Supporting Information

Beauchemin et al. 10.1073/pnas.1206683109

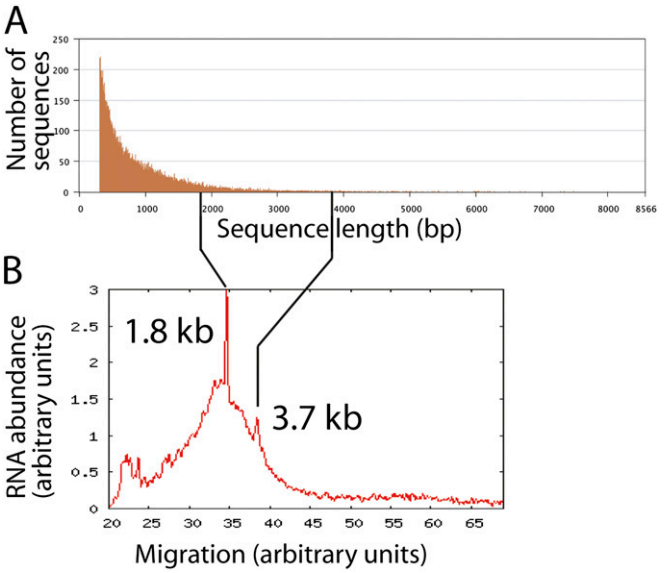

Fig. S1. Size distribution of sequences in the transcriptome and in the mRNA. (*A*) The size distribution plotted as a histogram of number of sequences for each contig length in the 74,655-sequence transcriptome. (*B*) The size distribution of the RNA sample used for sequencing as determined using a Bioanalyzer. The sizes of the two peaks of rRNA still visible in the analysis are shown.

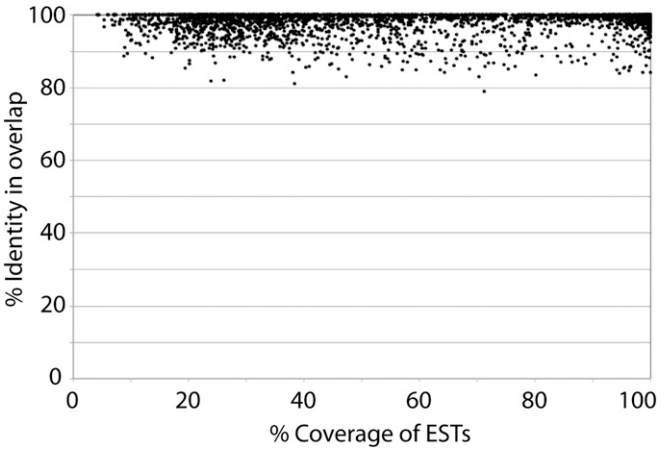

Fig. S2. Degree of sequence identity of *Lingulodinium* ESTs with the transcriptome. The degree of sequence identity as a function of the proportion of the EST sequence covered is shown by a comparison of the transcriptome sequences with 2,111 GC-rich *Lingulodinium* ESTs in GenBank. Each point represents a Sanger EST with a corresponding sequence in the transcriptome. Because of the short average length of the transcriptome sequences, there are many ESTs that are incompletely covered by the transcriptome contigs, and several ESTs have matches with more than one contig.

