Polarella glacialis genomes encode tandem repeats of single-exon genes with functions critical to adaptation of dinoflagellates

Timothy G. Stephens[1], Raúl A. González-Pech[1], Yuanyuan Cheng[2,†], Amin R. Mohamed[3], Debashish Bhattacharya[4], Mark A. Ragan[1] and Cheong Xin Chan[1,5,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

[2]UQ Genomics Initiative, The University of Queensland, Brisbane, QLD 4072, Australia

[3]Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture and Food, Queensland Bioscience Precinct, Brisbane, QLD 4067, Australia

[4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, U.S.A.

[5]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia

[†]Present address: School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Camperdown, NSW 2006, Australia

*Corresponding author (c.chan1@uq.edu.au)

Abstract

Dinoflagellates are diverse, ecologically important phytoplankton in marine and freshwater environments. Here, we present two draft *de novo* diploid genome assemblies of the free-living dinoflagellate *Polarella glacialis*, isolated from the Arctic and Antarctica. For each genome, we predicted >50,000 high-quality genes supported by full-length transcriptome data. About 68% of the total genome sequence is repetitive, and includes long terminal repeats that likely contribute to intra-species structural divergence and distinct genome sizes (3.0 and 2.7 Gbp). In each genome, ~40% of genes are encoded unidirectionally and ~25% are single exonic; these include tandemly repeated genes that encode functions related to cold adaptation and photosynthesis. Multi-genome comparison unveiled genes specific to *P. glacialis* and a common ancestral origin of ice-binding domains in cold-adapted dinoflagellates. Our results provide insights into how dinoflagellate genomes may enhance the transcriptional efficiency of critical genes as a mechanism of environmental adaption and niche specialisation.

Introduction

Dinoflagellates are a diverse group of phytoplankton that are ubiquitous in marine and fresh waters. Mostly photosynthetic, dinoflagellates form the base of food webs. They critically sustain global aquatic ecosystems *via* primary production and cycling of organic carbon and nitrogen. Some dinoflagellate lineages are symbiotic or parasitic. For instance, members of the family Symbiodiniaceae are crucial symbionts in corals and other coral reef animals (Baker 2003, LaJeunesse et al. 2018), and parasitic dinoflagellates can cause death in economically important crustaceans such as crabs and lobsters (Stentiford et al. 2005). Most dinoflagellates, however, are free-living. Bloom-forming taxa may cause “red tides”, which produce toxins that pose serious human health risks (Grattan et al. 2016). Some taxa have

specialised to inhabit extreme environments, such as those found in brine channels of polar sea ice (Montresor et al. 1999, Taylor et al. 2008, Gómez 2012, Le Bescot et al. 2016).

So far, available genome data of dinoflagellates are largely restricted to symbiotic or parasitic species (Shoguchi et al. 2013, Gornik et al. 2015, Lin et al. 2015, Aranda et al. 2016, Liu et al. 2018, Shoguchi et al. 2018, John et al. 2019). These species were chosen for sequencing because their genomes are relatively small, i.e. 0.12-4.8 Gbp. In comparison, genomes of other free-living dinoflagellates are much larger, ranging from 5 Gbp in the psychrophile *Polarella glacialis*, to over 200 Gbp in *Prorocentrum* sp. based on DAPI-staining of DNA content (LaJeunesse et al. 2005).

Repeat content has been estimated at >55% in the genome sequences of some free-living dinoflagellates (Allen et al. 1975, Jaeckisch et al. 2011); single-exon genes have also been described (Le et al. 1997). As most dinoflagellate lineages are free-living, whole genome sequences of these taxa are critical to understand the molecular mechanisms that underpin their successful diversification in specialised environmental niches.

Polarella glacialis, a psychrophilic free-living species, represents an excellent system for genomic studies of dinoflagellates for three reasons. First, it is closely related to Symbiodiniaceae (both lineages are in Order Suessiales). Second, *P. glacialis* has been reported only in polar regions. Studying the *P. glacialis* genome can thus provide a first glimpse into molecular mechanisms that underlie both the evolutionary transition of dinoflagellates from a free-living to a symbiotic lifestyle, and the adaptation to extreme environments. Third, the estimated genome size of *P. glacialis* is still in the smaller range (~7 Gbp (LaJeunesse et al. 2005)) of all dinoflagellate taxa; this presents a technical advantage.

Here, we report draft *de novo* genome sequences from two *P. glacialis* isolates: CCMP1383 and CCMP2088. The former is a xenic culture first isolated from brine in the upper sea ice in McMurdo Sound (Ross Sea, Antarctica) in 1991 (Montresor et al. 1999), and the latter is a xenic culture first isolated from a water sample collected adjacent to ice in northern Baffin Bay in 1998 (Montresor et al. 2003). These genomes represent the first generated from any free-living, psychrophilic dinoflagellates. Incorporating full-length transcriptome data, we systematically investigated distinct features in these genomes including repeat content, and gene structure and intra-species genome divergence. Our results reveal remarkable difference in genome sizes between these two isolates of the same species, and provide evidence of tandemly repeated, single-exon genes in shaping the evolution of dinoflagellate genomes.

Results

Genomes of *Polarella glacialis*

Draft genome assemblies for two *Polarella glacialis* isolates (CCMP1383 and CCMP2088) were generated using a combination of Illumina short-read and PacBio long-read data (Table 1 and Supplementary Table S1). Both genomes appear diploid based on their bimodal distributions of *k*-mer counts observed from the sequence data, which closely (model fit >92%) matches the standard theoretical diploid model (Supplementary Figure S1). These assemblies represent diploid genomes, reported for the first time in any dinoflagellates. The CCMP1383 assembly had fewer and more-contiguous scaffolds (33,494; N50 length 170 Kbp; Table 1) compared to the CCMP2088 assembly (37,768; N50 length 129 Kbp; Table 1); this is likely due to more long-read data generated for the former (Supplementary Table S2). Both assemblies are much more contiguous than their corresponding assemblies generated using only short-read data (N50 length < 73 Kbp; Supplementary Table S1). For CCMP1383 and CCMP2088 respectively, the total assembly sizes are 2.98 Gbp and 2.76 Gbp (Table 1

and Supplementary Table S1), and are very similar to independent genome-size estimates of 3.02 Gbp and 2.65 Gbp (Supplementary Table S3). These genomes are smaller than previously estimated (~7 Gbp (LaJeunesse et al. 2005)). However, they remain larger than those of Symbiodiniaceae (Shoguchi et al. 2013, Lin et al. 2015, Aranda et al. 2016, Liu et al. 2018, Shoguchi et al. 2018) (between 1.1 and 1.5 Gbp), and smaller than the 4.8 Gbp-genome of the parasitic *Hematodinium* sp. (Gornik et al. 2015) (Table 1 and Supplementary Table S1). These results reaffirm the tendency of DNA-staining or flow-cytometry to overestimate genome sizes of dinoflagellates (Shoguchi et al. 2013, Lin et al. 2015, Aranda et al. 2016, Liu et al. 2018, Shoguchi et al. 2018).

Table 1. Assembled genomes of *P. glacialis* compared to key publicly available dinoflagellate genomes. A more-comprehensive summary including all other available genomes is shown in Supplementary Table S1.

	Polarella glacialis		**Symbiodiniaceae**				**Parasitic**	
	CCMP1383	**CCMP2088**	***Symbiodinium microadriaticum***	***Breviolum minutum***	***Cladocopium goreaui***	***Fugacium kawagutii***	***Amoebophrya ceratii***	***Hematodinium* sp.**
%G+C	45.91	46.15	50.51	43.46	44.83	45.72	55.92	47.31
Total number of scaffolds	33,494	37,768	9,695	21,899	41,289	16,959	2,351	869,500
Total assembled bases (Gbp)	2.98	2.76	0.81	0.61	1.03	1.05	0.0877	4.77
N50 length of scaffolds (bp)	170,304	129,205	573,512	125,226	98,034	268,823	83,970	17,235
Maximum scaffold length (bp)	2,170,995	1,500,384	3,144,590	810,747	8,337,000	5,159,000	536,776	186,000
Estimated genome size (Gbp)	3.02	2.65	1.10	1.5	1.19	1.07	0.12	4.8

The non-repetitive regions from both assembled genomes are almost identical. In a comparison between the non-repetitive regions of CCMP2088 against the genome of CCMP1388, 98.6% of the regions share 99.4% sequence identity; likewise in CCMP1383,

98.2% of compared regions share 99.2% sequence identity with the CCMP2088 genome. Remarkably, the genome of the Antarctic isolate (CCMP1383) is approximately 230 Mbp larger than that of the Arctic isolate (CCMP2088). These results reveal, for the first time, structural divergence of genomes in dinoflagellates even within a single species, potentially explained by the uneven expansion of repetitive elements (see below). The two genome assemblies are reasonably complete; a similar proportion of core conserved eukaryote genes were recovered, i.e. 332 (72.49%) and 337 (73.58%) of the 458 CEGMA (Parra et al. 2009) genes in CCMP1383 and CCMP2088, respectively (Figure 1A and Supplementary Table S4). These numbers are comparable to those recovered in published Symbiodiniaceae genomes (Liu et al. 2018), e.g. 350 in *C. goreaui* and 348 in *F. kawagutii*, analysed using the same approach (Figure 1A).

The high extent of genome-sequence similarity between the two geographically distinct *P. glacialis* isolates suggests that they have either recently been transferred from one polar region to the other, or are being actively transported between the two locations, allowing for mixing of the two populations. To identify if *P. glacialis* is being actively transported between polar regions we interrogated the TARA Oceans database for the presence of this species in the broad sampled sites. Despite *P. glacialis* sequences being reported in 67 of the 68 TARA sample locations, an exhaustive search did not recover clear evidence of any sequence that is representative of *P. glacialis* (see Methods). While we cannot dismiss the presence of *P. glacialis* at low, undetectable levels, we find no evidence at this time to support the presence of *P. glacialis* in the waters outside of the polar regions.

Polarella glacialis genomes are highly repetitive

Both *P. glacialis* genomes reveal high content of repetitive elements that encompass ~68% (by length) of the assembled sequences (Figure 1B and Supplementary Figure S2); most of

these elements are simple and unknown repeats (i.e. unclassified *de novo* repeats; covering ~13.5% of each assembled genome; Figure 1B). The proportion of repeats in *P. glacialis* genomes is more than two-fold higher than that reported in Symbiodiniaceae (e.g. 27.9% in *Symbiodinium microadriaticum*, 16% in *Fugacium kawagutii*) (Liu et al. 2018). This observation is not unexpected, because even before high-throughput sequencing technology was available, the genome of the biotechnologically important dinoflagellate *Crypthecodinium cohnii* was estimated to contain 55-60% repeat content (Allen et al. 1975). A genome survey of *Alexandrium ostenfeldii* estimated the repeat content at ~58% (Jaeckisch et al. 2011). In comparison, the genome surveys of *Heterocapsa triquetra* (McEwan et al. 2008) and *Prorocentrum minimum* (Ponmani et al. 2016) estimated their repeat content at only ~5% and ~6%, respectively. These values are likely underestimates because only 0.0014% of the *H. triquetra* genome was surveyed, and >28% of the *P. minimum* genome data is putatively of bacterial origin (Ponmani et al. 2016).

The prevalence of repeats in *P. glacialis* genomes may explain their larger genome sizes compared to symbiotic dinoflagellates (Shoguchi et al. 2013, Lin et al. 2015, Aranda et al. 2016, Liu et al. 2018, Shoguchi et al. 2018), and may represent a genome signature of free-living dinoflagellates. These repeats are more conserved in *P. glacialis* (Kimura substitution level centred around 5; Figure 1B and Supplementary Figure S2) than those reported in Symbiodiniaceae (Kimura substitution level 10-30 (Liu et al. 2018)). We also recovered a substantial proportion of long-terminal repeat (LTR) elements (~12%) in *P. glacialis* genomes; these elements were largely absent (<0.7%) in Symbiodiniaceae (Liu et al. 2018). Transposable elements (such as LTRs) commonly comprise up to 80% of the genomes of plants and are induced by genome shock and polyploidization, resulting in genome restructuring (Galindo-Gonzalez et al. 2017). The abundance of LTRs in *P. glacialis* and the role of LTRs in genome restructuring may explain in part the difference in genome sizes

between the two isolates. These results suggest that repetitive elements and LTRs are key contributors that drive genome evolution of *P. glacialis*, both as a free-living and a cold-adapted dinoflagellate species. Because available dinoflagellate genomes (e.g. of Symbiodiniaceae) thus far have been generated largely using Illumina short-read data, we cannot dismiss the possibility that misassembly of these genomes (an inevitable artefact with short-read data) may have caused the under-estimation of repeat content (and the apparent absence of LTRs) in these genomes.

In an independent analysis of simple repeats (see Methods), 25.01% and 24.17% respectively of the CCMP1383 and CCMP2088 genomes are found to be composed of simple repeats. The most prominent simple repeat is the trinucleotide (TTG)n (in all six reading-frames; see Methods) that covered 19.1% and 18.5% of the CCMP1383 and CCMP2088 genome assemblies, respectively. The proportion of (TTG)n, observed as possible 3-mers of TTG, TGT or GTT (each ~7-8%) in the assembled genomes, is very similar to that observed in the sequence-read data (Figure 1C and Supplementary Table S5). Therefore, this observed prevalence of (TTG)n is unlikely due to assembly artefacts.

