Integrated genomic and transcriptomic analysis of the peridinin dinoflagellate *Amphidinium carterae* plastid

Richard G. Dorrell[1,2*], R.Ellen R. Nisbet[1*], Adrian C. Barbrook[1], Steven J.L. Rowden[1], and Christopher J. Howe[1+]

[1]Department of Biochemistry, University of Cambridge

[2]Present address : Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Université Paris 75005 Paris, France

*contributed equally to this work

+to whom correspondence should be addressed: ch26@cam.ac.uk

The plastid genomes of peridinin-containing dinoflagellates are highly unusual, possessing very few genes, which are located on small chromosomal elements termed "minicircles". These minicircles may contain genes, or no recognisable coding information. Transcripts produced from minicircles may undergo unusual processing events, such as the addition of a 3' poly(U) tail. To date, little is known about the genetic or transcriptional diversity of non-coding sequences in peridinin dinoflagellate plastids. These sequences include empty minicircles, and regions of non-coding DNA in coding minicircles. Here, we present an integrated plastid genome and transcriptome for the model peridinin dinoflagellate *Amphidinium carterae*, identifying a previously undescribed minicircle. We also profile transcripts covering non-coding regions of the *psbA* and *petB/atpA* minicircles. We present evidence that antisense transcripts are produced within the *A. carterae* plastid, but show that these transcripts undergo different end cleavage events from sense transcripts, and do not receive 3' poly(U) tails. The difference in processing events between sense and antisense transcripts may enable the removal of non-coding transcripts from peridinin dinoflagellate plastid transcript pools.

Introduction

Much is known about the organisation and expression of plastid genomes (Green 2011; Barbrook, et al. 2018). The plastid genomes of photosynthetic eukaryotes retain fewer than

250 genes (Green 2011; Muñoz-Gómez 2017); these are typically organised as a single, circular chromosome, although some may have alternative linear or branched forms (Oldenburg and Bendich 2004; Barbrook, et al. 2010; Janouskovec, et al. 2013; Del Cortona, et al. 2017). Genes are usually arranged in operons, and are co-transcribed before being cleaved into mature mRNAs (Barkan 2011; Luro, et al. 2013; Hotto, et al. 2015). The plastid transcript processing machinery is also involved in the degradation of antisense transcripts (Hotto, et al. 2015; Castandet, et al. 2016). which may otherwise anneal to and inhibit the function of the corresponding sense transcripts (Sharwood, et al. 2011; Hotto, et al. 2012). These antisense transcripts are generated through transcription from specific promoters located on the plastid template strand (Hotto, et al. 2012; Zhelyazkova, et al. 2012), through inefficient transcript termination of polymerases transcribing genes that are in opposing transcriptional orientation (Rott, et al. 1996; Sharwood, et al. 2011), and through RNA-dependent RNA polymerase activity in the plastid (Zandueta-Criado and Bock 2004).

Perhaps the most bizarre example of plastid genome organisation is in the peridinin-containing plastids of dinoflagellate algae (Barbrook, et al. 2018). Dinoflagellates are ecologically abundant, and include free-living photosynthesisers, heterotrophs, and endosymbionts of other organisms (Lima-Mendez, et al. 2015; Suggett, et al. 2017). The majority of the phototrophic species possess a plastid derived from the endosymbiosis of red algae, containing the light-harvesting pigment peridinin (Haxo, et al. 1976; Dorrell, Klinger, et al. 2017). The genome of this plastid contains twelve-coding protein genes, two ribosomal RNA genes, and an erratic small number of transfer RNA genes (Barbrook, et al. 2014; Mungpakdee, et al. 2014), and is fragmented into small circular DNA molecules, termed "minicircles" (Zhang, et al. 1999; Howe, et al. 2008). Each minicircle contains one gene only, although minicircles with multiple genes are known in individual species (Nelson, et al. 2007; Barbrook, et al. 2012; Dorrell, Klinger, et al. 2017), alongside a non-coding "core region" that is broadly conserved across all minicircles in a species (Zhang, et al. 2002; Nisbet, et al. 2004; Mungpakdee, et al. 2014; Barbrook, et al. 2018). In addition "empty" minicircles and microcircles, which lack coding information, are known in individual dinoflagellate species (Zhang, et al 1999; Hiller 2001; Nisbet, et al. 2004).

The organisation of the peridinin dinoflagellate plastid genome has influenced the diversity of transcripts produced. Both coding and empty minicircle sequences are transcribed (Wang, et al. 2005; Nisbet, et al. 2008) through a "rolling circle" mechanism, which can yield transcripts containing multiple copies of plastid minicircle sequence (Dang and Green 2010;

Barbrook, et al. 2012). However, the predominant bands identified in northern blots of dinoflagellate plastid transcripts correspond to monocistronic mRNAs, which are presumably generated through the processing of these long precursors (Barbrook, et al. 2001; Nisbet, et al. 2008; Dang and Green 2009). The post-transcriptional processing events observed include cleavage to generate mature transcript 5' and 3' ends (Nelson, et al. 2007; Dang and Green 2010; Barbrook, et al. 2012; Dorrell, Hinksman, et al. 2016), extensive substitutional editing (Zauner, et al. 2004; Mungpakdee, et al. 2014; Klinger, et al. 2018), and the frequent addition of a 3' poly(U) tail to plastid transcripts (Wang and Morse 2006). Both of the last two processes appear to have evolved within dinoflagellates and their closest evolutionary relatives (e.g. *Chromera*, apicomplexans), independently to other algal chloroplast lineages (Janouskovec, et al. 2010; Dorrell and Howe 2012; Dorrell, et al. 2014; Nisbet, et al. 2016).

We have previously generated extensive plastid sequence data for the model peridinin dinoflagellate *Amphidinium carterae*, including minicircle sequences for all fourteen known dinoflagellate plastid protein-coding and rRNA genes (Barbrook and Howe 2000; Barbrook, et al. 2001; Nisbet, et al. 2004; Barbrook, et al. 2006; Barbrook, et al. 2018), together with detailed transcript sequence maps generated through RT-PCR and northern blotting of the coding sequences produced from specific plastid minicircles (Nisbet, et al. 2008; Barbrook, et al. 2012; Dorrell, Klinger, et al. 2017).

Here, we characterise the diversity of non-coding genomic elements and transcripts in the *Amphidinium* plastid. Using integrated genomic and transcriptomic data, we identify a novel, highly transcribed minicircle with no obvious coding content; and profile non-coding transcripts generated from the coding *psbA* and *petB/atpA* minicircles, focussing on transcripts covering minicircle core regions. We also identify the presence of antisense transcripts. Notably, we find that antisense transcripts do not undergo similar terminal cleavage events to those found in mature mRNAs, and lack poly(U) tails. We propose that the differential processing of sense and antisense transcripts in dinoflagellate plastids might indirectly enable the degradation of non-coding sequence degradation during plastid transcript processing.

Results

Coding and non-coding diversity of the Amphidinium carterae plastid genome

First, we wished to produce a definitive inventory of the coding and empty minicircle sequences present in the *Amphidinium carterae* plastid. For this, we performed next-generation sequencing of genomic DNA isolated from *A. carterae* CCMP1314, which had been separated on a caesium chloride gradient (to obtain plastid-enriched DNA) and treated with Plasmid-Safe DNAse (to select for minicircle DNA) to yield a plastid-enriched fraction (Fig. 1) (Griffith 1991; Lang and Burger 2007). We generated 10,120,479 reads, which we could assemble into 21 minicircle contigs (Table S1, Sheet 1).

We identified all of the previously identified coding minicircles of the *A. carterae* plastid: *psbA, psbB, psbC, psbD/psbE/psbI, petD, petB/atpA, atpB,* and the 23S minicircle and 16S minicircle (previously annotated as Minicircle 4) (Barbrook and Howe 2000; Barbrook, et al. 2001; Nisbet, et al. 2004; Barbrook, et al. 2006), with between 30 and 116-fold average read coverage (Fig. 1A). We additionally identified all four of the previously identified "empty" minicircles (Minicircles 1-3 and 5) in *A. carterae* (Nisbet, et al. 2004; Nisbet, et al. 2008). These minicircles had comparable (47-80-fold) average read depths to the coding minicircles identified, indicating they form a substantial component of plastid DNA (Fig. 1A). We confirmed that these four empty minicircles lack obvious protein homologues in other sequenced dinoflagellate genomes, and in transcriptome libraries from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP), supporting the idea that they lack coding function (Fig. S1A) (Barbrook, et al. 2014; Mungpakdee, et al. 2014; Dorrell, Klinger, et al. 2017; Klinger, et al. 2018). We could not identify distinct equivalents of the "microcircles" (empty minicircles of < 1kbp size) previously sequenced from *A. carterae*, in the genomic dataset, which reflects the fact that these frequently contain sequences identical and indistinguishable from fragments of coding and empty minicircles (Nisbet, et al. 2004; Barbrook, et al. 2018).

Alongside this, we identified one entirely novel minicircle, referred henceforth as "empty minicircle 6". This corresponded to a 2344 bp empty minicircle sequence, containing 47.7 % GC bases, a value comparable to all existing *A. carterae* minicircle sequences, and a core region including a 9-bp AGAGAAAAA motif conserved in all other minicircle sequences from this strain (Barbrook, et al. 2006). We confirmed the presence of this minicircle through PCR and Sanger sequencing from independently isolated *A. carterae* CCMP1314 gDNA (Fig. S1B). This minicircle was not found to contain any recognisable coding, rRNA or tRNA sequences,

either by BLAST or structural searches, nor was it found to contain any regions of sequence similarity to other *Amphidinium* CCMP1314 minicircles (apart from the core region). Instead, it contained an ORF of 163 codons, which is highly similar (>90%) in amino acid sequence to an ORF found on a minicircle previously identified in the related strain *Amphidinium* sp. CS-21 (Hiller, 2001), and also similar to a shorter (≥67 aa, C-terminal incomplete) ORF inferred to be encoded by the *Amphidinium massartii* MMETSP transcriptome (Dorrell, Klinger, et al. 2017) (Fig. S1B). However, the novel ORF generates no significant matches in BLAST searches of protein and translated nucleotide databases outside of the *Amphidinium* genus (Fig. S1B).We identified transcription of this minicircle in the *A. carterae* CCMP1314 MMETSP library (Fig. S1A), and through RT-PCR of random-hexamer primed cDNA, but could not identify transcripts from RT-PCRs of oligo-d(A) primed cDNA, using 12 minicircle-specific primers, indicating that this minicircle does not contain any poly(U) sites (Table S2, section 1, lines 1-18).

Transcriptomic mapping of the Amphidinium carterae plastid

Next, we wished to determine how abundant the transcripts were from the coding and non-coding sequences contained within the *A. carterae* plastid. We generated a plastid transcriptome for *A. carterae* CCMP1314 by Illumina sequencing of adaptor-conjugated RNA (Table S1, sheet 2). Each plastid gene was highly represented in the RNA sequence library, with averages of between 3130-fold coverage (for the 16S rRNA) and 581.8-fold coverage (for the photosystem II subunit *psbI*) for each of the recognised genes located in the *A. carterae* plastid genome (Barbrook, et al. 2001; Nisbet, et al. 2004; Barbrook, et al. 2006) (Figs. 1B, 1C). We observed very little relative difference in the transcript abundance of each plastid gene (Fig. 1C). This is in contrast to what is observed in other alveolate plastid transcriptomes (e.g. *Chromera velia* and *Karenia mikimotoi;* (Janouskovec, et al. 2013; Dorrell, Hinksman, et al. 2016)), where transcripts encoding the photosystem II reaction centre subunits *psbA* and *psbD* are far more abundant than those from other plastid genes. We notably found higher transcript read coverage on the novel empty minicircle 6 (332.9) compared to other empty minicircles (81.6-148.7; Fig. 1C), consistent with the fact that this was the only empty minicircle detectable in the *A. carterae* MMETSP transcriptome (Fig. S1C).

We noted varying patterns of read coverage for different minicircles, consistent with different patterns of transcript maturation (Fig. 1B; Fig. S2). For example, the *psbE, ORF1* and *psbI* genes, which are adjacent to one another in the *psbD/psbE/ORF1/psbI* minicircle,

showed similar levels of read coverage with no clear decrease in coverage over the intergenic region, reinforcing previous evidence that they are cotranscribed and processed to form a polycistronic transcript (Barbrook, et al. 2012) (Fig. 1B). We similarly observed largely continuous read coverage between *petD* and *ORF2* in the polycistronic *petD/ORF2/ORF3/ORF4* minicircle, consistent with previous RT-PCR data indicating that these genes are cotranscribed and possess a common poly(U) site, located downstream of ORF2 (Barbrook et al., 2012) (Fig. S2).

In contrast, *psbD* which has been shown to be processed to a monocistronic transcript, was covered by a clearly defined pattern of associated reads (Fig. 1B) (Nisbet, et al. 2008; Barbrook, et al. 2012). We similarly observed defined gaps in read coverage between *petB* and *atpA* in the polycistronic *petB/ atpA* minicircle; and between *petD/ORF2* and *ORF3* in the *petD/ORF2/ORF3/ORF4* minicircle consistent with previous northern and RT-PCR data indicating these genes are represented by separate mRNA populations (Barbrook et al., 2001; Barbrook et al., 2012) (Fig. S2).

Elsewhere, the mRNA sequencing data provided windows into the processing events associated with previously unexplored *Amphidinium* plastid genes (fig. S2). These included substantial read coverage within the *petD* and *atpB* 5' UTRs, despite the fact that the 5' end of each gene is marked by a defined minimum in transcript coverage; these might plausibly correspond to previously undetected plastid ORFs, or regions encoding small RNAs. Moreover, we detected maximal coverage of over 6000 and 16000-fold respectively in the 23S and 16S minicircles, despite other presumed coding regions of both minicircles having substantially below 1000-fold read coverage (fig. S2). Alongside previous data suggesting that the 23S and 16S sequences of dinoflagellate plastids are highly divergent and lack otherwise conserved regions, these peaks might be consistent with internal cleavage of ribosomal RNAs into topographically separate elements (Barbrook et al., 2006; Dang and Green, 2009). Verification of these peaks by northern blotting or RT-PCR, will be essential to determine if they reflect genuine transcript processing events, or reflect secondary structures or other features within plastid mRNA pools.

Terminal processing of non-coding and core-containing transcripts of plastid minicircles

For all minicircles studied, we noted that there were very few reads covering non-coding sequences such as core regions. For example, in the case of the *psbD-psbE-ORF1-psbI* minicircle, which had an average 1333-fold read coverage over the entire sequence, and a maximum 3778-fold read coverage within the *psbD* CDS, the average read coverage for the

280 bp core region (Barbrook and Howe 2000) was 146-fold, with a minimum of 3-fold coverage within the core interior (Fig. 1B). We found similar scenarios in all other minicircles studied (Fig. 1B). This supports evidence from previous studies that transcripts covering minicircle core regions are present at only very low abundance in dinoflagellate plastid RNA pools (Nisbet, et al. 2008; Dang and Green 2010; Barbrook, et al. 2012).