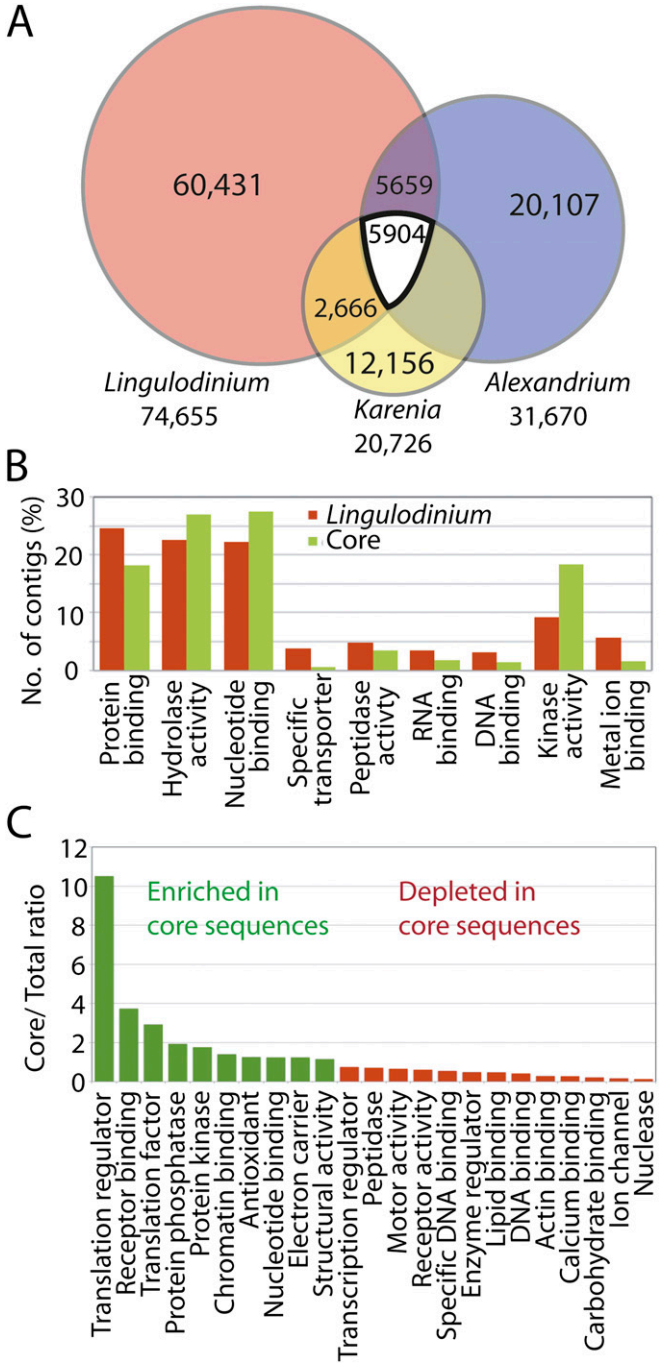

Fig. S3. Characterization of sequences common to *Lingulodinium*, *Alexandrium*, and *Karenia*. (*A*) The transcriptome was mapped onto *Alexandrium* and *Karenia* EST unigenes using tBLASTn at a cutoff of e^{-20}. (*B*) Compared with the full *Lingulodinium* transcriptome, the "core" dinoflagellate sequences are enriched in the kinase activity category by gene ontology classification. (*C*) The ratio between the number of sequences in the core dinoflagellate dataset and the full transcriptome was calculated for each gene ontology molecular function classification. Groups found enriched in the core dataset are shown in green and include sequences involved in translation and posttranslation control.

