buels *et al*., 2016; PubMed ID: 27072794), using the *E. huxleyi* CCMP3266 sGenome as reference (Data Set S4). Transcript IDs and nucleotide sequences can be retrieved for gene G26682 from Data Set S2 (column 2) and Data Set S1, respectively.

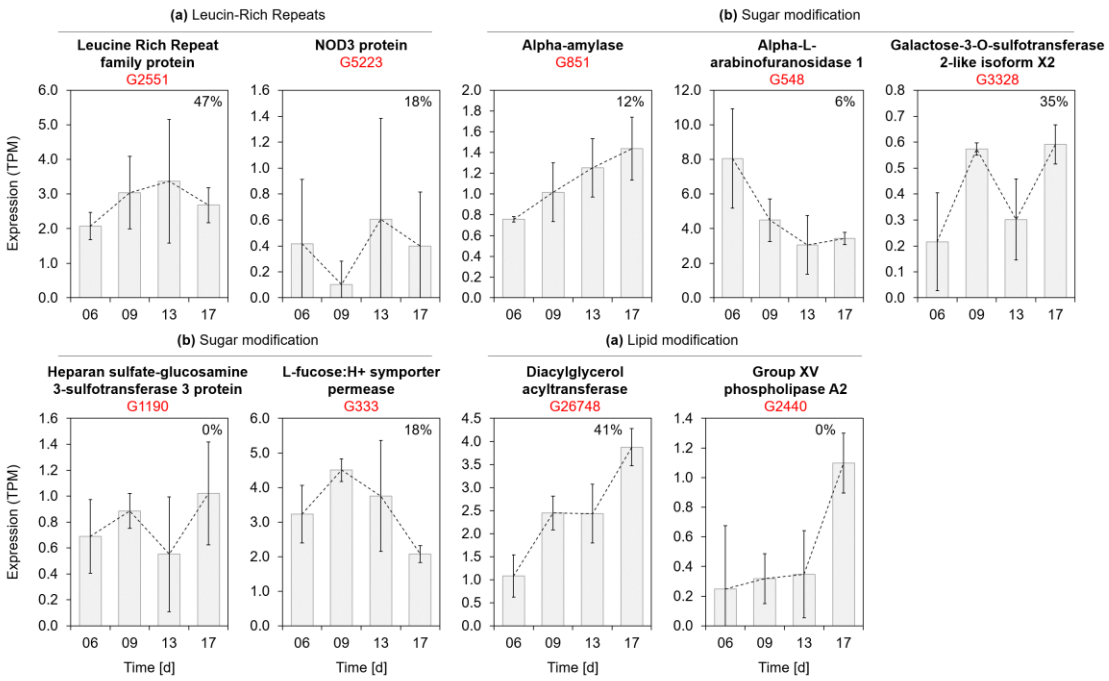


Figure S7 Expression profiles of variable *E. huxleyi* CCMP3266 genes putatively involved in microbial interactions. The depicted genes were absent from the CCMP1516 reference genome, and variably detected in 17 other *E. huxleyi* transcriptomes. Percentage values indicate the number of *E. huxleyi* transcriptomes in which the genes were detected (Data Set S6, column 7). Expression is shown as transcripts per million (TPM), with error bars indicating standard deviations of biological triplicates. Transcript IDs and nucleotide sequences can be retrieved from Data Set S6 (column 2) and Data S1, respectively.