We have previously characterised the mature coding transcripts produced from each minicircle both by northern blotting, and RT-PCR (Barbrook et al., 2001; Barbrook et al., 2012). We now wished to investigate what terminal processing events are associated with non-coding and core-containing transcripts. First, we hybridised northern blots of total cellular RNA using single-stranded RNA probes complementary to different regions of the *psbA* and *petB/ atpA* minicircles (Figs. 2A, 2B; Table S5, probes labeled "sense" in column B), including non-coding regions (the *petB* 5' UTR, and the *psbA* 5' and 3' UTR) and coding regions of each minicircle. To visualise the non-coding transcript diversity of each minicircle, we used much larger quantities of total cellular RNA (30 µg) than previously hybridised.

The most intense hybridisation obtained corresponded in size to the monocistronic mRNAs of each gene (transcripts labelled a-c; *psbA* : 1100 nt, *petB*: 700 nt, *atpA*: 1500 nt; Figs. 2A, B). We additionally identified lower abundance bands of less than one minicircle length (transcripts labeled d-f; *psbA*: 700 nt, *petB*: 750 nt, *atpA*: 1000-350 nt). We could additionally detect, in overexposed *psbA* blots, a 2300 nt transcript (band labeled g, Fig. 2A); and 2700 and 5400 nt transcripts in overexposed *petB/atpA* blots (band labeled h, k; Fig. 2B). As the length of the *psbA* and *petB/atpA* minicircles are respectively 2311 bp and 2713 bp, these transcripts are likely to correspond to transcripts containing one, or two complete minicircle sequences.

Alongside this, we performed circular RT-PCR, a technique that allows the precise identification of 5' and 3' terminus positions from individual transcripts (Barbrook, et al. 2012). For this, cDNA synthesis was carried out using primers specific to core regions of the *psbA* and *petB/atpA* minicircles, and PCR reactions were then performed using different combinations of primers, designed against different regions of minicircle sequence, to identify core-containing transcripts present (Materials and Methods) (Fig. S3A; Table S2, section 2, lines 19-41).

26 core-containing transcripts were identified through this approach (Fig. S3B; Table S3, sections 1-2, lines 5-57). Twenty of the transcripts that were identified (13 from the *psbA*

minicircle, 7 from *petB/atpA*) were less than one minicircle in length. These transcripts terminated at the 3' end within the minicircle core region, and none contained a complete coding sequence. The remaining six transcripts (5 from *petB/atpA*, and one *psbA* transcript) were of greater than one minicircle length, thus corresponding to multi-copy transcripts (Fig. S3B; Table S3, sections 1-2, transcripts marked "multi-copy" in column A). The multi-copy *psbA* transcript extended through a complete copy of *psbA* downstream of the core region, terminating in the *psbA* 3' UTR (Table S3, section 1, line 13); and all five of the multi-copy *petB/atpA* transcripts contained a complete downstream copy of *petB,* although in each case they terminated at the 3' end within the *atpA* CDS, hence did not have a complete downstream *atpA* sequence (Table S3, section 2, lines 44-48).

Two of the multi-copy transcripts from the *petB/atpA* minicircle had similar 5' ends to those identified for monocistronic, polyuridylylated *atpA* mRNAs, located 340 and 345 bp upstream of the *atpA* CDS (Fig. S3B, asterisked *petB/atpA* multi-copy transcripts ii and iii; Table S3, section 2, lines 45-46). Similar 5' end positions for monocistronic *atpA* transcripts have also been identified in previous circular RT-PCR studies (Barbrook, et al. 2012). Thus, some core-containing transcripts undergo similar cleavage events generating 5' ends as monocistronic mRNAs.

Core-containing transcripts can receive poly(U) tails

Given the presence of mature 5' ends on some core-containing transcripts, we wished to determine whether core-containing transcripts ever receive 3' poly(U) tails. As we did not detect polyuridylylated core-containing transcripts by circular RT-PCR, we performed an oligo-d(A) primed RT-PCR to investigate directly their presence for the *petB/atpA* and *psbA* minicircles. First, cDNA was synthesised using primers containing an oligo-d(A) sequence at its 5' end, and a gene-specific sequence at the 3' end of two minicircles (encoding *petB/atpA* and *psbA*; Table S2, section 3, lines 45-47) (Barbrook, et al. 2012). Each primer would therefore specifically anneal to transcripts containing a 3' poly(U) tail (through complementary pairing to the oligo-d(A) region), but should preferentially anneal to transcripts of the gene corresponding to the gene-specific region. PCR amplifications of the generated cDNA were performed using combinations of primers flanking the cDNA synthesis site, to identify transcripts that contained a second copy of the minicircle CDS (Fig. 2C; Table S2, section 3, lines 48-54).

For all three genes, products were identified consistent with the presence of polyuridylylated multi-copy transcripts (Fig. 2C; lanes 1, 3, 5). Products were not identified for any PCR under reverse transcriptase negative conditions, confirming that these products were not due to residual gDNA contamination (Fig. 2C; lanes 2, 4, 6). To confirm that the cDNA primers employed annealed specifically to polyuridylylated transcripts of one gene, PCRs were performed using template generated with the *psbA* cDNA primer and the *petB* PCR primers, and using template generated with either the *petB* or *atpA* primer and the *psbA* PCR primers (Fig. 2C; lanes 9-11). Products were not identified in any case, confirming that the cDNA primers used were specific to the intended template, and were not annealing to other transcripts in the RNA samples at detectable levels. Thus, poly(U) tails can be added to RNA transcripts of more than one minicircle length.

Presence of antisense transcripts in peridinin dinoflagellate plastids

We wished to test whether antisense transcripts were present in the *A. carterae* plastid. The RNA library used for sequencing was ligated to a custom RNA adapter prior to cDNA synthesis, allowing the precise identification of transcript termini (Fig. S4A) (Scotto-Lavino, et al. 2006; Dorrell, Hinksman, et al. 2016). As this RNA adapter was non-palindromic, it can be used to discriminate sense and antisense transcripts, i.e. by comparing the orientation of the ligation site to the the corresponding minicircle sequence (Fig. S4A). We therefore inspected the next-generation sequence data for evidence of antisense transcripts. We found 955 reads that contained a detectable transcript 5' end, of which 67 (7.0%) corresponded to transcripts in the reverse orientation of the minicircle sequence, i.e. possible plastid antisense transcripts (Georg, et al. 2010; Dorrell, Hinksman, et al. 2016) (Fig. S4A; Table S4). We detected only negligible numbers of plastid transcript 3' end ligation sites through this approach (data not shown).

We additionally searched for antisense transcripts using RNA-ligase mediated 5' RACE. This technique uses the same RNA adapter ligation previously used for the next-generation sequencing experiments. However, in this case, cDNA was synthesised from adapter ligated *A. carterae* RNA using primers with the same sequence as the coding strands of the *psbA* and *petB/atpA* minicircles (Fig. S4A; Table S2, section 4, lines 57-59), which would anneal to antisense transcripts. These cDNA products were used as template for PCRs, using primers with the same sequence as the coding strand of the *psbA* and *atpA* genes, and a PCR primer with the same sequence as the RNA adapter used (Fig. S4A; Table S2, section 4, lines 61-63).

Through this approach, we detected an antisense transcript for the *atpA* gene, which terminated at the 5' end at a position 115 bp inside the 5' end of the *atpA* (Fig. S4C). This product was detected independently three times, using different RNA samples, but could not be identified in ligation- or reverse-transcriptase negative controls, confirming that it was not an artifact of mis-annealing of the adapter-specific primer on similar regions of minicircle sequence, or the result of genomic contamination in the adapted RNA libraries.

Finally, we performed RT-PCRs using the antisense cDNA preparations, the primer used for cDNA synthesis, and a PCR primer designed to be similar to the spliced-leader (SL) sequence, a short motif associated with the 5' end of most dinoflagellate nuclear transcripts (Lin, et al. 2010; Gavelis, et al. 2015). Products were not detected, suggesting that it is unlikely that these transcripts were generated within the dinoflagellate nucleus.

Antisense transcripts are not complementary to sense transcripts

In plants, and in the fucoxanthin-containing dinoflagellate *Karenia mikimotoi,* sense and antisense transcripts from individual plastid genes have been documented to possess different consensus terminus positions (Zghidi-Abouzid, et al. 2011; Dorrell, Hinksman, et al. 2016). We tested whether this is true also in *A. carterae* by detailed profiling of the predominant terminus positions associated with sense and antisense transcripts over the *A. carterae psbA* and *petB/atpA* minicircles. First, we performed circular RT-PCR using cDNA synthesis primers complementary to the template strand of each minicircle, and PCRs using different combinations of primers, against different regions of each minicircle sequence, to obtain an estimate for the full diversity of antisense transcripts that are present (Fig. 3A; Table S2, section 5, lines 65-83). We additionally hybridised northern blots with single-stranded RNA probes with the same sequence as regions of each minicircle coding strand, to identify the antisense transcripts associated with regions of each minicircle complementary to previously investigated sense transcripts (Figs. 2A, B; Fig. 3B; Table S5, probes labeled "antisense" in column B).

We identified multiple potential antisense transcripts from each minicircle through each approach. The vast majority of antisense transcripts identified by circular RT-PCR terminated at positions distinct from the corresponding sense transcripts from each minicircle (Fig. 3A; Table S3, sections 3, 4, lines 59-95). For example, we could not identify any antisense transcripts that terminated at either ends at positions complementary to the *psbA, petB* or *atpA* poly(U) sites, and only found two that might possess termini complementary to

consensus mRNA 5' end positions (Fig. 3A, asterisked transcripts; Table S3, section 3, lines 73-74).

For the *psbA* minicircle, two transcripts, of 750 nt and 950 nt length, were detected by northern blot using probes for the 5' UTR and 5' end of the CDS (Fig. 3B; transcripts labeled i-ii). These may correspond to two antisense transcripts, one 748 nt length, and one 924 nt length, identified through circular RT-PCR (Fig. 3A). For the *petB/atpA* minicircle the predominant bands were identified for *atpA* (of 250 and 500 nt length) with the 5' end probe (Fig. 3B; transcripts labelled iii-iv). These may correspond to 261 nt and 508 nt length antisense transcripts identified by circular RT-PCR, covering regions internal to the *atpA* CDS (Figs. 3A, 3B; transcripts labelled iii, iv). A faint band of approximately 1250 nt length was identified on overexposure of the blots hybridised to both the 5' and 3' *atpA* probes, which may correspond to a 1228 nt transcript identified from circular RT-PCR to extend from the core into the region complementary to the 5' end of *atpA* (Figs. 3A, 3B; transcript labelled v).

In no case did we identify northern hybridization in an antisense transcript blot of the size corresponding to a mature mRNA (*i.e.,* an 1100 nt *psbA* transcript, a 700 nt *petB* transcript, or a 1500 nt *atpA* transcript), or a visualized non-coding or core-containing transcript (*e.g.,* transcripts corresponding to one or two complete minicircle lengths) from the sense strand blots (Figs. 2A, B; Fig. 3B). Moreover, we did not identify any hybridisation in the *psbA* blots that might correspond to the two antisense transcripts identified through circular PCR with terminus positions complementary to sense transcripts (expected transcript size 400-500 nt; Figs. 3A, 3B). Thus, sense and antisense transcripts undergo different terminal processing events in the *A. carterae* plastid.

Antisense transcripts lack poly(U) tails

We wished to determine whether poly(U) tail addition in peridinin dinoflagellates is preferentially associated with sense transcripts over antisense transcripts, as occurs in the fucoxanthin-containing species *Karenia mikimotoi*. To do this, cDNA was synthesised from *A. carterae* total cellular RNA using an oligo-d(A) primer previously shown to anneal to plastid poly(U) tails from a wide range of plastid genes in dinoflagellate and related species (Table S2, section 6, line 86) (Barbrook, et al. 2012; Dorrell and Howe 2012; Dorrell, et al. 2014; Richardson, et al. 2014). PCRs were then performed using the same oligo-d(A) primer as a PCR primer, paired with a PCR primer with the same sequence as the template strands of

the *psbA* and *petB/atpA* minicircles, to identify polyuridylylated antisense transcripts (Table S2, section 6, lines 88-95).

Products were not identified using any of the template strand primers, indicating that polyuridylylated antisense transcripts were not present (Fig. 4; lanes 2-3, 7-8). Products were not detected even following a second round of PCR amplification of the primary PCR product. In contrast, polyuridylylated sense transcripts were amplified from each cDNA preparation, by PCR with the oligo-d(A) primer, and PCR primers with the same sequence as the coding strand of the *psbA*, *petB* and *atpA* genes, confirming that the oligo-d(A) primed cDNA synthesis reactions were successful (Fig. 4; lanes 4, 9, 10). In addition, products covering each of the regions of sequence tested were amplified from each RNA samples, using gene-specific cDNA synthesis and PCR primers (Fig. 6, lanes 5-6; 11-12). Thus, antisense transcripts of the *psbA* and *petB/ atpA* minicircles in the *A. carterae* plastid do not possess 3' poly(U) tails.

Discussion

We have produced a comprehensive study of the plastid genome and transcriptome of the model peridinin-containing dinoflagellate species *Amphidinium carterae*. We show that existing surveys of the coding and non-coding diversity of this genome (Barbrook, et al. 2018) are largely complete, although we document a previously unidentified non-coding minicircle from next-generation sequencing (Fig. 1). Unusually, transcripts from this minicircle are at somewhat higher abundance than other empty minicircles, and it contains an ORF also detectable in other members of the genus *Amphidinium* (Fig. S1; Fig. S2); whether this is consistent with it possessing a coding or other function remains to be determined. Otherwise, we demonstrate that the peridinin dinoflagellate plastid is characterised by relatively uniform transcript abundance across all protein coding regions (Fig. 1; Fig. S2). This is consistent with differences in gene expression being mediated by post-transcriptional or translational regulation, as has previously been inferred to occur both at specific dinoflagellate plastid loci, and in the dinoflagellate nucleus (Okamoto and Hastings 2003; Wang, et al. 2005).

Our data indicate that core-containing, and non-coding regions of minicircle sequence are present at much lower abundance in plastid transcript pools than coding sequences (Figs. 2, 4), consistent with information from previous studies (Nisbet, et al. 2008; Dang and Green 2010; Barbrook, et al. 2012). The low read coverage over minicircle core regions might reflect limited transcription through these sequences, for example if the minicircle core acts as an effective transcriptional terminator, or alternatively might reflect very efficient cleavage or degradation of multi-copy transcripts by the plastid transcript processing machinery (Nisbet, et al. 2008; Dang and Green 2010). Notably, we infer the presence of both mature 5' end processing sites, and 3' poly(U) tails on some core-containing transcripts of more than one minicircle length (Fig. 2C; Fig. S3). Polyuridylylated polycistronic transcripts of less than one minicircle length have additionally previously been identified from the multigene *petB/atpA* and *psbD/psbE/psbI A. carterae* minicircles (Nisbet, et al. 2008; Barbrook, et al. 2012). These long transcripts with mature 5' and 3' termini in peridinin plastids might represent processing precursors of mature mRNAs, or might equally represent mis-processed transcripts generated as a result of inefficient end cleavage or transcript termination in peridinin plastids.

