The plastid genome in Cladophorales green algae is encoded by hairpin plasmids

Andrea Del Cortona[1,2,3,4]*, Frederik Leliaert[1,5], Kenny A. Bogaert[1], Monique Turmel[6], Christian Boedeker[7], Jan Janouškovec[8], Juan M. Lopez-Bautista[9], Heroen Verbruggen[10], Klaas Vandepoele[2,3,4] & Olivier De Clerck[1]

[1]Department of Biology, Phycology Research Group, Ghent University, 9000 Ghent, Belgium
[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium
[3]VIB Center for Plant Systems Biology, 9052 Ghent, Belgium
[4]Bioinformatics Institute Ghent, Ghent University, 9052 Ghent, Belgium
[5]Botanic Garden Meise, Meise, Belgium
[6]Institut de biologie intégrative et des systèmes, Département de biochimie, de microbiologie et de bio-informatique, Université Laval, Québec (QC) Canada
[7]School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand
[8]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom
[9]Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL35484-0345, USA
[10]School of BioSciences, University of Melbourne, Victoria 3010, Australia

Chloroplast genomes, relics of an endosymbiotic cyanobacterial genome, are circular double-stranded DNA molecules[1]. They typically range between 100-200 kb and code for circa 80-250 genes[2]. While fragmented mitochondrial genomes evolved several times independently during the evolution of eukaryotes[3], fragmented plastid genomes are only known in dinoflagellates, where genes are present on several minicircles[4]. Here we show that the chloroplast genome of the green alga *Boodlea composita* (Cladophorales) is fragmented into hairpin plasmids. Short and long read high-throughput sequencing of DNA and RNA demonstrated that the chloroplast genome is fragmented with individual genes encoded on 1-7 kb, GC-rich DNA contigs, each containing a long inverted repeat with one or two protein-coding genes and conserved non-coding regions putatively involved in replication and/or expression. We propose that these contigs correspond to linear single-stranded DNA molecules that fold onto themselves to form hairpin plasmids. The *Boodlea* chloroplast genes are highly divergent from their corresponding orthologs. This highly deviant chloroplast genome coincided with an elevated transfer of chloroplast genes to the nucleus. Interestingly, both in Cladophorales and dinoflagellates a similar set of chloroplast genes is conserved in these two unrelated lineages. A chloroplast genome that is composed only of linear DNA molecules is unprecedented among eukaryotes and highlights unexpected variation in the plastid genome architecture.

Cladophorales are an ecologically important group of marine and freshwater green algae, which includes several hundreds of species. These macroscopic multicellular algae have giant, multinucleate cells containing numerous chloroplasts (Figure 1a-c). Currently, and in stark contrast to other algae[5-8], little is known about the gene content and structure of the chloroplast genome in the Cladophorales, since most attempts to amplify common chloroplast genes have failed[9,10]. Cladophorales contain abundant plasmids within chloroplasts[11,12], representing a Low Molecular Weight (LMW) DNA fraction. Pioneering work revealed that these plasmids are single-stranded DNA (ssDNA) molecules about 1.5-3.0 kb in length that fold in a hairpin configuration and lack

similarity to the nuclear DNA[11-14]. Some of these hairpin plasmids contain putatively transcribed sequences with similarity to chloroplast genes (*psaB*, *psbB*, *psbC* and *psbF*)[13].

With the goal to determine the nature of the Cladophorales chloroplast genome, we sequenced and assembled the DNA from a chloroplast-enriched fraction of *Boodlea composita*, using Roche 454 technology (Extended Data Figure 1, Supplementary Note 1.1). Rather than assembling into a typical circular chloroplast genome, 21 chloroplast protein-coding genes were found on 58 short contigs (1,203-5,426 bp): *atpA*, *atpB*, *atpH*, *atpI*, *petA*, *petB*, *petD*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbJ*, *psbK*, *psbL*, *psbT* and *rbcL*. All but the *rbcL* gene belong to the major thylakoid transmembrane protein complexes (ATP synthase, cytochrome b6f, Photosystem I, and Photosystem II). These contigs contained inverted repeats at their 5' and 3' termini (Extended Data Figures 2e and 3, Supplementary Note 1.1) and, despite high coverage by sequence reads, they could not be extended by iterative contig extension (Supplementary Notes 1.2 and 1.3). The inverted repeats were also found on contigs with no sequence similarity to known proteins, raising the number of contigs of chloroplast origin to 136 (Supplementary Note 1.1). These contigs are further referred to as chloroplast 454 contigs.