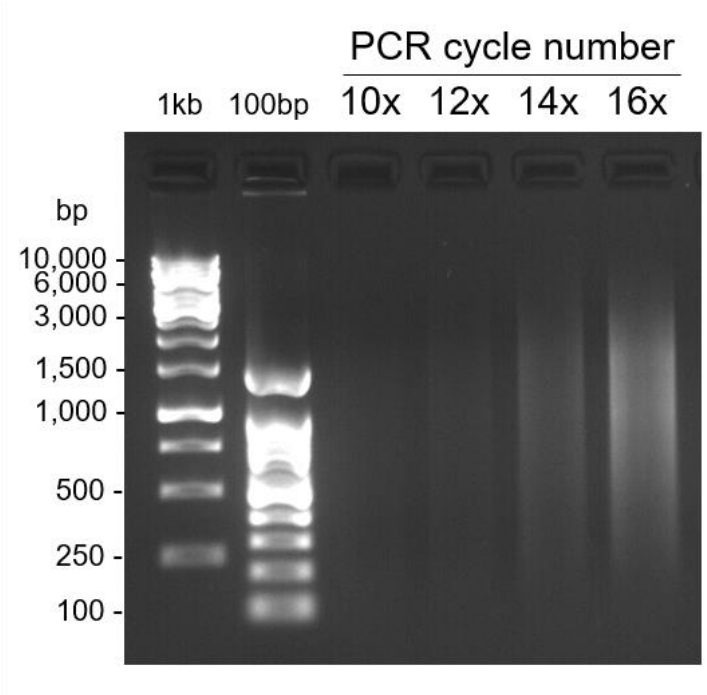


Fig. S8 Optimization of PCR cycle number for cDNA amplification, as part of the PacBio Iso-Seq library preparation protocol. The agarose gel shows homogenous replication of cDNA in an expected size range of ≈ 250 - 3,000 bp. A PCR cycle number of 12 was chosen for large-scale cDNA production to minimize over-amplification biases. Amplified cDNA was controlled on a 1.5% TAE agarose gel (100 V, 40 min), using 5 µl PCR product and 1 µl loading dye. As size marker, 2.5 µg of Thermo Scientific GeneRuler 1 kb DNA ladder and 2.5 µg of a GeneDireX 100 bp DNA ladder were used.

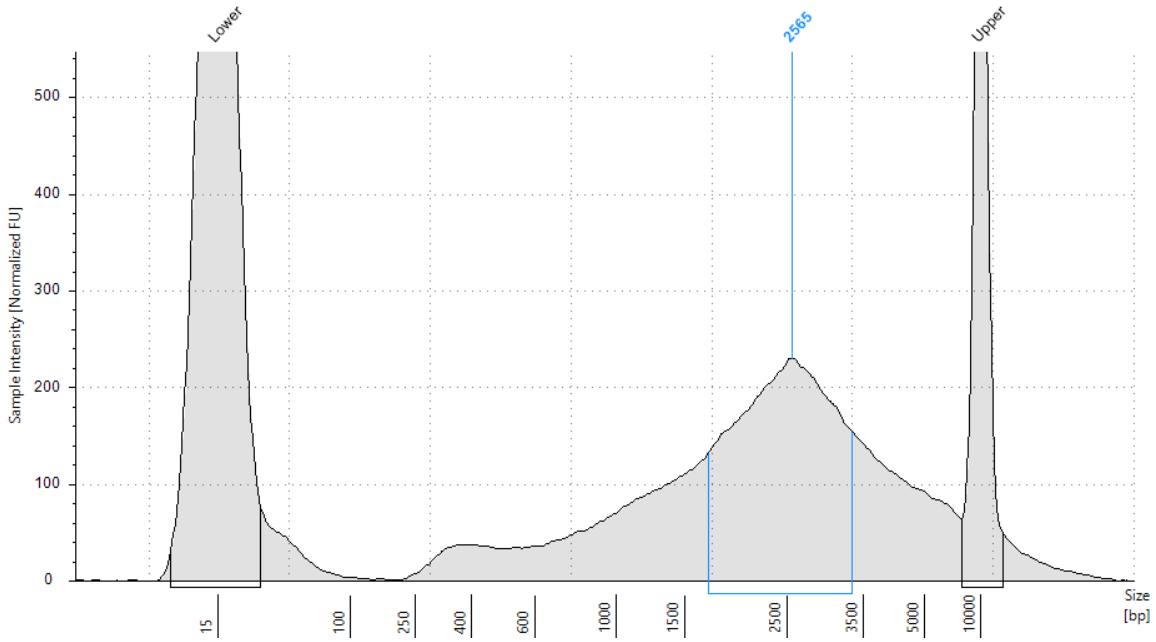


Fig. S9 Electropherogram of PacBio SMRTbell template sequencing library. The library had a concentration of 6.3 ng/µl and exhibited a peak at 2,565 bp (blue). A prominent lower peak (left), and an upper peak (right) are internal electronic standards. Data was generated using a TapeStation 4200 instrument with a D5000 HS ScreenTape.