Finally, we have identified antisense transcripts from the *A. carterae psbA* and *petB/atpA* minicircles (Figs. 4, 5; Fig. S4), similar to those previously identified in the plastids of plants

(Georg, et al. 2010; Castandet, et al. 2016), apicomplexans (Bahl, et al. 2010; Nisbet, et al. 2016), and fucoxanthin-containing dinoflagellates (Dorrell, Hinksman, et al. 2016). It remains to be determined how these transcripts are generated. We do not identify 3' poly(A) tails, nor 5' spliced leader sequences on any antisense transcript, which might be applied if they were products that had been expressed from previously plastid gene fragments, which have been relocated to the nucleus (Lin, et al. 2010; Owari, et al. 2014; Gavelis, et al. 2015). Moreover, the antisense transcripts do not have complementary terminus positions to those associated with mature mRNAs for each minicircle (Figs. 2, 3), which suggests that they are not the products of a direct RNA-dependent RNA polymerase activity on sense plastid transcripts (Zandueta-Criado and Bock 2004). Thus, we tentatively propose that the antisense transcripts are generated through the transcription of plastid minicircle coding DNA strands. This may be through the activity of specific promoters located in the reverse orientation, or through transcription initiation events that are not dependent on specific primary sequence motifs, with the plastid RNA polymerase recruited to features such as stem loops or single-stranded nicks in minicircle sequence (Zhang, et al. 2002; Dang and Green 2009; Leung and Wong 2009; Barbrook, et al. 2018). Verifying this will require detailed mapping of promoter sequences in dinoflagellate plastids, for example through high-throughput techniques such as dRNA-seq (Zhelyazkova, et al. 2012).

It remains to be determined whether antisense transcripts possess specific functions in dinoflagellate plastids. In plants, the accumulation of antisense plastid transcripts appears to vary in response to thermal stress (Georg, et al. 2010; Castandet, et al. 2016). Plastid antisense transcripts might therefore have regulatory effects, constraining the processing and expression of sense transcripts in response to environmental and physiological changes, similar to the functions of miRNAs in nuclear genomes (Fujii et al., 2005). Alternatively, antisense transcripts might have purely deleterious effects on transcript processing and translation efficiency, and be actively removed from plastid transcript pools, as has also been documented in plants (Sharwood, et al. 2011; Castandet, et al. 2013; Hotto, et al. 2015). In this latter regard, we note that the antisense transcripts in the *A. carterae* plastid appear to be present at low abundance (Fig.3; Fig. S4), as is the case in fucoxanthin dinoflagellate and apicomplexan plastids (Dorrell, Hinksman, et al. 2016; Nisbet, et al. 2016) and do not receive 3' poly(U) tails (Fig. 4). Previous studies have suggested that the poly(U) tail confers 3' end stability to plastid transcripts in dinoflagellates and related species (Dang and Green 2009; Barbrook, et al. 2012; Janouskovec, et al. 2013; Dorrell, et al. 2014). The specific addition of a poly(U) tail to sense transcripts during transcript processing might

therefore enable antisense transcripts to be preferentially degraded, leaving a plastid transcript pool enriched in mature mRNAs.

Concluding Remarks

Control of dinoflagellate plastid gene expression and physiology is clearly complex, relying on both transcriptional control and post-transcriptional processing events (Nassoury, et al. 2005; Wang, et al. 2005), and different processing events may delineate mature mRNAs from non-coding transcripts (Richardson, et al. 2014; Dorrell, Hinksman, et al. 2016). Understanding the functional significance of antisense transcripts, and their associated processing events, will depend on being able to manipulate dinoflagellate plastid genomes, e.g. through the introduction of transgenic minicircles producing antisense transcripts, which is now possible (Nimmo, et al., under review). Genetic manipulation of plastid genome content and gene expression in dinoflagellates may provide fresh insights into the unusual transcript processing events associated with this lineage.

Methods

Cultures and nucleic acid isolation

Amphidinium carterae CCMP 1314 was cultured in f/2 medium, which was prepared with Ultramarine Synthetica artificial sea water (Waterlife) and buffered with 500 μg/ ml tricine to pH 8, under a 12:12 light: dark cycle, at 30 μ E illumination, at 20°C, without shaking.

DNA from *A. carterae* was purified by phase separation with phenol: chloroform, followed by ethidium bromide-caesium chloride gradient centrifugation, following previous methodology (Barbrook and Howe 2000). The fraction corresponding to minicircle DNA was removed with a needle. Purified and cleaned DNA (1 μg) was treated with 10U Plasmid-Safe ATP-dependent DNase (Lucigen) overnight. The DNA was purified using a phenol-chloroform extraction retaining the aqueous phase, precipitated with ethanol and sodium acetate. The pellet was washed with 70% ethanol, air dried and resuspended in TE buffer, pH 8.0.

RNA was extracted using Trizol (Invitrogen), following previous methodology (Barbrook, et al. 2012), treated with DNase I (Qiagen), and cleaned with an RNeasy column (Qiagen), following the manufacturer's instructions. The integrity of cleaned RNA samples was confirmed following electrophoresis on RNase-free TBE-agarose gels, and RNA samples were confirmed to be free of residual DNA contamination via two rounds of PCR, in the absence of reverse transcription, using PCR primers against consensus regions of the nuclear 18S and ITS1 sequences (Gachon, et al. 2013). All nucleic acid concentrations were confirmed with a Nanodrop spectrophotometer.

Genome sequencing

Isolated *Amphidinium* DNA was sequenced using a NextSeq500 machine (Illumina) in single-read mode, running 150 cycles, by the sequencing facility in the Biochemistry Department, University of Cambridge. 10,120,479 reads of 151 bp length were generated. The quality of these reads was checked using Fast QC (Blankenberg, et al. 2010). Reads were then trimmed from base 5 to 135. Multiple *de novo* assemblies were made using Unicycler (Wick, et al. 2017), utilising the trimmed sequences. Inspection of the alignments revealed there was contamination from *A. carterae* mitochondrial DNA and from *Bacillus oceanisediminis*. All assemblies of 3300 bp and smaller were examined. The Fast QC, trimming and assemblies were performed using the Galaxy web interface ((Blankenberg, et al. 2010). This library has been deposited in NCBI SRA (ID: PRJNA524783).

Twenty-one minicircle contigs were identified through this approach, including four that mapped as circular. The sequence of one novel potential minicircle was confirmed by PCR and Sanger DNA sequencing, using a custom set of twelve gene-specific primers (Table S2, section 1, lines 1-16). The orthology and evolutionary conservation of minicircle sequences in different dinoflagellate species was assessed by tBLASTx search, with threshold evalue 1×10^{-05}, against this and all other dinoflagellate species sequenced as part of MMETSP, followed by verification through a reciprocal BLASTx search against nr, following previous methodology (Dorrell, Klinger, et al. 2017). Potential coding functions were assessed in novel minicircle sequences by BLAST search against the nr database; and potential tRNA sequences were searched for using tRNAscan-SE and Aragorn (Lowe and Eddy 1997; Laslett and Canback 2004; Nelson, et al. 2007). The novel minicircle sequence has been deposited in GenBank (ID: MK598758-MK598759), and is additionally provided in Table S1, sheet 4.

Transcriptome sequencing

Transcriptome sequencing was performed using adapter-ligated total cellular RNA, generated by incubation of 1 µg DNA-free total cellular RNA from *Amphidinium carterae*, with 1 µg of a custom RNA adapter (GCUGAUGGCGAUGAGCACUGGGUUGCAA) using T4 RNA ligase (Promega) as previously described (Dorrell, Hinksman, et al. 2016). The ligation products were cleaned with an RNeasy column (Qiagen) and eluted in 30 µl DEPC-treated water. 10 µl of the eluted product was used as template for synthesis with a Maxima H Minus double stranded cDNA synthesis kit (Thermo), per the manufacturer's instructions. The product of the cDNA synthesis reaction was cleaned using a MinElute cleanup column (Qiagen), and eluted in a further 30 µl DEPC-treated water.

Double stranded cDNA was quantified using a Qubit fluorometer (Invitrogen) following the manufacturer's instructions. A sequencing library was generated from 100 ng purified product using a NexteraXT tagmentation kit (Illumina). The library was sequenced over 500 cycles using a MiSeq sequencer. Reads were trimmed of sequencing adaptors using the Miseq reporter version 2.0.26. Low-quality sequences (defined as all sequence within each read following the first residue with a Phred score below 20) were removed from each read; quality control statistics and paired-end mapping of each read were performed using FastQC and custom bash scripts (De Wit, et al. 2012). Nucleotide composition histograms and quality control boxplots are shown for each read in Fig. S5. The resulting library was estimated to contain 61.9% duplicate reads, and 4.7% singletons. This library has been deposited in NCBI SRA (ID PRJNA518128).

To estimate read coverage against the *Amphidinium* plastid genome, a composite library consisting of all minicircle sequences previously reported from *A. carterae* CCMP1314/CCMP 1102/6, and the complete MMETSP transcriptome cDNA libraries of *A. carterae* CCMP1314 that had been cleaned of residual sequence contamination using a previously defined protocol (Marron, et al. 2016; Dorrell, et al. 2017) was assembled. To avoid duplication of plastid sequences within this library, the cDNA sequences corresponding to plastid-encoded proteins (Dorrell, Klinger, et al. 2017), as previously defined, were excluded from the MMETSP library. Each cleaned sequence paired-end read was searched against this library with BLASTn, and sequences that produced a top hit against a plastid minicircle sequence, over the complete length of the read sequence, and with a minimum threshold identity of 95%, were mapped to the minicircle sequence.

To assess the distribution of 5' end positions over minicircle sequences within the library, the paired-end read sequences were filtered for those that yielded a top hit against a minicircle sequence, and contained either a complete RNA adapter sequence, terminated at the 5' end with a region corresponding to the 3' terminal fragment of the RNA adapter, or terminated at the 3' end with a region corresponding to the complement of the 3' terminal fragment of the RNA adapter. The minimum length of RNA adapter used for this analysis corresponded to the last ten nt within the adapter sequence (TGGGTTGCAA), as this sequence was found (by BLASTn search) not to occur naturally in any of the screened minicircle sequence. Reads in which the RNA adapter was found in the reverse complement direction were reverse complemented, and each read was then trimmed at the 5' end to the end of the RNA adapter sequence. The trimmed reads were then searched again against the composite *Amphidinium* reference library. The position of the first residue (corresponding to the 5' ligation site of the transcript) and orientation (identifying sense/ antisense transcription) of each trimmed read that mapped to an *Amphidinium* minicircle over the full length of the read, with at least 95% identity, was recorded.

RT-PCR experiments

cDNA synthesis reactions were performed using Superscript II reverse transcriptase (Sigma), and PCRs were performed with GoTaq DNA polymerase (Invitrogen), following the manufacturers' instructions. Oligo-d(A) primed RT-PCR, circular RT-PCR and RNA-ligase mediated 5'RACE were performed according to previously defined protocols (Scotto-Lavino, et al. 2006; Barbrook, et al. 2012; Dorrell, Hinksman, et al. 2016). PCR products were Sanger

sequenced using an Applied Biosystems 3730xl DNA Analyser (Department of Biochemistry, University of Cambridge).

cDNA synthesis primers for sense transcripts were designed against the coding strands of minicircle sequence (Table S2, section 2), and for antisense transcripts against the non-coding strands of minicircle sequence (Table S2, sections 4, 5). Each cDNA synthesis primer was designed so that the final eight nt of the primer sequence was not found anywhere else, in any orientation, on any minicircle in the *A. carterae* plastid (Barbrook and Howe 2000; Barbrook, et al. 2001; Hiller 2001; Nisbet, et al. 2004), except in the desired annealing site, to minimise the possibility (for example) of mis-priming of primers designed for antisense transcripts to sense transcripts from the same minicircle.

To identify the full diversity of 5' and 3' termini associated with multi-copy and antisense transcripts by circular RT-PCR, five different PCR forward, and five different PCR reverse primers were designed against different regions of each minicircle sequence (Table S2, sections 2, 5). For example, for the *psbA* minicircle, a PCR reverse primer was designed specific to the 5' end of the *psbA* gene, which would preferentially amplify multi-copy transcripts with mature 5' termini. Three further reverse primers were designed specific to non-coding regions of the *psbA* minicircle, upstream of the *psbA* mature transcript 5' terminus position that would preferentially amplify multi-copy transcripts containing extensive UTR sequence. A final reverse primer was designed specific to the 3' end of *psbA* that would preferentially amplify transcripts with 5' termini located within the CDS. PCRs were then performed using each possible combination of forward and reverse primer, to amplify the ligated termini of transcripts covering minicircle core regions and antisense transcripts. Circular PCR was performed with the same thermal cycle scheme as in Barbrook et al. (2012): a 10 minute initial denaturation step at 95˚C, followed by 40 cycles consisting of 95˚C, 45s; 55˚C, 45s; and 72˚C, 2 minutes. Each circular RT-PCR was performed three times, using cDNA generated from independently isolated RNA samples, to maximise the capture of minicircle transcripts.

Northern blots

Probes for each northern blot were generated using a DIG Northern Starter kit (Roche). Probe template sequences were generated by fusion of selected minicircle amplicons to the T7 promoter region of pGem-tEasy vector (Promega), following a previously established protocol (Dorrell, et al. 2014; Nisbet, et al. 2016). Probes were designed that were specific to

the 5' and 3' UTRs of the *psbA* minicircle, as well as probes that were specific to the 5' and 3' ends of the *psbA* CDS (Table S5). For the *petB/atpA* minicircle, probes were designed that were specific to the *petB* CDS, the 5' and 3' ends of the *atpA* CDS and the *petB* 5' UTR. To facilitate the direct comparison of sense and antisense transcripts, the RNA probes were complementary in sequence to the probes previously designed for sense transcripts from each minicircle, and identical RNA electrophoresis and detection conditions were used for sense and antisense transcript blots.

Northern blots, using freshly isolated RNA, were made and probed following previously established protocols (Dorrell, et al. 2014; Dorrell, Hinksman, et al. 2016). Each northern blot was made using 30 µg total cellular RNA, as this has been shown to be adequate to detect very low abundance and multi-copy transcripts in *H. triquetra* (Dang and Green 2010). Northern probes were hybridized using a DIG labelling kit (Roche), following the manufacturer's instructions. Chemiluminescence of DIG-conjugated horseradish peroxidase was detected in an Agilent Infinity 1260 over periods of between two minutes and twelve hours, dependent on the relative abundance of the corresponding transcript. Each northern blot experiment was performed twice using independently isolated RNA samples, and consistent banding patterns were identified in each case.

Figure legends

Figure 1. Overview of the *Amphidinium carterae* plastid genome and transcriptome.

