# Identification of Polycistronic Transcriptional Units and Non-canonical Introns in Green Algal Chloroplasts Based on Long-read RNA Sequencing Data

**Xiaoxiao Zou** ( zouxiaoxiao@itbb.org.cn )
Institute of Tropical Bioscience and Biotechnology https://orcid.org/0000-0002-1758-9718
**Heroen Verbruggen**
University of Melbourne School of BioSciences
**Tianjingwei Li**
ITBB: Institute of Tropical Bioscience and Biotechnology
**Jun Zhu**
ITBB: Institute of Tropical Bioscience and Biotechnology
**Zuo Chen**
ITBB: Institute of Tropical Bioscience and Biotechnology
**Henqi He**
ITBB: Institute of Tropical Bioscience and Biotechnology
**Shixiang Bao**
ITBB: Institute of Tropical Bioscience and Biotechnology
**Jinhua Sun**
Environment and Plant Protection institute, CATAS

**Research article**

**Keywords:** chloroplast genome, polycistronic transcripts, gene fragmentation, freestanding ORF, group II intron, siphonous algae, PacBio, Iso-seq

**Posted Date:** December 1st, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-114353/v1

**Version of Record:** A version of this preprint was published at BMC Genomics on April 23rd, 2021. See the published version at https://doi.org/10.1186/s12864-021-07598-y.

# Abstract

**Background:** Chloroplasts are important semi-autonomous organelles in plants and algae. Unlike higher plants, the chloroplast genomes of green algal linage have distinct features both in organization and expression. Despite the architecture of chloroplast genome have been extensively studied in higher plants and several model species of algae, little is known about transcriptional features in green algal lineages.

**Results:** Based on full-length cDNA (Iso-Seq) sequencing, we identified widely co-transcribed polycistronic transcriptional units (PTUs) in the green alga *Caulerpa lentillifera.* In addition to clusters of genes from the same pathway, we identified a series of PTUs of up to nine genes whose function in the plastid is not understood. The RNA data further allowed us to confirm widespread expression of fragmented genes and conserved open reading frames, which are both important features in green algal chloroplast genomes. In addition, a newly fragmented gene specific to *C. lentillifera* was discovered, which may represent a recent gene fragmentation event in chloroplast genome.

Taking the accurate exon-intron boundary information, gene structural annotation was greatly improved across the siphonous green algae lineages. Our data also revealed a type of non-canonical Group II introns, with a deviant secondary structure and intronic ORFs lacking known splicing or mobility domains. These widespread introns have conserved positions in their genes and are excised precisely despite lacking clear consensus intron boundaries.

**Conclusion:** Our study fills important knowledge gaps in chloroplast genome organization and transcription in green algae, and providing new insights into expression of polycistronic transcripts, freestanding ORFs and fragmented genes in algal chloroplast genomes. Moreover, we revealed an unusual type of Group II intron with distinct features and conserved positions in Bryopsidales. Our data represents interesting additions to knowledge of chloroplast intron structure and highlights clusters of uncharacterized genes that probably play important roles in plastid.


# Background

Chloroplasts, the light-harvesting organelles in plants and green algae, are derived from a photosynthetic cyanobacterium through the process of endosymbiosis [1]. Chloroplasts have retained a reduced cyanobacteria-derived genome, which is generally a circular-mapping DNA molecule ca. 100–200 kb in size, although there are exceptions [2, 3]. During endosymbiosis and subsequent genome evolution, most cyanobacterial genes were transferred to the host nucleus or lost, and only a core set of genes encoding key proteins involved in photosynthesis, transcription and translation have been retained in most chloroplast genomes [4].

Chloroplasts have both prokaryotic and eukaryotic properties [5, 6], with gene expression reminiscent to prokaryotes, involving sigma70 promoters and genes organized into operons that are usually transcribed as polycistronic transcripts. Among the eukaryote-like features are the prevalence of introns, highly stable mRNAs, and a more complex regulation of gene expression [7]. Transcripts of chloroplast genes are post-

transcriptionally modified in some lineages, including polycistronic transcripts processing, intron splicing, RNA editing and the recently identified non-coding and antisense RNAs [6, 7].

The rapid uptake of high throughput sequencing has led to large numbers of plant and green algal chloroplast genomes being sequenced across the green lineage, advancing our knowledge of their structural diversity and evolutionary dynamics. In the green algal lineage (Chlorophyta), 168 chloroplast genomes are now available on NCBI. The order Bryopsidales, a group of marine seaweeds with a siphonous cell architecture [8], has become a model system for algal chloroplast genome evolution, with studies characterizing a range of genes of possible bacterial origin [9], the evolutionary dynamics of different groups of introns and non-standard open reading frames (ORFs) associated with mobile functions [10], and genome dynamics in relation to habitat features [11].

Bryopsidalean chloroplast genomes also feature a number of genes fragmented into two subsequent ORFs, either with an in-frame stop codon separating them or with a frame shift along the gene, or occasionally more widely spaced around an insertion not associated with group I or group II introns [12]. These genes have previously been considered pseudogenes, but sequence conservation would suggest they are not, and to date it is not known whether and how these genes are transcribed and possibly modified post-transcriptionally. Most of the recent work on green algal chloroplast genomes has been based on short-read sequencing (SRS) data assembled and annotated via largely automated methods, which may in some cases result in incomplete assembly and misannotation, particularly of features like exon-intron boundaries that not transfer well across species. Furthermore, while genome dynamics have been well-characterized, little is known about how genes are transcribed and modified post-transcriptionally.

Here we focused our study on the bryopsidalean species, and take *Caulerpa lentillifera* (Bryopsidales, Chlorophyta), an important edible alga with high nutritional and economic values [13–16], as a major object. The goal of this study is to fill these knowledge gaps in genomic organization and transcription in bryopsidalean green algae. First, we aim to take advantage of long-read sequencing to assemble a high-quality chloroplast genome, identify potential errors in SRS-based assemblies and evaluate the possibility of structural genome variants. Second, we aim to characterize the expression of polycistronic transcripts and freestanding (non-intronic) ORFs, and evaluate the transcriptional features of fragmented genes using full-length cDNA isoform sequencing (Iso-seq) technologies. Finally, we aim to take advantage of the full-length cDNA sequences to better understand and improve prediction of exon-intron boundaries in chloroplast genomes of siphonous green algae.

# Results

## Improved chloroplast genome assembly and annotation

About 3.28 Gbp of PacBio reads from the DNA library of a single genotypic isolate (V1) of *C. lentillifera* showed affinity to chloroplast genome and were assembled. The long reads and high average read

coverage (ca. 25000×), resulted in the chloroplast genome being assembled into a single contig without gaps or ambiguous regions (Fig. 1). The obtained circular chloroplast genome (referred to as Clcp-v1) was 126,969 bp in length.

When compared with the previously reported *C. lentillifera* chloroplast genome (119,402 bp, GenBank Accession No. MG753774.1), the average identity between the sequences is 99.89%, but our assembly is 7.5 kb larger, with 8 structural differences (range from 35 bp to 2.8 kb), 19 InDels and 33 SNVs. Among these differences, several InDels are related to copy number variation of tandem repeat sequences (Fig. 1 and Additional file 1: Table S1-S3). Analysis of the assembly graph of our long-read DNA sequencing data did not identify any structural heterogeneity, unlike the recent observations for another *Caulerpa* species [17]. The differences between the *C. lentillifera* chloroplast genomes Clcp-v1 and MG753774.1 rather appears to reflect strain-level variation.

Based on the results of automatic prediction by GeSeq [18] and the PacBio full-length cDNA isoform sequencing (Iso-seq) to guide annotation, a total of 146 genes were annotated in the *C. lentillifera* chloroplast genome, including 76 protein coding genes, 28 tRNA genes, 3 rRNA genes and 39 ORFs (≥ 300 bp in length) (Fig. 1).

Compared with the annotations in MG753774.1, which was conducted by automated prediction, the gene content of rRNA, tRNA and protein coding genes are almost the same between Clcp-v1 and MG753774.1. However, there are a few differences in gene structural annotation between the two sequences. Most of these differences were caused by incorrect annotation of the exon-intron boundaries, or the annotation of introns where no introns exist. Taking advantages of the full-length cDNA reads, the introns and intron boundaries of rRNA genes and protein-coding genes (such as *rbc*L, *atp*F and *ccs*A) were confirmed, and the misannotated introns in *rps*7, *ycf*20, *cys*A in the previously published genome are corrected in our Clcp-v1 genome (Additional file 2: Figure S1 and Figure S2). *C. lentillifera* has a similar gene content to other *Caulerpa* species, but our re-examination revealed that a few genes were missed in previous studies (Table 1 and Additional file 3: Table S4).