DinoSL in full-length transcripts of *Polarella glacialis*

To generate high-quality evidence to guide our gene-prediction workflow, we generated transcriptomes for both *P. glacialis* isolates, including full-length transcripts using PacBio IsoSeq technology (see Methods). Mature nuclear transcripts of dinoflagellates are known to contain a 22-nucleotide trans-spliced leader sequence (DinoSL: DCCGTAGCCATTTTGGCTCAAG, where D = T, A, or G) at the 5′-end (Zhang et al. 2007). Relic DinoSL sequences arise when transcripts with attached DinoSL are integrated back into the genome, expressed and trans-spliced with a new leader sequence. Successive rounds of transcript re-integration results in multiple relic DinoSLs on a single transcript. We

searched the full-length transcripts (435,032 in CCMP1383 and 1,266,042 in CCMP2088) for presence of DinoSL and relic DinoSL sequences (see Methods). DinoSL sequences were recovered in 13.54% and 50.39% of transcripts (hereinafter DinoSL-type transcripts) in CCMP1383 and CCMP2088 respectively (Supplementary Table S6. An earlier study (Zhang et al. 2009) reported a single transcript in CCMP2088 that has a non-canonical DinoSL sequence (ATCGTAGCCATGTTGGCTCAAG), but our exhaustive search against the transcripts in either isolate did not recover this sequence.

Although our experiment (see Methods) was designed to recover full-length transcripts (with complete 3′ and 5′ regions), it is possible that the adopted library preparation step is not 100% efficient, and that the lack of a DinoSL in so many transcripts may be due to mRNA degradation. This may, in the first instance, be reflected by the varying degrees of truncation of the DinoSL-type transcripts in our data. However, 70.45% of these transcripts in CCMP1383 and 69.18% of transcripts in CCMP2088 start at one of the two contiguous cytosine bases (i.e. at positions 2 and 8) of the DinoSL (Figure 2A; Supplementary Table S7). This preference for start sites at the double-cytosine positions suggests that the 5′ selection method we used (that purifies for the 5′ methylated cap site) is binding to these regions instead of the true 5′-cap. This in turn may happen because cytosines at these sites are methylated. Cytosine methylation has been described in genomes of eukaryotes including dinoflagellates, potentially as a mechanism for silencing of transposable elements and regulation of gene expression (Lohuis et al. 1998, Law et al. 2010, Zemach et al. 2010); a recent study of the *Breviolum minutum* genome revealed that cytosine methylation often occurred at CG dinucleotides (de Mendoza et al. 2018). The impact of methylation on recovery of splice leaders in dinoflagellates remains to be systematically investigated.

In CCMP1383 and CCMP2088, 0.68% and 2.31% of all full-length transcripts respectively were found to encode one relic DinoSL (immediately following their primary DinoSL), smaller proportions (0.020% and 0.048% respectively) encode multiple relic DinoSL sequences. We recovered 30 transcripts in CCMP2088 that have four putative relic DinoSL sequences; they shared >99% sequence identity among one another. Five of these 30 transcripts are shorter than the others, suggesting an alternative transcription 3′-termination site, thus a distinct isoform.

We further assessed the diversity of alternative splice-forms by clustering the full-length transcripts by sequence similarity using PASA (see Methods). Each resulted PASA "assembly" (Haas et al. 2003) represents a distinct alternative isoform, and overlapping "assemblies" constitute a transcriptional unit (Supplementary Table S8). We identified 30,463 and 22,531 alternative isoforms comprising 24,947 and 19,750 transcriptional units in CCMP1383 and CCMP2088 respectively. When focusing only on DinoSL-type transcripts, these numbers are 8714 and 6576, comprising 7146 and 5110 transcriptional units respectively (Supplementary Table S8). In both isolates, alternative exons are the most common events observed among all transcript isoforms (e.g. 45.85% of all inferred events in CCMP2088), followed by alternative donor (22.05%) and acceptor (20.43%) sites (Supplementary Table S9).

The addition of DinoSL sequences was proposed as a mechanism to split polycistronic pre-mRNA into monocistronic mature mRNA (Zhang et al. 2007). A little over half (50.45% in CCMP1383, 58.78% in CCMP2088) of these DinoSL-type transcriptional units are located within 5 Kb of one another (Figure 2B). Interestingly, among the DinoSL-type transcript isoforms, the two most-enriched Pfam domains are bacteriorhodopsin-like protein (PF01036) and cold-shock DNA-binding (PF00313) in both isolates (Supplementary Table S10). To

further assess the functional diversity of DinoSL-type transcripts, we independently sequenced 747,959 full-length transcripts from CCMP1383 specifically selected for DinoSL (Supplementary Table S6; see Methods). These transcripts comprised only 1187 isoforms (3.9% of total 30,463 isoforms; Supplementary Table S8). Similar functions are prevalent among these genes (Supplementary Table S10), thus lending support to our observation of functional bias in DinoSL-type transcripts. In addition, the frequency at which DinoSL-type transcripts are integrated back into the genome is likely dependent on their relative abundance in the nucleus. Therefore, transcripts containing relic DinoSLs are likely to be, or have been, highly expressed. In both isolates, the ice-binding (DUF3494) and the bacteriorhodopsin-like protein domains, both important for adaptation to cold (see below), are among the most-enriched features in transcripts encoded with a relic DinoSL.

Prediction of gene models in *Polarella glacialis* is likely impacted by RNA editing

Using a gene-prediction workflow customised for dinoflagellate genomes (Liu et al. 2018) (see Methods), we predicted 58,232 genes and 51,713 genes in the CCMP1383 and CCMP2088 genomes respectively (Table 2 and Supplementary Table S11). Of the 58,232 genes predicted in CCMP1383, 51,640 (88.68%) of their coded proteins were recovered among those in CCMP2088 (Figure 3). Likewise, of the 51,713 genes predicted in CCMP2088, 46,228 (89.39%) of their coded proteins were recovered among those in CCMP1383 (Figure 3). The difference in numbers of predicted genes and sequence dissimilarity observed between the two genomes could be explained in part by the presence of distinct transcript isoforms. Although transcriptome evidence can improve the quality of predicted genes, our results indicate that this evidence can also complicate prediction when a gene has multiple isoforms, more so when these isoforms are recovered unevenly between the two isolates.

Table 2: Predicted gene models in *P. glacialis* compared to key publicly available dinoflagellate genomes. A more-comprehensive summary including gene models from all available dinoflagellate genomes is shown in Supplementary Table S11.

	Polarella glacialis		**Symbiodiniaceae**				**Parasitic**
	CCMP1383	**CCMP2088**	***Symbiodinium microadriaticum***	***Breviolum minutum***	***Cladocopium goreaui***	***Fugacium kawagutii***	***Amoebophrya ceratii***
Genes							
Number of genes	58,232	51,713	49,109	47,014	35,913	26,609	19,925
Gene models supported by transcriptome (%)	93.95	94.35	76.30	77.20	67.02	64.40	24.37
G+C content of CDS (%)	57.84	57.78	57.67	50.80	56.70	54.95	60.77
Exons							
Number of exons per gene	11.64	10.84	21.8	19.6	10	8.7	3.39
Average length (bp)	105.67	108.71	109.5	100.8	175.9	199.5	577.8
Total length (Mb)	71.6	60.9	117.3	83.0	63.4	46.2	39.1
Introns							
Number of genes with introns (%)	73.79	75.60	98.2	84.4	92.9	94	71
Average length (bp)	1,408	1,296	504.7	500.4	575.1	619.4	337.1
Total length (Mb)	838	636.2	516.1	332.2	186.8	126.9	16.1
Intergenic regions							
Average length (bp)	21,625	20,922	3,633	1,993	10,627	23,042	1,525

Note: *Hematodinium* sp. is not shown as no predicted genes were reported.

Interestingly, almost all predicted proteins not recovered in the proteins of the counterpart isolate (i.e. 6003 of 6592 in CCMP1383 and 4660 of 5485 in CCMP2088) were recovered in the transcriptome of the counterpart isolate (Figure 3). These results indicate that some transcriptome evidence was not incorporated in ~10% of the predicted genes in each genome. We hypothesise that this is likely due to RNA editing in *P. glacialis*. RNA editing has been characterised in the nuclear-encoded genes of *Symbiodinium microadriaticum* (Liew et al. 2017), as well as organellar genes of other dinoflagellates (Lin et al. 2002, Klinger et al. 2018). RNA editing may introduce changes in the transcripts (e.g. base substitutions or indels) affecting the identification of open-reading frames (e.g. disruption by in-frame stop

codons or correction of premature stop codons) in the genome sequences, and thus impacting prediction of gene models.

Given the diploid genome assembly for each isolate, we assume that the number of predicted genes would approximate twice the number expected in a haploid genome (e.g. 50,000 genes in a diploid assembly versus 25,000 in a haploid assembly). In comparison, the reported number of predicted genes is generally greater among the six (haploid) genomes of Symbiodiniaceae (Table 2 and Supplementary Table S11); the numbers vary from 26,609 in *Fugacium kawagutii* (Liu et al. 2018) to 69,018 in *Symbiodinium tridacnidorum* (Shoguchi et al. 2018). The wide-ranging number of predicted genes in Symbiodiniaceae genomes may be due to the different gene-prediction workflows adopted in the earlier studies, and the extent and quality of transcriptome data used in guiding gene prediction (Chen et al. 2019). We expect such technical biases would intensify in highly divergent genomes, as is the case for Symbiodiniaceae (Chen et al. 2019). In addition, the proportion of genes in these genomes that are supported by transcriptome evidence is smaller (~72% averaged among six genomes) than that we observed in our predicted genes of *P. glacialis* (~94% for each isolate; Table 2). This result may be explained by the more-extensive transcriptome data we generated in this study (using both RNA-Seq short-read and Iso-Seq full-length transcripts) specifically to guide our gene-prediction workflow (see Methods), compared to the transcriptome data (based on RNA-Seq short-reads) used in the earlier studies. To what extent these gene numbers are comparable remains to be investigated. However, our gene-prediction workflow (see Methods), adapted from Liu et al. (2018), is customised for dinoflagellate genomes; genes were predicted based on a set of stringent criteria adopting both *ab initio* and evidence-based methods, and strong support from transcriptome evidence. Our predicted genes in *P. glacialis* represent the best-quality predictions of dinoflagellate genes to date.

Unidirectional tandem single-exonic genes in *P. glacialis*

In *P. glacialis*, longer intergenic regions have higher fraction of repeats covering those regions (Supplementary Figure S3). Roughly a third of the intergenic regions (35.86% in CCMP1383; 34.97% in CCMP2088) are ≤ 5 Kbp in length. The fraction of these regions covered by repetitive elements is 32.92% (CCMP1838) and 32.93% (CCMP2088; Supplementary Figure S3); these numbers are 59.65% and 59.05% (Supplementary Figure S3) among intergenic regions >5 Kbp. This observation suggests that expansion of repeats is greater in (and likely contributes to) longer intergenic regions in the genome. Approximately 50% of the analysed genes (26,580 in CCMP1383, 21,376 in CCMP2088) appear to have intergenic regions ≤ 5 Kbp (Figure 4A), indicating a tendency for these genes to occur in clusters. Remarkably, almost all of these clustered genes (24,276 and 19,544 respectively for those of CCMP1383 and CCMP2088; ~40% of total genes in each genome) were found to be encoded unidirectionally. These unidirectional gene clusters may represent a mechanism in *P. glacialis* to ensure transcriptional efficiency, with genes in close physical proximity potentially transcribed together. In some cases, these gene clusters encode the same or similar functions, e.g. 22 unidirectionally encoded genes in CCMP1383 (and 19 in CCMP2088) putatively code for major basic nuclear protein 2.

Among the predicted genes in both genomes, 4898 (CCMP1383; 8.4%) and 3359 (CCMP2088; 6.5%) are located (nested) within introns of multi-exon genes. Although most cases (71.02% in CCMP1383 and 74.90% in CCMP2088) represent one nested gene per multi-exon gene, in extreme cases we observed 18 (CCMP1383) and 24 (CCMP2088). Supplementary Figure S4 shows an example of 15 nested genes of CCMP1383 spanning three introns of the gene putatively encoding an alanine–tRNA ligase. Among the nested genes within each intron, five encode for fucoxanthin chlorophyll *a/c*-binding protein, and

four for light-harvesting complex protein. The validity of the nested gene structure was confirmed by expression evidence based on full-length transcripts.