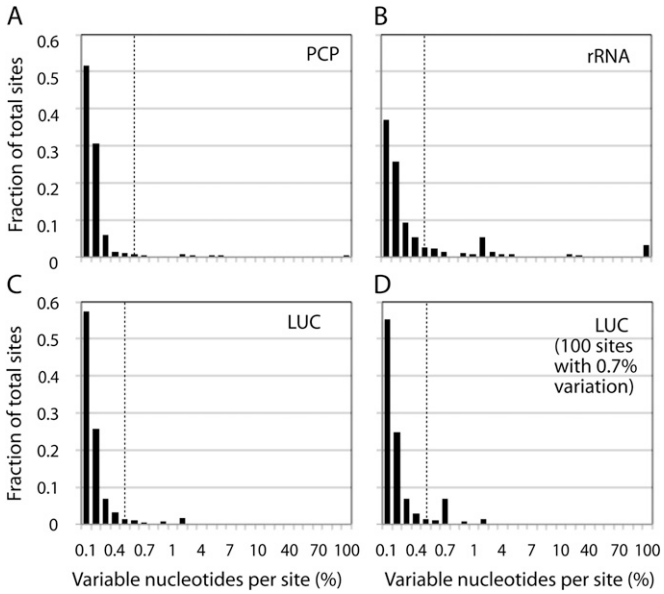

Fig. S4. Frequency spectrum of sequence variation in peridinin-chlorophyll a-protein (PCP) and luciferase (LUC) tandem array gene transcripts. The sequences of PCP (*A*), rRNA (*B*), and luciferase (*C*) were first used to align reads trimmed to remove low-quality or ambiguous bases from a dataset containing 89×10^6 reads. A plot of this data directly displays the nucleotide variation at each position (as in Fig. 2*A* for PCP). These data are then used to determine the frequency spectrum of polymorphic variation by counting the number of times each percent nucleotide variation is observed over the sequence. Bin sizes are 0.1% up to 1%, 1% from 1% to 10%, 10% from 10% up to 100%, and the data are reported as the fraction of total sites with a given level of variation. The dotted vertical line represents the same 0.5% variation shown in Fig. 2*A*, and the high number of sites with a sequence variation of up to 0.1% reflects, at least in part, an error in the sequencing reaction (a Q30 corresponds to an accuracy of 99.9%). The peak at 1% variation in *B* reflects an increased bin size (from 0.1 to 1%). (*D*) A hypothetical frequency spectrum of polymorphic variation in luciferase gene transcripts was constructed by arbitrarily adding 100 nucleotide positions with a 0.7% variation to the data before binning. The predicted peak at 0.7% variation observed in the frequency spectrum corresponds to what would be expected if one of the 146 gene copies differed from the others at 100 of the 4,000 positions in the gene sequence and transcripts from all gene copies accumulated equally in the transcriptome.