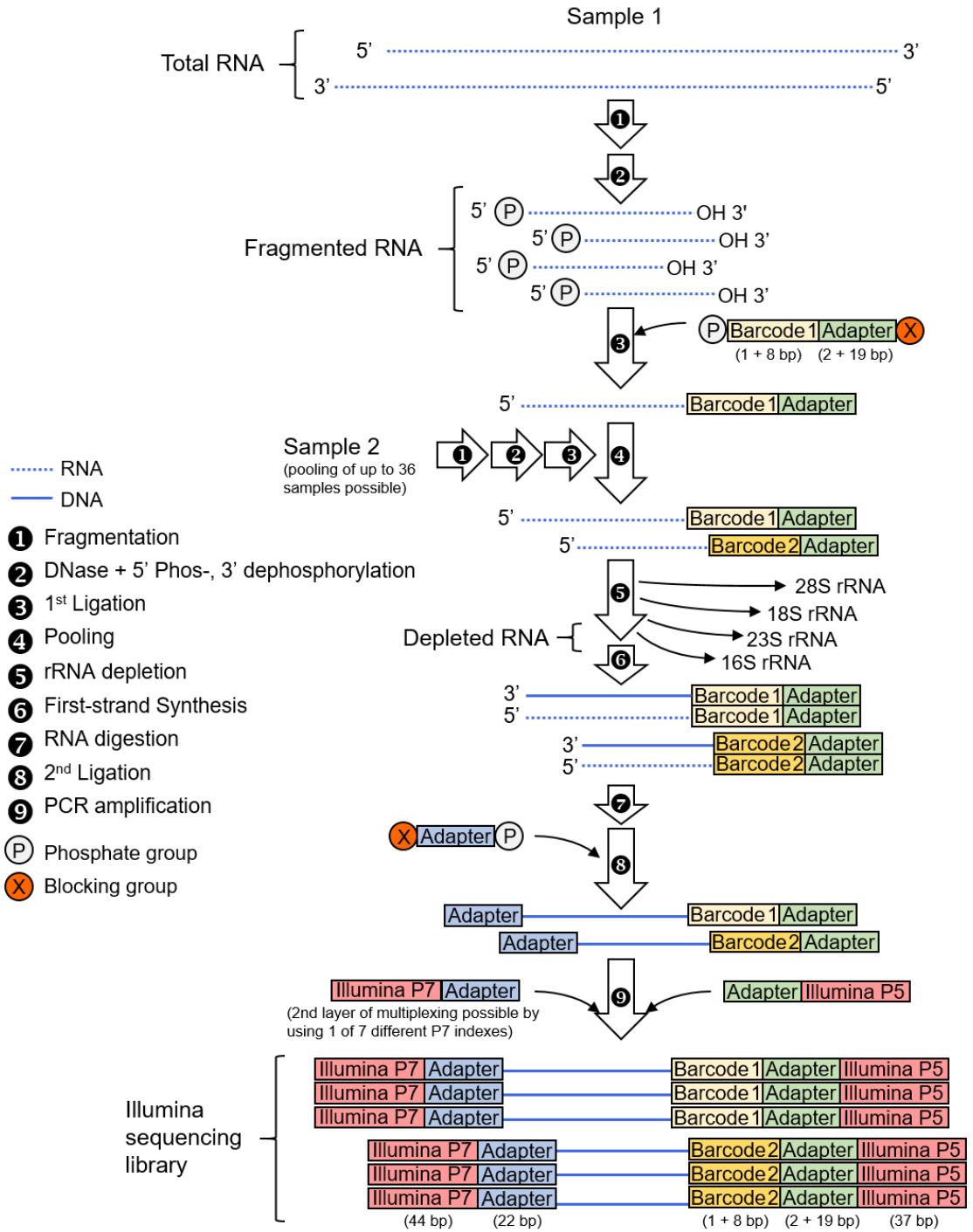


Fig. S10 Scheme of stranded, total RNA sequencing Illumina library preparation protocol. The protocol was developed by Avraham *et al*. 2016 (PubMed ID: 27442864), and the scheme was adapted from materials given therein.