Of particular interest, we recovered 15,263 (26.2%) and 12,619 (24.4%) single-exon genes in CCMP1383 and CCMP2088 respectively (Table 2). These proportions are higher than those in symbiodiniacean genomes (< 20% of genes; Table 2 and Supplementary Table S11) except that in the earlier assembled genome of *Fugacium kawaguii* (34.9% of genes)(Lin et al. 2015). Almost all single-exon genes in *P. glacialis* (99.08% of those in CCMP1383, 98.12% of those in CCMP2088) are supported by transcriptome evidence (including full-length transcripts that were selected for 3′-polyadenylation and 5′-cap sites). These results suggest that these genes are *bona fide P. glacialis* genes (i.e. not bacterial contaminants, nor artefacts of our gene-prediction workflow). Many of the Pfam domains enriched in the single-exon genes are also enriched in the predicted genes of *P. glacialis* compared with symbiodiniacean genes (see also below; Supplementary Table S12). Enriched features of *P. glacialis* such as bacteriorhodopsin-like protein (PF01036), peridinin-chlorophyll *a*-binding (PF02429) and DUF3494 (PF11999) are encoded as single-exon genes. A number of other domains are enriched in the single-exon genes of both *P. glacialis* isolates. The bacterial DNA-binding protein domain (PF00216), which is predominantly found in bacteria, is enriched and potentially has arisen in *P. glacialis* via lateral genetic transfer. The reverse transcriptase (PF00078) domain is also enriched and is likely involved in the activity of retrotransposons in the *P. glacialis* genomes.

What makes *Polarella Polarella*?

We compared the annotated functions of *P. glacialis* genes against those from Symbiodiniaceae genomes (Shoguchi et al. 2013, Aranda et al. 2016, Liu et al. 2018, Shoguchi et al. 2018). When comparing the annotated PFAM domains, we observed a

significant over-representation of DUF3494 (PF11999) and chlorophyll *a-b* binding (PF00504) domains in *P. glacialis* relative to the Symbiodiniaceae (Supplementary Table S13), as we previously observed in an independent analysis of transcriptomes (Stephens et al. 2018). In this study using high-quality gene models predicted from genome data, we further observed over-representation of pentatricopeptide repeat (PF13041, PF13812, PF01535), bacteriorhodopsin-like protein (PF01036), and peridinin-chlorophyll *a*-binding (PF02429) in *P. glacialis*. Interestingly, the peridinin-chlorophyll *a*-binding and bacteriorhodopsin-like domains predominantly are encoded in blocks of tandemly repeated single-exon genes, implicating hundreds of genes in *P. glacialis* (Figure 4B). All but one of these gene blocks (i.e. a contiguous region containing two or more genes) are unidirectionally encoded. Peridinin-chlorophyll *a*-binding protein was thought to be encoded as 5000 single-exon gene copies in tandem repeat blocks in the bloom-forming dinoflagellate of *Lingulodinium polyedra* (Le et al. 1997), and that these coding genes are likely monocistronic (Beauchemin et al. 2012); this protein may be universally important in free-living dinoflagellates, and potentially to a lesser extent among symbiotic lineages dinoflagellates (Reichman et al. 2003). The tendency of tandemly repeated genes to have fewer introns was also reported in the bloom-forming *Amphidinium carterae* (Bachvaroff et al. 2008). In combination with our other results (above), our observations suggest gene-family expansion through tandem duplication drives the genome evolution of *P. glacialis*, and potentially of other free-living dinoflagellates. The use of a DinoSL sequence to split polycistronic transcripts into mature RNAs (Zhang et al. 2007) may facilitate this mechanism.

Bacterial-derived rhodopsin, a transmembrane protein involved in bacterial phototrophy independent of chlorophyll through retinal binding, is encoded in diverse dinoflagellate lineages (Lin et al. 2010, Slamovits et al. 2011). The proton-pump type rhodopsins are known to create a proton gradient to drive synthesis of ATPase, in lieu of photosynthesis (Beja et al.

2001, Fuhrman et al. 2008). An earlier gene-expression analysis of the bloom-forming *Prorocentrum donghaiense* (Shi et al. 2015) revealed that proton-pump rhodopsins may compensate for photosynthesis under light-deprived conditions. These rhodopsins were also found to be highly expressed in diatoms under iron-deficient conditions (Marchetti et al. 2012). All genes in both *P. glacialis* isolates have top hits to the sequences coding for proton-pump rhodopsins in *Oxyrrhis marina*. These rhodopsins were previously found to be more abundantly expressed in *O. marina* than the sensory-type rhodopsins involved in light-harvesting for photosynthesis (Guo et al. 2014). We hypothesise that the over-representation of rhodopsin and other photosynthesis-related genes in *P. glacialis* is an adaptation to light-limited (and potentially iron-limited) conditions, as expected in their natural habitat of ice-brine channels.

A previous study based on transcriptome analysis (Stephens et al. 2018) revealed dark proteins (i.e. proteins that do not share sequence similarity to those of known function) that are conserved and/or lineage-specific in dinoflagellates. Using 437,829 protein sequences predicted from genome data of *P. glacialis* and six Symbiodiniaceae species (Supplementary Table S14), we constructed 43,465 putatively homologous protein sets that consist of 78.7% of the total protein sequences analysed. Of these sets, 9445 (21.73%) containing 8.95% of the clustered proteins (30,584 proteins; 6.99% of 437,829) were classified as dark (i.e. they were not annotation with a known function based on sequence-similarity searches against UniProt database; see Methods). The number of dark proteins (and hence dark genes) from each dataset (Supplementary Table S14) were largely congruent with the proportions of dark genes reported previously (Stephens et al. 2018). Of the 9445 dark homologous sets, 4443 (47.04%) contain sequences from only *P. glacialis*; 4371 (98.38% of the 4443) contain sequences from both isolates, thus they are unlikely to have arisen due to assembly artefacts. We consider a dark set as single-exonic if all its members are encoded in single exons, and a dark set as

multi-exonic if at least one member is encoded in multiple exons. Following this definition, most (2988; 67.3%) of the 4,443 *P. glacialis*-specific sets are multi-exonic, while 1,455 (32.7%) are single-exonic. Of the 1,455 single-exonic dark sets, 719 (49.4%) are supported by IsoSeq data and 1,451 (99.7%) by IsoSeq and/or RNA-Seq data. Therefore, these genes likely represent true genetic (and functional) innovation specific to *P. glacialis*.

Bacterial species associated with *P. glacialis* in the environment

Two bacterial genomes were generated as part of our effort in sequencing the genomes of *P. glacialis*. A CCMP1383 scaffold (5,892,869 bp) from the preliminary (short-read-only) assembly mapped at 93.8% sequence identity to the genome of *Paraglaciecola psychrophila* strain 170T (GenBank NC_020514, 5,413,691 bp; gamma-Proteobacteria), and another (3,839,769 bp) at 83.9% sequence identity to the genome of *Sphingorhabdus* sp. YGSM121 (GenBank NZ_CP022548, 3,864,176 bp; alpha-Proteobacteria). These two scaffolds show strong conserved synteny to the published genomes, with a few structural rearrangements (Supplementary Figure S5). *P. psychrophila* 170 was isolated from the Arctic (Yin et al. 2013), and *Sphingorhabdus* sp. YGSM121 from temperate sea sediment near South Korea (GenBank NZ_CP022548). Similarly, the three largest contaminant scaffolds (total 4,274,053 bp) in the preliminary CCMP2088 assembly have likely originated from the Arctic *Maribacter arcticus* (GenBank GCF_900167935.1, 4,211,145 bp; Bacteroidetes/Chlorobi) isolate. Therefore, these bacterial species are likely microbial associates of *P. glacialis* in the origin environment (i.e. part of the *P. glacialis* holobiont) where the two isolates were first isolated, but this notion remains to be investigated.

Single evolutionary origin of ice-binding domains in dinoflagellates

The Pfam domain DUF3494, a known ice-binding domain (Vance et al. 2019), has been shown to be over-represented in cold-adapted dinoflagellates (Stephens et al. 2018). In both

P. glacialis isolates (Supplementary Table S15), most putative ice-binding genes encode only the DUF3494 domain. They are encoded in single exons, and in unidirectional, tandemly repeated blocks, potentially as a mechanism to enhance the efficiency of expression of these genes. As the DUF3494 domain in many species has arisen via lateral genetic transfer (Vance et al. 2019), the presence of these genes in this configuration suggests that they might have arisen via the same mechanism in *P. glacialis*.

Figure 5 shows part of a phylogenetic tree reconstructed based on 1080 sequences of available DUF3494 domains encompassing archaea, bacteria, and eukaryotes; the complete tree is available as Supplementary Data S1. All DUF3493 sequences from the dinoflagellates (*P. glacialis*, *Heterocapsa arctica*, *Scrippsiella hangoei* and *Peridinium aciculiferum*), plus some sequences from the ice diatom *Fragilariopsis cylindrus*, form a strongly supported clade (bootstrap support [BS] = 100% based on ultrafast bootstrap approximation (Hoang et al. 2018)) (Figure 5). Within this dinoflagellate+diatom clade, the 169 DUF3494 sequences from *P. glacialis* (97 from CCMP1383, 72 from CCMP2088) form a strongly supported monophyletic clade (BS 100%), indicating that these domains in *P. glacialis* have an evolutionary history distinct from that in other dinoflagellates. In comparison, the domains in the ice diatom *F. cylindrus* were recovered in three distinct clades on the tree (two shown in Figure 5), indicating their different origins. As previously reported, DUF3494 domains in eukaryotes trace their origins to multiple events of lateral genetic transfer from bacteria and other eukaryotes (Sorhannus 2011, Arai et al. 2019). We also observed this pattern on the tree in Figure 5; the exact origin of these domains in dinoflagellates remains unclear, with potential sources of Proteobacteria, Bacteroidetes/Chlorobi, or Euryarchaeota that also gave rise to the domains in some fungal species. Fungi are also distributed in multiple clades on this tree (Supplementary Data S1). The DUF3494 domains we recovered from the bacterial genomes (above; see Methods) were grouped with their closely related species within the

corresponding phylum (i.e. Proteobacteria and Bacteroidetes/Chlorobi) in distinct clades, indicating that they are indeed prokaryotic. These results indicate that all ice-binding domains in dinoflagellates share a single common origin likely from a Proteobacteria or Bacteroidetes/Chlorobi source, and that those specific to *P. glacialis* have a distinct evolutionary history that may reflect niche specialisation.

Discussion

We generated two draft *de novo* diploid assemblies of *P. glacialis*, the first of any free-living psychrophilic dinoflagellates, and high-quality gene models supported by full-length transcriptomes. Genome features of *P. glacialis* provide a first glimpse of how genomes of dinoflagellates have evolved to adapt in a harsh environment. The difference in genome sizes between the two isolates highlights the extensive structural divergence of genomes within a dinoflagellate species. The abundance of repetitive elements and LTRs in the genomes suggests their important role in shaping the evolution of these genomes, potentially contributing to the genome-size difference. The exact molecular mechanisms and selective pressure that contribute to the larger genome size in the Antarctic isolate than in the Arctic isolate remains an open question, and can best be addressed using assembled genomes at chromosomal resolution. The trans-spliced DinoSL was thought to be a global signature of all transcripts in dinoflagellates, but our results reveal only a small proportion of full-length transcripts are encoded with DinoSL (and with relic DinoSLs) and remarkably, these transcripts mostly encode functions that are critical for adaptation to cold and to low-light conditions, both relevant to the natural habitat of *P. glacialis* in ice-brine channels. In addition, genes encoding these functions are unidirectionally encoded, often in a tandemly repeated single-exonic structure. This distinctive organisation of genes is likely a genome signature of free-living dinoflagellates, and may serve as a mechanism to enhance

transcription efficiency of genes encoding critical functions. The independently evolved ice-binding domains and the lineage-specific dark genes in *P. glacialis* highlight functional innovation in dinoflagellate genomes relevant to environmental adaptation and niche specialisation as successful psychrophiles in the extreme cold environment.

Materials & Methods

Cultures of *Polarella glacialis*

The cultures of *Polarella glacialis* isolates were acquired from the National Center for Marine Algae and Microbiota at the Bigelow Laboratory for Ocean Sciences, Maine, USA. Both cultures were maintained in f/2 medium without silica (Guillard et al. 1962) (100mL culture in 250mL conical flasks, 12h:12h light:dark cycle, 90 µmol photon·m^{-2}·s^{-1}, 4ºC). The cultures were treated with ampicillin (100 µg·mL^{-1}), kanamycin (50 µg·mL^{-1}) and streptomycin (50 µg·mL^{-1}) for 24 hours before cell harvest. For extraction of nucleic acids, the cells (50mL; >10^6 per mL) were harvested by centrifugation (3000 *g*, 5 min). The resulting cell pellet was rinsed with 0.22µm-filtered artificial seawater (Instant Ocean salt mixture, 33.3 g.L^{-1}; 1 mL), transferred to an 1.5mL-tube, and collected by further centrifugation (3000 *g*, 5 min). The supernatant (seawater) was removed, and the tube was immediately snap-frozen with liquid nitrogen and stored at –80ºC until DNA/RNA extraction.