Panel A shows all of the minicircle sequences (inferred via the presence of a core region; Barbrook and Howe, 2000), identified from CsCl-enriched *A. carterae* genomic DNA. Identification of orthologous sequences for each minicircle, and PCR verification of the novel empty minicircle sequence, are shown in fig. S1. **Panel B** shows an exemplar read coverage plot for the *psbD/psbE/ORF1/psbI* minicircle. This plot shows the total number of reads identified at each position in the minicircle sequence and is shown against a linearised complete minicircle sequence. Thick black arrows correspond to coding sequence, thick grey bars to core sequence, and thin black lines to other non-coding regions of the minicircle. High read coverage is observed over each coding sequence, contrasting with low read coverage in the core region. The intergenic region between *psbD* and *psbE* is demarcated by low read coverage, consistent with *psbD* primarily being processed as a monocistronic transcript in *Amphidinium*, whereas there is a continuous and high read coverage throughout the *psbE-ORF1* and *ORF1-psbI* intergenic regions, consistent with the presence of polycistronic transcripts covering the *psbE-ORF1-psbI* region (Barbrook, et al. 2012). Additional read coverage plots are supplied in fig. S2. **Panel C** tabulates the average read coverage for the complete sequences of each *Amphidinium* plastid minicircle with verified coding content, along with the core regions of each minicircle, and each confirmed open reading frame present.

Figure 2: Diversity of multi-copy and core-containing transcripts. Panels A and B show northern blots hybridised with single-stranded RNA probes complementary to different regions of coding strand sequences, respectively from the *psbA* **(A)** and *petB/atpA* **(B)** minicircles. Key bands are identified with arrows. Sizes of each band were calculated by comparison to a DIG-labelled RNA ladder separated on the same gel. Transcripts a-c represent the predominant, monocistronic mRNAs produced from each gene, as identified by comparison to transcript sizes estimated from previous northern blot and circular RT-PCR studies (Barbrook, et al. 2001; Barbrook, et al. 2012); and transcripts d-f represent lower abundance bands of less than one minicircle length. Corresponding circular RT-PCR experiments are shown in fig. S3. **Panel C** shows RT-PCRs employed to detect polyuridylylated multi-copy transcripts from the *A. carterae psbA* and *petB/atpA* minicircles. (Top) a putative transcript of a hypothetical minicircle is shown. The black arrow corresponds to all coding regions present within the minicircle, i.e. either one gene (for

monocistronic minicircles) or multiple gene sequences (for polycistronic circles); and the core region as a box. A cDNA synthesis primer (grey arrow) was designed containing a 5' oligo-d(A) region, and a 3' region complementary to the 3' UTR sequence region directly upstream of either the *psbA, petB* or *atpA* poly(U) site. PCRs (below) were performed using the cDNA and pairs of PCR primers that flank the cDNA primer annealing site. These reactions will specifically amplify polyuridylylated transcripts that additionally contain a second copy of minicircle sequence. The gel photograph (bottom) shows products as following. Lane **1**: RT-PCR to detect polyuridylylated multi-copy transcripts from the *petB/atpA* minicircle using a cDNA synthesis primer specific to the *petB* poly(U) site. Lane **2**: reverse transcriptase negative control for lane 1. Lane **3**: RT-PCR to detect polyuridylylated multi-copy transcripts using a cDNA synthesis primer specific to the *atpA* poly(U) site. Lane **4**: reverse transcriptase negative control for lane 3. Lane **5**: RT-PCR to detect polyuridylylated multi-copy transcripts from the *psbA* minicircle. Lane **6**: reverse transcriptase negative control for lane 5. Lane **7**: blank lane. Lane **8**: reaction positive control, using a genomic DNA template and PCR primers internal to the *psbA* CDS. Lane **9-10**: RT-PCRs using the *petB* and *atpA* poly(U) site cDNA synthesis primers, and PCR primers internal to the *psbA* CDS, confirming the minicircle specificity of cDNA synthesis.

Figure 3. Circular RT-PCR and northern blots of *psbA* and *petB/atpA* minicircle antisense transcripts. Panel A shows the diversity of antisense transcript termini mapped by circular RT-PCR for the *psbA* and *petB/atpA* minicircles. Transcripts are displayed on transcript diagrams against a linearised hypothetical sense transcript of each minicircle, shown as per Fig. 1. Transcript sequences that terminate at the same positions as mature mRNA sequences are labelled with asterisks. Transcript sequences that correspond, to bands identified in northern blots of antisense transcripts, are labelled with brackets. Antisense transcript termini identified through next-generation sequencing, and 5' RACE, are shown in fig. S4. **Panel B** shows the results of northern blots hybridised with single-stranded RNA probes complementary to different regions of template strand sequence to detect antisense transcripts, as Figs. 2A, 2B. Each antisense transcript band is labelled with a number corresponding to probable matching transcripts in the circular RT-PCR experiment above. Probes complementary to antisense transcripts covering the *psbA* CDS 3' end and 3' UTR, and the *petB* CDS and 5' UTR, failed to yield any distinct bands, the corresponding blots are not shown. As the *atpA* 3' CDS blot only produced very weak fluorescence, an overexposed blot image is shown.

Figure 4. Absence of poly(U) tails from antisense transcripts. This gel photograph shows the result of a series of RT-PCRs to test for poly(U) tails on antisense transcripts of the *psbA* and *petB/atpA* minicircles. Lanes 1, 13: blank lane. Lanes 2-3, 7-8: RT-PCRs performed with an oligo-d(A) primer for cDNA synthesis, and PCR with the same oligo-d(A) primer and a primer with the same sequence as the template strand of the *psbA* CDS (2) and UTR (3), and the *petB* (7) and *atpA* CDS (8), demonstrating the absence of polyuridylylated antisense transcripts extending over these regions. Lanes 4, 9-10: RT-PCR performed with oligo-d(A) primed cDNA as before, and PCR with oligo-d(A) and primers with the same sequence as the coding strands of the *psbA* (4), *petB* (9) and *atpA* CDS (10), confirming the presence of polyuridylylated sense transcripts in the RNA sample. Lanes 5, 6, 11, 12: positive controls for the presence of antisense transcripts over the *psbA* CDS (5) and UTR (6), and the *petB* (11) and *atpA* CDS (12), using a gene-specific cDNA synthesis and PCR primer with the same sequence as the coding strands of minicircle sequence, and the same template strand PCR primer as used in the corresponding oligo-d(A) primed PCR for each reaction.

Figure S1. Organisation and evolutionary conservation of *Amphidinium carterae* plastid minicircles. Panel A shows a heatmap of which orthologues of different *A. carterae* plastid gene and empty minicircle sequences could be identified in previously assembled dinoflagellate transcriptome datasets from the Marine Microbial Eukaryote Transcriptome Sequencing Project (Dorrell, Klinger, et al. 2017), by tBLASTx search with threshold evalue 1 x 10^{-05}. Cells are shaded in grey if an orthologue was detected; and black if an orthologue with a poly(U) tail, confirming plastid origin, was detected. Species are arranged per the topology constructed in (Dorrell, Klinger, et al. 2017). Of note, of the five empty minicircle queries, only the novel minicircle 6 was detected in the *A. carterae* MMETSP transcriptome, and none was identified to have orthologues in other dinoflagellate species. **Panel B** shows (left) a PCR gel photo, confirming the presence of three overlapping amplicons from the novel minicircle 6 [lane 1: primers forward 1 and reverse 4; lane 2: primers forward 2 and reverse 1; lane 3: primers forward 6 and reverse 3; as defined in Table S2); and (right) an alignment of the contained open reading frame sequence translation, and its orthologues from *Amphidinium carterae* sp. CS-21 and *Amphidinium massartii*.

Figure S2. Read coverage maps for *Amphidinium carterae* plastid minicircles. This figure shows the number of transcripts mapped to each position in each identified *A. carterae* plastid minicircle with defined coding content, shown as per Fig. 1B.

Figure S3 Circular RT-PCR of core-containing minicircle transcripts. Panel A shows a diagram of the circular RT-PCR protocol used to map the termini of transcripts from the *A. carterae psbA* and *petB/atpA* minicircles. A heterogeneous population of transcripts is treated with T4 RNA ligase, generating circularised RNA. Only transcripts that have undergone prior 5' cleavage are ligated. The circular RNA is reverse transcribed using cDNA synthesis primers specific to minicircle core regions. PCRs are then performed using the cDNA preparations using different combinations of primers, to identify the ligated end of transcript sequences. **Panel B** shows a transcript diagram of the diversity of multi-copy transcripts and core-containing transcripts of less than one minicircle length mapped for each minicircle. Transcripts are shown to scale against a hypothetical linearised transcript sequence for each minicircle. The consensus terminus positions associated with monocistronic polyuridylylated *psbA, petB* and *atpA* transcripts are additionally shown, and the boundary positions of each core region are shown with dashed vertical lines. The 5' ends of two multi-copy transcripts from the *petB/atpA* minicircle that are similar to the consensus 5' end position of monocistronic *atpA* mRNA are labelled with asterisks.

Figure S4. Identification of plastid antisense transcripts by RNA ligation. Panel A shows a schematic diagram of the protocol used for generation of the *A. carterae* cDNA library for next generation sequencing, and for 5' RACE. **(i)** Total cellular RNA is ligated to a custom RNA adapter sequence using T4 RNA ligase (Dorrell, Hinksman, et al. 2016). As the RNA adapter used for the initial ligation is non-palindromic, its orientation relative to the sequence of the transcript can be used to identify the orientation of the transcript (forward or reverse) relative to the corresponding minicircle sequence. **(ii, iii)** The adapter-ligated RNA is used as a coding for cDNA synthesis, generating cDNA containing the adapter sequences at transcript terminus points, and is then amplified by PCR. For next-generation sequencing, a random hexamer cDNA synthesis primer is used followed by random PCR amplification, to generate as diverse a range of transcripts as possible. For 5' RACE, the cDNA synthesis primer corresponds specifically to the minicircle coding strand, and thus should preferentially amplify antisense transcripts. Nested rounds of PCR are then performed using a primer specific to the non-coding strand of the minicircle, and a primer corresponding to the 5' RNA adapter, thus generating products containing an RNA adapter ligated to the 5' end of a minicircle antisense transcript. **Panel B** shows **(i)** the total number of transcript 5' ends, defined as reads containing at least the final 8 nt of the adaptor sequence either at the start or the end of the reads sequence, identified from the next generation sequencing library. The transcript 5' ends are divided into ends that appear to correspond to sense or

antisense transcripts based on their orientation, with the majority originating from sense transcripts. An exemplar map of 5' transcript termini, for the *Amphidinium* 23S minicircle, is shown in (**ii**). Notably, sense and antisense transcript termini do not occur at corresponding locations, indicating that they derive from separate processing events. **Panel C** shows an exemplar RNA ligase-mediated 5' RACE amplification of an *Amphidinium* plastid antisense transcript. This transcript end is ligated to an RNA adapter of known sequence (**i**) allowing the identification of the transcript relative to the genomic sequence. Alignment of the transcript end region (**ii**) against the genomic *atpA* sequence reveals a region identical in sequence to the *atpA* template strand, followed immediately by a sequence that aligns with the 3' end of the RNA adapter used. This confirms that the transcript terminus identified is a 5' end of a minicircle antisense transcript.

Figure. S5. Quality control statistics for the *Amphidinium carterae* RNA sequencing library. This figure shows **(A)** nucleotide composition histograms and **(B)** QC score boxplots for the forward **(i)** and reverse **(ii)** adaptor reads of each paired-end sequence pair. Both libraries show high read quality (median QC score ~40) for > 120 bp from the read 5' end. Scores are plotted using the NGS:QC pipeline implemented into the Galaxy server (Blankenberg, et al. 2010).

Table S1. Tabulated read coverage for *Amphidinium carterae* plastid minicircles. Sheet 1 provides fasta format sequences of the *Amphidinium* minicircles identified through next-generation sequencing of CsCl-purified genomic DNA. **Sheet 2** provides an overview of the number of reads from the adaptor-ligated *Amphidinium* RNA library mapped to minicircle sequence. **Sheet 3** provides detailed read coverage values for RNA sequence data, for each residue in each minicircle sequence. Each minicircle sequence is given with the core region corresponding to residues 1-280. **Sheet 4** provides GenBank flatfile form sequences of two variants of novel minicircle 6, assembled by PCR.

Table S2. Primers used for RT-PCR analysis. The annealing position of each primer is given, relative to a linearised sequence of the minicircle, in which the core region corresponds to residues 1-280.

Table S3. Tabulated circular RT-PCR sequences. This table provides 5' and 3' terminus positions for core-containing and antisense transcripts identified by circular RT-PCR. The terminus positions of each transcript, and the primers used for PCR amplification of the transcript, are given corresponding to a linearised sequence of the minicircle, in which the core region corresponds to residues 1-280. For reference, the positions of each coding

sequence, as well as the intervals of residues over which mature mRNA 5' ends and poly(U) sites have been identified in previous circular RT-PCR studies (Barbrook, et al. 2012), are supplied.

Table S4. Tabulated transcript 5' ends identified by next-generation sequencing. This table shows the position and orientation of 5' end termini within the adaptor-ligated *Amphidinium* RNA library against each published minicircle sequence. **Sheet 1** provides tabulated BLAST outputs of each of the 5' end sequences identified. **Sheet 2** provides an overview of the number and orientation of 5' termini mapped to minicircle. **Sheet 3** tabulates the number and orientation of 5' end termini identified for each residue in each minicircle sequence. Each minicircle sequence is given with the core region corresponding to residues 1-280.

Table S5. Tabulated northern probes. This table provides a linearised sequence of the T7 arm of the pGem-tEasy vector, and sequences of the template strand regions used for the generation of northern probes against sense transcripts, and the non-coding strand regions used for the generation of northern probes against antisense transcripts. The sequence intervals covered by each probe are given corresponding to a linearised sequence of the minicircle, in which the core region corresponds to residues 1-280.

References

Bahl A, Davis PH, Behnke M, Dzierszinski F, Jagalur M, Chen F, Shanmugam D, White MW, Kulp D, Roos DS. 2010. A novel multifunctional oligonucleotide microarray for *Toxoplasma gondii*. BMC Genom 11: 603.

Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RER, Howe CJ. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. Plant Mol Biol 79:347-357.

Barbrook AC, Howe CJ. 2000. Minicircular plastid DNA in the dinoflagellate *Amphidinium operculatum*. Mol Gen Genet 263:152-158.

Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. 2010. Organization and expression of organellar genomes. Phil Trans R Soc Biol 365:785-797.

Barbrook AC, Howe CJ, Nisbet RER. 2018. Breaking up is hard to do: the complexity of the dinoflagellate chloroplast genome. Perspectives Phycol, 0084.

Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. 2006. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. BMC Genom 7: 297.

Barbrook AC, Symington H, Nisbet RER, Larkum A, Howe CJ. 2001. Organisation and expression of the plastid genome of the dinoflagellate *Amphidinium operculatum*. Mol Genet Genom 266:632-638.

Barbrook AC, Voolstra CR, Howe CJ. 2014. The chloroplast genome of a *Symbiodinium* sp clade C3 isolate. Protist 165:1-13.

Barkan A. 2011. Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. Plant Physiol 155:1520-1532.

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Team G. 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics 26:1783-1785.

Castandet B, Hotto AM, Fei Z, Stern DB. 2013. Strand-specific RNA sequencing uncovers chloroplast ribonuclease functions. FEBS Lett 587:3096-3101.