The 21 chloroplast genes displayed high sequence divergence compared to orthologous genes in other photosynthetic organisms (Figure 1d). To verify if this divergence is a shared feature of the Cladophorales, we generated additional sequence data from nine other species within this clade (Supplementary Note 7). A maximum likelihood phylogenetic tree based on a concatenated alignment of the 19 chloroplast genes showed that despite high divergence, the Cladophorales sequences formed a monophyletic group within the green algae (Figure 1e). For some genes, the identification of start and stop codons was uncertain and a non-canonical genetic code was identified. The canonical stop codon UGA was found 11 times internally in six genes (*petA*, *psaA*, *psaB*, *psaC*, *psbC* and *rbcL*; Supplementary Note 3), next to still acting a stop codon. Dual meaning of UGA as both stop and sense codons has recently been reported from a number of unrelated protists[15-17]. Interestingly, a different non-canonical genetic code has been described for Cladophorales nuclear genes, where UAG and UAA codons are reassigned to glutamine[18], which implies two independent departures from the standard genetic code in a single organism.

In order to confirm the transcription of the divergent chloroplast genes, we generated two deep-coverage RNA-seq libraries: a total-RNA library and a mRNA library enriched for nuclear transcripts (Supplementary Note 4). Following *de novo* assembly, we identified chloroplast transcripts by a sequence similarity search. Transcripts of 20 chloroplast genes were identical to the genes encoded by the chloroplast 454 contigs (Extended Data Figure 3, Supplementary Note 5). The high total-RNA to mRNA ratio observed for reads that mapped to the chloroplast 454 contigs corroborated that these genes were not transcribed in the nucleus (Extended Data Figure 3, Supplementary Notes 1.2 and 5). Moreover, complete congruence between RNA and DNA sequences excluded the presence of RNA editing. Additional transcripts of 66 genes that have been located in the chloroplast in other Archaeplastida were identified (Extended Data Table 1). Although their subcellular origin was not determined experimentally, they are probably all nuclear-encoded, based on high mRNA to total-RNA reads ratio and their presence on High Molecular Weight (HMW) DNA reads (see below).

The failure to assemble a circular chloroplast genome might be due to repetitive elements that impair the performance of short-read assemblers[19]. To overcome assembly artefacts and close putative gaps in the chloroplast 454 contigs, we applied Single-Molecule Real-Time (SMRT) sequencing (Pacific Biosciences) to the HMW and LMW DNA fractions (Extended Data Figure 1). Hypothetical gaps between the chloroplast 454 contigs could not be closed with long HMW

DNA reads, nor did a hybrid assembly generate a circular chloroplast genome (Supplementary Note 1.4). As a consequence, we conclude that the chloroplast genome is not a single large molecule. 22 HMW DNA reads harboured protein-coding genes commonly present in Archaeplastida chloroplast genomes (Extended Data Table 1). All but three of these genes (which likely correspond to carry-over LMW DNA) contained introns, were absent in the chloroplast 454 contigs, and had a high mRNA to total-RNA ratio of mapped reads, altogether suggesting that they are encoded in the nucleus (Extended Data Figure 4, Extended Data Table 1).