Table 1
General characteristics of *C. lentillifera* chloroplast genome and comparison with other *Caulerpa* species

| Species | Accession Number | Length (kb) | Overall GC content (%) | No. of genes | | | ORFs [b] | Reference |
|---|---|---|---|---|---|---|---|---|
| | | | | Protein coding[a] | rRNA genes | tRNA genes | | |
| *C. lentillifera* | This study | 126.97 | 32.43 | 76 | 3 | 28 | 39 | This study |
| *C. lentillifera* | MG753774.1 | 119.40 | 32.68 | 76 (75) | 3 | 28 | 45 | [21] |
| *C. manorensis* | NC_037367.1 | 140.60 | 34.62 | 76 (75) | 3 | 28 | 36 | [12] |
| *C. cliftonii* | NC_031368.1 | 131.14 | 37.64 | 76 (74)[c] | 3 | 27 | 34 | [11] |
| *C. racemosa* | NC_032042.1 | 176.52 | 33.64 | 76 (75) | 3 | 27 | 34 | [44] |
| *C. okamurae* | KX809677.1 | 148.27 | 34.54 | 76(73)[d] | 3 | 27 | 28 | - |
| *C. verticillata* | NC_039523.1 | 148.11 | 33.52 | 76 (75) | 3 | 28 | 57 | [10] |

[a]. The numbers outside and inside parentheses represent corrected and previously reported gene numbers, respectively. *til*S was included for counting; fragmented genes (such as *rpo*Ba and *rpo*Bb) were counted once.

[b]. Open reading frames (ORFs) > = 300 bp were included in the count; Previously annotated ORFs that related to *til*S were excluded. For data consistency, the ORFs of *C. racemosa* (NC_032042.1) and *C. lentillifera* (MG753774.1) were reannotated.

[c]. The mis-annotated *pet*L in previous report was included in corrected result of *C. cliftonii*.

[d]. The miss annotated *rpl*32 and *pet*L in previous report were included in corrected result of *C. verticillata*.

# Chloroplast genes are widely co-transcribed

Our full-length Iso-seq data provided evidence for the nature and configuration of polycistronic transcripts in *C. lentillifera*. The 121,614 Iso-seq reads that mapped to the chloroplast covered ca. 87% of the Clcp-v1 genome. Among the 5,812 reads covering at least one intact gene/exon and with the same transcriptional direction, 236 cistronic transcriptional types could be identified. Due to these overlapping cistronic transcripts, which may represent different stages of post-transcriptional processing of a polycistronic transcript, we finally concatenated adjacent overlapping cistronic transcripts into the same group, and defined these groups as polycistronic transcriptional units (PTUs). In the full PTU maps, 16 such PTUs covering 43 named protein-coding genes and 29 ORFs were recovered. Among them, the ribosomal

protein operon (*rpl*23-*rpl*2-*rps*19-*rps*3-*rpl*16-*rpl*14-*rpl*5-*rps*8-*inf*A) contains the largest number of protein coding genes (Fig. 2 and Additional file 4: Table S5).

Most protein-coding genes joined in a PTU are functionally related and among the 16 PTUs in *C. lentillifera*, four corresponded to previously identified conserved gene clusters across Bryopsidales [12], providing further evidence for the importance of their co-occurrence in PTUs (Fig. 2).

Prediction of putative sigma70 promoters (Ps70P) showed that 30 protein coding genes and 20 ORFs had at least one Ps70P site within 500 bp upstream from their first codon. Among them, 12 protein coding genes or ORFs were the first gene of PTUs, showing a rough correspondence between predicted promoters and the start positions of most PTUs. However, we observed a high proportion of genes harboring Ps70P that are not the first gene in a PTU (Additional file 5: Table S6), suggesting that these genes can also be transcribed via their own promoters.

# Expression of freestanding ORFs

Of the 39 putative protein-coding ORFs (24 freestanding and 15 intronic) that we identified in the *C. lentillifera* chloroplast genome, the large majority (35 ORFs) were supported by transcripts in our PacBio Iso-seq data (Additional file 6: Table S7). Similar to other reported bryopsidalean species, most of the ORFs were distributed across the genome but organized into clusters of two or more genes, with the largest one containing 9 ORFs within a 10-kb region. Functional annotation for these ORFs showed that 28 ORFs were found to harbor known structural or functional domains (blastp E-values < 1e-10), while 11 ORFs (including 7 novel ORFs that might be specific to *C. lentillifera*) did not show any significant similarity with known proteins in the nr database (Table 2 and Additional file 6: Table S7). Putative homing endonucleases were the most common domains found in the ORFs, including eight LAGLIDADG homing endonuclease, five HNH endonuclease and one GIY-YIG homing endonuclease. Nine ORFs were found to harbor DNA methyltransferase or methylase domains (Table 2).

Table 2
Putative function of freestanding ORFs in chloroplast genome of *C. lentillifera*

| Putative function | | Number of ORFs |
|---|---|---|
| Homing endonucleases | LAGLIDADG homing endonuclease | 8 |
| | HNH endonuclease | 5 |
| | GIY-YIG homing endonuclease | 1 |
| Methyltransferase/methylase | N-6 DNA methylase | 6 |
| | DNA cytosine methyltransferase | 1 |
| | DNA adenine methylase | 2 |
| Other function | | 5 |
| Unknown | | 11 |
| Total | | 39 |

Our Iso-seq data showed that many of these ORFs with unknown function are co-transcribed. For example, of the 9 consecutive ORFs found mentioned above, the last 6 (ORF33-ORF38) were co-transcribed as a single PTU, and the former 3 (ORF30-ORF32) were co-transcribed along with three ATP synthase genes and *rps*2 (Fig. 2 and Additional file 4: Table S5).

# Expression of fragmented genes

Because previous studies had revealed that two protein coding genes, RNA polymerase b-subunit (*rpo*B) and tRNA (Ile)-lysidine synthase (*til*S) were fragmented in several bryopsidalean chloroplast genomes, we wanted to investigate whether these were transcribed and possibly post-transcriptionally modified. In addition to *til*S and *rpo*B, our *C. lentillifera* genome showed a third fragmented gene, the chloroplast envelope membrane protein (*cem*A). While *rpo*B and *til*S were fragmented into two pieces across all the published *Caulerpa* species, *cem*A fragmentation was only observed in *C. lentillifera*. Frame shifts were present in all these genes in all *Caulerpa* species, leading them to be divided into two adjacent ORFs (labeled as a and b in this study), respectively (Fig. 3).

No single PacBio RNA read covered the entire region encompassing *rpo*Ba and *rpo*Bb, but a number of reads covered either only *rpo*Ba, or from *rpo*Bb (5' truncated) to *cys*T/*cyf*1, suggesting that *rpo*Ba and *rpo*Bb are probably translated from different mRNA molecules. The subunits of the two other fragmented genes (*cem*Aa and *cem*Ab, *til*Sa and *til*Sb) were observed to be co-transcribed in PTUs, but some shorter reads (some are 5' or 3' truncated) covered either fragment a or b, indicating that they might also be cleaved into different transcripts during the post-transcriptional process (Additional file 7: Figure S3). We were unable to detect Shine-Dalgarno (SD)-like sequences (a sequence located upstream of the start codon to initiate translation [19, 20]) in the 5' untranslated region of any of the shorter transcripts, so it is unclear whether the two subunits could be separately translated. To verify whether post-transcriptional

RNA editing may be used to overcome the frame shifts within the fragmented genes, we carefully compared the sequences between chloroplast genome and the aligned Iso-seq reads. No RNA editing site was found, suggesting it is unlikely that these fragmented genes are restored to a single continuous reading frame by RNA editing.

Because the RNA library for Iso-seq was constructed by polyA-enrichment, which may not fully reflect the transcriptional state of chloroplast genes, we further validated the transcription patterns of the three fragmented genes by RT-PCR. In accordance with Iso-seq results, the two pieces of *cem*A and *til*S were confirmed to be co-transcribed. There was extremely weak amplification of the *rpo*Ba to *rpo*Bb section, but the bands representing amplicons of separate *rpo*Ba or *rpo*Bb fragments were very strong in comparison. This suggests that while the two pieces of *rpo*B could occasionally be co-transcribed, most transcripts exist as *rpo*Ba and *rpo*Bb separately (Fig. 4).

# Atypical group II introns with widely conserved features

Automated predictions (MFannot and Geseq) and PacBio Iso-seq guided annotation resulted in a few clear differences of predicted introns and exon-intron boundaries (Table 3). Because *atp*F and *ccs*A both contained introns with intron boundaries clearly differing between both annotation methods, we collected these genes across Bryopsidales (40 *atp*F and 39 *ccs*A sequences) and tried to manually reannotate them using our Iso-seq-informed intron boundaries as a reference (Additional file 8: Table S8). Interestingly, after adjusting the exon-intron boundaries of *atp*F and *ccs*A in these species, the reading frame of all these genes lined up very well and the amino acids sequences near the exon-intron boundaries showed much stronger conservation, with the sequence identities of amino acids (calculated for the entire gene) significantly improving from 50.42–54.37% for *atp*F and from 42.90–47.83% for *ccs*A (Additional file 9: Figure S4-S5). Furthermore, the analyzed introns seemed to occupy the same positions across the Bryopsidales, and the conservation could be also supported by intron-less *atp*F or *ccs*A in other species of Bryopsidales, and even in other class of Chlorophyta (Fig. 5 and Additional file 9: Figure S4-S5).