Castandet B, Hotto AM, Strickler SR, Stern DB. 2016. ChloroSeq, an optimized chloroplast RNA-seq bioinformatic pipeline, reveals remodeling of the organellar transcriptome under heat stress. G3 6:2817-2827.

Dang Y, Green BR. 2010. Long transcripts from dinoflagellate chloroplast minicircles suggest "rolling circle" transcription. J Biol Chem 285:5196-5203.

Dang Y, Green BR. 2009. Substitutional editing of *Heterocapsa triquetra* chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA. Gene 442:73-80.

De Wit P, Pespeni MH, Ladner JT, Barshis DJ, Seneca F, Jaris H, Therkildsen NO, Morikawa M, Palumbi SR. 2012. The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Mol Ecol Resour 12:1058-1067.

Del Cortona A, Leliaert F, Bogaert KA, Turmel M, Boedeker C, Janouskovec J, Lopez-Bautista JM, Verbruggen H, Vandepoele K, De Clerck O. 2017. The plastid genome In Cladophorales green algae is encoded by hairpin plasmids. Curr Biol, 27: 3771-3782

Dorrell RG, Drew J, Nisbet RE, Howe CJ. 2014. Evolution of chloroplast transcript processing in Plasmodium and its chromerid algal relatives. PLoS Genet 10:e1004008.

Dorrell RG, Gile G, McCallum G, Méheust R, Bapteste EP, Klinger CM, Brillet-Guéguen L, Freeman KD, Richter DJ, Bowler C. 2017. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. Elife 6: 23717.

Dorrell RG, Hinksman GA, Howe CJ. 2016. Diversity of transcripts and transcript processing forms in plastids of the dinoflagellate alga *Karenia mikimotoi*. Plant Mol Biol 90:233-247.

Dorrell RG, Howe CJ. 2012. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. Proc Natl Acad Sci USA 109:18879-18884.

Dorrell RG, Howe CJ. 2015. Integration of plastids with their hosts: lessons learnt from dinoflagellates Proc Natl Acad Sci USA 112: 10247-10254.

Dorrell RG, Klinger CM, Newby RJ, Butterfield ER, Richardson E, Dacks JB, Howe CJ, Nisbet RE, Bowler C. 2016. Progressive and biased divergent evolution underpins the origin and diversification of peridinin dinoflagellate plastids. Mol Biol Evol 34,:361-379.

Fujii H, Chiou TJ, Lin SI, Aung K, Zhu JK. 2005. A miRNA Involved in Phosphate-Starvation Response in *Arabidopsis*. Curr Biol 15: 2038-2043.

Gachon CMM, Heesch S, Kuepper FC, Achilles-Day UEM, Brennan D, Campbell CN, Clarke A, Dorrell RG, Field J, Gontarek S, et al. 2013. The CCAP KnowledgeBase: linking protistan and cyanobacterial biological resources with taxonomic and molecular data. Systemat and Biodiver 11:407-413.

Gavelis GS, White RA, Suttle CA, Keeling PJ, Leander BS. 2015. Single-cell transcriptomics using spliced leader PCR: Evidence for multiple losses of photosynthesis in polykrikoid dinoflagellates. BMC Genom 16:528.

Georg J, Honsel A, Voss B, Rennenberg H, Hess WR. 2010. A long antisense RNA in plant chloroplasts. New Phytol 186:615-622.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J 66:34-44.

Griffith OM. 1991. Cesium chloride gradients in carbon fiber fixed angle rotor for purification of viruses, organelles, and plasmid DNA. Am Biotechnol Lab 9:24, 26.

Haxo FT, Kycia JH, Somers GF, Bennett A, Siegelman HW. 1976. Peridinin chlorophyll A proteins of dinoflagellate *Amphidinium carterae* (Plymouth 450). Plant Physiol 57:297-303.

Hiller RG. 2001. 'Empty' minicircles and petB/atpA and psbD/psbE (cytb(559) alpha) genes in tandem in *Amphidinium carterae* plastid DNA. Febs Lett 505:449-452.

Hotto AM, Castandet B, Gilet L, Higdon A, Condon C, Stern DB. 2015. *Arabidopsis* chloroplast mini-ribonuclease III participates in rRNA maturation and intron recycling. Plant Cell 27:724-740.

Hotto AM, Germain A, Stern DB. 2012. Plastid non-coding RNAs: emerging candidates for gene regulation. Trends Plant Sci 17:737-744.

Howe CJ, Nisbet RER, Barbrook AC. 2008. The remarkable chloroplast genome of dinoflagellates. J Exp Bot 59:1035-1045.

Janouškovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci USA 107:10949-10954.

Janouškovec J, Sobotka R, Lai DH, Flegontov P, Koník P, Komenda J, Ali S, Prásil O, Pain A, Oborník M, et al. 2013. Split photosystem protein, linear-mapping topology, and growth of structural complexity in the plastid genome of *Chromera velia*. Mol Biol Evol 30:2447-2462.

Klinger CM, Paoli L, Newby RJ, Wang MY, Carroll HD, Leblond JD, Howe CJ, Dacks JB, Bowler C, Cahoon AB, et al. 2018. Plastid transcript editing across dinoflagellate lineages shows lineage-specific application but conserved trends. Genom Biol Evol 10:1019-1038.

Lang BF, Burger G. 2007. Purification of mitochondrial and plastid DNA. Nat Protoc 2:652-660.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11-16.

Leung SK, Wong JTY. 2009. The replication of plastid minicircles involves rolling circle intermediates. Nucleic Acids Res 37:1991-2002.

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F, et al. 2015. Ocean plankton. Determinants of community structure in the global plankton interactome. Science 348:1262073.

Lin SJ, Zhang HA, Zhuang YY, Tran B, Gill J. 2010. Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. Proc Natl Acad Sci USA 107:20033-20038.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955-964.

Luro S, Germain A, Sharwood RE, Stern DB. 2013. RNase J participates in a pentatricopeptide repeat protein-mediated 5' end maturation of chloroplast mRNAs. Nucleic Acids Res 41:9141-9151.

Marron AO, Ratcliffe S, Wheeler GL, Goldstein RE, King N, Not F, de Vargas C, Richter DJ. 2016. The evolution of silicon transport in eukaryotes. Mol Biol Evol 33:3226-3248.

Mungpakdee S, Shinzato C, Takeuchi T, Kawashima T, Koyanagi R, Hisata K, Tanaka M, Goto H, Fujie M, Lin S, et al. 2014. Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. Genom Biol Evol 6:1408-1422.

Muñoz-Gómez SAM-F, F.G. Durnin, K. Colp, M. Grisdale, C.J. Archibald, J.M. Slamovits, C.H. 2017. The new red Algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. Curr Biol 27:1677-1684

Nassoury N, Wang Y, Morse D. 2005. Brefeldin a inhibits circadian remodeling of chloroplast structure in the dinoflagellate gonyaulax. Traffic 6:548-561.

Nelson MJ, Dang YK, Filek E, Zhang ZD, Yu VWC, Ishida K, Green BR. 2007. Identification and transcription of transfer RNA genes in dinoflagellate plastid minicircles. Gene 392:291-298.

Nimmo I, Barbrook A, Chen J, Geisler K, Smith A, Aranda M, Purton S, Waller R, Nisbet R, Howe C. 2019. Genetic transformation of the dinoflagellate chloroplast. Manuscript under review.

Nisbet RE, Kurniawan DP, Bowers HD, Howe CJ. 2016. Transcripts in the *Plasmodium* apicoplast undergo cleavage at tRNAs and editing, and include antisense sequences. Protist 167:377-388.

Nisbet RER, Hiller RG, Barry ER, Skene P, Barbrook AC, Howe CJ. 2008. Transcript analysis of dinoflagellate plastid gene minicircles. Protist 159:31-39.

Nisbet RER, Koumandou VL, Barbrook AC, Howe CJ. 2004. Novel plastid gene minicircles in the dinoflagellate *Amphidinium operculatum*. Gene 331:141-147.

Okamoto OK, Hastings JW. 2003. Novel dinoflagellate clock-related genes identified through microarray analysis. J Phycol 39:519-526.

Oldenburg DJ, Bendich AJ. 2004. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. J Mol Biol 335:953-970.

Owari S, Hayashi A, Ishida K-i. 2014. Subcellular localization of minicircle DNA in the dinoflagellate *Amphidinium massartii.* Phycol Res 62:1-8.

Richardson E, Dorrell RG, Howe CJ,. 2014. Genome-wide transcript profiling reveals the coevolution of chloroplast gene sequences and transcript processing pathways in the fucoxanthin dinoflagellate *Karlodinium veneficum*. Mol Biol Evol31: 2376-2386:

Rott R, Drager RG, Stern DB, Schuster G. 1996. The 3' untranslated regions of chloroplast genes in *Chlamydomonas reinhardtii* do not serve as efficient transcriptional terminators. Mol Gen Genet 252:676-683.

Scotto-Lavino E, Du G, Frohman MA. 2006. Amplification of 5 ' end cDNA with 'new RACE'. Nature Protocol 1:3056-3061.

Sharwood RE, Halpert M, Luro S, Schuster G, Stern DB. 2011. Chloroplast RNase J compensates for inefficient transcription termination by removal of antisense RNA. RNA 17:2165-2176.

Suggett DJ, Warner ME, Leggat W. 2017. Symbiotic dinoflagellate functional diversity mediates coral survival under ecological crisis. Trends Ecol Evol 32:735-745.

Wang YL, Jensen L, Hojrup P, Morse D. 2005. Synthesis and degradation of dinoflagellate plastid-encoded psbA proteins are light-regulated, not circadian-regulated. Proc Natl Acad Sci USA 102:2844-2849.

Wang YL, Morse D. 2006. Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate *Lingulodinium*. Nucleic Acids Res 34:613-619.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.

Zandueta-Criado A, Bock R. 2004. Surprising features of plastid *ndhD* transcripts: addition of non-encoded nucleotides and polysome association of mRNAs with an unedited start codon. Nucleic Acids Res 32:542-550.

Zauner S, Greilinger D, Laatsch T, Kowallik KV, Maier UG. 2004. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate *Ceratium horridum*. Febs Lett 577:535-538.

Zghidi-Abouzid O, Merendino L, Buhr F, Ghulam MM, Lerbs-Mache S. 2011. Characterization of plastid *psbT* sense and antisense RNAs. Nucleic Acids Res 39:5379-5387.

Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. Nature 400:155-159.

Zhang ZD, Cavalier-Smith T, Green BR. 2002. Evolution of dinoflagellate unigenic minicircles and the partially concerted divergence of their putative replicon origins. Mol Biol Evol 19:489-500.

Zhelyazkova P, Sharma CM, Förstner KU, Liere K, Vogel J, Börner T. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. Plant Cell 24:123-136.

A

Bar chart. Y-axis: "Mean fold read depth" (0–140). Black bars labeled "Coding": 16S/Minicircle 4, 23S, psbA, petD/ORF2/ORF3/..., psbD/psbE/psbI, petB/atpA, psaA, psbC, psaB, atpB, psbB. Gray bars labeled "Empty": Minicircle 6, Minicircle 5, Minicircle 1, Minicircle 3, Minicircle 2. Legend: ▪ Empty (gray), ▪ Coding (black).

B

psbD/psbE/ORF1/psbI: 1333.2

Coverage plot (Y-axis 0–4000). Gene map below: core: 279.2 | *psbD*: 2298.2 | *psbE*: 778.6 | *psbI*: 581.8 / *ORF1*: 1121.2

C

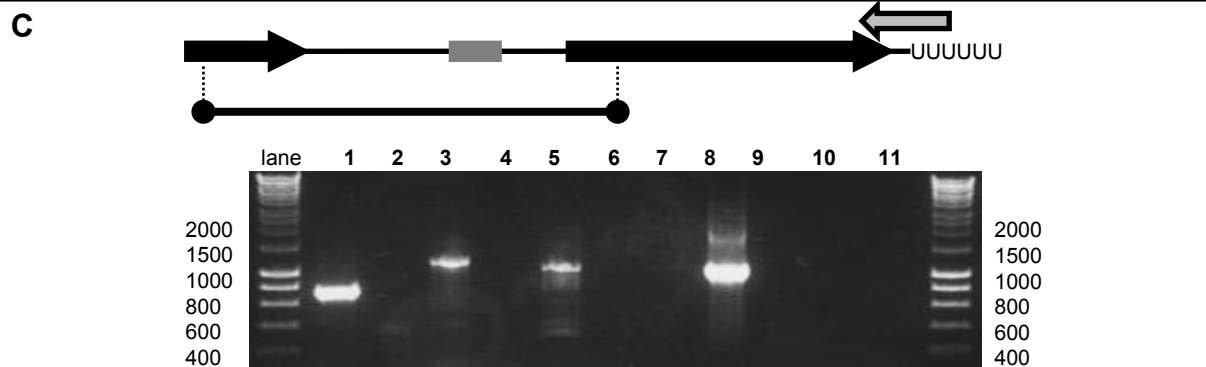| Minicircle	CDS	Core	Minicircle	CDS	Core
16S/ minicircle 4 — 3130.3		96.6	psaB — 753.9		14.6
23S — 2707.6		6.5	*psaB*	939.1	
atpB — 480.7		151.7	psbA — 720.4		0.0
atpB	671.5		*psbA*	811.3	
petB-atpA — 855.0		8.9	psbB — 640.7		74.8
petB	1250.4		*psbB*	950.3	
atpA	1038.6		psbC — 421.9		71.4
petD-ORF2-ORF3-ORF4 — 740.5		185.7	*psbC*	659.3	
petD	1469.6		psbD-psbE-ORF1-psbI — 1333.2		146.3
ORF2	1411.7		*psbD*	2298.2	
ORF3	415.0		*psbE*	786.6	
ORF4	183.0		*ORF1*	1121.2	
			psbI	581.8	
Empty minicircle 1 — 81.6		25.8			
Empty minicircle 2 — 82.4		14.2	Empty minicircle 5 — 132.0		53.4
Empty minicircle 3 — 148.7		78.5	Empty minicircle 6 — 332.9		36.7

A

Ladder	5' UTR	5' CDS	3' CDS	3' UTR

2661
1821
1517
1049
575
483
310

1100 → (a)
1100 → (a)
1100 — (a)
700 — (d)

psbA 3' UTR overexposed

2300 → (g)

B

Ladder	*petB* 5' UTR	*petB* CDS	*atpA* 5' CDS	*atpA* 3' CDS

2661
1821
1517
1049
575
483
310

750 → (e)
750 → (e)
700 → (b)
1500 → (c)
1500 → (c)
1000-350 (f)

atpA 3' CDS overexposed

5400 → (k)
2700 → (h)

C

UUUUUU

lane	1	2	3	4	5	6	7	8	9	10	11

2000
1500
1000
800
600
400

2000
1500
1000
800
600
400

A)

Column headers: psbA, psbB, psbC, psbD, psbE, psbI, psaA, psaB, petB, petD, atpA, atpB, orf1, orf3, orf4, Minicircle 4 / 16S, 23S, Empty minicircle 6, Empty minicircle 1, Empty minicircle 2, Empty minicircle 3, Empty minicircle 5