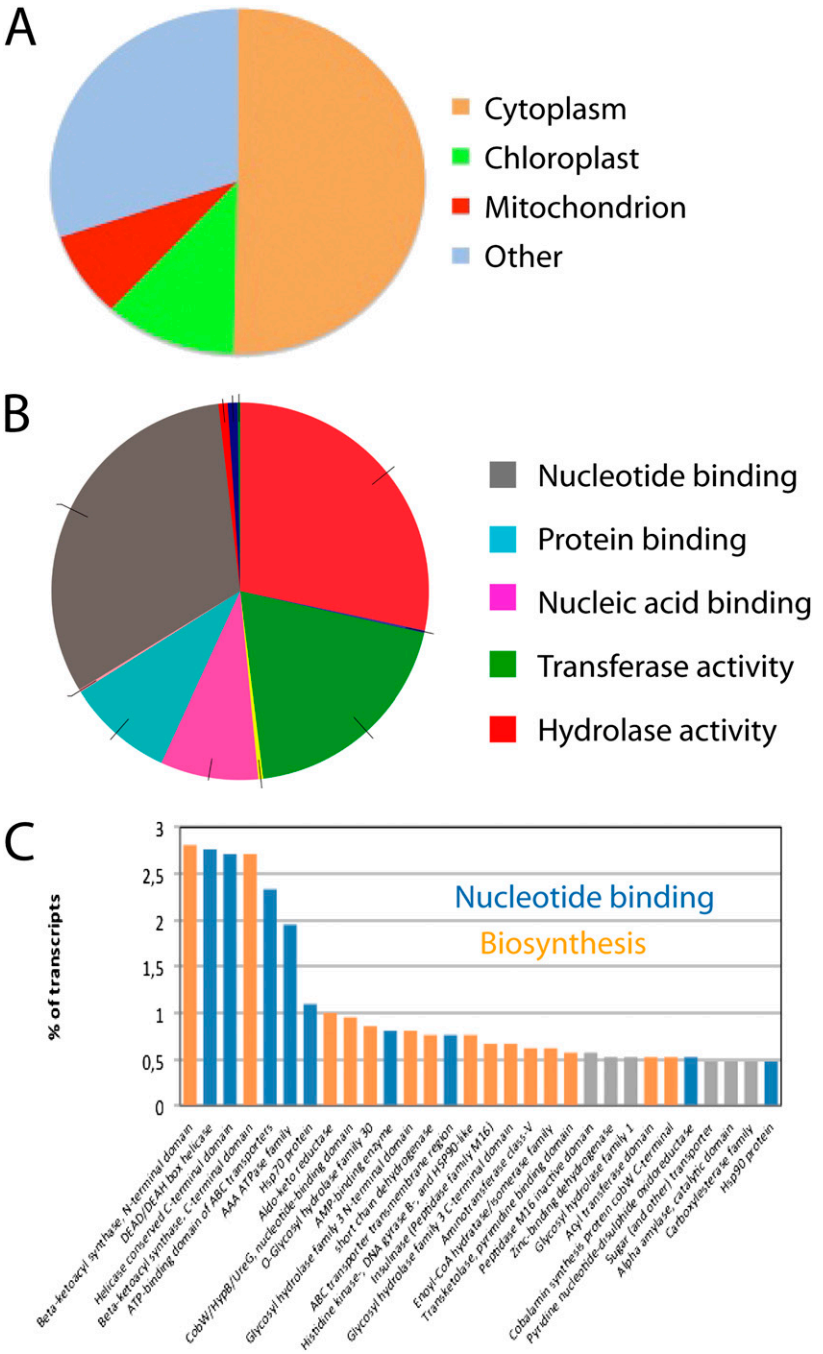

Fig. S5. Characteristics of the bacterial-like sequence in the transcriptome. (*A*) Of 2,354 sequences identified as putative bacterial sequences, 414 had an annotated match to GenBank. The annotated sequences were classified into the gene ontology compartment categories cytoplasm, mitochondrion, chloroplast, or all other membrane-bound compartments. (*B*) Functional classification shows enrichment of sequence in nucleotide-binding and enzymatic gene ontology categories. (*C*) The 30 most abundant protein family domains include principally nucleotide-binding and biosynthetic functions.

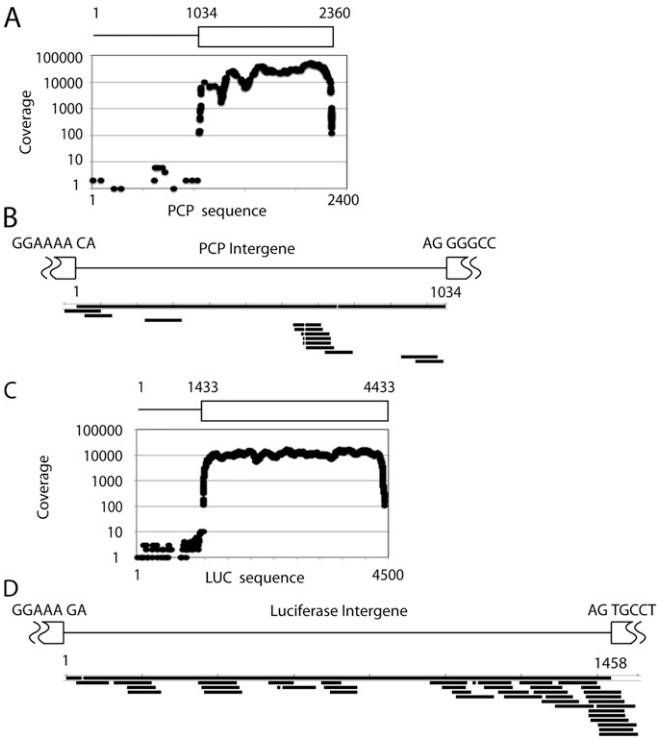

Fig. S6. Detection of reads aligning to intergenic spacers. (*A*) A PCP genomic sequence (GenBank accession no. U93077) containing the intergenic spacer (line) and coding sequence (box) was used to align 0.58×10^6 reads trimmed to remove low-quality or ambiguous bases from a dataset of 89×10^6 reads. The coverage (number of reads) is shown at each position. (*B*) Detail of the read assembly to the 1034-bp intergenic spacer region. Sequences corresponding to the polyadenylation site (*Left*) and transsplice site (*Right*) were determined by comparing the assembly with the genomic sequence. (*C*) The genomic sequence of the luciferase coding sequence (AF085332) and intergenic spacer (AF003893) were used to align reads trimmed to remove low-quality or ambiguous bases from a dataset containing 89×10^6 reads. The alignment of 0.43×10^6 reads with the intergenic spacer (line) and coding region (box) is displayed as coverage (number of reads) at each position. (*D*) A detail of the reads assembled only to the intergenic spacer is shown together with the nucleic acid sequences at the site of transsplicing (*Right*) and the polyadenylation site (*Left*).