Extraction of genomic DNA and total RNA

Genomic DNA was extracted following the 2xCTAB protocol with modifications. The cells were suspended in a lysis extraction buffer (400 µL; 100 mM Tris-Cl pH 8, 20 mM EDTA pH 8, 1.4 M NaCl), before silica beads were added. In a freeze-thaw cycle, the mixture was vortexed at high speed (2 min), and immediately snap-frozen in liquid nitrogen; the cycle was repeated 5 times. The final volume of the mixture was made up to 2% w/v CTAB (from 10%

w/v CTAB stock; kept at 37 °C). The mixture was treated with RNAse A (Invitrogen; final concentration 20 µg/mL) at 37°C (30 min), and Proteinase K (final concentration 120 µg/mL) at 65°C (2 h). The lysate was then subjected to standard extractions using equal volumes of phenol:chloroform:isoamyl alcohol (25:24:1 v/v; centrifugation at 14,000 g, 5 min, RT), and chloroform:isoamyl alcohol (24:1 v/w; centrifugation at 14,000 g, 5 min, RT). DNA was precipitated using pre-chilled isopropanol (gentle inversions of the tube, centrifugation at 18,000 *g*, 15 min, 4 °C). The resulting pellet was washed with pre-chilled ethanol (70% v/v), before stored in Tris-HCl (100 mM, pH 8) buffer. Total RNA was extracted using RNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. The concentration of DNA or RNA was determined with NanoDrop (Thermo Scientific), and a sample with A230:260:280 ≈ 1.0:2.0:1.0 was considered appropriate for sequencing.

Generation of genome data

For generation of short-read sequence data, samples of genomic DNA were sent for sequencing using the Illumina technology, using the HiSeq2500 (Australian Genome Research Facility, Melbourne), and HiSeq4000 (Translational Research Institute and the Australian Genome Research Facility, Brisbane) platforms (Supplementary Table S16). For each isolate, two paired-end TruSeq libraries (inserts of ~250bp and 600bp for HiSeq2500; ~350bp and ~600bp for HiSeq4000), and three mate-pair Nextera libraries (inserts of ~2, 5 and 10Kb) were generated for sequencing (in 2x150 bases). In total, we generated 399.3 Gbp of Illumina short-read sequencing data for CCMP1383, and 746.7 Gbp for CCMP2088 (Supplementary Table S16). All sequencing data are available upon request.

For generation of long-read data, samples of genomic DNA were sent for SMRT sequencing using the PacBio Sequel platform available at the Queensland University of Technology Central Analytical Research Facility, and at the Ramaciotti Centre of Genomics (University

of New South Wales, Sydney). In total, 15 SMRT cells were sequenced for CCMP1383 producing 9.3 million subreads (74.6 Gbp), and 7 SMRT cells for CCMP2088 generating 4 million subreads (35.9 Gbp); see Supplementary Table S2 for detail.

Generation of transcriptome data (RNA-Seq)

For generation of RNA-Seq data, total RNA samples were sent for sequencing at the Australian Genome Research Facility (Brisbane) using the Illumina HiSeq4000 platform. Illumina paired-end (2x150 bp reads) RNA-Seq data was generated for both CCMP1383 (55.4 Gbp) and CCMP2088 (61.7 Gbp); see Supplementary Table S17 for details.

Generation of full-length transcript data (PacBio IsoSeq)

Using the extracted total RNA samples (above), a full-length cDNA library was constructed for each of CCMP1383 and CCMP2088 using the TeloPrime Full-Length cDNA Amplification Kit (Lexogen, Vienna) following the kit manual. Two cDNA synthesis reactions were carried out in parallel for each sample, with 2 µg of total RNA used as starting material in each reaction. Double-stranded cDNA resulting from the two reactions was combined before performing PCR amplification using the TeloPrime PCR Add-on Kit (Lexogen, Vienna). For each sample, 16 parallel PCRs were carried out using 22 amplification cycles and 2 µL of double-stranded cDNA per reaction as template; the PCR products were pooled together and then split into two fractions, which were purified using 1x and 0.5x AMPure PB beads (Pacific Biosciences, California), respectively, and pooled at equal molarity. For sample CCMP1383, a total of 2.52 µg of purified full-length cDNA was obtained and was used for PacBio SMRTbell library preparation with the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences, California); the library was sequenced on 4 SMRT cells v2 LR using 20-hour movies on a Sequel platform at the Institute for Molecular Bioscience Sequencing Facility (University of Queensland, Brisbane). For CCMP2088, 1.95

ng of cDNA were obtained and submitted to The Ramaciotti Centre for Genomics (University of New South Wales, Sydney) for SMRTbell library preparation and sequencing on a PacBio Sequel System, also using 4 SMRT Cells v2 LR and 20-hour movies.

In addition to the cDNA libraries described above, a spliced-leader-specific transcript library was generated for CCMP1383. Four parallel PCR reactions were performed with the TeloPrime PCR Add-on Kit (Lexogen, Vienna) using 12 amplification cycles, conserved spliced leader fragment (5′-CCGTAGCCATTTTGGCTCAAG-3′) as forward primer, TeloPrime PCR 3′ primer as reverse primer, and 2 µL of double-stranded cDNA synthesised using the TeloPrime Full-Length cDNA Amplification Kit (Lexogen, Vienna) as template. The PCR products were pooled and purified (same method as above), resulting in 987 ng of cDNA. SMRTbell library construction was carried out using the PacBio SMRTbell Template Prep Kit 1.0, followed by sequencing on 2 SMRT cells v2 LR using 20-hour movies on the Sequel at the Institute for Molecular Bioscience Sequencing Facility (University of Queensland, Brisbane). Data yield from each SMRT cell is detailed in Supplementary Table S18.

Processing of sequencing data

Adaptor sequences were removed and low quality bases trimmed from paired-end reads using Trimmomatic v0.35 (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:30 MINLEN:50) (Bolger et al. 2014), overlapping read pairs (250 bp insert size) were merged using FLASH v1.2.11 (max-overlap 85) (Magoc et al. 2011). Mate-pair reads were processed using the NextClip v1.3 pipeline (Leggett et al. 2014) using the preliminary CLC assembly as a reference. Only the category A, B and C mate-pair reads were retained from the NextClip analysis. Further trimming of the mate-pair reads to remove low quality regions and adapters

was performed using Trimmomatic v0.35 (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:20 MINLEN:25).

Trimmed paired-end reads were mapped using bowtie2 (Langmead et al. 2012) against the initial CLC assembly and the mean insert size and standard deviation were computed using Picard tools v2.6.0 "CollectInsertSizeMetrics". Mate-pair sequences were aligned only against scaffolds from the initial assembly with a length >15 Kbp using bbmap (rcs=f pairedonly=t ambig=toss). The different approach taken for the mate-pair reads was to discard ambiguously mapped reads (i.e. reads that map equally well to multiple locations), as they have a more pronounced effect on inset size estimation with mate-pair data. The maximum insert size set during the alignment stage of both mate-pair and paired-end libraries was double the maximum expected insert size.

Illumina RNA-Seq data was trimmed for adapters and low-quality regions using Trimmomatic v0.35 (LEADING:10 TRAILING:10 SLIDINGWINDOW:4:30 MINLEN:50). PacBio IsoSeq data from each SMRT cell was polished (--polish) using the circular consensus sequencing tool (ccs v3.1.0 https://github.com/PacificBiosciences/unanimity/blob/develop/doc/PBCCS.md); only polished reads with a quality >0.99 were retained. Primers were removed from the polished reads using lima v1.8.0 (https://github.com/pacificbiosciences/barcoding) in IsoSeq mode. Only reads with the correct 5-prime/3-prime primer configuration were retained. The PacBio IsoSeq tool v3.1.0 was used to remove concatemers (using the refine option) from the primer trimmed reads (https://github.com/PacificBiosciences/IsoSeq3/blob/master/README_v3.1.md).

Genome-size estimation using *k*-mers

The *k*-mers frequency distribution in the trimmed paired-end reads (including merged) was used to estimate genome size and to assess ploidy. Genome size estimation was conducted following the approach described in Liu et al. (2018). The enumeration of *k*-mers was performed using Jellyfish (Marcais et al. 2011) at k = 17, 19, 21, 23, 25, 27, 29 and 31. For diploid genomes, a bimodal *k*-mer-count distribution is expected, genome size estimated from the first peak represents the diploid state, and that estimated from the second peak represents the haploid state. The standard theoretical model of a diploid genome in GenomeScope (Vurture et al. 2017) was used (*k*-mer size 21) to verify the diploidy observed in the sequence data (Supplementary Figure S1).

De novo genome assemblies

Initial *de novo* assemblies for CCMP1383 and CCMP2088 were generated independently using CLC Genomics Workbench (v7.5) (default parameters), incorporating all trimmed paired-end reads (using merged reads where applicable). The initial CLC assemblies was further processed using the Redundans package (retrieved 10 March 2019) (Pryszcz et al. 2016) using the trimmed paired-end and mate-pair reads.

Final genome assemblies for each isolate were generated with MaSuRCA v3.2.8 (Zimin et al. 2017) using untrimmed paired-end reads, trimmed mate-pair reads, and PacBio reads (>5 Kbp). For each isolate, the parameters of estimated assembly size in MaSuRCA was set based on the average estimated haploid genome size (in Supplementary Table S3), ploidy was set to two. Scaffolds < 1Kb were discarded from the final assembly.

Identification and removal of archaeal, bacterial and viral sequences

Identification and removal of contaminant sequences in the genome assemblies was assessed using similar method to Aranda et al. (Aranda et al. 2016). Genome scaffolds were compared using BLASTN against a database of archaeal, bacterial and viral genome sequences retrieved from RefSeq (release 88). Scaffolds were retained, and considered non-contaminant, if ≤10% of their length was covered by BLAST hits with a bit score >1000 and $E \leq 10^{-20}$.

Identification and removal of organellar sequences

The coding sequences from the plastid genome of *Cladocopium* sp. C3 (formerly *Symbiodinium* subtype C3) were used to identify putative plastid sequences from the assembly (Barbrook et al. 2014). A scaffold was considered putative plastidic if it shared significant sequence similarity (BLASTN) to one of the above sequences, covering >75% the sequence length at $E \leq 10^{-10}$.

The complete CDS of *cox1*, *cox3* and *cob* from *Breviolum minutum* (LC002801.1 and LC002802.1) were retrieved (as no complete sequences yet exist for *Polarella glacialis*) and used to identify putative mitochondrial scaffolds. A scaffold was considered putative mitochondrial if it shared significant similarity (BLASTN, max_target_seqs 10000) to one of the above sequences, covering >75% of the sequence length at $E \leq 10^{-10}$.

Customised gene prediction workflow tailored for dinoflagellate genomes

An *ab initio* gene prediction approach similar to Liu et al. (2018) was applied to the genomes of *P. glacialis*. For each genome assembly, a *de novo* repeat library was first derived using RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler/). All repeats

(including known repeats in RepeatMasker database release 20170127) were masked using RepeatMasker v4.0.7 (http://www.repeatmasker.org/).

We used transcriptome data generated in this study to guide gene prediction of assembled genomes. For RNA-Seq data, we assembled the reads using Trinity (Grabherr et al. 2011) independently in "de novo" mode (v2.6.6) and "genome-guided" mode (v2.8.4). The combined Trinity assemblies were trimmed using SeqClean (https://sourceforge.net/projects/seqclean/). The RNA-Seq and polished PacBio IsoSeq transcripts were combined into gene assemblies using PASA v2.3.3 (Haas et al. 2003) that was customised (available at http://smic.reefgenomics.org/download/) to recognise an additional donor splice site (GA). TransDecoder v5.2.0 (Haas et al. 2003) was used to predict open reading frames on the PASA assembled transcripts. Complete proteins (CDS with both start and stop codons) predicted by TransDecoder that had valid genome coordinates and more than one exon were retained for further analysis.

These proteins were searched (BLASTP, $E \leq 10^{-20}$) against a customised protein database that consists of RefSeq proteins release 88 and other predicted Symbiodiniaceae and *Polarella* proteins (Supplementary Table S19). Only nearly full-length proteins were included in the subsequent analysis; we defined nearly full-length proteins as sequences with a BLAST hit that covered >80% of both the query and subject sequences.

The nearly full-length gene models were checked for TEs using HHblits v2.0.16 (Remmert et al. 2011) (-p 80 -e 1E-5 -E 1E-5) searching against the JAMg transposon database (https://sourceforge.net/projects/jamg/files/databases/), as well as with Transposon-PSI (http://transposonpsi.sourceforge.net/). Gene models containing TEs were removed from the gene set, and redundancy reduction was conducted using CD-HIT v4.6.8 (Li et al. 2006)

(ID = 75%; -c 0.75 -n 5). The remaining gene models were processed using the Prepare_golden_genes_for_predictors.pl (http://jamg.sourceforge.net/) script from the JAMg pipeline (altered to recognise GA donor splice sites). This script produces a set of "golden genes", which were used as a training set for the gene-prediction tools AUGUSTUS v3.3.1 (Stanke et al. 2006) and SNAP version 2006-07-28 (Korf 2004). We used a customised code of AUGUSTUS (available at http://smic.reefgenomics.org/download/) so it recognises GA donor splice sites, and trained it to predict both coding sequences and untranslated regions; SNAP was trained for both GT and GC donor splice sites. Soft-masked genomes were passed to GeneMark-ES (Lomsadze et al. 2005) for training and gene prediction.