Row labels (top to bottom):
Amphidinium_carterae_CCMP1314
Amphidinium_massartii_CS259
Togula_jolla_CCCM_725
Gymnodinium_catenatum_GCC744
Scrippsiella_trochoidea_CCMP3099
Scrippsiella_hangoei_like_SHHI4
Peridinium_aciculiferum_PAER2
Brandtodinium_nutriculum_RCC3387
Pelagodinium_beii_RCC1491
Prorocentrum_minimum_CCMP1329
Prorocentrum_minimum_CCMP2233
Heterocapsa_arctica_CCMP445
Heterocapsa_triquestra_CCMP_448
Polarella_glacialis_CCMP_1383
Polarella_glacialis_CCMP2088
Symbiodinium_sp_CCMP421
Symbiodinium_sp_C1
Symbiodinium_sp_C15
Symbiodinium_sp_Mp
Symbiodinium_sp_D1a
Symbiodinium_sp_CCMP2430
Ceratium_fusus_PA161109
Lingulodinium_polyedra_CCMP_1738
Protoceratium_reticulatum_CCMP1899
Gonyaulax_spinifera_CCMP409
Pyrodinium_bahamense_pbaha01
Gambierdiscus_australes_CAWD_149
Alexandrium_margalefi_AMGDE01CS322
Alexandrium_tamarense_CCMP1771
Alexandrium_monilatum_CCMP3105
Alexandrium_andersonii_CCMP2222
Alexandrium_catenella_OF101

Key

Transcript detected

Transcript with poly(U) tail detected

B)

lane 1 2 3

Ladder values (left): 2000, 1500, 1000, 800, 600, 400
Ladder values (right): 2000, 1500, 1000, 800, 600, 400

Sequence alignment:

Identity
1. Amphidinium_carterae
2. Amphidinium_CS-21
3. Amphidinium_massartii

Positions 1–70:
1. Amphidinium_carterae: MLLRSYLAKHSLRRYVNSLGLPTSFLDPGTVLVSSSSFLLRKPKYGLVRLASLNPSLAINVVQVNALNHAAL
2. Amphidinium_CS-21: MLLRSYLAKHSLRRYVNSLGLPISFL-PGTVLVSSSSFLLRKPKYGLVRLPLSTRLFHPMSVQVNALNHAAL

Positions 80–140:
1. Amphidinium_carterae: LSNSLVRSIEDGCNPTIVFQSVMGT---LKRSNRVLGVKLELRALVLAPFARHKVWNYGYGKTNKAVARIKD
2. Amphidinium_CS-21: LSNSLVRSIEGGCNPTIVFQSVMGT---LKRSNRVLGVKLELRALVLAPFARHRVWNYGYGKTNKAVARIKD
3. Amphidinium_massartii: MAAHTARSIQDGCNPTVVFQTTLGTGRATSRSNKILRIKLELRASVLPSFNRHRVWKYGYGRTGKHS

Positions 150–166:
1. Amphidinium_carterae: VGKASLVTYMGLISVSTAISSG
2. Amphidinium_CS-21: VGKASLVTYMGLISVSTAISSG

Empty minicircle	Average read coverage/ bp
Minicircle 1	81.6
Minicircle 2	82.3

Empty minicircle	Average read coverage/ bp
Minicircle 3	148.7
Minicircle 5	132.0

A
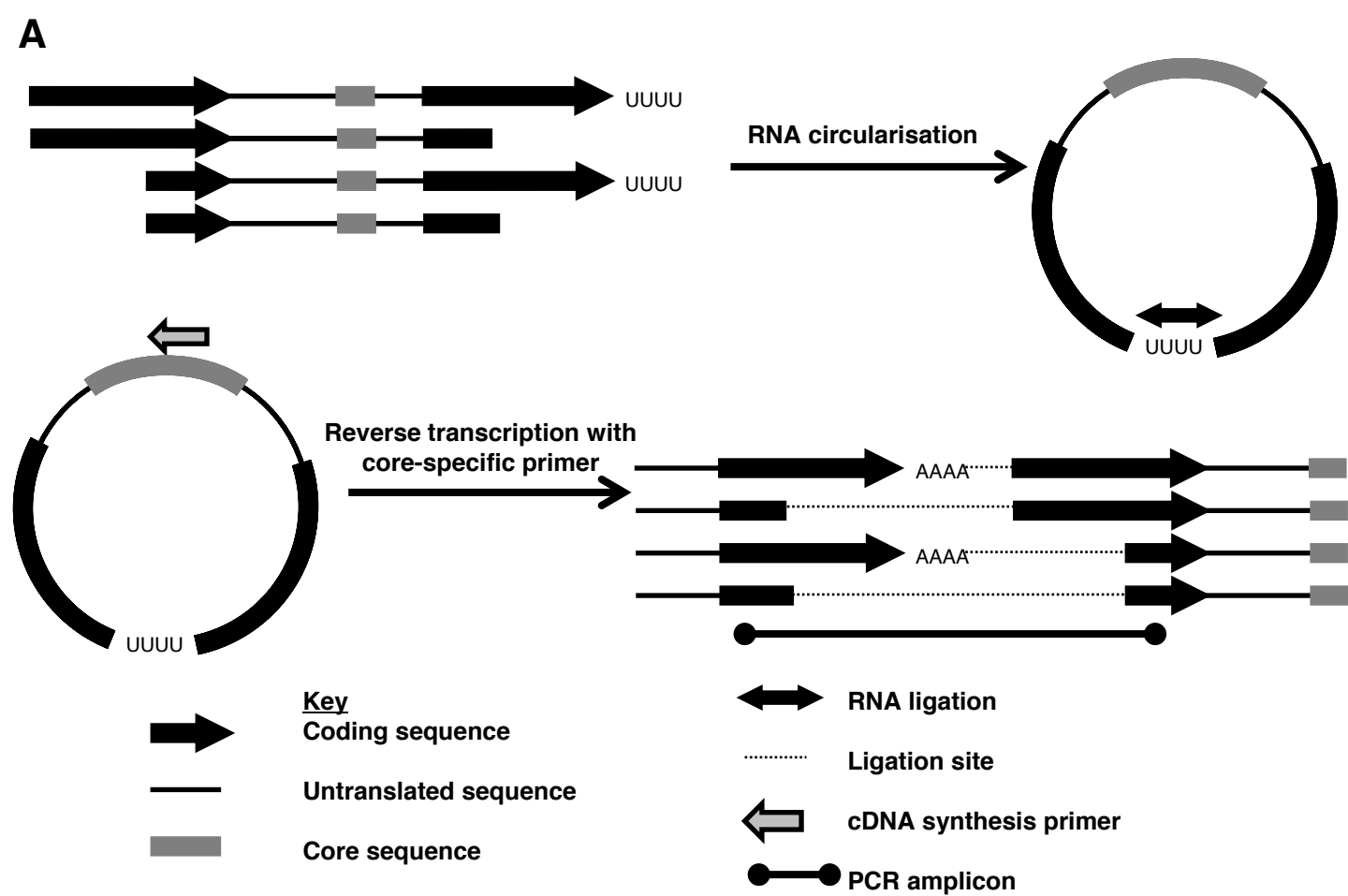
B

i) *psbA*

Multi-copy transcript 1 (3132 nt)

Core-containing transcript 1 (1109 nt)
Core-containing transcript 2 (1110 nt)
Core-containing transcript 3 (1007 nt)
Core-containing transcript 4 (1152 nt)
Core-containing transcript 5 (1152 nt)
Core-containing transcript 6 (1083 nt)
Core-containing transcript 7 (841 nt)
Core-containing transcript 8 (838 nt)
Core-containing transcript 9 (706 nt)
Core-containing transcript 10 (693 nt)
Core-containing transcript 11 (512 nt)
Core-containing transcript 12 (401 nt)
Core-containing transcript 13 (385 nt)

psbA mRNA
Minicircle sequence

core   *psbA*   core   *psbA*   core
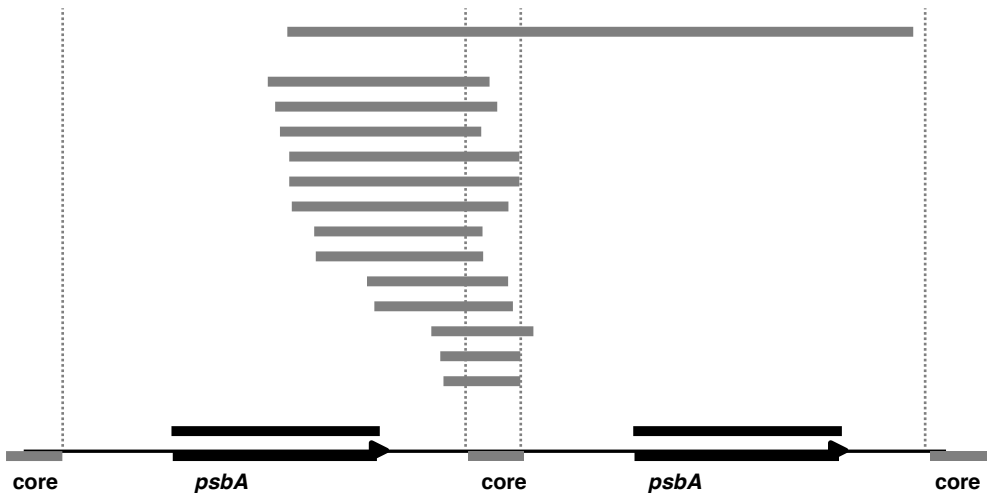
ii) *petB/atpA*

Multi-copy transcript 1 (3488 nt)
Multi-copy transcript 2 (3879 nt)
Multi-copy transcript 3 (3489 nt)
Multi-copy transcript 4 (3481 nt)
Multi-copy transcript 5 (3481 nt)

Core-containing transcript 1 (1833 nt)
Core-containing transcript 2 (1565 nt)
Core-containing transcript 3 (1566 nt)
Core-containing transcript 4 (1266 nt)
Core-containing transcript 5 (1833 nt)
Core-containing transcript 6 (1482 nt)
Core-containing transcript 7 (1225 nt)

atpA mRNA
petB mRNA
Minicircle sequence

core   *petB*   *atpA*   core   *petB*   *atpA*

Key
Coding sequence
Untranslated sequence
Core sequence

RNA ligation
Ligation site
cDNA synthesis primer
PCR amplicon

RNA circularisation

Reverse transcription with core-specific primer
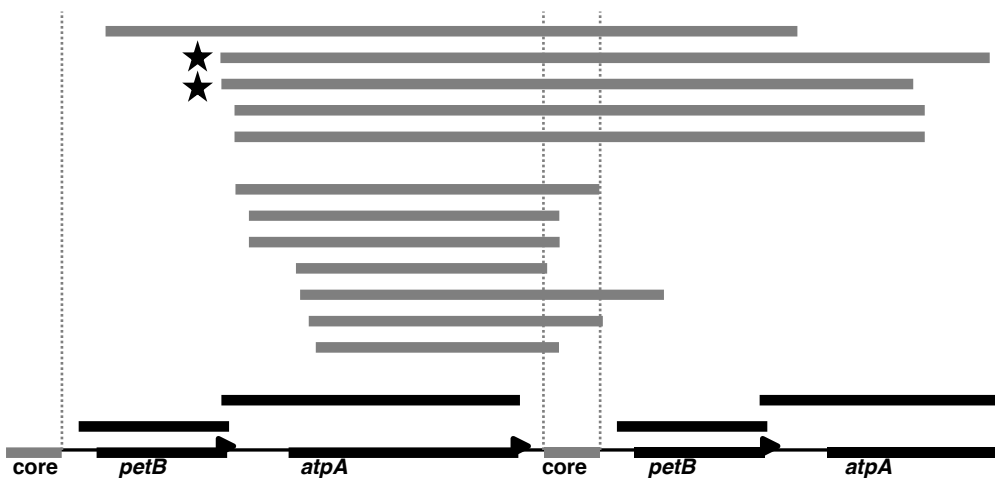
A

i) RNA ligation
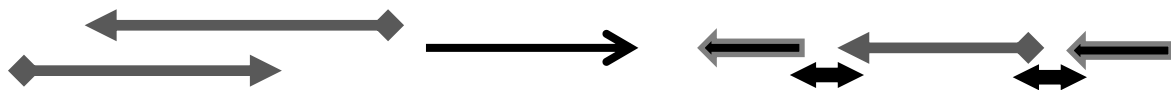
ii) cDNA synthesis

iii) PCR amplification

Key

Transcript sequence

RNA adapter sequence

RNA ligation

cDNA synthesis

PCR amplicons for next-generation sequencing or cloning
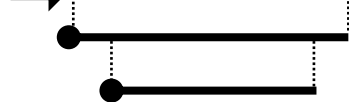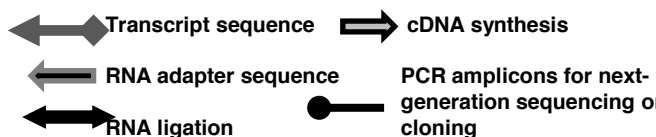
B

Minicircle	Sense	Antisense	% sense
Plastid_psbD-psbE-ORF1-psbI	20	17	45.9
Plastid_petB-atpA	10	2	16.7
Plastid_23S	374	34	8.3
Plastid_psaB	5	3	37.5
Plastid_psbA	11	2	15.4
Plastid_petD-ORF2-ORF3-ORF4	4	4	50.0
Plastid_psbC	0	3	100.0
Plastid_psbB	6	0	0
Plastid_atpB	5	2	28.6
Plastid_16S	448	0	0
Empty minicircle 110	4	0	0
Empty minicircle 3	1	0	0
Total	**888**	**67**	**7.0**

C

i) RNA adapter sequence

1   10   20   28
G C U G A U G G C G A U G A G C A C U G G G U U G C A A

5' end → 3' end
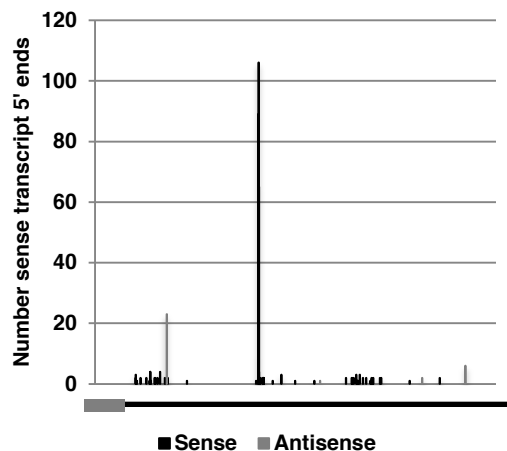
ii) Alignment of *A. carterae atpA* gene and antisense 5' RACE products

forward genomic
ACCTGTCTCTTCCAGTGTTCATTGCGTATGATGATCTGGCAAAGCAGGCCTCCGCTTATCGTGAGA

reverse
N L S L P V F I A Y D D L A K Q A S A Y R E

TGGACAGAGAAGGTCACAAGTAACGCATACTACTAGACCGTTTCGTCCGGAGGCGAATAGCACTCT

5' RACE product 1
5' RACE product 2   ACCTGTCTCTTCCAGTGCTCATCGCCATCAGCAACCCAGTGCTCATCGCCATCAGC
5' RACE product 3   ACCTGTCTCTTCCAGTGCTCATCGCCATTAGCAACCCAGTGCTCATCACCATCAGC
RNA adapter (reversed) TTGCCTCTTCCAGTGCTCATCGCCATCAGCAACCCAGTGCTCATCGCCATCAGC