Table 3
Comparison of intron annotation between automated predictions and Iso-seq alignment of chloroplast genes in *C. lentillifera*.

| Intron contained genes | Number of introns | | | Intron type [c] |
|---|---|---|---|---|
| | Iso-seq data | FMannot | Geseq | |
| rrn16 | 3 | 0 | 0 | Intron 1: undetermined (LHE)<br>Intron 2: undetermined (LHE)<br>Intron 3: undetermined (LHE) |
| rrn23 | 5 | 0 | 0 | Intron 1: group I<br>Intron 2: group I<br>Intron 3: group I<br>Intron 4: group I<br>Intron 5: undetermined (LHE) |
| *psb*A | 4 | 4 | 4[a] | Intron 1: group II<br>Intron 2: group II<br>Intron 3: group I<br>Intron 4: group II |
| *psb*D | 1 | 1 | 1 | Intron 1: group I |
| *rbc*L | 1 | 1 | 1 | Intron 1: group II |
| *atp*F | 1 | 0 | 0 [b] | Intron 1: group II (derived domain V) |
| *ccs*A | 1 | 0 | 0 | Intron 1: group II |
| *rpo*C2 | 0 | 0 | 1 | - |
| *rpo*C1 | 0 | 0 | 1 | - |

[a] Boundaries of *psb*A intron 2 cannot be accurately predicted.

[b] *atp*F was mis-predicted by Geseq.

[c] LHE indicate a LAGLIDADG homing endonuclease domain within the intron.

At least five motifs were identified in the 40 *atp*F and 37 *ccs*A intron sequences with high statistical support (E-value < $10^{-140}$). Among them, motifs 1, 4 and 5 were the most common (Fig. 6a and Additional file 10: Figure S6), with motif 1 located towards the 3' end of almost all of these introns, and motif 5 located upstream of motif 1, with the average interval between these two motifs 63 bp (Additional file 10:

Figure S6). Motif 5 was composed of a highly conservative element (CYGAAAGG) and AT-rich flanking sequence.

Intron type prediction showed that most of the introns in *atp*F or *ccs*A are putative group II introns, similar to most other introns of protein coding genes in the chloroplast genome of *C. lentillifera* (Table 3 and Additional file 8: Table S8). Secondary structure analysis showed that motif 1 overlapped with intron domain V (most were derived type) and includes the highly conserved 5'-AGC-3' trinucleotide (Fig. 6b).

However, there are several distinct features of the introns in *atp*F and *ccs*A in comparison with other group II introns (mainly compared with those of *psb*A and *rbc*L here) in Bryopsidales. Firstly, intron boundaries of *ccs*A and *atp*F were highly variable, while the more canonical group II introns of *psb*A and *rbc*L had conserved boundary nucleotide patterns (5’ GUGYG…AY 3’, Fig. 6c). Secondly, the bulge in domain V, which is another catalytically important site conserved among most group II introns is much more variable in *atp*F and *ccs*A introns (Fig. 6b and Additional file 10: Figure S7); Thirdly, although intronic ORF that contains a reverse transcriptase (RT) and/or intron maturase (IM) domain are common in group II introns of the Bryopsidales (Cremen *et al.*, 2018), none of the 86 intronic ORFs ( > = 150 bp) in *atp*F and *ccs*A were predicted to contain conserved domains of known function (Additional file 11: Table S9).

# Discussion

Using long-read sequencing, we obtained an intact chloroplast genome and a well-defined gene structural annotation of *Caulerpa lentillifera*. The new genome is 7.5 kb larger than the previously reported genome sequence [21] for the genome, and in addition to SNVs and a few possible structural variations between the two versions, there are several indels relating to copy number variation of tandem repeat sequences (Additional file 1: Table S2). While these differences could be due to intraspecific differences between the isolates, limitations of using only short sequence reads in the previous work may have also contributed to the differences, as short reads can fail to assemble repetitive sequences [22]. Since the raw reads for the previously published genome are not available, we cannot determine the exact reason for the differences between the two sequences at present. Long-read sequences also permit identifying heteroplasmy within individuals, as recently shown in the chloroplast genome of a related species by nanopore sequencing [17]. Our PacBio long-read data did not reveal any evidence of such structural variations, and in our opinion the prevalence and nature of heteroplasmy across the siphonous green algae requires further work based on long-read methods that deliver highly accurate reads.

Several bioinformatic tools for automated feature annotation of chloroplast genomes have been developed, but relatively little work has been done to compare their predictions to experimentally determined RNA sequences. Our Iso-seq work shows that the majority of genes encoding proteins and rRNA were accurately predicted by MFannot and GeSEq. However, we found that the Iso-seq data-guided annotation could greatly improve the annotation of introns and exon-intron boundaries. Taking our exon-intron boundary information as a reference, we were able to greatly improve structural annotations of *atp*F and *ccs*A across the Bryopsidales, and corrected intron structures facilitated the analysis of the

unusual characteristics of these introns. Several common features of the *atp*F and *ccs*A introns were identified, such as the conserved domain V motif and other common motifs upstream from it. Domain V, which is one of the six conserved domains radiating from a “central wheel” of group II introns, is the most conserved element and important component in catalytic reactions of group II introns [23, 24]. It was clear that the 2-nt bulge (AY) and the catalytic triad (AGC or CGC for some introns) at the stem of domain V are most important for chemical catalysis of excision [25, 26]. Although the catalytic triad are still conserved retained across all the analyzed group II introns of Bryopsidales, the bulge of domain V in *atp*F and *ccs*A introns are relative variable, indicating the splicing mechanisms of these introns might be different from typical group II introns. Previous work mainly based on land plants and *Euglena* showed that most group II introns are degenerated in their RNA structures or have lost the intron encoded proteins [27]. Our results indicate that the introns in *atp*F and *ccs*A have several obvious differences from canonical group II introns, including the absence of consensus intron boundary sequences, ORFs lacking homology to splicing or mobility, and deviant overall structure making it difficult to accurately determine the secondary domains other than domain V. However, our Iso-seq data showed that the introns in *atp*F and *ccs*A were spliced predictably, suggesting that an effective mechanism has evolved to recognize and splice these atypical introns in bryopsidalean chloroplasts.

The fragmentation of several protein-coding genes has been a puzzling feature of green algal chloroplast genomes. In this study, three protein-coding genes were found to be fragmented in the *C. lentillifera* plastid genome, with *cem*A shown to be fragmented in addition to the previously reported *til*S and *rpo*B, which are known to be fragmented across Bryopsidales [12] and some other green algal lineages (e.g. [25, 28, 29]. Considering that *cem*A is not fragmented in other *Caulerpa* species, it likely represents a recent event. This observation, along with reports of some other fragmented genes such as *rpo*C1 and *rpo*C2 in *Chlamydomonas* species [30], suggests that gene fragmentation may be fairly common in green algal chloroplast genomes. The fragmented genes in *Caulerpa* retained high sequence conservation following the fragmentation, a clear indication that they are not pseudogenes. Our Iso-seq data and RT-PCR results provide clear evidence for transcription of these genes. They also indicate that the two pieces of both *cem*A and *til*S are co-transcribed in transcriptional units, but the presence of shorter transcripts covering either fragment of these genes suggests that the transcripts may be divided into two portions by RNA processing mechanisms. Our results for *rpo*B contrast with the other genes, rather showing that while the two fragments were occasionally found on a single transcript, they were more commonly transcribed separately. A careful comparison between chloroplast genome and the aligned Iso-seq reads showed no evidence for RNA editing, thus it seems unlikely that the frame shifts in these fragmented genes were modified to restore normal reading frames. Ribosomal frameshifting [31] could be a hypothetical alternative mechanism to correct the frameshifts in fragmented genes at the level of translation, but the fact that various types of gene fragmentation exist in Bryopsidalean lineages [12], including some with longer inserts between the fragments, would suggest this is unlikely and that it is more likely that the two pieces of these fragmented genes are translated separately and combine after translation. We did not find SD-like sequences (translation initiation signals of bacteria and some chloroplast mRNAs) upstream of the translation initiation sites of the gene fragments, so it remains to be confirmed whether the

transcriptional products of fragments a and b are separately translated and perform their normal functions by forming protein complexes of both subunits. Nevertheless, gene fragmentation (or gene fission) as well as gene fusion are important mechanisms that contribute to the evolution of gene architecture and origination of new genes. Gene fusion/fission was major contributor to evolution of multi-domain proteins in bacteria and creation of new genes in *Drosophila* [32, 33], and the mechanism of the origin of gene fission has been revealed as a two-step process consisting of duplication and degeneration in *Drosophila.* Recently, gene fragmentation was found to be very prominent in mitochondrial genomes of Diplonemids, where the resulting modules (gene fragments) are transcribed separately, which might contribute to a gradual increase in the complexity of a given cellular machinery [34]. What drives gene fragmentation in chloroplast genomes as well as the mechanisms and consequences of this process in these organelles remain open questions.