Conversely, 22 chloroplast genes (that is, the 21 protein-coding genes identified in chloroplast 454 contigs as well as the 16S rRNA gene) were found in the LMW DNA reads (Extended Data Table 1). An orthology-guided assembly of chloroplast 454 contigs and the LMW DNA reads (Supplementary Note 1.7) resulted in 34 contigs between 1,179 and 6,925 bp in length, henceforth referred to as "chloroplast genome" (Extended Data Table 2, Supplementary Note 2). Several of these contigs are long near-identical palindromic sequences, including full-length coding sequences (CDSs), and a somewhat less conserved tail region (Figure 1f-h, Extended Data Figure 2a). The remaining contigs appear incomplete but are often indicative of the presence of similar palindromic structures (Extended Data Figure 2). Such palindromes allow regions of the single-stranded LMW DNA molecules to fold into hairpin-like secondary structures, which had been inferred from denaturing gel analysis and visualized by electron microscopy[11]. These hairpin plasmids are directly evidenced in several genes by long LMW DNA reads and are consistent with the structure of all chloroplast contigs (Extended Data Table 2). The 16S rRNA gene was split across two distinct hairpin plasmids and much reduced compared to algal and bacterial homologs (Extended Data Figure 5), similar to what is found in the chloroplast genomes of peridinin-pigmented dinoflagellates and nematode mitochondrial genomes[20,21]. We could not detect the 23S rRNA gene nor the 5S rRNA gene.

Non-coding DNA regions (ncDNA) of the hairpin plasmids showed high sequence similarity among all molecules of the *Boodlea* chloroplast genome. Within the ncDNA we identified six conserved motifs, 20 to 35 bp in length (Extended Data Figure 6), which lack similarity to known regulatory elements. Motifs 1, 2 and 5 were always present upstream of the start codon of the chloroplast genes, occasionally in more than one copy. Although their distances from the start codon were variable, their orientations relative to the gene were conserved, indicating a potential function as a regulatory element of gene expression and/or replication of the hairpin plasmids. A sequence similarity search revealed that these motifs are also present in 1,966 LMW DNA reads lacking genes. This evidence supports earlier findings of abundant non-coding LMW DNA molecules in the Cladophorales[11,13] and is consistent with the expectation that recombination and cleavage of repetitive DNA will produce a heterogeneous population of molecules, as observed in dinoflagellates plastids[4] (Extended Data Figure 2).

In contrast, a very small fraction of the HMW DNA reads (15 corrected reads) displayed the ncDNA motifs and these were found exclusively in long terminal repeat retrotransposons (RT-LTRs, Supplementary Note 6). Some RT-LTRs were also abundant in the 454 contigs (Supplementary Note 1.1). These observations are suggestive of DNA transfer between nuclear RT-LTRs and the chloroplast DNA, an event that may be responsible for the origin of the hairpin plasmids. Hypothetically, an invasion of nuclear RT-LTRs in the Cladophorales ancestor may have resulted in an expansion of the chloroplast genome and its subsequent fragmentation into hairpin plasmids. An RT-LTR invasion could also have accounted for chloroplast gene transfer to the nucleus. Current LMW DNA in Cladophorales may thus represent non-autonomous retro-elements, which require an independent retrotranscriptase for their replication. Linear hairpin

plasmids characterized as retroelements have been reported in the mitochondria of the ascomycete *Fusarium oxysporum* in addition to a canonical mitochondrial genome[22,23].

We collected several lines of evidence that *Boodlea composita* lacks a typical large circular chloroplast genome. The chloroplast genome is instead reduced and fragmented into linear hairpin plasmids. Thirty-four hairpin plasmids were identified, harbouring 21 protein-coding genes and the 16S rRNA gene, which are highly divergent in sequence compared to orthologs in other algae. The exact set of *Boodlea* chloroplast genes remains elusive, but at least 19 genes coding for chloroplast products appear to be nuclear-encoded, of which nine are always chloroplast-encoded in related green algae. This suggests that chloroplast genome fragmentation in the Cladophorales has been accompanied with an elevated transfer of genes to the nucleus, similarly to the situation in peridinin-pigmented dinoflagellates[4]. Indeed, the two algal groups have converged on a very similar gene distribution: chloroplast genes code only for the subunits of photosynthetic complexes (and also for Rubisco in *Boodlea*), whereas the expression machinery is fully nucleus-encoded (Extended Data Table 1). Other non-canonical chloroplast genome architectures have recently been observed, such as a monomeric linear chromosome in the alveolate microalga *Chromera velia*[24] and three circular chromosomes in the green alga *Koshicola spirodelophila*[25], but these represent relatively small deviations from the paradigm. The reduced and fragmented chloroplast genome in the Cladophorales is wholly unprecedented and will be of significance to understanding processes driving organellar genome reduction, endosymbiotic gene transfer, and the minimal functional chloroplast gene set.