A)

Quality score distribution
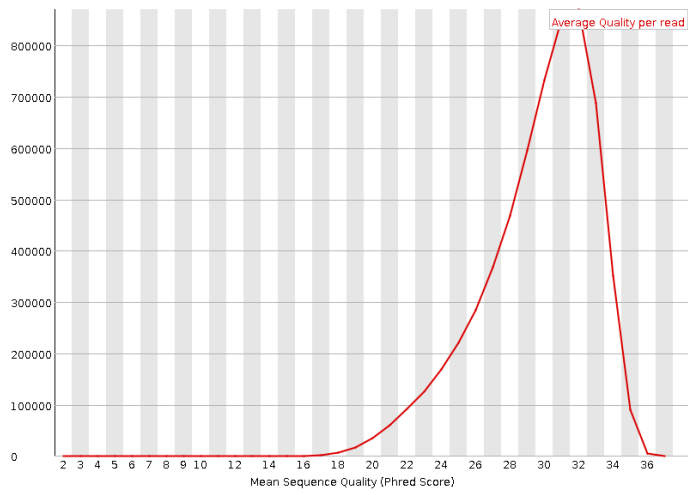
GC content distribution
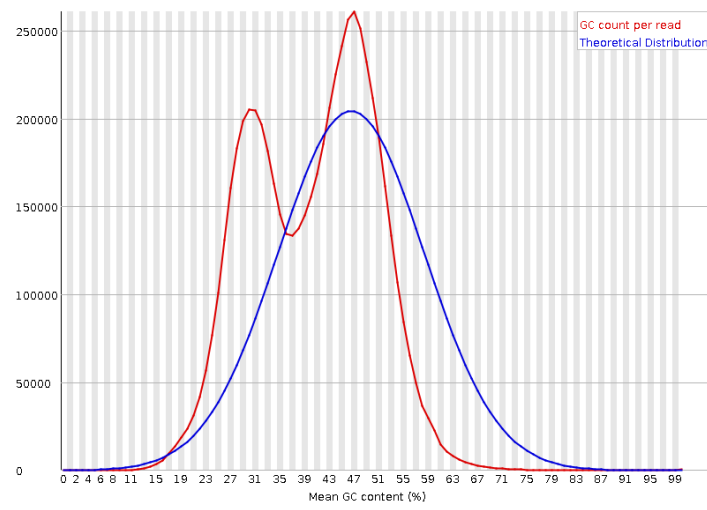
B)

Nucleotide distribution

QC score boxplots

i) Forward reads

ii) Reverse reads

N
T   C
G   A

Quartiles
Medians

Table S1. Tabulated read coverage for *Amphidinium carterae* plastid minicircles.

Sheet 1 provides fasta format sequences of the Amphidinium minicircles identified through next-generation sequencing of CsCl-purified genomic DNA.

Sheet 2 provides an overview of the number of reads from the adaptor-ligated Amphidinium RNA library mapped to minicircle sequence.

Sheet 3 provides detailed read coverage values for RNA sequence data, for each residue in each minicircle sequence. Each minicircle sequence is given with the core region corresponding to residues 1-280.

Sheet 4 provides GenBank flatfile form sequences of two variants of novel minicircle 6, assembled by PCR.

Table S2. Primers used for RT-PCR analysis. The annealing position of each primer is given, relative to a linearised sequence of the minicircle, in which the core region corresponds to residues 1-280.

---

1. PCR of novel minicircle 6

Name	Sequence	Annealing site
Forward_1	AAGGTAGAGAACATCAAACAGG	364 F
Forward_2	GTCTACCGAGTACCCACCC	724 F
Forward_3	GAGAGTCCCGTCAAATGG	1138 F
Forward_4	CTCGTCAGGCTTGCATC	1491 F
Forward_5	GTGGTACTTGTTCTATGCGTTG	1904 F
Forward_6	CACCATAGGGACATGAATGAG	2085 F
Reverse_1	CACGACTGTGGAATGCAC	103 R
Reverse_2	CGCACTCCATCGAACC	495 R
Reverse_3	ACGCGGTAACATTGACC	866 R
Reverse_4	TACGGAGGAGGAAACTGG	1443 R
Reverse_5	TGGTTGGATTACAGCCG	1585 R
Reverse_6	AAGTGTAAGGAGAAGTAGGGC	1862 R

---

2. Circular RT-PCR of multi-copy transcripts

Minicircle			core-specific cDNA primer		Annealing site	
psbA			AGTCTCCCGATTGTCTATTCTC		41 R	

	PCR forward primers	Annealing site	Position	PCR reverse primers	Annealing site	Position
1	CGAGTCAGAGGCATCAAAC	228 F	Core	CTTTAGACTGCGGTGTGAAC	563 R	5' UTR
2	TACATTGAGTAGGCATCTTTAATAGC	512 F	5' UTR	AGTTAGAGCGAATAAGGCTTG	858 R	CDS 5' end
3	CTGGGGTTCTTTCGTTCAAAC	860 F	CDS 5' end	GATACCAATTACAGGCCAAGC	1670 R	CDS 3' end
4	ACGCTCATAACTTCCCTCTTG	1825 F	CDS 3' end	ATCGTTAATCAGAAAGCCTAGTC	1918 R	3' UTR
5	CCTCCTACCGAAAGTCAATTC	2238 F	3' UTR	ATTGACTTTCGGTAGGAGGC	2256 R	3' UTR

Minicircle	positive control cDNA primer	Annealing site	Position	core-specific cDNA primer	Annealing site	Position
petB/ atpA	GCATTGCTGTGGAATAGAC	2417 R		CCTTTCCGTATCCTTCATTCG	2679 R	

	PCR forward primers	Annealing site		PCR reverse primers	Annealing site	
1	GAAAATCCAGGTCATATCATAGGAG	133 F	Core	GCAACTCAAGACGCTCTTCAC	498 R	*petB* CDS 5' end
2	GCAGACGATATCCTCTCTAAG	507 F	*petB* CDS 5' end	CAAACACTGTACCCAACGAAG	963 R	*petB* CDS 3' end
3	CCTTCTCCTTACTCATTTCCTAATG	1058 F	*petB* CDS 3' end	ACAAGGCCATATACGACATC	1276 R	*atpA* CDS 5' end
4	TCAGTCTGTCTGCGAACCAC	1529 F	*atpA* CDS 5' end	CTTCTGACCCACAGGGACAT	1715 R	*atpA* CDS 3' end
5	GGTCTTCTTGGGTTATTTCC	2480 F	*atpA* CDS 3' end	CCTTTCCGTATCCTTCATTCG	2679 R	*atpA* 3' UTR
control	GGTCTTCTTGGGTTATTTCC	2480 F	*atpA* CDS 3' end	ACAAGGCCATATACGACATC	1276 R	*atpA* CDS 5' end

---

3. RT-PCR to detect polyuridylylated multi-copy transcripts

Gene	cDNA primer	Annealing site
psbA	AAAAAAARATAAAGGGG	1870/ 1872 R
petB	AAAAAAAWAAGAATAGAAGT	1123/ 1226 R
atpA	AAAAAAAAAAAAAAAAAATATACAC	2592 R

Gene	PCR forward primer	Annealing site	PCR reverse primer	Annealing site
psbA	CAAGCCTTATTCGCTCTAACT	838 F	ATCGTTAATCAGAAAGCCTAGTC	1918 R
petB	GCAGACGATATCCTCTCTAAG	507 F	ACAAGGCCATATACGACATC	1276 R
atpA	TCAGTCTGTCTGCGAACCAC	1529 F	CCTTTCCGTATCCTTCATTCG	2679 R

---

4. 5' RACE of antisense transcripts

RNA adapter	GCUGAUGGCGAUGAGCACUGGGUU	Transcript	cDNA primer	Annealing site
Adapter primer	GCTGATGGCGATAGC	psbA antisense	CAAGCCTTATTCGCTCTAACT	838 F
Adapter primer	GATGAGCACTGGGTTGC	atpA antisense	TCAGTCTGTCTGCGAACCAC	1529 F

Gene	Gene-specific PCR primer 1	Annealing site	Genes-specific PCR primer 2	Annealing site
psbA	CTGGGGTTCTTTCGTTCAAAC	860 F	CTTCTAACGCAATCGGTGTCC	1075 F
atpA	CAGCGTGAACTAATTATTGGTG	1599 F	ACGAGAAGGTTCTATCCGTCTATG	1675 F

---

5. Circular RT-PCR of antisense transcripts

	psbA	Annealing site	Position	petB/ atpA	Annealing site	Position
cDNA synthesis primers						
1	CGAGTCAGAGGCATCAAAC	228 F	5' UTR	GCAGACGATATCCTCTCTAAG	507 F	*petB* CDS 5' end
2	CTGGGGTTCTTTCGTTCAAAC	860 F	CDS 5' end	ACGAGAAGGTTCTATCCGTCTATG	1675 F	*atpA* CDS 5' end
3	CCTCTCTTGGTGTTGCTACTATG	1678 F	CDS 3' end	GTAGGTATCTCGGTTACACG	2190 F	*atpA* CDS 3' end
4	GACTAGGCTTTCTGATTAACGAT	1896 F	3' UTR	CGAATGAAGGATACGGAAAGG	2659 F	*atpA* 3' UTR
PCR forward primers						
1	AGTCTCCCGATTGTCTATTCTC	41 R	core	GCAACTCAAGACGCTCTTCAC	498 R	*petB* CDS 5' end
2	AGTTAGAGCGAATAAGGCTTG	858 R	CDS 5' end	CACCAATAATTAGTTCACGCTG	1620 R	*atpA* CDS 5' end
3	GATACCAATTACAGGCCAAGC	1670 R	CDS 3' end	GCATTGCTGTGGAATAGAC	2417 R	*atpA* CDS 3' end
4	ATCGTTAATCAGAAAGCCTAGTC	1918 R	3' UTR	TCGTTCAACCACACTTTATACAGAA	2607 R	*atpA* 3' UTR
PCR reverse primers						
1	TACATTGAGTAGGCATCTTTAATAGC	512 F	5' UTR	ATCATCCAAGCGGCAACT	588 F	*petB* CDS 5' end
2	CTTCTAACGCAATCGGTGTCC	1075 F	CDS 5' end	TCCCTGTGGGTCAGAAG	1699 F	*atpA* CDS 5' end
3	ACGCTCATAACTTCCCTCTTG	1825 F	CDS 3' end	GGTCTTCTTGGGTTATTTCC	2480 F	*atpA* CDS 3' end
4	CCTCCTACCGAAAGTCAATTC	2238 F	3' UTR	GAAAATCCAGGTCATATCATAGGA	133 F	core

---

6. RT-PCR to detect polyuridylylated antisense transcripts

oligo-d(A) cDNA synthesis primer		GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAAAAA	

Amplicon	PCR gene-specific primer	Annealing site	Gene-specific cDNA synthesis primer	Annealing site
Antisense psbA	GCTCGTGCATTACCTCGATAC	1821 R	CAAGCCTTATTCGCTCTAACT	838 F
Antisense psbA	CTTTAGACTGCGGTGTGAAC	563 R	GACTAGGCTTTCTGATTAACGAT	1896 F
Antisense petB	AAGGTGTGAGCCTGATAGAAC	1033 R	GCAGACGATATCCTCTCTAAG	507 F
Antisense atpA	CTTCTGACCCACAGGGACAT	1715 R	ACGAGAAGGTTCTATCCGTCTATG	1675 F
Sense psbA	CAAGCCTTATTCGCTCTAACT	838 F	n/a	
Sense petB	GCAGACGATATCCTCTCTAAG	507 F	n/a	
Sense atpA	ACGAGAAGGTTCTATCCGTCTATG	1675 F	n/a	

Table S3. Tabulated circular RT-PCR sequences. This table provides 5' and 3' terminus positions for core-containing and antisense transcripts identified by circular RT-PCR.

The terminus positions of each transcript, and the primers used for PCR amplification of the transcript, are given corresponding to a linearised sequence of the minicircle, in which the core region corresponds to residues 1-280.

For reference, the positions of each coding sequence, as well as the intervals of residues over which mature mRNA 5' ends and poly(U) sites have been identified in previous circular RT-PCR studies (Barbrook, et al. 2012), are supplied.

1. Core-containing psbA transcripts

	Minicircle length	2311	Core	1-281

	5' end	3' end	mRNA 5' end	poly(U) site
psbA	834	1856	600-829	1870-1872

	5' end	3' end	R primer	F primer	Poly(U)	Size	Notes
multi-copy transcript 1	1407	2228 (2)	1670 R	1825 F	N	3132	
core-containing transcript 1	1310	108 (2)	1670 R	2238 F	N	1109	
core-containing transcript 2	1347	146 (2)	1670 R	1825 F	N	1110	
core-containing transcript 3	1371	67 (2)	1918 R	2238 F	N	1007	
core-containing transcript 4	1416	257 (2)	1670 R	2238 F	N	1152	
core-containing transcript 5	1416	257 (2)	1669 R	1825 F	N	1152	
core-containing transcript 6	1430	202 (2)	1670 R	2238 F	N	1083	
core-containing transcript 7	1542	72 (2)	1670 R	2238 F	N	841	
core-containing transcript 8	1549	76 (2)	1670 R	1825 F	N	838	
core-containing transcript 9	1806	201 (2)	1918 R	2238 F	N	706	
core-containing transcript 1	1843	225 (2)	1918 R	2238 F	N	693	
core-containing transcript 1	2127	328 (2)	2256 R	228 F	N	512	
core-containing transcript 1	2172	262 (2)	2256 R	228 F	N	401	
core-containing transcript 1	2188	262 (2)	2256 R	228 F	N	385	

2. Core-containing petB/atpA transcripts

	Minicircle length	2713	Core	1-281

	5' end	3' end	mRNA 5' end	poly(U) site
petB	456	1115	310-424	1122-1126
atpA	1206	2582	1080-1088	2591

	5' end	3' end	R primer	F primer	Poly(U)	Size	Notes
atpA mRNA 1	1086	2591	1276 R	2480 F	U24	1529	
atpA mRNA 2	1086	2591	1276 R	2480 F	U32	1537	
atpA mRNA 3	1087	2591	1276 R	2480 F	U26	1530	
atpA mRNA 4	1087	2591	1276 R	2480 F	U35	1539	
atpA mRNA 5	1089	2591	1276 R	2480 F	U26	1528	
multi-copy transcript 1	501	1276 (2)	963 R	1058 F	N	3488	
multi-copy transcript 2	1080	2246 (2)	1276 R	1529 F	N	3879	5' end similar to mature mRNA 5' end position
multi-copy transcript 3	1085	1861 (2)	1276 R	1529 F	N	3489	5' end similar to mature mRNA 5' end position
multi-copy transcript 4	1151	1919 (2)	1276 R	1529 F	N	3481	
multi-copy transcript 5	1151	1919 (2)	1276 R	1529 F	N	3481	
core-containing transcript 1	1157	277 (2)	1276 R	2480 F	N	1833	
core-containing transcript 2	1224	76 (2)	1276 R	2480 F	N	1565	
core-containing transcript 3	1224	77 (2)	1276 R	2480 F	N	1566	
core-containing transcript 4	1461	14 (2)	1715 R	2480 F	N	1266	
core-containing transcript 5	1483	603 (2)	1715 R	2480 F	N	1833	
core-containing transcript 6	1526	295 (2)	1715 R	2480 F	N	1482	
core-containing transcript 7	1562	74 (2)	1715 R	2480 F	N	1225	