Our Iso-seq data allowed us to experimentally verify polycistronic mRNAs and post-transcriptional isoforms, which are important for understanding the mechanisms of plastid genome expression. Although transcriptional and post-transcriptional regulation of chloroplast genes have been well studied in higher plants [6, 35, 36], little is known about the situation in algae. Unlike higher plants, it has been assumed that transcript processing may be less important in controlling plastid gene expression in algae, because nearly all genes seemed to be transcribed as monocistronic RNAs in the unicellular green alga *Chlamydomonas reinhardtii* [6]. However, our analysis revealed that more than half of the protein-coding genes are co-transcribed with adjacent genes, forming polycistronic transcripts of up to 9 genes in *Caulerpa*, so the observations of *Chlamydomonas* can certainly not be extrapolated to other green algae. In addition, because we used very strict criteria to consider genes as co-transcribed on PTUs, several genes flanking our PTUs but that were not entirely covered by Iso-seq reads were not counted as part of the PTUs, so the extent of gene co-transcription is probably even larger. Our work, along with observations of co-transcribed chloroplast gene clusters in *C. reinhardtii* based on RNA-seq coverage analysis [37, 38], provide clear evidence for polycistronic transcription in algae. Unsurprisingly, genes on the same PTU in *C. lentillifera* were often functionally related. This observation extends to the co-transcribed clusters of unknown ORFs, which often shared conserved functional domains of the same class within the PTU (Additional file 4: Table S5). Four of the six conserved gene clusters in Bryopsidales [12] were found to be co-transcribed in PTUs, underlining the strong evolutionary conservation of co-transcription and the importance of the co-occurrence of these genes. The remaining two conserved gene clusters observed in Bryopsidales (*psa*M-*psb*30-*psb*K-*psb*N-*trn*M and *psb*E-*psb*F-*psb*L-*psb*J) were not observed as PTUs in this study. This is because our full-length RNA data did not contain any reads of the genes in question, perhaps due to their lower levels of expression or faster degradation.

# Conclusion

Our study is the first to experimentally determine and examine the genomic organization and transcriptional units in green algae using long-read sequencing technology, providing new insights into structural variations, expression of polycistronic transcripts, freestanding ORFs and fragmented genes in algal chloroplast genomes. Drastic improvements in the detection of exon-intron boundaries using Iso-

Seq data permitted a detailed investigation of the structural annotation of intron-containing genes across Bryopsidales, revealing atypical Group II introns with distinct features and conserved positions in the *atp*F and *ccs*A genes. Our results also further our knowledge of gene fission in chloroplast genomes and form a valuable resource for further organellar transcriptomics studies.

## Methods

# Nucleic acid isolation and sequencing

A single genotypic isolate (V1) of *C. lentillifera* was collected from our marine culture base in Changjiang, Hainan province and cultured in sterilized seawater. To minimize the contamination of environmental microbes, the thallus of V1 was treated with a combination of various antibiotics following Brawley et al. [39] for two weeks.

For genome sequencing, DNA of the *C. lentillifera* isolate V1 was extracted using a Plant Genomic DNA Kit (DP305, Tiangen Inc, Beijing, China) following the manufacturer's instructions. A 20-kb insert SMRTbell library was prepared and sequenced by Novogene (Beijing, China) using the PacBio Sequel platform (Pacific Biosciences, CA, USA).

For PacBio full-length cDNA isoform sequencing (Iso-Seq), *C. lentillifera* samples were treated under multiple conditions, such as high temperature (30 °C), low temperature (18 °C), high salinity (50 PSU), low salinity (15 PSU), high light (260 μmol photons·$m^{-2}$·$s^{-1}$), shading, desiccation and normal conditions (25 °C, 32PSU and 10 μmol photons·$m^{-2}$·$s^{-1}$). Total RNA of each treatment was isolated with the RNAprep pure plant kit (Tiangen Inc, Beijing, China) and treated with RNase-free DNaseI (RT411, Tiangen Inc, Beijing, China) following the manufacturer's instructions. The quality of RNA was checked on a Bioanalyzer 2100 system (Agilent, Palo Alto, CA, USA). Then extracted RNAs were pooled evenly, and Iso-Seq libraries were constructed with SMARTer™ PCR cDNA Synthesis Kit (Clontech, CA, USA) and BluePippin™ Size-Selection System, and then sequenced on the PacBio Sequel platform.

# Chloroplast genome assembly

The PacBio raw reads were assembled by Falcon (pb-assembly 0.0.8) [40]. From this draft assembly, seven contigs which had high sequence similarity with chloroplast genomes of *Caulerpa* were identified as candidate chloroplast sequences by blastn using related chloroplast genomes as query. The PacBio raw reads were mapped to the draft assembly by minimap2 [41] and all reads mapping with less than 0.2 divergence and alignment length exceeding 2000 bp were selected for *de novo* assembly in Canu 1.9 [42] with parameters corOutCoverage = 5000, genomeSize = 150 k, rawErrorRate = 0.200, correctedErrorRate = 0.035, minReadLength = 20000, minOverlapLength = 8000. As a result, a single contig was obtained, and inspected by using Bandage [43] to evaluate the presence of structural variations in the assembled genome. Finally, the obtained contig was polished with arrow and pilon by using default parameters, and set as a circular molecule with the start point at the beginning of the 16S rRNA gene. The obtained circular *C. lentillifera* chloroplast genome is referred to as Clcp-v1.

# Chloroplast genome annotation and exon-intron boundaries

Initial annotations of *C. lentillifera* chloroplast DNA were performed by GeSeq [18] using the published chloroplast genomes of *Caulerpa* species (NC_037367.1, NC_032042.1, KX809677.1, NC_031368.1, NC_039523.1 and MG753774.1) as BLAT reference sequences [10–12, 21, 44], and selected the options to perform tRNAscan-SE v2.0.6 and ARAGORN v1.2.38 to detect tRNA genes, and HMMER profile search to detect chloroplast CDS and rRNA. ORFs with a minimum size set at 300 bp were identified using ORFfinder from NCBI.

Then, the predicted rRNA genes, tRNA genes, protein-coding genes, putative open reading frames and additional features such as the exon-intron boundaries of intron-containing genes, were compared manually to our PacBio Iso-seq data and modified where necessary. The generated chloroplast genome sequence of *C. lentillifera* in this study is available in GenBank under accession number MT271684, and all related PacBio raw data have been deposited on the Sequence Read Archive (SRA) database of NCBI with the number PRJNA658421.

Genes for which predicted exon-intron boundaries differed from those based on Iso-seq data (mostly *atp*F and *ccs*A) were investigated in more detail, and in order to compare their features to other group II introns, two additional protein coding genes (*rbc*L and *psb*A) harboring group II introns in *C. lentillifera* were also analyzed. Additional sequences of these four genes across the Bryopsidales were sourced from Genbank. The exon-intron boundaries were compared through multiple sequence alignment, and corrected manually by referring to the intron boundary information of *atp*F, *ccs*A, as well as *rbc*L and *psb*A of *C. lentillifera* supported by our Iso-seq data. Identification of motifs in intronic sequences was performed by MEME [45], in addition, in order to detect more intronic ORFs that may relate to splicing or mobility of the group II introns in *atp*F and *ccs*A, the minimum size of ORF predictions was set at 150 bp. Secondary structure predictions of relative intronic sequences from *atp*F, *ccs*A and other genes that contained group II introns were performed by RNAfold [46]. Consensus secondary structures for the alignments of group II introns from these genes, were carried out by RNAz [47] based on both thermodynamic stability and structural conservation.

# Polycistronic transcripts and sigma70 promoter prediction

Our full-length Iso-seq transcripts were mapped to Clcp-v1 with Gmap [48], with parameters –no-chimeras-cross-species–expand-offsets 1-B5-K50000-f samse-n 1. Polycistronic transcripts, introns and RNA editing sites were identified through the mapped data and visualized using IGV (Integrative Genomics Viewer, version 2.5.2, https://software.broadinstitute.org/software/igv/download).

To calculate how many types of co-transcribed genes can be classified in *C. lentillifera* chloroplast genome, we detected gene clusters (two or more adjacent genes) oriented in the same direction and occurring together in at least one Iso-seq long read. To increase credibility of the results, we used even stricter criteria: only those genes (or exons of intron-containing genes) completely covered by PacBio Iso-seq reads with the same transcriptional direction, were considered as part of a cistronic transcriptional

type. Then the adjacent and overlapping cistronic transcripts, which may represent different stages of post-transcriptional processing of a polycistronic transcript, were combined into full PTU maps.

For the prediction of putative sigma70 promoters in *C. lentillifera* chloroplast genome, three tools, BPROM [49], CNNPromoter_b [50] and PATLOC (Pattern Locator, Institute of Bioinformatics, University of Georgia, U.S.A., https://www.cmbl.uga.edu/software/patloc.html) were used with default parameters. For the results of PATLOC, only predicted sigma70 promoters in the intergenic regions were kept for further analysis.