UniProt-SwissProt (retrieved 27/06/2018) proteins and other predicted Symbiodiniaceae and *Polarella* proteins (Supplementary Table S19) were combined to produce a set of gene models using MAKER v2.31.8 (altered to recognise GA donor splice sites) (Holt et al. 2011) in protein2genome mode; the custom repeat library was used by RepeatMasker as part of MAKER prediction. Two sets of predicted protein coding genes, one derived using the RNA-Seq data and one using the IsoSeq data, were constructed using PASA (--ALT_SPLICE -N 2) and TransDecoder (ORF prediction guided by Pfam database release 31). Gene models constructed using the IsoSeq data were assumed to be full-length and an extra step was taken to correct predicted proteins produced by TransDecoder that were five-prime partial. If a protein had an in-frame start codon within either the first 30 position or the first 30% of the sequence, that position was then considered as the start of that sequence. Sequence not satisfying these criteria were left unchanged. A primary set of predicted genes was produced using EvidenceModeler v1.1.1 (Haas et al. 2008), which had been altered to recognise GA donor splice sites. This tool combined the gene models from PASA RNA-Seq, PASA IsoSeq (with corrected start positions where applicable), AUGUSTUS, MAKER protein2genome and GeneMark-ES, into a single set of evidence-based predictions. EvidenceModeler was

allowed to predict genes within introns of other genes if the intron was >10,000 bp (--search_long_introns 10000).

Unlike Liu et al. (2018), we did not incorporate gene predictions from the SNAP program into the EvidenceModeler stage of the prediction workflow. This was done because SNAP produced an excessive number of overlapping genes that were on encoded on opposite strands. As genes encoded in this manner were not found to be supported in the transcriptome, we decided to exclude the results of this program from our predictions. We did not provide the location of putative repetitive elements to EvidenceModeler either, as multi-copy genes are often classified as repeats by RepeatModeler and would have been excluded from our final gene set. The weightings used for integration of gene models with EvidenceModeler were: PASA IsoSeq (with corrected start sites) 15, PASA RNA-Seq 10, Maker protein2genome 4, AUGUSTUS 1 and GeneMark-ES 1. EvidenceModeler gene models were considered high-confidence if they had been constructed using evidence from either PASA inputs or from ≥2 other prediction methods.

The transcriptome support shown in Supplementary Table S11 was calculated for each *P. glacialis* isolate by searching the high-confidence EvidenceModeler genes against a database of all RNA-Seq and IsoSeq transcripts (from the same isolate) using BLASTN. Genes were considered to have transcriptome support if they had a hit with >90% identity that covered >50% of the gene.

Functional annotation of predicted genes

Protein domains were searched using pfam_scan.pl (v1.6; Pfam database release 31) at *E*-value < 0.001 following earlier studies (Gonzalez-Pech et al. 2017, Shoguchi et al. 2018, Stephens et al. 2018). Where required, proteins were queried using BLASTP against

SwissProt and TrEMBL databases (UniProt release 2018_02) independently. Only the top 20 hits from each search were retained if $E \leq 10^{-5}$.

Pfam domains in *P. glacialis* was assessed for enrichment against a background set using Fisher's exact test, with correction for multiple testing using the Benjamini and Hochberg method (R Core Team 2015). GO enrichment was conducted using the topGO R (v2.34.0) (Alexa et al. 2010) package, applying the Fisher's Exact test with the 'elimination' methods to correct for the hierarchical structure of GO terms. The background used consisted of the available Symbiodiniaceae genomes: *Symbiodinium microadriaticum* (Aranda et al. 2016), *Breviolum minutum* (Shoguchi et al. 2013), *Cladocopium goreaui* (Liu et al. 2018), *Fugacium kawagutii* (Liu et al. 2018), *Symbiodinium tridacnidorum* (Shoguchi et al. 2018) and *Cladocopium* sp. C92 (Shoguchi et al. 2018).

Analysis of completeness of assembled genomes and predicted proteins

Completeness of the predicted genes in *P. glacialis* was assessed using BUSCO v3.1.0 (--mode proteins) (Simao et al. 2015) with the alveolate_stramenophiles_ensembl, Eukaryota_odb9 and protists_ensembl datasets (retrieved 22 September 2017), BLASTP searches ($E \leq 10^{-5}$) using the same three BUSCO datasets and BLASTP searches ($E \leq 10^{-5}$) using the protein orthologs from the Core Eukaryotic Genes dataset (Parra et al. 2009) (Supplementary Table S4).

Completeness of the assembled genomes of *P. glacialis* was assessed using BUSCO v3.1.0 (--mode proteins) (Simao et al. 2015) and TBLASTN searches ($E \leq 10^{-5}$) using the same three BUSCO datasets and TBLASTN searches ($E \leq 10^{-5}$) using the protein orthologs from the Core Eukaryotic Genes dataset (Parra et al. 2009) (Supplementary Table S4). The modified version of Augustus used for gene prediction was used for the BUSCO analysis as well.

Identification of *P. glacialis* sequences in the TARA database

The Ocean Microbial Reference Gene Catalog was retrieved from ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz. Genes classified as being from "Dinophyceae" or that were from the kingdom "undef" were extracted and searched against the genome of both *P. glacialis* isolates using BLASTN. Genes were retained if they had a hit to the genome that covered >75% of their length and with >95% identity. The retained hits were searched against the nr database from NCBI using the online BLASTN tool (20/05/2019).

Comparison of predicted proteins and genome-sequence similarity between *P. glacialis* isolates

Comparison between the protein sequences of CCMP1383 and CCMP2088 was conducted using BLASTP ($E \leq 10^{-5}$; Figure 3). For each isolate, protein sequences that do not share similarity to those of the counterpart isolate were identified. For these proteins, the corresponding coding gene sequences were searched (BLASTN) against the transcripts of the counterpart isolate; we consider a shared sequence similarity of >90% identity covering >50% of the query as significant.

Sequence similarity between the genomes of the *P. glacialis* isolates was assessed using non-repeat regions of the genome. Repeat features predicted using RepeatModeler and RepeatMasker were excluded from the analysis; regions between repeats that were ≤10 bp of each other were also removed. From the remaining non-repetitive regions, only those ≥100bp and with ≤10 ambiguous ("N") bases were used as query in a BLASTN (-dust no, $E \leq 10^{-10}$) search against the genome of the other isolate. The top hit of each sequence was retained for further analysis.

Inference of homologous protein sets among Suessiales

Putatively homologous protein sets were constructed using OrthoFinder v2.3.3 (inflation 1.5) (Emms et al. 2019) with sequence similarity computed with DIAMOND v0.9.24 (Buchfink et al. 2014). We defined dark homologous protein sets using the same criteria as in (Stephens et al. 2018) but excluding hits from any sequences with functions described as "Uncharacterized".

Functional classification of rhodopsin

Predicted proteins of *P. glacialis* with top hits described as "Rhodopsin" in UniProt were retrieved. The specific type for each identified *P. glacialis* rhodopsin was identified using a BLASTP ($E \leq 10^{-5}$) search against the known proton-pump (ABV22426, ADY17811) and sensory type (ADY17810, KF651052, KF651053, KF651054, KF651055) sequences from *Oxyrrhis marina.* The top hit for each query sequence was used to assign type.

Phylogenetic inference of DUF3494 domains

A comprehensive set of DUF3494 domain-encoding proteins was collected from the transcriptomes of *Heterocapsa arctica* CCMP445, *Peridinium aciculiferum* PAER_2, *Scrippsiella hangoei* like-SHHI_4 and *Scrippsiella hangoei* SHTV5 (retrieved from Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al. 2014)) for comparison against those predicted in both *P. glacialis* isolates. Predicted DUF3494-encoding proteins from the bacteria associated with *P. glacialis* were found with pfam_scan.pl (see above). DUF3494 domain regions were extracted from the proteins if they covered >50% the length of the Pfam DUF3494 domain HMM. DUF3494 domains from the Pfam_Full dataset (retrieved 14 April 2019) were retrieved. Identical sequences within each dataset were removed using cd-hit (-c 1.00 -n 5)(Li et al. 2006). All DUF3494 domains and domain regions were aligned using MAFFT v7.407 (--localpair --maxiterate 1000) (Katoh et

al. 2013), from which a Maximum Likelihood tree was constructed using IQ-TREE v1.6.10 (-m MFP -msub nuclear -bb 2000 -nm 2000) (Nguyen et al. 2015, Kalyaanamoorthy et al. 2017, Hoang et al. 2018). Support of nodes in the inferred tree was determined using 2000 ultrafast bootstraps (Hoang et al. 2018).

Analysis of simple repeats and multi-copy genes

The *de novo* repeat families identified by RepeatModeler during gene prediction were scrutinised for the presence of multi-copy genes. Unclassified repeat consensi (type unknown) were compared using BLASTN ($E \leq 10^{-5}$) against the final gene models. Queries (repeat consensi) with >80% of their sequence, or the sequence of the subject (predicted genes), being covered in the BLAST hit were retained. This strategy (considering cover of both query and subject) is designed to capture cases where either the whole repeat is contained within a gene (repetitive exon) or a whole gene in contained within a larger repeat.

To specifically assess the presence of simple repeats in the assembled genomes, RepeatMasker was re-run on each genome using the default library (RepeatMasker database release 20170127) searching for just simple repeats (-noint). Repeats of type (TTG)n, (TGT)n, (GTT)n, (AAC)n, (ACA)n, and (CAA)n are all derived from the same pattern and thus are considered interchangeable for the purposes of this study. Overlapping repeats of these types were merged and their total length was reported as the coverage of the TTG repeat. 3-mers were extracted from the cleaned genome assembly using kmercountexact.sh from the bbmaps tool suite (Supplementary Table S5). The quality trimmed and merged genome reads were sampled at 5% before 3-mers were extracted (reformat.sh samplerate=0.05, 3-mers extracted using kmercountexact.sh). This was done to prevent 3-mer counts from exceeding the maximum value a 32 bit integer can store.

Analysis of spliced leader sequences

Polished PacBio IsoSeq sequences that contained the dinoflagellate spliced leader sequence (CCGTAGCCATTTTGGCTCAAG) were identified using BLASTN (-max_target_seqs 1000000 -task blastn-short -evalue 1000). Only sequences with hits that start ≤5 bp from their 5′-end, ended ≥20 bp along the DinoSL sequence, had zero gap openings and a maximum of one mismatch were considered to contain the spliced leader sequence. Relic DinoSL sequences were identified by BLASTN (-max_target_seqs 1000000 -task blastn), using the full DinoSL and a relic sequence joined together as the query (Slamovits et al. 2008). Multiple relic DinoSL were identified using the full DinoSL and multiple relic DinoSL sequences joined together. Sequences were considered to contain a relic DinoSL if they had a hit that started within the first 11 bases of the relic DinoSL query sequence (allows for truncation of the transcript), within the first 5 bases of the transcript, and finished within 5 bases of the end of the relic DinoSL.

Acknowledgements

T.G.S. was supported by an Australian Government Research Training Program (RTP) Scholarship. This project was supported by two Australian Research Council grants (DP150101875 awarded to M.A.R., C.X.C. and D.B., and DP190102474 awarded to C.X.C. and D.B.), and the computational resources of the National Computational Infrastructure (NCI) National Facility systems through the NCI Merit Allocation Scheme (Project d85) awarded to C.X.C. and M.A.R.

Author contributions

T.G.S., M.A.R., and C.X.C. conceived the study; T.G.S., R.A.G.P., C.X.C., M.A.R. and D.B. designed the analyses and interpreted the results; C.X.C maintained the dinoflagellate

cultures; C.X.C. and A.R.M. extracted biological materials for sequencing; Y.C. generated the long-read libraries for genome and full-length transcriptome sequencing; T.G.S. conducted all computational analyses, prepared all figures and tables, and prepared the first draft of the manuscript; all authors prepared, wrote, reviewed, commented on and approved the final manuscript.

Data availability

The assembled genomes, predicted gene models and proteins from both *P. glacialis* isolates are available at: https://cloudstor.aarnet.edu.au/plus/s/Nx08JEMt7FjK3zY.

Figure Legends

Figure 1. Genomes of *Polarella glacialis* and repeat content. (A) Recovery of conserved core eukaryote genomes from CEGMA in the assembled *P. glacialis* genomes of CCMP1383 and CCMP2088 compared to the assembled genomes of *Cladocopium goreaui* and *Fugacium kawagutii* (Liu et al. 2018). (B) Interspersed repeat landscape and proportion of distinct repeat classes in the assembled genome of CCMP1383, relative to sequence divergence in Kimura substitution level. (C) Percentage of identified 3-mers in the assembled genome and the sequence data for CCMP1383 for the ten most-abundant 3-mers.

Figure 2. DinoSL-type full-length transcripts in *P. glacialis*. (A) Percentage of DinoSL-type transcripts of *P. glacialis* based on the identified start position along the DinoSL sequence, shown for positions 1 through 12. (B) Distribution of distances (in bp) between DinoSL-type transcriptional units shown for transcriptomes of CCMP1383 and CCMP2088.