Table S1. Number of genes found for various Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways

	Essential enzymes, *n*		
Pathway	Expected	Found	Worst E value
Glycolysis*	10	10	e^{-87}
TCA cycle	9	9	e^{-107}
Oxidative phosphorylation†	5	5	e^{-102}
Carbon fixation	11	11	e^{-71}
Purine synthesis	18	17	e^{-56}
Pyrimidine synthesis	12	11	e^{-58}
Fatty acid synthesis	6	4	e^{-84}
Fatty acid oxidation	4	4	e^{-138}
F, Y, W synthesis	16	15	e^{-50}
S, G, T synthesis	9	9	e^{-64}
R, P synthesis	11	11	e^{-56}
A, D, N, E, Q synthesis	5	5	e^{-118}
C, M synthesis	9	8	e^{-117}
V, L, I synthesis	8	7	e^{-97}
K synthesis	9	6	e^{-94}
H synthesis	9	9	e^{-55}
Total	151	141	

Amino acids are given using the single letter code. TCA, tricarboxylic acid.
*Glucokinase replaces hexokinase in the first reaction.
†Only the core proteins of the different complexes were analyzed. Most of the subunits in each of the five complexes were well represented.

Table S2. Number of KEGG pathway sequences found in mammals, plants, alveolata, and diatoms for replication, transcription, splicing, and translation

		Mammals/ plants		Alveolata*			Diatoms
Process	Subunit	Homsa	Arath	Plafa	Linpo	Tetth	Thaps
DNA replication	DNA polymerase						
	α Complex	4	4	4	4	4	4
	δ Complex	4	4	2	2	2	2
	ε Complex	4	2	2	2	2	3
	MCM complex	6	6	6	6	6	6
	RPA	3	2	1	1	1	2
	Clamp/loader	4	4	4	4	4	4
	Other						
	Helicase	1	1	0	1	1	0
	RNaseH1	3	3	1	0	3	1
	Fen1	1	1	1	1	1	1
	DNA ligase	1	1	1	1	1	1
Transcription	RNA polymerase I, II, and III						
	Core	10	9	10	10	9	10
	Specific	13	12	6	6	6	10
	Common	5	5	4	5	4	4
	Basal transcription factors						
	TFIIA	2	2	0	0	0	0
	TFIIB	1	1	1	0	0	0
	TFIID	15	10	1	1	3	4
	TBP	1	1	1	0	1	1
	TFIIE	2	2	0	0	0	1
	TFIIF	3	2	0	0	0	0
	TFIIH (NER)	10	10	9	3	5	8
Translation	Ribosome						
	B	0	7	0	0	1	1
	B/A/E	47	87	40	51	34	48
	A/E	25	25	20	24	21	23
	E	12	12	10	10	8	12
	aa-tRNA synthesis						
	Enzymes	23	23	22	22	22	22
	Basic translation factors						
	Initiation	42	57	29	29	33	37
	Elongation	10	16	8	9	8	9
	Release	11	11	5	3	7	4
Splicing	Splicesome						
	General	9	8	8	9	9	7
	U1	8	7	5	5	5	4
	U2	12	10	7	8	9	10
	U4/U6	7	7	6	7	7	7
	U5	8	8	6	7	7	7
	U5/U4/U6	5	5	4	2	4	5
	Prp19 complex	9	8	7	5	7	7
	Prp19 related	9	8	8	7	5	8
	EJC/TREX	6	5	4	3	4	5
	Common	3	3	1	1	2	1
Translation related	Ribosome biogenesis						
	90S preribosome	18	18	6	12	15	14
	Nucleus	14	12	9	10	12	12
	Nucleolus	17	15	13	14	14	14
	Cytoplasm	7	6	5	6	6	6
	mRNA transport						
	Nucleus	11	9	9	7	7	8
	Cytoplasm	6	6	2	3	2	3
	NPC	34	22	3	5	8	13
	SMNC	9	2	0	0	0	1
	eIFs	14	10	9	9	7	11
	EJC	16	12	6	9	6	6
	TREX	6	6	0	0	2	3

Table S2. Cont.