# RT-PCR validation of fragmented gene expression

To further evaluate the hypothesis that both pieces of some fragmented genes (*rpo*Ba and *rpo*Bb, *til*Sa and *til*Sb, *cem*Aa and *cem*Ab) were transcribed on a single mRNA, we carried out an RT-PCR experiment. Total RNA samples were isolated using E.Z.N.A. Plant RNA Kit (R6827-01, OMEGA Bio-tek, GA, USA), and contaminated genomic DNA was removed with RNase-Free DNase  Set (E1091-01, OMEGA Bio-tek, GA, USA) according to the manufacturer's instructions. cDNAs were synthesized with random hexamer primers by using the RevertAid First Stand cDNA Synthesis Kit (K1622, Thermo Scientific, MA, USA). Gene specific primers that covering different region of the fragmented genes were designed by using Primer Premier (version 5.0, Premier Biosoft International, USA), and are listed in Additional file 12: Table S10.

# Declarations

# Acknowledgements

We are very grateful to Peng Li, Fengshou Wang and other staff working in the marine culture base for their help in preparation of the *C. lentillifera* isolate V1, and the platform provided by Overseas Intelligence Project of Hainan Province (China-ASEAN Cooperation Research on Tropical Seaweed Resources).

# Authors' Contribution

XZ, JS and SB planned and designed research, and acquired funding. XZ, SB and ZC collected specimens and prepared samples for sequencing. XZ, JS, TJL, JZ and QH performed experiment and data analyses. XZ, JS and HV guided the experimental design, data processing and interpretation, all authors interpreted results. XZ and HV wrote the manuscript. All authors reviewed and approved the final manuscript.

# Funding

This research was supported by Special Project on Blue Granary Science and Technology Innovation under the National Key R&D Program (No. 2018YFD0901503), Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of Tropical Agricultural Sciences (No. 1630052016011), Financial Fund of the Ministry of Agriculture and Rural Affairs of China (No. NFZX2018).

# Availability of data and materials

The datasets supporting the findings of this article are available in the Sequence Read Archive (SRA) database of NCBI with the number PRJNA658421 (https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA658421).

# Competing interests

The authors declare that no competing interests exist.

# Ethics approval and consent to participate

Not applicable.

# Consent for publication

Not applicable.

## References

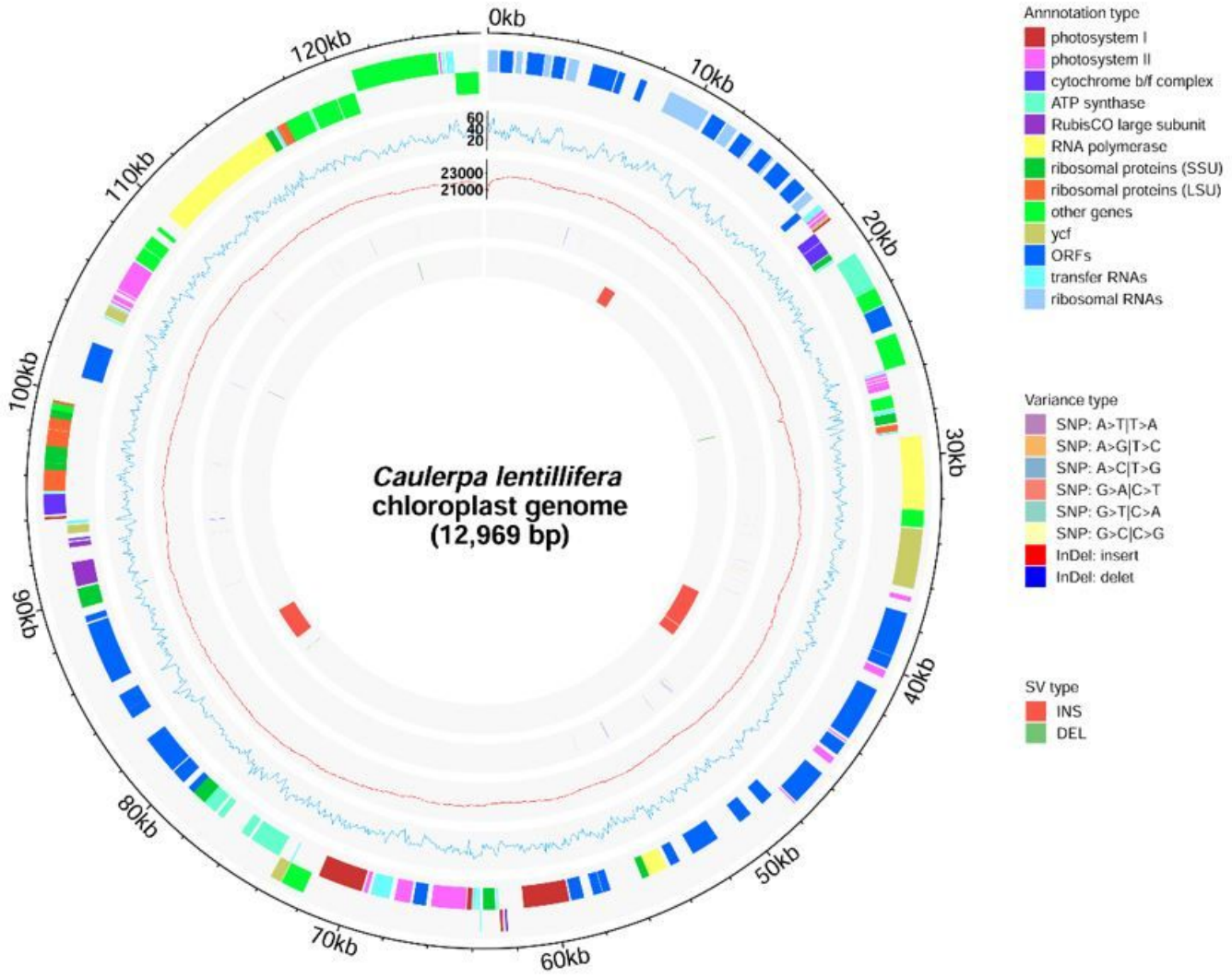

1. Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365:729-48.
2. Del Cortona A, Leliaert F, Bogaert KA, Turmel M, Boedeker C, Janouškovec J, Lopez-Bautista JM, Verbruggen H, Vandepoele K, De Clerck O. The plastid genome in Cladophorales green algae is encoded by hairpin chromosomes. Current Biology. 2017;27:3771-82.e6.
3. Bauman N, Akella S, Hann E, Morey R, Schwartz AS, Brown R, Richardson TH. Next-generation sequencing of *Haematococcus lacustris* reveals an extremely large 1.35-megabase chloroplast genome. Genome Announcements. 2018;6:e00181-18.
4. Green BR. Chloroplast genomes of photosynthetic eukaryotes. The Plant Journal. 2011;66:34-44.
5. Mayfield SP, Yohn CB, Cohen A, Danon A. Regulation of chloroplast gene expression. Annual Review of Plant Physiology and Plant Molecular Biology. 1995;46:147-66.
6. del Campo EM. Post-transcriptional control of chloroplast gene expression. Gene Regul Syst Bio. 2009;3:31-47.
7. Yagi Y, Shiina T. Recent advances in the study of chloroplast gene expression and its evolution. Front Plant Science. 2014;5.
8. Coneva V, Chitwood DH. Plant architecture without multicellularity: quandaries over patterning and the soma-germline divide in siphonous algae. Front Plant Science. 2015;6:287-.

9. Leliaert F, Lopez-Bautista JM. The chloroplast genomes of *Bryopsis plumosa* and *Tydemania expeditiones* (Bryopsidales, Chlorophyta): compact genomes and genes of bacterial origin. BMC Genomics. 2015;16:204.
10. Cremen MCM, Leliaert F, West J, Lam DW, Shimada S, Lopez-Bautista JM, Verbruggen H. Reassessment of the classification of Bryopsidales (Chlorophyta) based on chloroplast phylogenomic analyses. Molecular Phylogenetics and Evolution. 2019;130:397-405.
11. Marcelino VR, Cremen MCM, Jackson CJ, Larkum AAW, Verbruggen H. Evolutionary dynamics of chloroplast genomes in low light: a case study of the endolithic green alga *Ostreobium quekettii*. Genome Biology and Evolution. 2016;8:2939-51.
12. Cremen MCM, Leliaert F, Marcelino VR, Verbruggen H. Large diversity of nonstandard genes and dynamic evolution of chloroplast genomes in siphonous green algae (Bryopsidales, Chlorophyta). Genome Biology and Evolution. 2018;10:1048-61.
13. Ratana-arporn P, Chirapart A. Nutritional evaluation of tropical green seaweeds *Caulerpa lentillifera* and *Ulva reticulata*. The Kasetsart Journal. 2006;40 (Suppl.) 75-83
14. Mary A, Matias JR. Rediscovery of naturally occurring seagrape *Caulerpa lentillifera* from the Gulf of Mannar and its mariculture. Current Science. 2009;97:1418-20.
15. Titlyanov EA, Titlyanova TV, Pham VH. Stocks and the use of economic marine macrophytes of Vietnam. Russian Journal of Marine Biology. 2012;38:285-98.
16. Marquez GPB, Santiañez WJE, Trono GC, Montaño MNE, Araki H, Takeuchi H, Hasegawa T. Seaweed biomass of the Philippines: Sustainable feedstock for biogas production. Renewable and Sustainable Energy Reviews. 2014;38:1056-68.
17. Sauvage T, Schmidt WE, Yoon HS, Paul VJ, Fredericq S. Promising prospects of nanopore sequencing for algal hologenomics and structural variation discovery. BMC Genomics. 2019;20:850.
18. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6-W11.
19. McCarthy JEG, Brimacombie R. Prokaryotic translation: the interactive pathway leading to initiation. Trends Genetics. 1994;10:402-7.
20. Hirose T, Sugiura M. Functional Shine-Dalgarno-like sequences for translational initiation of chloroplast mRNAs. Plant and Cell Physiology. 2004;45:114-7.
21. Gao D, Huang C, Yao J, Li Y, Tan W, Sun Z. Characterization of the whole chloroplast genome *Caulerpa lentillifera* J. Agardh (Bryopsidales, Chlorophyta). Mitochondrial DNA Part B. 2018;3:1198-9.
22. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nature Methods. 2011;8:61-5.
23. Michel F, Umesono K, Ozeki H. Comparative and functional anatomy of group II catalytic introns — a review. Gene. 1989;82:5-30.
24. Toor N, Hausner G, Zimmerly S. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. RNA. 2001;7:1142-52.