Online Content. Methods, along with any additional Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Supplementary Information is available in the online version of the paper.

Acknowledgments We thank Ellen Nisbett, Christopher Howe, Bram Verhelst, Sven Gould, and John W. La Claire II for help and advice. This work was supported by UGent BOF/01J04813 the Australian Research Council (DP150100705) to H.V., the National Science Foundation (GRAToL 10136495) to J.L.B.

Author Contributions F.L., M.T., C.B., H.V., K.V. and O.D.C conceived the study. F.L. and C.B. provided chloroplast 454 data. A.D.C., K.V. and O.D.C provided the RNAseq and PacBio data. F.L. and J.M.L-B. provided genome data of additional Cladophorales species. A.D.C., F.L., K.A.B., J.J., K.V. and O.D.C analysed genomic data (including genomes assemblies, annotations, and phylogenetic analyses). A.D.C., F.L., K.A.B., M.T., J.J., H.V., K.V. and O.D.C wrote, and all authors edited and approved, the manuscript.

Sequence Information Sequence data have been deposited to the NCBI Sequence Read Archive as BioProject PRJNA384503. The annotated chloroplast contigs of *Boodlea composita* and additional Cladophorales species were deposited to GenBank under accession numbers ***–***. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to O.D.C. (olivier.declerck@ugent.be).