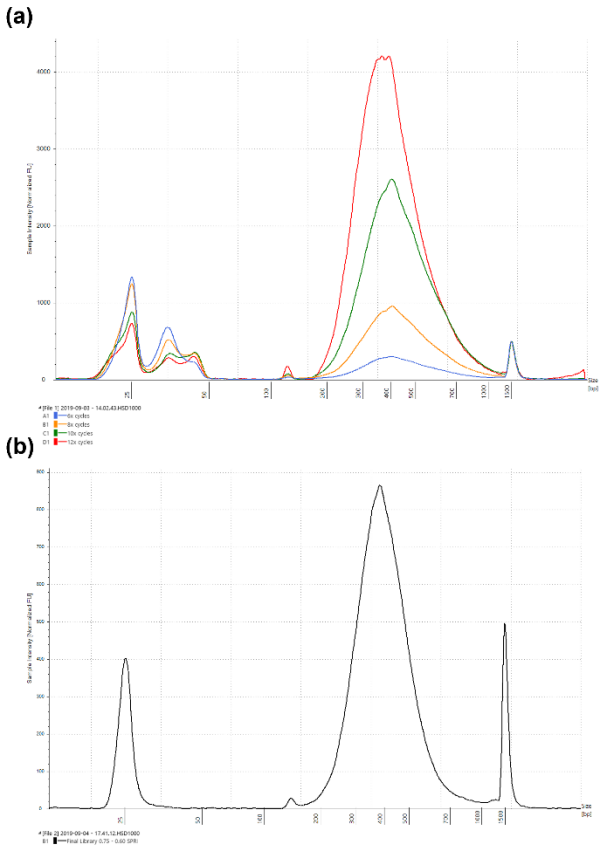


Fig. S11 Optimization of PCR cycle numbers during Illumina library preparation to minimize cDNA over-amplification. (a) Electropherogram of PCR products generated with 6 (blue), 8 (orange), 10 (green) and 12 (red) PCR amplification cycles. A cycle number of 8 was chosen for final library preparation. The upper internal electronic standard forms a peak of 1,500 bp (right). The lower internal standard (25 bp, left) is overlaid by PCR primers of the same size. (b) The final library was purified using 0.6x and 0.75x RNAClean XP SPRI beads for right- and left-sided clean-up, respectively. The clean-up removed PCR primers and narrowed the final sequencing library to a size between ≈ 200 and ≈ 700 bp, with a peak at 383 bp. Data were generated using a TapeStation 4150 instrument with a D1000 HS ScreenTape.

Table S1 Metrics summary of PacBio Iso-Seq sequencing output, CCS read generation and CCS read classification. Values were taken from the SMRT Link software (v8.0). Productivity values indicate the number of ZMW sequencing wells that were empty (P0), that contained a single template (P1), or that harbored multiple templates (P2). Only wells with a single template were processed. P1 values > 33% indicate a productive sequencing run. The PacBio company specifies a minimum CCS read quality of Q30 (> 99.9 % accuracy).

Raw data report		
Productivity (%)	P0	25.1
	P1	50.0
	P2	24.8
Polymerase reads (total bases)		19,107,415,731
Polymerase reads (counts)		490,061
Polymerase read length (mean)		38,990
Polymerase read length (N50)		68,738
Insert length (mean)		2,505
Insert length (N50)		2,719
Subread length (mean)		1,486
Subread length (N50)		1,891
Unique molecular yield		1,034,664,493
CCS Report (≥ Q20)		
Yield (bp)		622,651,701
CCS reads		336,459
CCS read length (mean, bp)		1,850
CCS read quality (median)		Q35
CCS Read classification		
Reads with 5’ and 3’ primers		308,251
Non-concatamer reads with 5’ and 3’ primers		306,154
Non-concatamer reads with 5’ and 3’ primers and poly(A) tail (= FLNC reads)		306,061
Mean length of FLNC reads (bp)		1,720

Table S2 Summary of Illumina sequencing output. Illumina sequencing was conducted using a NextSeq 500 instrument in paired-end mode (FW read 1: 88 bp; RV read 2: 78 bp), and a four-lane high output v2.5 sequencing cassette (150 cycles). Density (K/mm2): density of bridge-amplified clusters of DNA fragments in thousands per mm^2 (Illumina specified optimal density: 170 - 220 K/mm^2); Cluster PF (%): relative number of clusters passing the internal Illumina filter, which removes overlapping clusters; Reads (M): total amount of sequencing reads in million; Reads PF (M): reads passing the internal Illumina filter (Illumina specified max. output per lane: 100 M); % > Q30: percentage of reads with average Q-score of 30 or higher (Q30 = probability of 1 in 1000 incorrect bases); Yield total (Gb): reads PF multiplied by read length; % aligned to PhiX: percentage of the reads aligned to the PhiX genome (1% PhiX control DNA was added to the sequencing library); Error rate: percentage of incorrectly called bases (calculated from reads that aligned to the PhiX genome).