25. Toor N, Keating KS, Taylor SD, Pyle AM. Crystal structure of a self-spliced group II intron. Science. 2008;320:77-82.
26. Robart AR, Chan RT, Peters JK, Rajashankar KR, Toor N. Crystal structure of a eukaryotic group II intron lariat. Nature. 2014;514:193-7.
27. Zimmerly S, Semper C. Evolution of group II introns. Mobile DNA. 2015;6:7.
28. Brouard J-S, Otis C, Lemieux C, Turmel M. The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. Genome Biology and Evolution. 2010;2:240-56.
29. Turmel M, Otis C, Lemieux C. Dynamic evolution of the chloroplast genome in the green algal classes Pedinophyceae and Trebouxiophyceae. Genome Biology and Evolution. 2015;7:2062-82.
30. Turmel M, Brouard J-S, Gagnon C, Otis C, Lemieux C. Deep division in the chlorophyceae (chlorophyta) revealed by chloroplast phylogenomic analyses. Journal of Phycology. 2008;44:739-50.
31. Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. 2016;44:7007-78.
32. Wang W, Yu H, Long M. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. Nat Genet. 2004;36:523-7.
33. Pasek S, Risler J-L, Brézellec P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics. 2006;22:1418-23.
34. Kaur B, Záhonová K, Valach M, Faktorová D, Prokopchuk G, Burger G, Lukeš J. Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. Nucleic Acids Res. 2020;48:2694-708.
35. Sugita M, Sugiura M. Regulation of gene expression in chloroplasts of higher plants. Plant Molecular Biology. 1996;32:315-26.
36. Grabsztunowicz M, Koskela MM, Mulo P. Post-translational modifications in regulation of chloroplast function: Recent advances. Front Plant Science. 2017;8.
37. Cavaiuolo M, Kuras R, Wollman F-A, Choquet Y, Vallon O. Small RNA profiling in *Chlamydomonas*: insights into chloroplast RNA metabolism. Nucleic Acids Res. 2017;45:10783-99.
38. Gallaher SD, Fitz-Gibbon ST, Strenkert D, Purvine SO, Pellegrini M, Merchant SS. High-throughput sequencing of the chloroplast and mitochondrion of *Chlamydomonas reinhardtii* to generate improved de novo assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. The Plant Journal. 2018;93:545-65.
39. Brawley SH, Blouin NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson HV, Jenkins JW, Blaby-Haas CE, Helliwell KE, Chan CX *et al*. Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis* (Bangiophyceae, Rhodophyta). Proceedings of the National Academy of Sciences of the United States of America. 2017;114:E6361-E70.
40. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A *et al*. Phased diploid genome assembly with single-molecule real-time

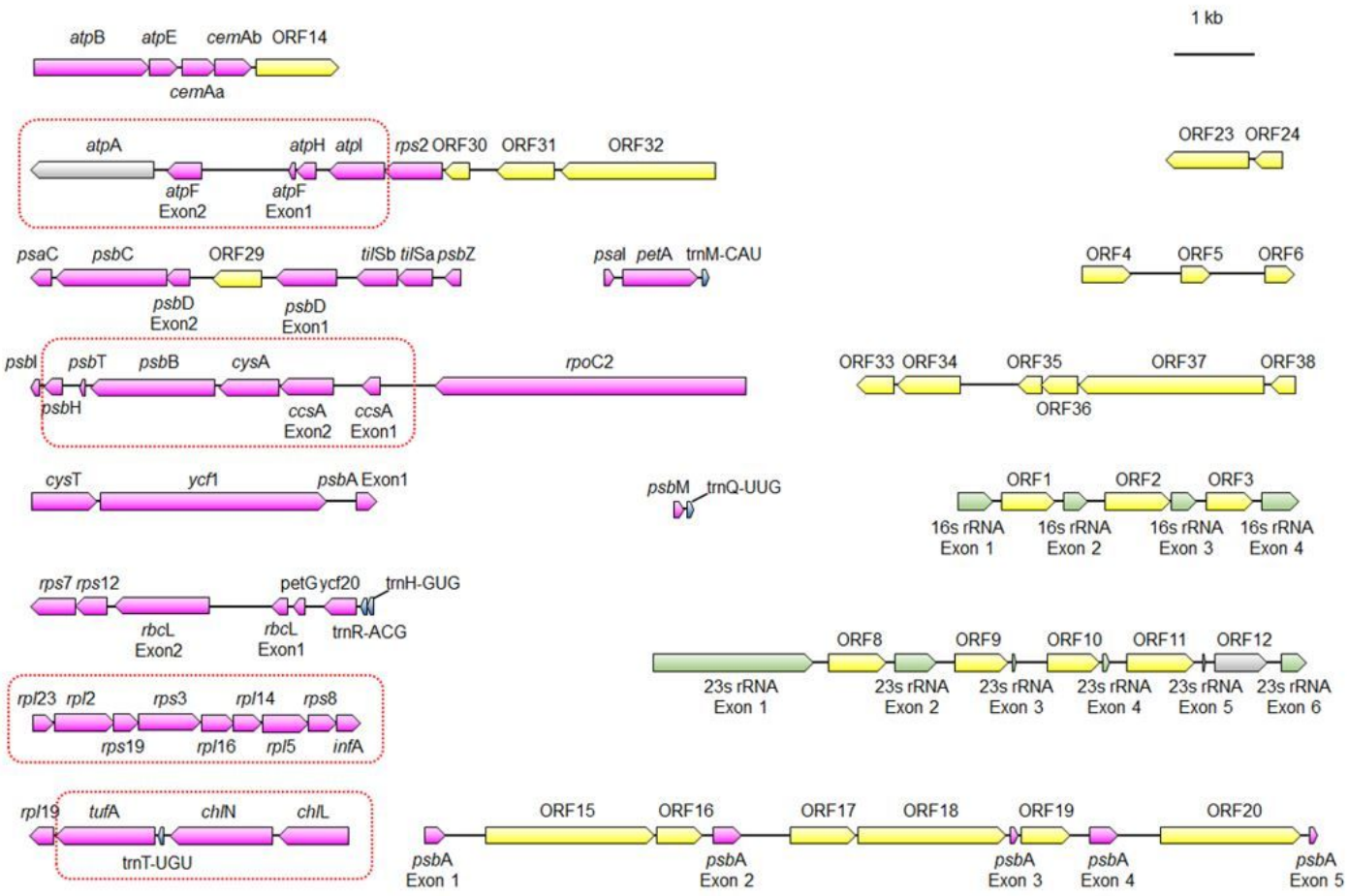
sequencing. Nature Methods. 2016;13:1050-4.
41. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094-100.
42. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722-36.
43. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics. 2015;31:3350-2.
44. Lam DW, Lopez-Bautista JM. Complete chloroplast genome for *Caulerpa racemosa* and compartive analyses of siphonous green seaweeds plastomes. Cymbella. 2016;2:23-32.
45. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202-W8.
46. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26-.
47. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. Pacific Symposium on Biocomputing. 2010:69-79.
48. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859-75.
49. Solovyev V, Salamov A. Automatic annotation of microbial genomes and metagenomic sequences. In: Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies Li RW, editor.: Nova Science Publishers, Inc.; 2011. p. 62–78.
50. Umarov RK, Solovyev VV. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS One. 2017;12:e0171410.