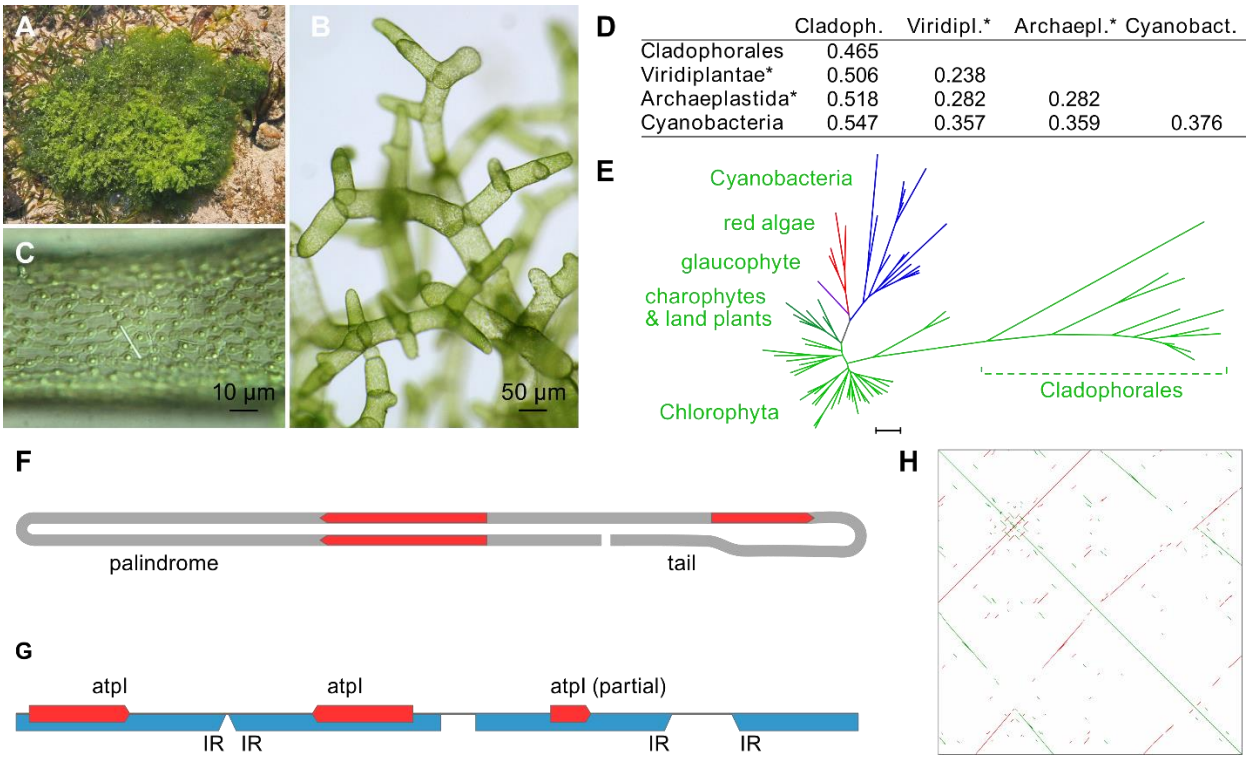


Figure 1 | *Boodlea composita* and its chloroplast hairpin plasmids. a. Specimen in natural environment. **b.** Detail of branching cells. **c.** Detail of chloroplasts dotted with pyrenoids and forming a parietal network. **d.** Maximum amino acid pairwise sequence distance estimated between the chloroplast genes of the taxa in the phylogenetic tree shown in Figure 1e (* excluded Cladophorales). **e.** Maximum likelihood phylogenetic tree inferred from a concatenated amino acid alignment of 19 chloroplast genes (Supplementary Note 7.2). Taxon names and bootstrap support values are provided in the tree shown in Figure S7.2. The scale represent 0,3 substitution per amino acid position. **f.** Schematic representation of the predicted native conformation of chloroplast hairpins plasmids, with a near-perfect palindromic region and a less conserved tail). Red arrows represent CDSs. **g.** Schematic representation of a group A read (see Extended Data Figure 2). Red arrows represent CDSs, blue arrows represent major inverted repeats. **h.** Dotplot illustrating the complexity of repetitive sequences. The read is plotted in both the x and the y axes; green lines indicate similar sequences, while red lines indicate sequences similar to the respective reverse complements.