Process	Subunit	Mammals/ plants		Alveolata*			Diatoms
		Homsa	Arath	Plafa	Linpo	Tetth	Thaps
	mRNA surveillance pathway						
	Nucleus	34	26	9	15	10	17
	Cytoplasm	15	12	8	10	7	9
Total		540	545	323	349	342	396

aa, aminoacyl, Arath, *Arabidopsis thaliana*; eIFs, eukaryotic initiation factors; EJC, exon junction complex; Homsa, *Homo sapiens*; Plafa, *Plasmodium falciparum*; Linpo, *Lingulodinium polyedra*; MCM, mini chromosome maintenance; NER, nucleotide excision repair; NPC, nuclear pore complex; RPA, replication protein A; SMNC, survival motor neuron complex; TBP, TATA-binding protein; Tetth, *Tetrahymena thermophila*; TFII, transcription factor II; Thaps, *Thalassiosira pseudonana*; TREX, transcription/export.
*Apicomplexans, dinoflagellates, and ciliates.

Table S3. Nuclear- and plastid-encoded reference sequences from GenBank used for comparison of synonymous (dS) and nonsynonymous (dN) mutations

Gene name	Accession no.	Length (bp)	dS	dN
p43	AY423581	1,429	14 (0.05)	33 (0.04)
Phosphoribulokinase	AY772247	1,461	64 (0.22)	32 (0.03)
Histone-like protein	AF482694	511	13 (0.17)	18 (0.08)
Luciferase	AF085332	4,000	25 (0.03)	45 (0.02)
GAPDH (plastid isoform)	AF028560	1,433	150 (0.47)	35 (0.04)
Actin	AY423582	1,407	59 (0.23)	15 (0.02)
RuBisCo	GONR15B	1,912	136 (0.36)	33 (0.03)
Carbonic anhydrase	EU044834	1,636	18 (0.06)	23 (0.02)
Cyclin	AY618995	1,825	6 (0.02)	30 (0.03)
Cellulase	GQ258705	1,425	40 (0.12)	30 (0.03)
Glucose phosphate isomerase	DQ812892	1,875	10 (0.03)	18 (0.01)
Fructose-1,6-bisphophatase	DQ508159	1,235	12 (0.05)	32 (0.04)
Sedoheptulose-1,7-bisphophatase	DQ508153	1,492	58 (0.18)	44 (0.05)
Superoxide dismutase	AF289824	744	21 (0.13)	5 (0.01)
Peridinin-chlorophyll a-protein	JO692699	1,127	7 (0.025)	11 (0.013)
Luciferin-binding protein	GONLBPA	2,217	37 (0.08)	62 (0.04)
psaA	DQ264850	2,506	72 (0.14)	273 (0.16)
psaB	DQ264852	2,174	41 (0.09)	146 (0.09)
psbA	DQ264844	1,074	10 (0.04)	32 (0.04)
psbB	DQ264845	1,559	29 (0.08)	103 (0.09)
psbC	DQ264846	1,418	43 (0.17)	112 (0.14)
psbD	DQ264847	1,236	11 (0.04)	55 (0.06)
atpA	DQ264853	1,609	20 (0.06)	69 (0.07)
atpB	DQ264857	771	17 (0.10)	82 (0.14)
petB	DQ264849	842	21 (0.13)	29 (0.06)
petD	DQ264848	545	13 (0.11)	48 (0.13)

Column dS shows the total number of synonymous changes (ratio of synonymous substitutions per synonymous site) while column dN shows the total number of nonsynonymous changes (ratio nonsynonymous substitutions per nonsynonymous site).