3. Antisense psbA transcripts

	Minicircle length	2311	Core	1-281

	5' end	3' end	mRNA 5' end	poly(U) site
psbA	834	1856	600-829	1870-1872

	5' end	3' end	cDNA primer	R primer	F primer	Poly(U)	Size	Notes
antisense transcript 1	146 (2)	1347	1896 F	2238 F	1670 R	N	1110	
antisense transcript 2	72 (2)	1542	1896 F	2238 F	1670 R	N	841	
antisense transcript 3	1868	1071	1678 F	1825 F	1670 R	N	797	
antisense transcript 4	1867	1520	1678 F	1825 F	1670 R	N	347	
antisense transcript 5	1678	544	860 F	1075 F	858 R	N	1134	
antisense transcript 6	1196	715	860 F	1075 F	858 R	N	481	3' end similar to mature mRNA 5' end position
antisense transcript 7	1188	765	860 F	1075 F	858 R	N	423	3' end similar to mature mRNA 5' end position
antisense transcript 8	1092	168	860 F	860 F	858 R	N	924	
antisense transcript 9	916	168	860 F	860 F	858 R	N	748	

4. Antisense petB/ atpA transcripts

	Minicircle length	2713	Core	1-281

	5' end	3' end	mRNA 5' end	poly(U) site
petB	456	1115	310-424	1122-1126
atpA	1206	2582	1081-1088	2591

	5' end	3' end	cDNA primer	R primer	F primer	Poly(U)	Size	Notes
antisense transcript 1	832 (2)	2437	507 F	588 F	2607 R	N	1108	
antisense transcript 2	301 (2)	2086	2659 F	133 F	2417 R	N	928	
antisense transcript 3	294 (2)	2084	2659 F	133 F	2417 R	N	923	
antisense transcript 4	204 (2)	1218	2659 F	133 F	1620 R	N	1699	
antisense transcript 5	192 (2)	1677	2659 F	133 F	2417 R	N	1228	
antisense transcript 6	2040	1532	1675 F	1699 F	1620 R	N	508	
antisense transcript 7	1770	1509	1675 F	1699 F	1620 R	N	261	
antisense transcript 9	916	168	860 F	860 F	858 R	N	706	

Table S4. Tabulated transcript 5' ends identified by next-generation sequencing.

This table shows the position and orientation of 5' end termini within the adaptor-ligated Amphidinium RNA library against each published minicircle sequence. Each minicircle sequence is given with the core region corresponding to residues 1-280.

Sheet 1 provides tabulated BLAST outputs of each of the 5' end sequences identified.

Sheet 2 provides an overview of the number and orientation of 5' termini mapped to minicircle.

Sheet 3 tabulates the number and orientation of 5' end termini identified for each residue in each minicircle sequence.

Table S5. Tabulated northern probes. This table provides a linearised sequence of the T7 arm of the pGem-tEasy vector, and sequences of the template strand regions used for the generation of northern probes against sense transcripts, and the non-coding strand regions used for the generation of northern probe
The sequence intervals covered by each probe are given corresponding to a linearised sequence of the minicircle, in which the core region corresponds to residues 1-280.

Probe		Start	End	Sequence
T7 arm of probe seque				TAATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCCGCGGGATT
psbA				
5' UTR	Sense	852	510	AGCGAATAAGGCTTGTCATAATAGATATGGAAAAAGGTTTACGCTGGCCTCACTAACGAGAGTTCGTCTTCTTACCGAGGCACCGAGGTACCAGGAAGGTAAACGCGGATACACCATTACAGCTCACCACTCGACTACTTGCCTACCCCTGCCACTCCACCGCCACAAGCACTCCGTAAGACGGGGACTGAATACACAGACCATTTGATAGGATCCCCGAA
5' CDS	Sense	1077	838	AAGAAGGAATAACAGCACCAGAGATGATGTTGTTACCATAGATAAGGGAACCTGCAACTGGCTCACGGATACCATCAATATCGACTGCAGGAGCAAGGAAGAAAGCAGTGATATAAGCTACGGTAGCAAGGGAGAGAAGTGGGAAGACGAGGAGACCGAACCAACCGATATAAAGACGGTTAGAGGAAGAAGTGATTGTTTGAACGAAAGAACCCCAGGAGTT
3' CDS	Sense	1845	1616	CAAGAGGGAAGTTATGAGCGTTACGCTCGTGCATTACCTCGATACCAAGATCAGCACGGTTGAGGATATCAGCCCAAGAGTTAATGTAGTGACCAGATTCGTCAAGAATAGACTGGTTGAAGTTGAAACCGTTAAGGTTGAATGCCATAGTAGCAACACCAAGAGAGGTGAACCAGATACCAATTACAGGCCAAGCTGCAAGGAAGAAGTGAAGAGAACGGGA
3' UTR	Sense	2256	1896	ATTGACTTTCGGTAGGAGGCTCAAAGAAGGAAAGGCACTATTACCTAACCAGAGATAACTACAAGCGGAGATTCCAAGCCTTTATGGACTCATTCTAGCATCTAAGAGGTGTTGCAGGATAACCATACAACAGTTGTACCTGTTTACCCGGGTATTGATACGTCGCCACAACGATCTCTACTTAGGTAGGTAGGTCAATACGTCGTTCGAAAGGCGCCCGAA
5' UTR	Antisense	510	852	TTACATTGAGTAGGCATCTTTAATAGCTTGTAGGGTTCACACCGCAGTCTAAAGGTAGAAAGGTCCAACAATCAGGTCATGTGGCCCAGAAACCACGCTGGTAACCTGAGACCCTACCTCCCTTCGGGGATCCTATCAAATGGTCTGTGTATTCAGTCCCCGTCTTACGGAGTGCTTGTGGCGGTGGAGTGGCAGGGGTAGGCAAGTAGTCGAGTGGTGA(
5' CDS	Antisense	838	1077	CAAGCCTTATTCGCTCTAACTCCTGGGGTTCTTTCGTTCAAACAATCACTTCTTCCTCTAACCGTCTTTATATCGGTTGGTTCGGTCTCCTCGTCTTCCCACTTCTCTCCCTTGCTACCGTAGCTTATATCACTGCTTTCTTCCTTGCTCCTGCAGTCGATATTGATGGTATCCGTGAGCCAGTTGCAGGTTCCCTTATCTATGGTAACAACATCATCTCTG(
3' CDS	Antisense	1616	1845	CAACAACTCCCGTTCTCTTCCACTTCTTCCTTGCAGCTTGGCCTGTAATTGGTATCTGGTTCACCTCTCTTGGTGTTGCTACTATGGCATTCAACCTTAACGGTTTCAACTTCAACCAGTCTATTCTTGACGAATCTGGTCACTACATTAACTCTTGGGCTGATATCCTCAACCGTGCTGATCTTGGTATCGAGGTAATGCACGAGCGTAACGCTCATAACTTC
3' UTR	Antisense	1896	2256	GACTAGGCTTTCTGATTAACGATGATGTAATATTATTAAGAACGATGGCTAGTAATAATGCACGAAGATTAAGAACGATTAGATTACCTTCCTTATTATTGGGAACGTTTGAACGATGAATTAAACTCACAGTAGGCCCTTCGGGCGCCTTTCGAACGACGTATTGACCTACCTACCTAAGTAGAGATCGTTGTGGCGACGTATCAATACCCGGGTAAACAGGTA
petB/ atpA				
5' UTR	Sense	498	206	GCAACTCAAGACGCTCTTCACACCAATCGTAAATGAAACCCATGTAGAGAATGCTTGTTAACGAATAGGCATGAGGGTTCCTGGTATACTTCTGGCCTAACCAGGAGGCATTTGTTAATTAATTCGAGACCATATTGTATCTCTCGACTCCCAACAGTGACTCTGCACTCATGGTACTGTGTTGATGGGGAGCCTTTCATCTTATCTTTCTATTAAGTGTCGTT
petB	Sense	961	504	AACACTGTACCCAACGAAGGGAACACGTTATTGAGAGCCTCAGGAGTCGCAGTTACGATCTTGCACGCCCAATAGCCAACCTGATCCCATGGTAGGGAGTAACCTGTGACACCGAAGGACACAGTACAGATGGCTAGGATAACGCCGGAGATCCATGTGAGTTCCCTTGGCTTCTTGAAACCCGCAGTCAGATAAACACGACTAACGTGAAGAAGTAGGAC
atpA 5'	Sense	1715	1531	CTTCTGACCCACAGGGACATAGACGGATAGAACCTTCTCGTACTTAAGGTTAACGATGGTATCGAGGCAAATAGAGGTCTTACCGGTCTGACGATCACCAATAATTAGTTCACGCTGACCTCGACCAATTGGAATCATTGCATCGATTGATACGATACCAGTAGCGAGTGGTTCGCAGACAGACT
atpA 3'	Sense	2525	2190	TAGGAAACACCAAGACGGTAGGCTAAGGAAATAACCCAAGAAGACCTTCACAAGAGAGTCAGATGGAATCCTAGCCAATACACCAGAACCTGCAAGCGATAGTACTGCGCATTGCTGTGGAATAGACATTGGAGATGCAACATCCTGCTTAAGAACCTCACGAATCCTCCTACCATTAGCTAGAGCCCTCGCTGTGTCCTCACCGAGATCGAGGCAAACTG
5' UTR	Antisense	206	498	TCCCGATCTCACAAGTCTCCATTAAGGACTAGAGCTTGAAAGACAAATGAACGACTAGATAAATGTAGAAACGACACTTAATAGAAAGATAAGATGAAAGGCTCCCCATCAACACAGTACCATGAGTGCAGAGTCACTGTTGGGAGTCGAGAGATACAATATGGTCTCGAATTAATTAACAAATGCCTCCTGGTTAGGCCAGAAGTATACCAGGAACCCTCATG(
petB	Antisense	504	961	ATTGCAGACGATATCCTCTCTAAGTTCGTTCCTTCTCATGTTAACATCTTCTATTGCTTCGGCGGTATCGTGTTGACATGTTTCATCATCCAAGCGGCAACTGGTTTTGCCATGACTCTCTACTACAGGCCTAACGTGGTTGAGGCACTCTCGAGTGTCACGTATATCACAAATGAGGTATCATTCGGTTGGTTGGTTCGATCCATCCACCGTACCTCTTCTC
atpA 5'	Antisense	1531	1715	AGTCTGTCTGCGAACCACTCGCTACTGGTATCGTATCAATCGATGCAATGATTCCAATTGGTCGAGGTCAGCGTGAACTAATTATTGGTGATCGTCAGACCGGTAAGACCTCTATTTGCCTCGATACCATCGTTAACCTTAAGTACGAGAAGGTTCTATCCGTCTATGTCCCTGTGGGTCAGAAG
atpA 3'	Antisense	2190	2525	GTAGGTATCTCGGTTACACGTGTCGGTTCTGCGGCACAACAGGACCAGATGAAGATGGTTGCGGGTCGACTTAAGCTTACTCTTGCACAGTTCGTTGAGTTGGAGGCTTTCTCTCAGTTTGCCTCTGATCTCGGTGAGGACACAGCGAGGGCTCTAGCTAATGGTAGGAGGATTCGTGAGGTTCTTAAGCAGGATGTTGCATCTCCAATGTCTATTCCACA

AGGGAGGTAGGGTCTCAGGTTACCAGCGTGGTTTCTGGGCCACATGACCTGATTGTTGGACCTTTCTACCTTTAGACTGCGGTGTGAACCCTACAAGCTATTAAAGATGCCTACTCAATGTAA
TAGAGCGAATAAGGCTTG
AGTTGTTG
GGGCCTACTGTGAGTTTAATTCATCGTTCAAACGTTCCCAATAATAAGGAAGGTAATCTAATCGTTCTTAATCTTCGTGCATTATTACTAGCCATCGTTCTTAATAATATTACATCATCGTTAATCAGAAAGCCTAGTC
GCTGTAATGGTGTATCCGCGTTTACCTTCCTGGTACCTCGGTGCCTCGGTAAGAAGACGAACTCTCGTTAGTGAGGCCAGCGTAAACCTTTTTCCATATCTATTATGACAAGCCTTATTCGCT
GTGCTGTTATTCCTTCTT
CCTCTTG
ACAACTGTTGTATGGTTATCCTGCAACACCTCTTAGATGCTAGAATGAGTCCATAAAGGCTTGGAATCTCCGCTTGTAGTTATCTCTGGTTAGGTAATAGTGCCTTTCCTTCTTTGAGCCTCCTACCGAAAGTCAAT

TCTACATTTATCTAGTCGTTCATTTGTCTTTCAAGCTCTAGTCCTTAATGGAGACTTGTGAGATCGGGA
CGAGGACCATCAAGCCAGAAGAGGTACGGTGGATGGATCGAACCAACCAACCGAATGATACCTCATTTGTGATATACGTGACACTCGAGAGTGCCTCAACCACGTTAGGCCTGTAGTAGAGAGTCATGGCAAAACCAGTTGCCGCTTGGATGATGAAACATGTCAACACGATACCGCCGAAGCAATAGAAGATGTTAACATGAGAAGGAACGAACTTAGAGAGGATATCGTCTGCAAT

GAGAGAAAGCCTCCAACTCAACGAACTGTGCAAGAGTAAGCTTAAGTCGACCCGCAACCATCTTCATCTGGTCCTGTTGTGCCGCAGAACCGACACGTGTAACCGAGATACCTAC
CCTATTCGTTAACAAGCATTCTCTACATGGGTTTCATTTACGATTGGTGTGAAGAGCGTCTTGAGTTGC
GGCTTGATGGTCCTCGTCCTACTTCTTCACGTTAGTCGTGTTTATCTGACTGCGGGTTTCAAGAAGCCAAGGGAACTCACATGGATCTCCGGCGTTATCCTAGCCATCTGTACTGTGTCCTTCGGTGTCACAGGTTACTCCCTACCATGGGATCAGGTTGGCTATTGGGCGTGCAAGATCGTAACTGCGACTCCTGAGGCTCTCAATAACGTGTTCCCTTCGTTGGGTACAGTGTT

AGCAATGCGCAGTACTATCGCTTGCAGGTTCTGGTGTATTGGCTAGGATTCCATCTGACTCTCTTGTGAAGGTCTTCTTGGGTTATTTCCTTAGCCTACCGTCTTGGTGTTTCCTA