# Figures



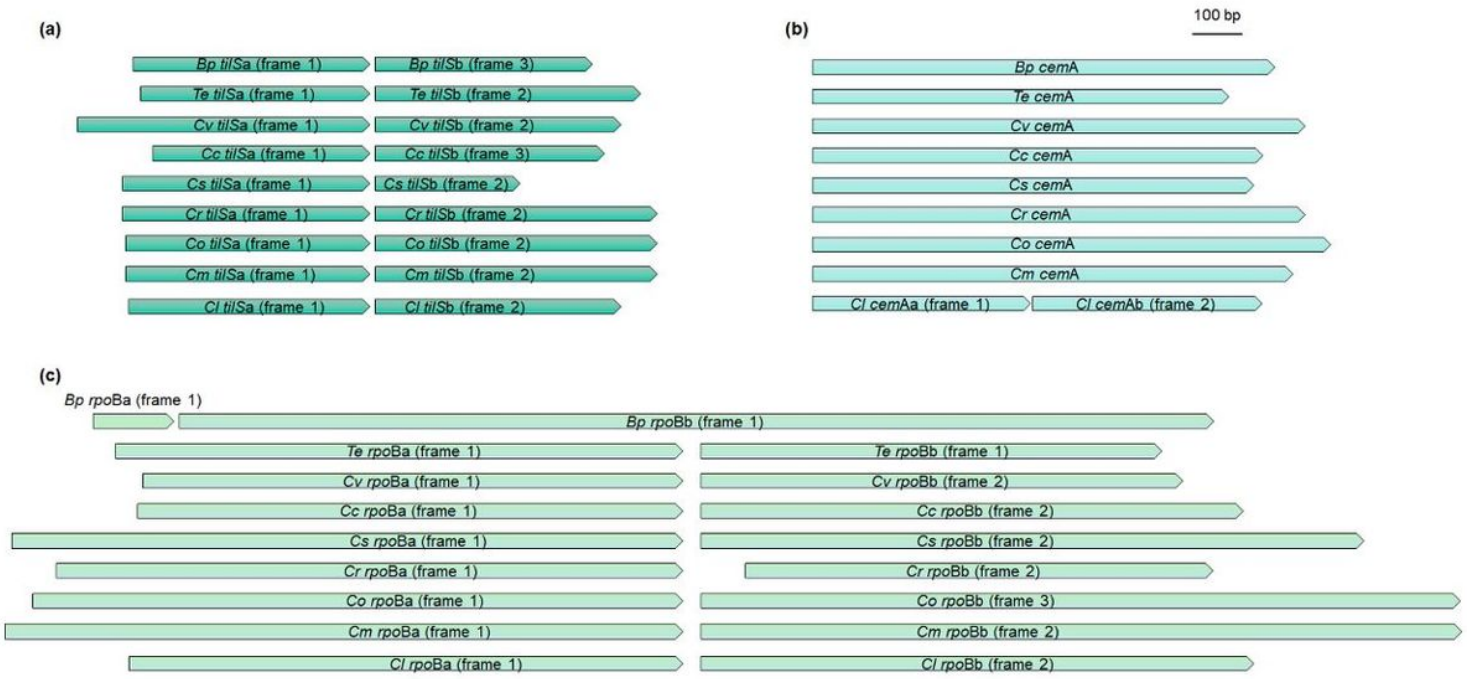

## Figure 1

Chloroplast genome map of Caulerpa lentillifera (Clcp-v1) and comparison with the previous version, MG753774.1. The outermost circle is positions of Clcp-v1 sequences. Annotation of genes in Clcp-v1 is shown in the second circle, the genes present outside of this circle are transcribed in a clockwise direction, while those inside are transcribed counterclockwise. Genes are colored according to the functional categories listed in the legend. The third and fourth circles indicate the GC content, and mapping depth and coverage of PacBio long-read sequencing data of Clcp-v1, respectively. The fifth and innermost circles indicate the distribution of SNPs and indels, as well as structural variations compared with MG753774.1, respectively. Variation types are colored as shown in the legend.



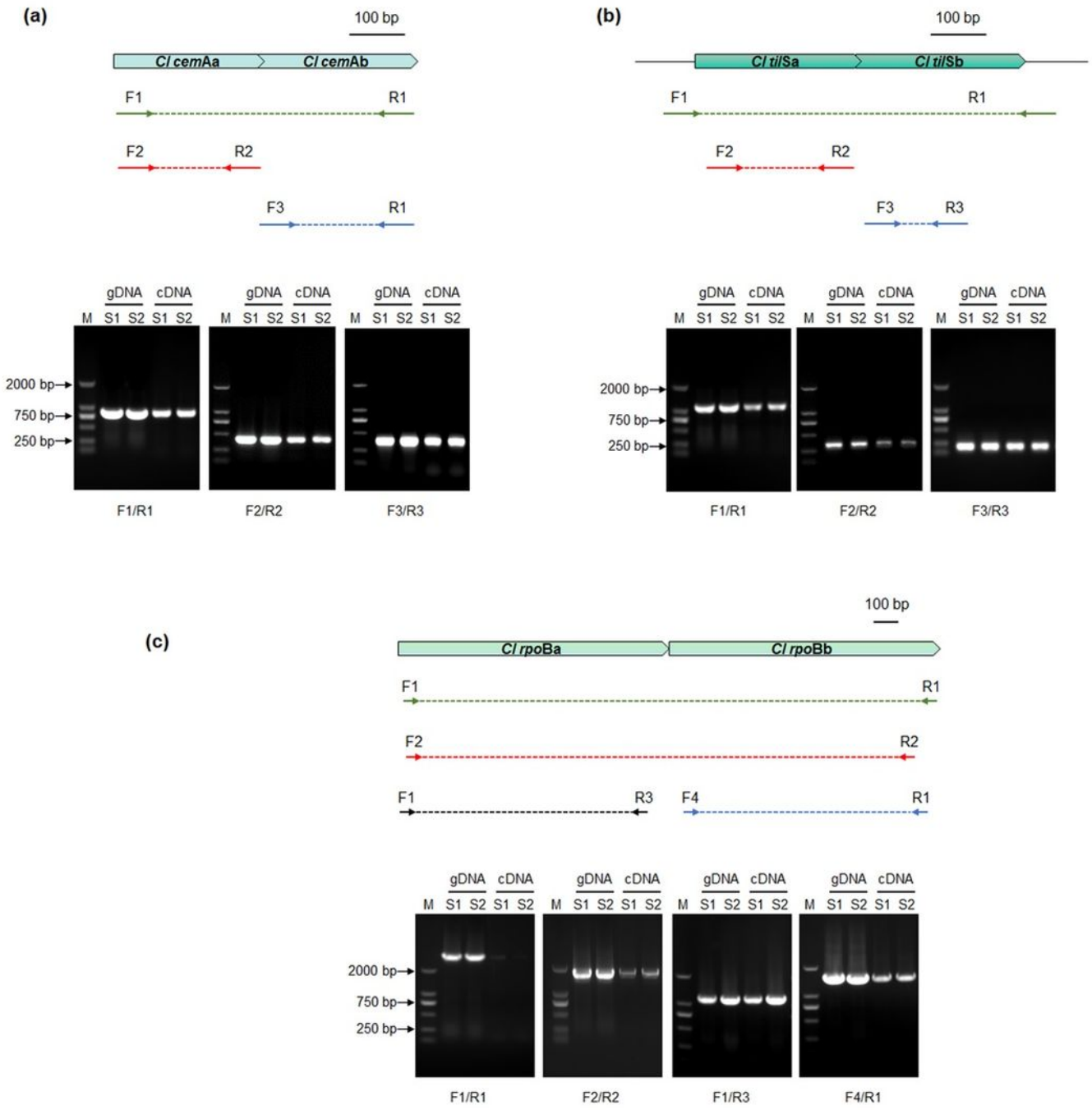

## Figure 2

The identified polycistronic transcriptional units (PTUs) in the C. lentillifera chloroplast. The protein-coding genes, ORFs, rRNA and tRNA genes are indicated in pink, yellow, green or blue, respectively; atpA and ORF12 are shown in gray because the Iso-seq reads did not cover the entire gene. Genes in the dotted boxes are related to the conserved gene clusters across all Bryopsidales identified by Cremen et al. (2018). The gene order, including intronic ORFs within each PTU, is based on the physical genome location.