Figure 3. Comparison of predicted gene models between the two *P. glacialis* genomes. The comparison of predicted proteins in CCMP1383 against those in CCMP2088 is shown, incorporating evidence from the corresponding transcriptome data.

Figure 4. Intergenic regions and tandemly repeated genes. (A) Distribution of the sizes of intergenic regions (in bp; ≤ 30,000 bp) shown for the assembled *P. glacialis* genomes of CCMP1383 and CCMP2088. (B) Number of tandemly repeated and/or single-exonic genes in CCMP1383 and CCMP2088, shown for genes encoding bacteriorhodopsin and peridinin chlorophyll *a*-binding proteins.

Figure 5. Evolutionary history of ice-binding domains in *P. glacialis* and dinoflagellates. Only a small part of the 1080-taxon maximum likelihood protein tree is shown. Support values, based on 2000 ultrafast bootstrap approximations, are shown at the internal nodes. Only values >50% are shown. The unit of branch length is the number of substitutions per site.

Supplementary Material

Supplementary Figure S1: GenomeScope 21-mer profile for (A) CCMP1383 and (B) CCMP2088.

Supplementary Figure S2: Interspersed repeat landscape and proportion of distinct repeat classes in the assembled genome of CCMP2088, relative to sequence divergence in Kimura substitution level.

Supplementary Figure S3: Relationship between length of intergenic regions and their coverage by repeats for the predicted genes from (A) CCMP1383 and (B) CCMP2088. The red trend line was constructed using a moving average with a window size of 250.

Supplementary Figure S4: An example of a genome region containing genes nested within the long introns of a putative alanine–tRNA ligase (from scaffold CCMP1383_scf7180000588947). The EvidenceModeler predicted genes, mapped IsoSeq transcripts and mapped RNA-Seq transcripts are shown in the green, red and blue boxes.

Supplementary Figure S5: Conserved synteny between the two sequenced bacterial scaffolds and the published (A) *Paraglaciecola psychrophila* strain 170T (GenBank NC_020514) and (B) *Sphingorhabdus* sp. YGSM121 (GenBank NZ_CP022548) genomes. Syntenic regions between the two sequences are shown with ribbons; red representing direct and green represents inverted regions.

Supplementary Tables S1 through S19

Supplementary Data S1

References

Alexa, A. and J. Rahnenführer (2010). "topGO: enrichment analysis for Gene Ontology." R package version 2.22.0.

Allen, J. R., M. Roberts, A. R. Loeblich, 3rd and L. C. Klotz (1975). "Characterization of the DNA from the dinoflagellate *Crypthecodinium cohnii* and implications for nuclear organization." Cell **6**(2): 161-169.

Arai, T., D. Fukami, T. Hoshino, H. Kondo and S. Tsuda (2019). "Ice-binding proteins from the fungus Antarctomyces psychrotrophicus possibly originate from two different bacteria through horizontal gene transfer." The FEBS Journal **286**(5): 946-962.

Aranda, M., Y. Li, Y. J. Liew, S. Baumgarten, O. Simakov, M. C. Wilson, J. Piel, H. Ashoor, S. Bougouffa, V. B. Bajic, T. Ryu, T. Ravasi, T. Bayer, G. Micklem, H. Kim, J. Bhak, T. C. LaJeunesse and C. R. Voolstra (2016). "Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle." Sci Rep **6**: 39734.

Bachvaroff, T. R. and A. R. Place (2008). "From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*." PLoS One **3**(8): e2929.

Baker, A. C. (2003). "Flexibility and specificity in coral-algal symbiosis: diversity, ecology, and biogeography of *Symbiodinium*." Annual Review of Ecology, Evolution, and Systematics **34**(1): 661-689.

Barbrook, A. C., C. R. Voolstra and C. J. Howe (2014). "The chloroplast genome of a *Symbiodinium* sp. clade C3 isolate." Protist **165**(1): 1-13.

Beauchemin, M., S. Roy, P. Daoust, S. Dagenais-Bellefeuille, T. Bertomeu, L. Letourneau, B. F. Lang and D. Morse (2012). "Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic." Proceedings of the National Academy of Sciences **109**(39): 15793-15798.

Beja, O., E. N. Spudich, J. L. Spudich, M. Leclerc and E. F. DeLong (2001). "Proteorhodopsin phototrophy in the ocean." Nature **411**(6839): 786-789.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Buchfink, B., C. Xie and D. H. Huson (2014). "Fast and sensitive protein alignment using DIAMOND." Nature Methods **12**: 59.

Chen, Y., T. G. Stephens, D. Bhattacharya, R. A. González-Pech and C. X. Chan (2019). "Evidence that inconsistent gene prediction can mislead analysis of algal genomes." bioRxiv: 690040.

de Mendoza, A., A. Bonnet, D. B. Vargas-Landin, N. Ji, H. Li, F. Yang, L. Li, K. Hori, J. Pflueger, S. Buckberry, H. Ohta, N. Rosic, P. Lesage, S. Lin and R. Lister (2018). "Recurrent acquisition of cytosine methyltransferases into eukaryotic retrotransposons." Nature Communications **9**(1): 1341.

Emms, D. M. and S. Kelly (2019). "OrthoFinder: phylogenetic orthology inference for comparative genomics." bioRxiv.

Fuhrman, J. A., M. S. Schwalbach and U. Stingl (2008). "Proteorhodopsins: an array of physiological roles?" Nat Rev Microbiol **6**(6): 488-494.

Galindo-Gonzalez, L., C. Mhiri, M. K. Deyholos and M. A. Grandbastien (2017). "LTR-retrotransposons in plants: Engines of evolution." Gene **626**: 14-25.

Gómez, F. (2012). "A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata)." Systematics and Biodiversity **10**(3): 267-275.

Gonzalez-Pech, R. A., M. A. Ragan and C. X. Chan (2017). "Signatures of adaptation and symbiosis in genomes and transcriptomes of *Symbiodinium*." Sci Rep **7**(1): 15021.

Gornik, S. G., Febrimarsa, A. M. Cassin, J. I. MacRae, A. Ramaprasad, Z. Rchiad, M. J. McConville, A. Bacic, G. I. McFadden, A. Pain and R. F. Waller (2015). "Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate." Proc Natl Acad Sci U S A **112**(18): 5767-5772.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. D. Zeng, Z. H. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman and A. Regev (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature Biotechnology **29**(7): 644-U130.

Grattan, L. M., S. Holobaugh and J. G. Morris, Jr. (2016). "Harmful algal blooms and public health." Harmful Algae **57**(Pt B): 2-8.

Guillard, R. R. L. and J. H. Ryther (1962). "Studies of marine planktonic diatoms: I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran." Canadian Journal of Microbiology **8**(2): 229-239.

Guo, Z., H. Zhang and S. Lin (2014). "Light-promoted rhodopsin expression and starvation survival in the marine dinoflagellate *Oxyrrhis marina*." PLoS One **9**(12): e114941.

Haas, B. J., A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith, L. I. Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg and O. White (2003). "Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies." Nucleic Acids Research **31**(19): 5654-5666.

Haas, B. J., S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R. Buell and J. R. Wortman (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments." Genome Biology **9**(1).

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh and L. S. Vinh (2018). "UFBoot2: Improving the Ultrafast Bootstrap Approximation." Mol Biol Evol **35**(2): 518-522.

Holt, C. and M. Yandell (2011). "MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects." BMC Bioinformatics **12**: 491.

Jaeckisch, N., I. Yang, S. Wohlrab, G. Glockner, J. Kroymann, H. Vogel, A. Cembella and U. John (2011). "Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*." PLoS One **6**(12): e28012.

John, U., Y. Lu, S. Wohlrab, M. Groth, J. Janouškovec, G. S. Kohli, F. C. Mark, U. Bickmeyer, S. Farhat, M. Felder, S. Frickenhaus, L. Guillou, P. J. Keeling, A. Moustafa, B. M. Porcel, K. Valentin and G. Glöckner (2019). "An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome." Science Advances **5**(4): eaav1110.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler and L. S. Jermiin (2017). "ModelFinder: fast model selection for accurate phylogenetic estimates." Nat Methods **14**(6): 587-589.

Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." Mol Biol Evol **30**(4): 772-780.

Keeling, P. J., F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. A. Amaral-Zettler, E. V. Armbrust, J. M. Archibald, A. K. Bharti, C. J. Bell, B. Beszteri, K. D. Bidle, C. T. Cameron, L. Campbell, D. A. Caron, R. A. Cattolico, J. L. Collier, K. Coyne, S. K. Davy, P. Deschamps, S. T. Dyhrman, B. Edvardsen, R. D. Gates, C. J. Gobler, S. J. Greenwood, S. M. Guida, J. L. Jacobi, K. S. Jakobsen, E. R. James, B. Jenkins, U. John, M. D. Johnson, A. R. Juhl, A. Kamp, L. A. Katz, R. Kiene, A. Kudryavtsev, B. S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A. M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M. A. Moran, S. Murray, G. Nadathur, S. Nagai, P. B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M. C. Posewitz, K. Rengefors, G. Romano, M. E. Rumpho, T. Rynearson, K. B. Schilling, D. C. Schroeder, A. G. Simpson, C. H. Slamovits, D. R. Smith, G. J. Smith, S. R. Smith, H. M. Sosik, P. Stief, E. Theriot, S. N. Twary, P. E. Umale, D. Vaulot, B. Wawrik, G. L. Wheeler, W. H. Wilson, Y. Xu, A. Zingone and A. Z. Worden (2014). "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing." Plos Biology **12**(6): e1001889.

Klinger, C. M., L. Paoli, R. J. Newby, M. Y.-W. Wang, H. D. Carroll, J. D. Leblond, C. J. Howe, J. B. Dacks, C. Bowler, A. B. Cahoon, R. G. Dorrell and E. Richardson (2018). "Plastid Transcript Editing across Dinoflagellate Lineages Shows Lineage-Specific Application but Conserved Trends." Genome Biology and Evolution **10**(4): 1019-1038.

Korf, I. (2004). "Gene finding in novel genomes." BMC Bioinformatics **5**: 59.

LaJeunesse, T. C., G. Lambert, R. A. Andersen, M. A. Coffroth and D. W. Galbraith (2005). "*Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates." Journal of Phycology **41**(4): 880-886.

LaJeunesse, T. C., J. E. Parkinson, P. W. Gabrielson, H. J. Jeong, J. D. Reimer, C. R. Voolstra and S. R. Santos (2018). "Systematic Revision of Symbiodiniaceae Highlights the Antiquity and Diversity of Coral Endosymbionts." Curr Biol **28**(16): 2570-2580 e2576.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Law, J. A. and S. E. Jacobsen (2010). "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nat Rev Genet **11**(3): 204-220.

Le Bescot, N., F. Mahe, S. Audic, C. Dimier, M. J. Garet, J. Poulain, P. Wincker, C. de Vargas and R. Siano (2016). "Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding." Environmental Microbiology **18**(2): 609-626.

Le, Q. H., P. Markovic, J. W. Hastings, R. V. Jovine and D. Morse (1997). "Structure and organization of the peridinin-chlorophyll a-binding protein gene in Gonyaulax polyedra." Mol Gen Genet **255**(6): 595-604.

Leggett, R. M., B. J. Clavijo, L. Clissold, M. D. Clark and M. Caccamo (2014). "NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries." Bioinformatics **30**(4): 566-568.

Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." Bioinformatics **22**(13): 1658-1659.

Liew, Y. J., Y. Li, S. Baumgarten, C. R. Voolstra and M. Aranda (2017). "Condition-specific RNA editing in the coral symbiont *Symbiodinium microadriaticum*." PLoS Genet **13**(2): e1006619.

Lin, S., S. Cheng, B. Song, X. Zhong, X. Lin, W. Li, L. Li, Y. Zhang, H. Zhang, Z. Ji, M. Cai, Y. Zhuang, X. Shi, L. Lin, L. Wang, Z. Wang, X. Liu, S. Yu, P. Zeng, H. Hao, Q. Zou, C. Chen, Y. Li, Y. Wang, C. Xu, S. Meng, X. Xu, J. Wang, H. Yang, D. A. Campbell, N. R. Sturm, S. Dagenais-Bellefeuille and D. Morse (2015). "The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis." Science **350**(6261): 691-694.

Lin, S., H. Zhang, D. F. Spencer, J. E. Norman and M. W. Gray (2002). "Widespread and extensive editing of mitochondrial mRNAS in dinoflagellates." J Mol Biol **320**(4): 727-739.

Lin, S., H. Zhang, Y. Zhuang, B. Tran and J. Gill (2010). "Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates." Proc Natl Acad Sci U S A **107**(46): 20033-20038.

Liu, H., T. G. Stephens, R. A. Gonzalez-Pech, V. H. Beltran, B. Lapeyre, P. Bongaerts, I. Cooke, M. Aranda, D. G. Bourne, S. Foret, D. J. Miller, M. J. H. van Oppen, C. R. Voolstra, M. A. Ragan and C. X. Chan (2018). "*Symbiodinium* genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis." Commun Biol **1**: 95.