	Density (K/mm^2)	Cluster PF (%)	Reads (M)	Reads PF (M)	% > Q30	Yield Total (Gb)	% aligned to PhiX	Error rate
FW read1								
Lane 1	184.5	85.0	119.7	101.7	93.3	8.8	1.003	0.30
Lane 2	182.9	85.1	118.6	101.1	93.3	8.8	1.013	0.32
Lane 3	184.7	85.3	119.8	102.2	93.3	8.9	1.000	0.32
Lane 4	182.9	85.5	118.6	101.4	93.6	8.8	0.998	0.31
RV read2								
Lane 1	184.5	85.0	119.7	101.7	90.7	7.8	0.979	0.31
Lane 2	182.9	85.1	118.6	101.0	90.4	7.8	0.986	0.31
Lane 3	184.7	85.3	119.8	102.2	90.8	7.9	0.977	0.31
Lane 4	182.9	85.5	118.6	101.4	90.6	7.8	0.972	0.31

Table S3 Nucleotide sequences of three manually curated, nucleus-, plastid- and mitochondrion-encoded *E. huxleyi* CCMP3266 rRNA sequences. Each rRNA sequence contains multiple, neighboring ribosomal subunits (marked in colors). The three manually curated rRNA sequences were constructed to replace erroneously assembled rRNA fragments in the hybrid transcriptome (Data Set S1).

Organelle	Position of rRNA-encoding loci in the genome of *E. huxleyi* CCMP1516	Curated rRNA sequences of *E. huxleyi* CCMP3266
Nucleus	18S: NW_005194754.1:234,501..236,303* 5.8S: NW_005194754.1:234,083..234,243* 28S: NW_005194754.1:230,532..233,788*	>rRNA_nucleus_18S_5S_28S_curated [illegible]
Plastid	16S: JN022705.1:54..1,536 (copy 1) JN022705.1:19,326..20,808 (copy 2) 23S: JN022705.1:1,814..4,690 (copy 1) JN022705.1:16,172..19,048 (copy 2) 5S: JN022705.1:4,722..4,839 (copy 1) JN022705.1:16,023..16,140 (copy 2)	>rRNA_plastid_16S_23S_5S_curated [illegible]
Mitochondrion	23S: JN022704.1:44..2,749 16S: JN022704.1:2,793..4,366	>rRNA_Mitochondrion_23S_16S_curated [illegible]

*The genome assembly of *E. huxleyi* CCMP1516 contains several copies of the nucleus-encoded 18S, 5.8S and 28S rRNA gene. Position NW_005194754.1:230,532..236,303 is visualized in Fig. S7.

Data Set S1 Hybrid transcriptome of *E. huxleyi* CCMP3266 (FASTA format). The FASTA header matches the "TransID.SPAdes", "GeneID" and "TransID.TSA" columns of Data Set S2. Data Set S1 can be downloaded from Zenodo: https://zenodo.org/record/5702921

Data Set S2 *E. huxleyi* CCMP3266 hybrid transcriptome annotation table (tsv format). Column 1 - 4: CCMP3266 gene and transcript IDs; column 5 - 6: gene and transcript length; column 7: longest transcript per gene; column 8: transcripts with protein-coding ORF; column 9 - 10: Illumina short-read counts; column 11 - 13: results of differential gene expression analysis; column 14: PacBio CCS long-read counts; column 15 - 27: blastx/blast2GO functional annotations. Data Set S2 can be downloaded from Zenodo: https://zenodo.org/record/5702921

Data Set S3 *E. huxleyi* CCMP1516 reference genes used for transcriptome completeness estimates (FASTA format). The FASTA file contains nucleotide sequences of *E. huxleyi* CCMP1516 core genes supported by expressed sequence tags (ESTs). The set of genes was compiled from data given by (Read *et al*., 2013; PubMed ID: 23760476). Data Set S3 can be downloaded from Zenodo: https://zenodo.org/record/5702921

Data Set S4 *E. huxleyi* CCMP3266 sGenome (FASTA format). Data Set S4 can be downloaded from Zenodo: https://zenodo.org/record/5702921

Data Set S5 *E. huxleyi* CCMP3266 sGenome gene annotation file (GFF3 format). Data Set S5 can be downloaded from Zenodo: https://zenodo.org/record/5702921

Data S6 *E. huxleyi* CCMP3266 novel genes that were absent from the CCMP1516 reference genome (tsv format). Column 1 (GeneID) and column 2 (TransID.SPAdes) include CCMP3266 gene and transcript identifiers, which can be used to retrieve nucleotide sequences from the hybrid transcriptome (Data Set S1). Column 3 - 4: gene and transcript length; column 5: gene expression levels determined by mapping Illumina QC reads to the sGenome (RPK normalized); column 6: gene expression levels determined by mapping PacBio CCS reads to

the sGenome (read counts); column 7: number of publically available *E. huxleyi* transcriptomes (n = 17; available at TSA database) that produced a significant BLAT hit; column 8 - 15: blastx/blast2GO functional annotations. Data Set S6 can be downloaded from Zenodo: https://zenodo.org/record/5702921