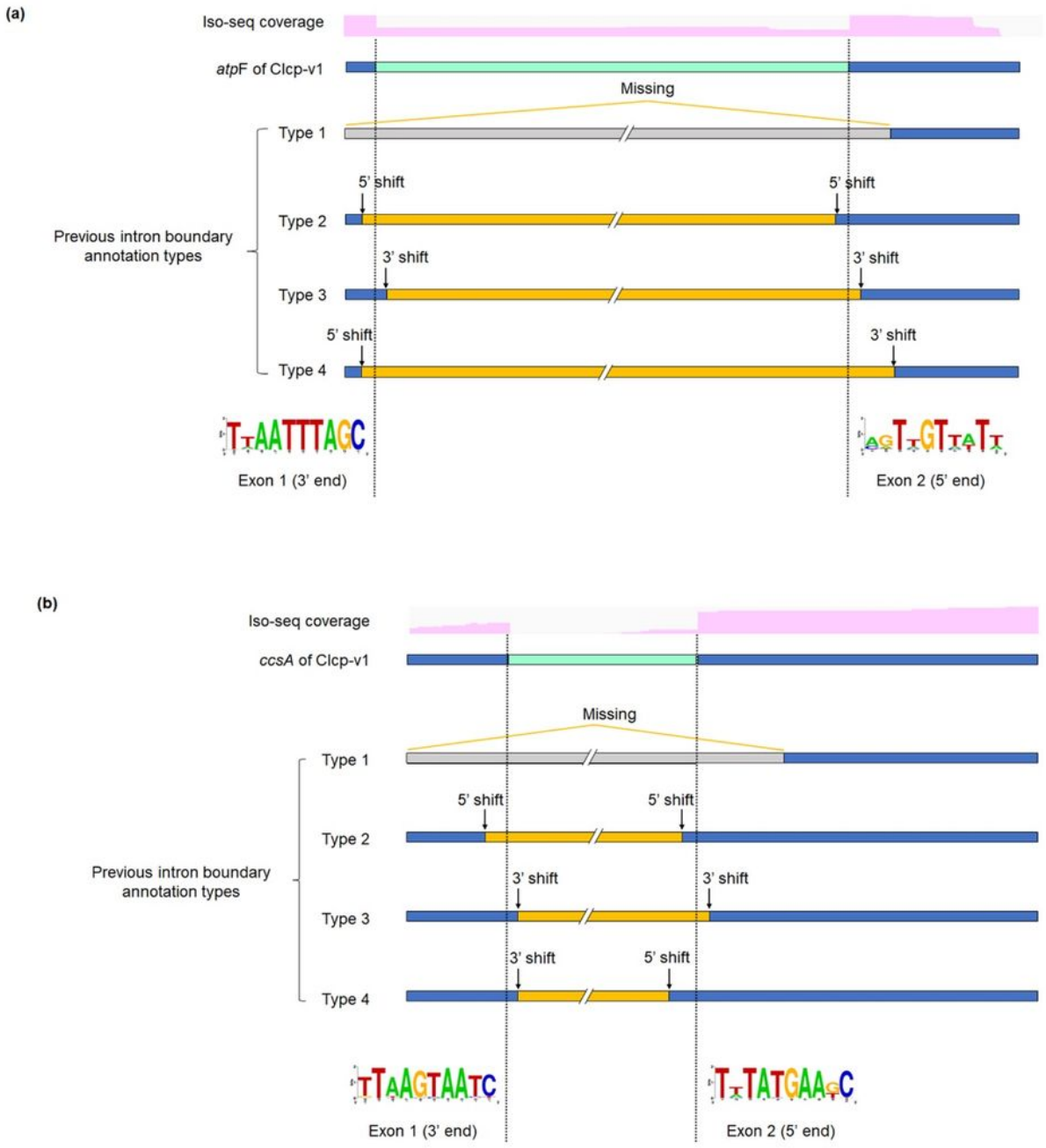

## Figure 3

Alignments of the fragmented genes. Fragmentation of tilS (a), cemA (b), and rpoB (c) in Caulerpa species are shown. Tydemania expeditionis (Te) and Bryopsis plumosa (Bp) were selected as representatives of suborder Halimedineae and Bryopsidineae, respectively. Cv: C. verticillata, Cc: C. cliftonii, Cs: C. serrulata, Cr: C. racemosa, Co: C. okamurae, Cm: C. manorensis, Cl: C. lentillifera.



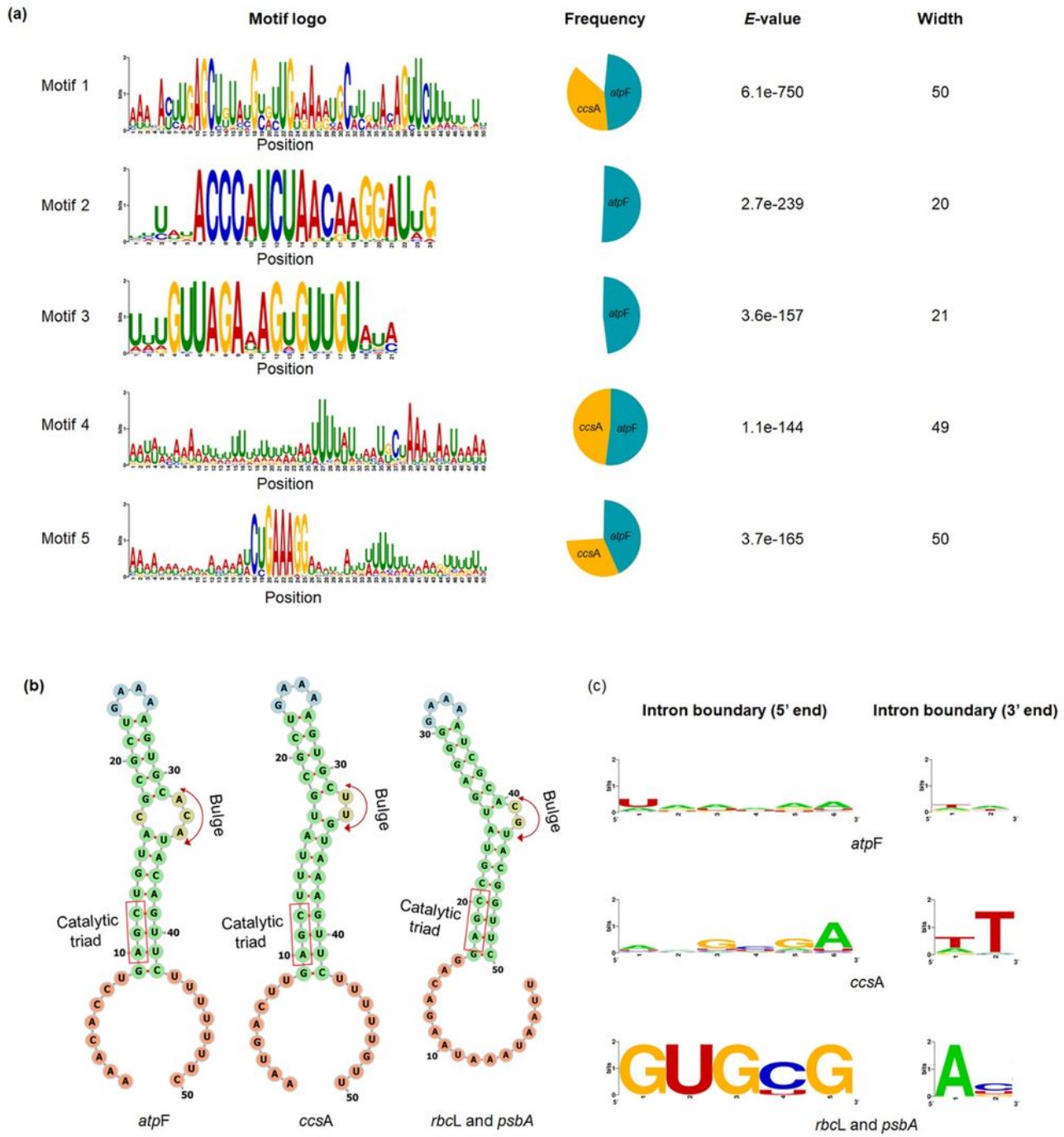

## Figure 4

Transcriptional characteristics analysis of fragmented genes by RT-PCR, cemA(a), tilS(b), rpoB (c). The positions of primer pairs were shown as arrows with different color. Electrophoresis of PCR products on 1% agarose gel. gDNA (genomic DNA) was used as positive control. S1 and S2 are two different individuals of C. lentillifera.




## Figure 5

Comparison conserved intron positions after adjusting with previous exon-intron boundaries annotation types. Comparison exon-intron boundaries of atpF(a) and ccsA (b). The yellow and grey boxes indicate the position of previously annotated introns and mis-annotated gene portions in current databases, respectively. Sequences logos at the 3’ boundary of exon 1 and 5' boundary of exon 2 in atpF and ccsA, respectively, are shown at the lower part.




## Figure 6

Motif discovery and distinct features of the introns from atpF and ccsA. The intron features were comparison with other group II introns (mainly from rbcL and psbA) in the Bryopsidales. (a) The top five motifs that were found from the 40 atpF and 37 ccsA intron sequences by MEME. Height of letters in motif logos indicates the occurrence of nucleotides at specific position. The frequency pie chart represents the frequency of occurrences of each motif across the 77 intron sequences, and the E-value

indicates the statistical significance of the motifs. Width represents the number of nucleotides in a particular motif. (b) Comparison of the consensus secondary structures of domain V among atpF ccsA, and other group II introns (rbcL and psbA). The conserved trinucleotide 5'-AGC-3', and the 2-nt bulge in domain V, which were the two catalytically important sites of group II introns, were highlighted by boxes and double arrow curves. (c) Comparison of the intron boundaries (5’ and 3’ end) among atpF ccsA, and other group II introns (rbcL and psbA).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BMCAdditionalfile1.docx
- BMCAdditionalfile10.docx
- BMCAdditionalfile11.xlsx
- BMCAdditionalfile12.xlsx
- BMCAdditionalfile2.docx
- BMCAdditionalfile3.xlsx
- BMCAdditionalfile4.xlsx
- BMCAdditionalfile5.xlsx
- BMCAdditionalfile6.xlsx
- BMCAdditionalfile7.docx
- BMCAdditionalfile8.xlsx
- BMCAdditionalfile9.docx