Lohuis, M. R. and D. J. Miller (1998). "Hypermethylation at CPG-motifs in the dinoflagellates *Amphidinium carterae* (Dinophyceae) and *Symbiodinium microadriaticum* (Dinophyceae): Evidence from restriction analyses, 5-azacytidine and ethionine treatment." Journal of Phycology **34**(1 %@ 0022-3646): 152-159.

Lomsadze, A., V. Ter-Hovhannisyan, Y. O. Chernoff and M. Borodovsky (2005). "Gene identification in novel eukaryotic genomes by self-training algorithm." Nucleic Acids Res **33**(20): 6494-6506.

Magoc, T. and S. L. Salzberg (2011). "FLASH: fast length adjustment of short reads to improve genome assemblies." Bioinformatics **27**(21): 2957-2963.

Marcais, G. and C. Kingsford (2011). "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." Bioinformatics **27**(6): 764-770.

Marchetti, A., D. M. Schruth, C. A. Durkin, M. S. Parker, R. B. Kodner, C. T. Berthiaume, R. Morales, A. E. Allen and E. V. Armbrust (2012). "Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability." Proc Natl Acad Sci U S A **109**(6): E317-325.

McEwan, M., R. Humayun, C. H. Slamovits and P. J. Keeling (2008). "Nuclear genome sequence survey of the Dinoflagellate *Heterocapsa triquetra*." J Eukaryot Microbiol **55**(6): 530-535.

Montresor, M., C. Lovejoy, L. Orsini, G. Procaccini and S. Roy (2003). "Bipolar distribution of the cyst-forming dinoflagellate *Polarella glacialis*." Polar Biology **26**(3): 186-194.

Montresor, M., G. Procaccini and D. K. Stoecker (1999). "*Polarella glacialis*, gen. nov., sp. nov. (Dinophyceae): Suessiaceae are still alive!" Journal of Phycology **35**(1): 186-197.

Nguyen, L. T., H. A. Schmidt, A. von Haeseler and B. Q. Minh (2015). "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." Mol Biol Evol **32**(1): 268-274.

Parra, G., K. Bradnam, Z. Ning, T. Keane and I. Korf (2009). "Assessing the gene space in draft genomes." Nucleic Acids Res **37**(1): 289-297.

Ponmani, T., R. Guo and J.-S. Ki (2016). "Analysis of the genomic DNA of the harmful dinoflagellate *Prorocentrum minimum*: a brief survey focused on the noncoding RNA gene sequences." Journal of Applied Phycology **28**(1): 335-344.

Pryszcz, L. P. and T. Gabaldon (2016). "Redundans: an assembly pipeline for highly heterozygous genomes." Nucleic Acids Research **44**(12).

R Core Team (2015). "R: a language and environment for statistical computing."

Reichman, J. R., T. P. Wilcox and P. D. Vize (2003). "PCP gene family in *Symbiodinium* from *Hippopus hippopus*: Low levels of concerted evolution, isoform diversity, and spectral tuning of chromophores." Mol Biol Evol **20**(12): 2143-2154.

Remmert, M., A. Biegert, A. Hauser and J. Soding (2011). "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." Nat Methods **9**(2): 173-175.

Shi, X., L. Li, C. Guo, X. Lin, M. Li and S. Lin (2015). "Rhodopsin gene expression regulated by the light dark cycle, light spectrum and light intensity in the dinoflagellate *Prorocentrum*." Front Microbiol **6**: 555.

Shoguchi, E., G. Beedessee, I. Tada, K. Hisata, T. Kawashima, T. Takeuchi, N. Arakaki, M. Fujie, R. Koyanagi, M. C. Roy, M. Kawachi, M. Hidaka, N. Satoh and C. Shinzato (2018). "Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes." BMC Genomics **19**(1): 458.

Shoguchi, E., C. Shinzato, T. Kawashima, F. Gyoja, S. Mungpakdee, R. Koyanagi, T. Takeuchi, K. Hisata, M. Tanaka, M. Fujiwara, M. Hamada, A. Seidi, M. Fujie, T. Usami, H. Goto, S. Yamasaki, N. Arakaki, Y. Suzuki, S. Sugano, A. Toyoda, Y. Kuroki, A. Fujiyama, M. Medina, M. A. Coffroth, D. Bhattacharya and N. Satoh (2013). "Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure." Curr Biol **23**(15): 1399-1408.

Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics **31**(19): 3210-3212.

Slamovits, C. H. and P. J. Keeling (2008). "Widespread recycling of processed cDNAs in dinoflagellates." Curr Biol **18**(13): R550-552.

Slamovits, C. H., N. Okamoto, L. Burri, E. R. James and P. J. Keeling (2011). "A bacterial proteorhodopsin proton pump in marine eukaryotes." Nat Commun **2**: 183.

Sorhannus, U. (2011). "Evolution of antifreeze protein genes in the diatom genus *Fragilariopsis*: evidence for horizontal gene transfer, gene duplication and episodic diversifying selection." Evolutionary bioinformatics online **7**: 279-289.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern (2006). "AUGUSTUS: ab initio prediction of alternative transcripts." Nucleic Acids Res **34**(Web Server issue): W435-439.

Stentiford, G. D. and J. D. Shields (2005). "A review of the parasitic dinoflagellates *Hematodinium* species and *Hematodinium*-like infections in marine crustaceans." Dis Aquat Organ **66**(1): 47-70.

Stephens, T. G., M. A. Ragan, D. Bhattacharya and C. X. Chan (2018). "Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions." Sci Rep **8**(1): 17175.

Taylor, F. J. R., M. Hoppenrath and J. F. Saldarriaga (2008). "Dinoflagellate diversity and distribution." Biodiversity and Conservation **17**(2): 407-418.

Vance, T. D. R., M. Bayer-Giraldi, P. L. Davies and M. Mangiagalli (2019). "Ice-binding proteins and the 'domain of unknown function' 3494 family." FEBS J **286**(5): 855-873.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski and M. C. Schatz (2017). "GenomeScope: fast reference-free genome profiling from short reads." Bioinformatics **33**(14): 2202-2204.

Yin, J., J. Chen, G. Liu, Y. Yu, L. Song, X. Wang and X. Qu (2013). "Complete genome sequence of *Glaciecola psychrophila* strain 170T." Genome Announc **1**(3).

Zemach, A. and D. Zilberman (2010). "Evolution of eukaryotic DNA methylation and the pursuit of safer sex." Curr Biol **20**(17): R780-785.

Zhang, H., D. A. Campbell, N. R. Sturm and S. Lin (2009). "Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements." Mol Biol Evol **26**(8): 1757-1771.

Zhang, H., Y. Hou, L. Miranda, D. A. Campbell, N. R. Sturm, T. Gaasterland and S. Lin (2007). "Spliced leader RNA trans-splicing in dinoflagellates." Proc Natl Acad Sci U S A **104**(11): 4618-4623.

Zimin, A. V., D. Puiu, M. C. Luo, T. Zhu, S. Koren, G. Marcais, J. A. Yorke, J. Dvorak and S. L. Salzberg (2017). "Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm." Genome Res **27**(5): 787-792.

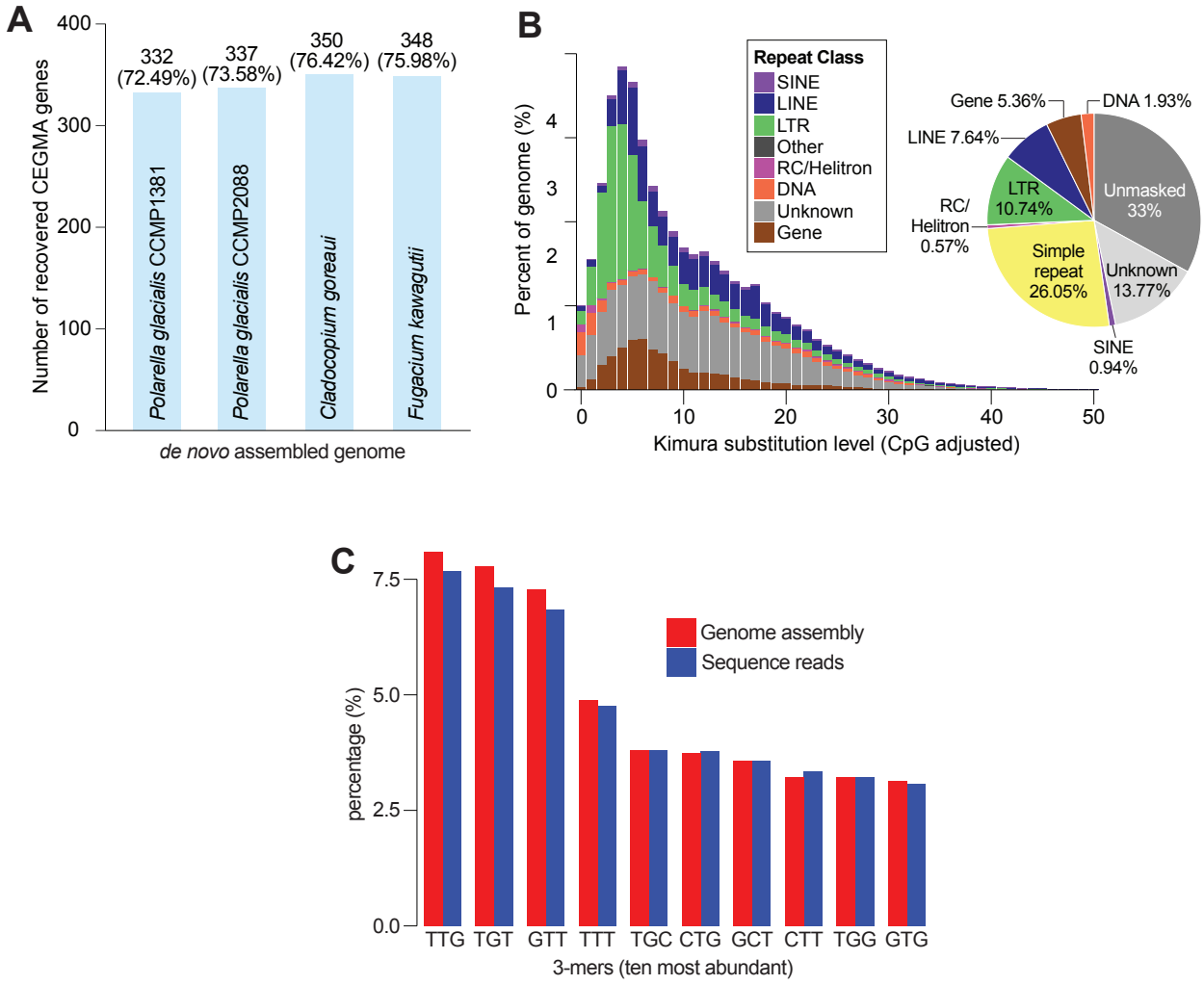


Figure 1. Genomes of *Polarella glacialis* and repeat content. (A) The recovery of conserved core eukaryote genomes from CEGMA in the assembled *P. glacialis* genomes of CCMP1383 and CCMP2088 compared to the assembled genomes of *Cladocopium goreaui* and *Fugacium kawagutii* (Liu et al. 2018). (B) Interspersed repeat landscape and proportion of distinct repeat classes in the assembled genome of CCMP1383, relative to sequence divergence in Kimura substitution level. (C) Percentage of identified 3-mers in the assembled genome and the sequence data for CCMP1383, showing for the ten most abundant 3-mers.

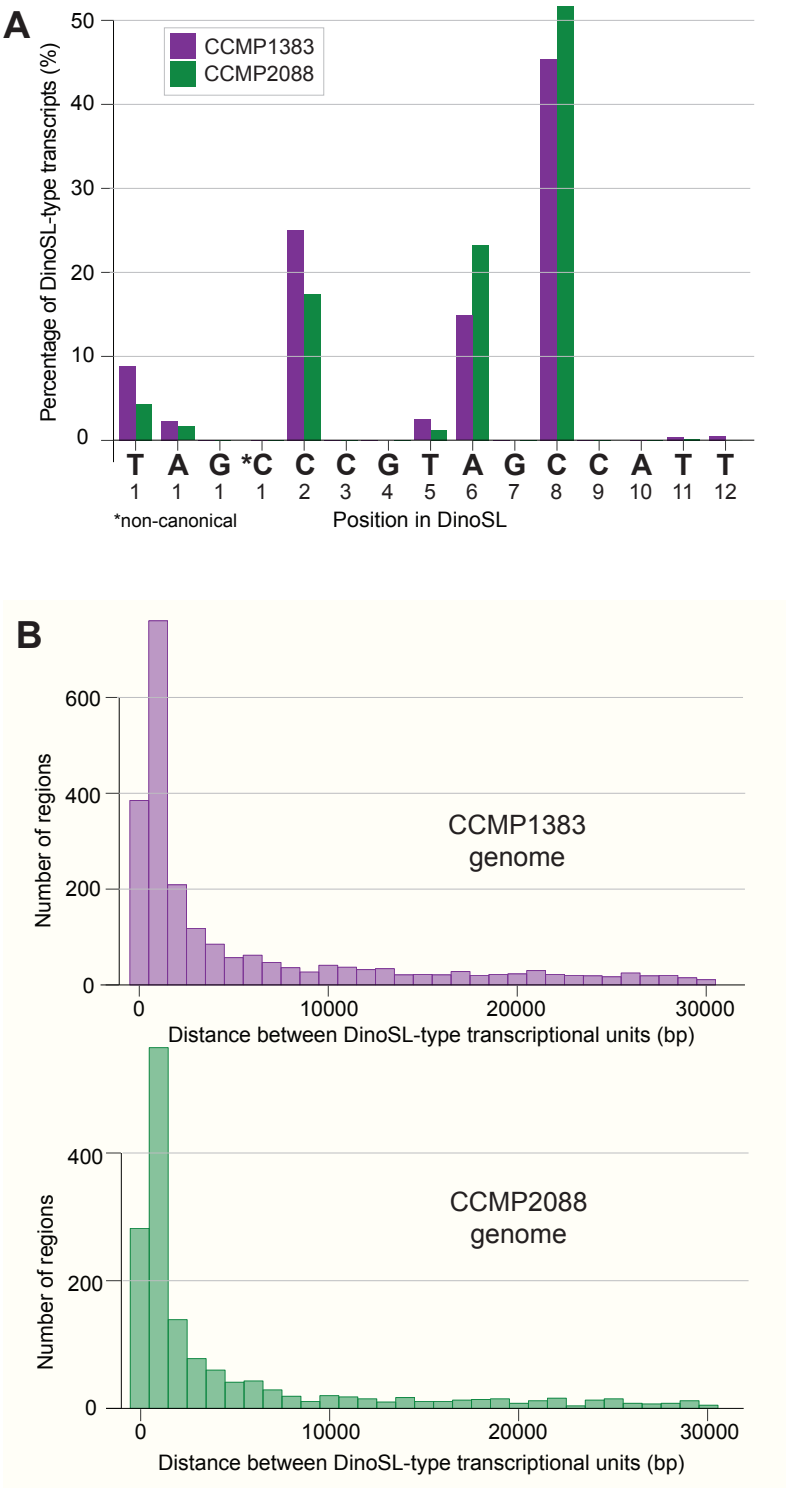


Figure 2. DinoSL-type full-length transcripts in *P. glacialis*. (A) Percentage of DinoSL-type transcripts of *P. glacialis* based on the identified start position along the DinoSL sequence, shown for positions 1 through 12. (B) Distribution of distances (in bp) between DinoSL-type transcriptional units shown for transcriptomes of CCMP1383 and CCMP2088.

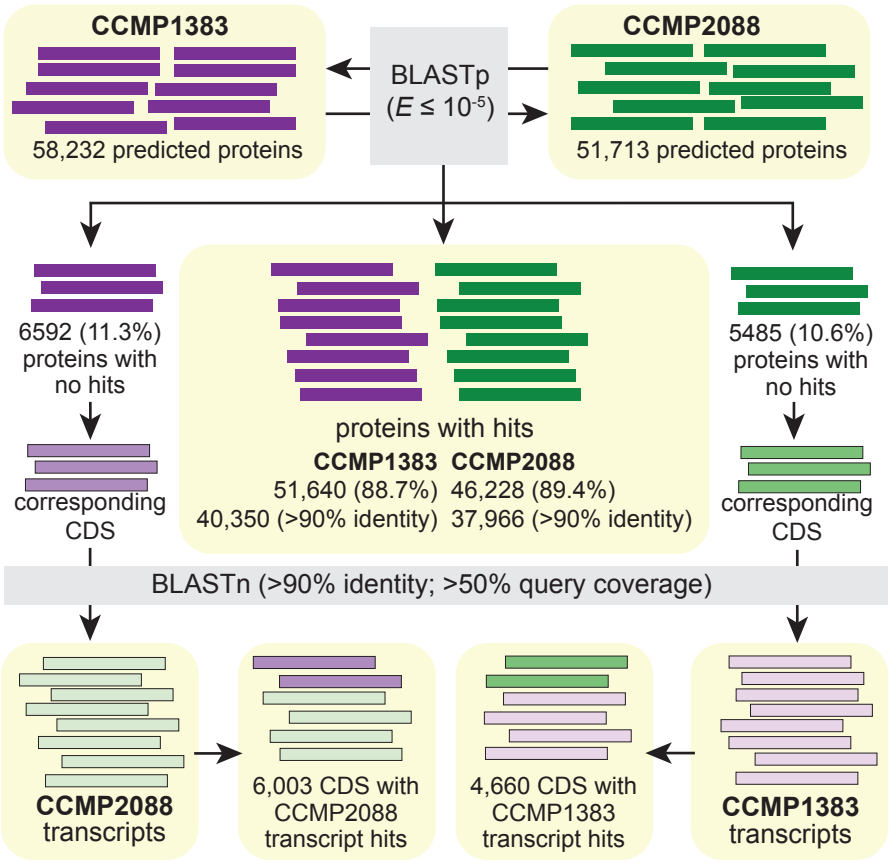
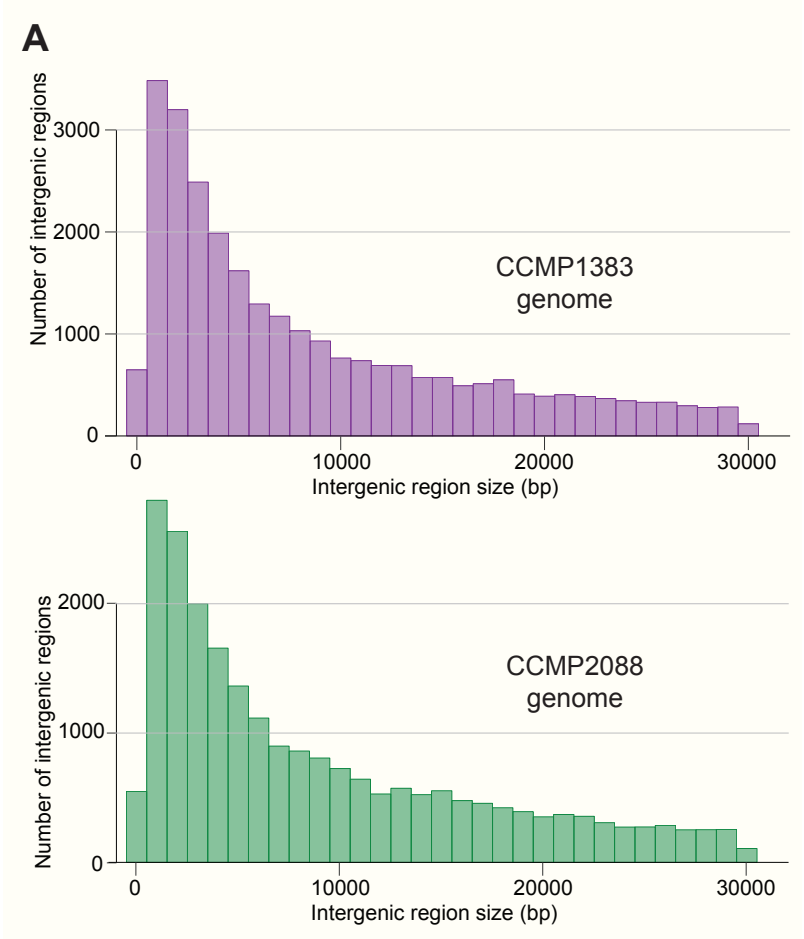


Figure 3. Comparison of predicted gene models between the two *P. glacialis* genomes. The comparison of predicted proteins in CCMP1383 against those in CCMP2088 is shown, incorporating evidence from the corresponding transcriptome data.



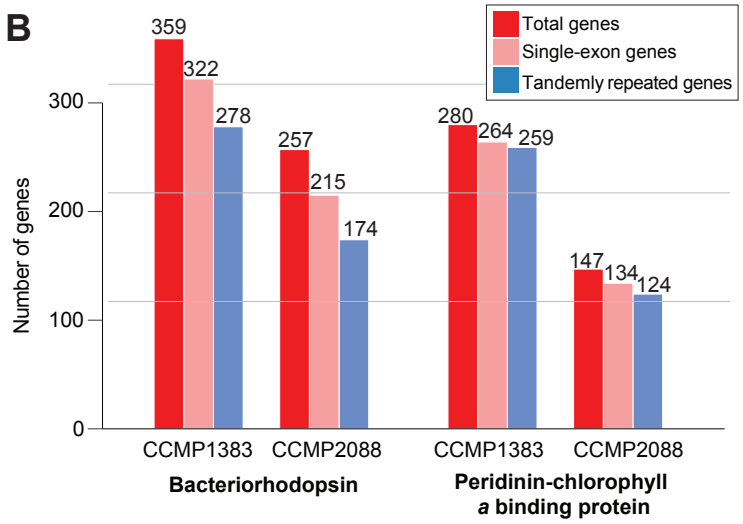


Figure 4. Intergenic regions and tandemly repeated genes. (A) Distribution of the sizes of intergenic regions (in bp; ≤ 30,000 bp) shown for the assembled *P. glacialis* genomes of CCMP1383 and CCMP2088. (B) Number of tandemly repeated and/or single-exonic genes in CCMP1383 and CCMP2088, shown for genes encoding bacteriorhodopsin and peridinin chlorophyll *a*-binding proteins.

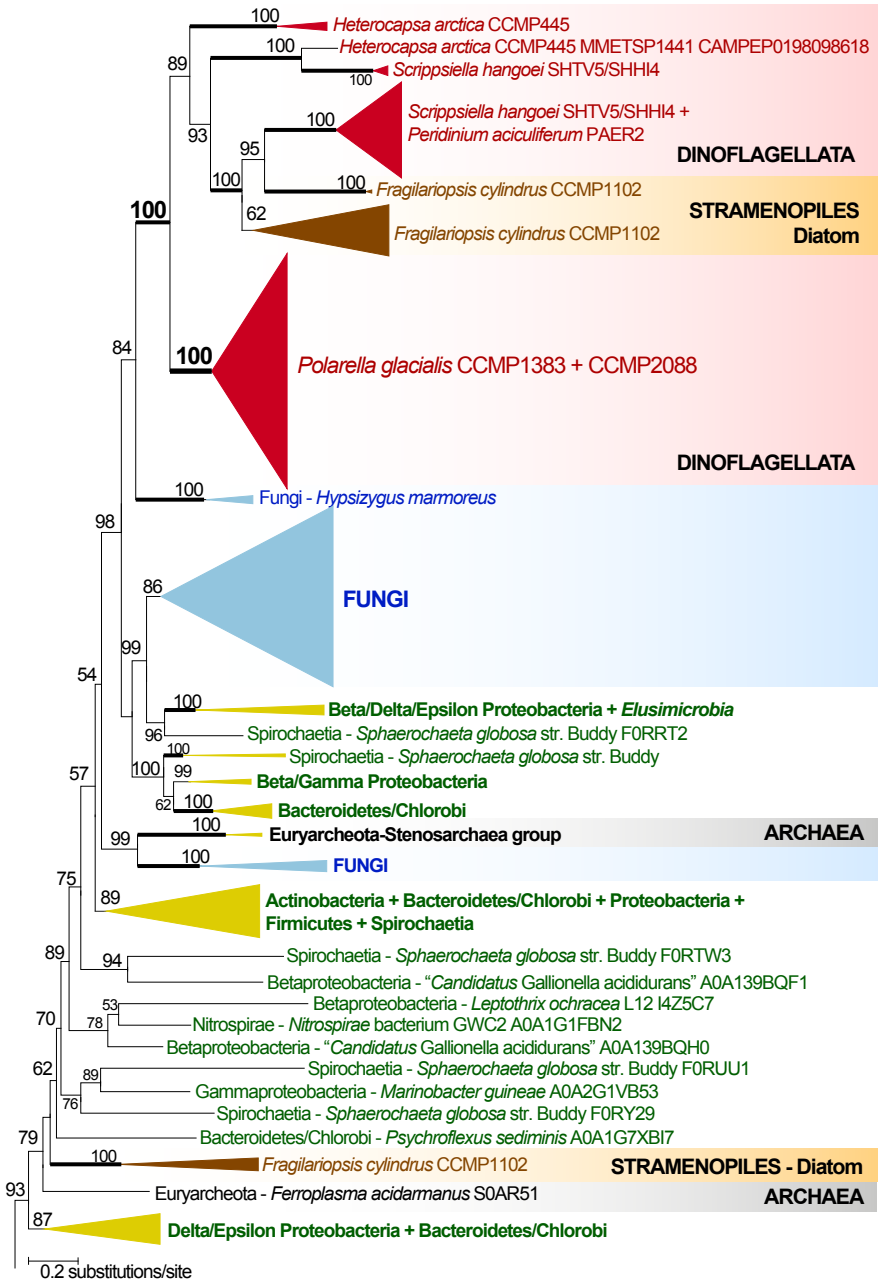


Figure 5. Evolutionary history of ice-binding domains in *P. glacialis* and dinoflagellates. Only a small part of the 1080-taxon maximum likelihood protein tree is shown. Support values, based on 2000 ultrafast bootstrap approximations, are shown at the internal nodes. Only values >50% are shown. The unit of branch length is the number of substitutions per site.