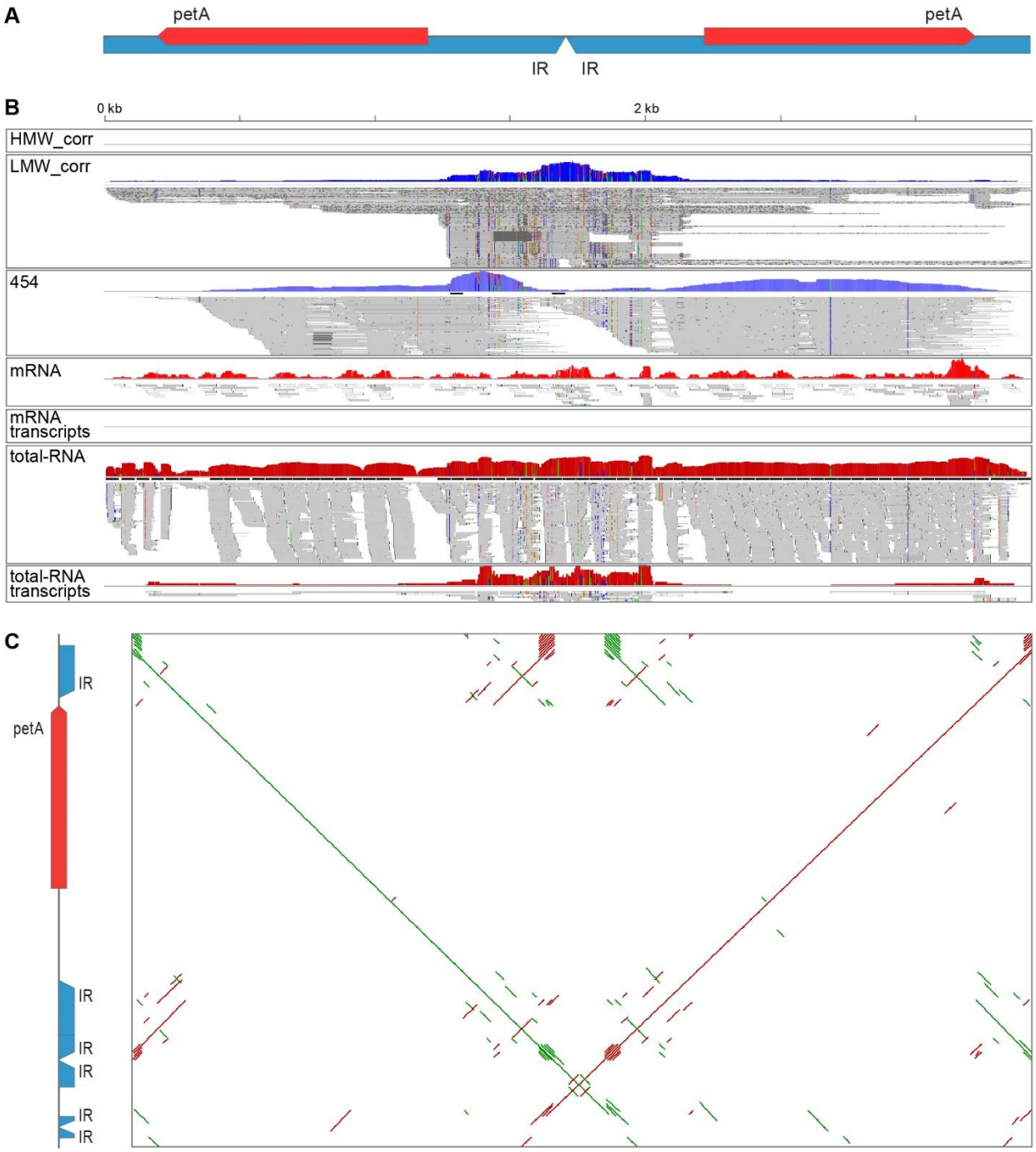


Figure 2 | LMW DNA reads containing chloroplast genes are expressed, enriched in the total-RNA fraction and congruent to the respective chloroplast 454 contigs. a. Representation of *petA* LMW DNA read (3,398 bp), a representative of group B contigs and reads (see Extended data Figure 2). The red arrows indicate CDSs, the blue arrows indicate inverted repeats. **b.** Corresponding Genome Browser track, from top to bottom: corrected HMW DNA coverage [0], corrected LMW DNA read coverage [range 0-541], 454 read coverage [range 0-567], mRNA library read coverage [range 0-17], assembled mRNA transcripts mapped [0], total-RNA library read coverage [range 0-7,936], and assembled total-RNA transcripts mapped [range 0-17]. **c.**

Dotplot showing congruence between *petA* LMW DNA read (x axis) and the corresponding *petA*-containing chloroplast 454 contig (y axis, 2,012 bp). Green lines indicate similar sequences, while red lines indicate sequences similar to the respective reverse complements.

Online Materials and Methods

Algal material and chloroplast isolation. A clonal culture of *Boodlea composita* (FL1110) is maintained in the algal culture collection of the Phycology Research Group, Ghent University. The specimen was grown in enriched sterilized natural seawater[26] at 22°C under 12:12 (light:dark) cool white fluorescent light at 60 µmol photons m^{-2} s^{-1}. A chloroplast-enriched fraction was obtained following the protocol of Palmer et al.[27].

Genomic DNA extraction, quantification and sequencing. Total genomic DNA from fresh *Boodlea* culture was isolated by using a modified CTAB extraction protocol[28], where the first step was to grind the sample in liquid nitrogen with mortar and pestle prior to extraction. This first step was skipped for DNA extraction from the chloroplast-enriched fraction. HMW and LMW DNA bands were size-selected using a BluePippin™ system (Sage Science, USA). The HMW DNA band was isolated with a cut-off range of 10 kb to 50 kb, while the LMW DNA band was isolated with a cut-off range of 1.5 kb to 2.5 kb. The quantity, quality and integrity of the extracted DNA were assessed with Qubit (ThermoFisher Scientific, USA), Nanodrop spectrophotometer (ThermoFisher Scientific), and Bioanalyzer 2100 (Agilent Technologies, USA).

DNA from the chloroplast-enriched fraction was sequenced with Roche 454 GS FLX at GATC Biotech, Germany. The HMW and LMW DNA fractions were sequenced on two SMRT cells on a PacBio RS II (VIB Nucleomics Core facilities, Leuven, Belgium) using PacBio P5 polymerase and C3 chemistry combination (P5-C3). For the HMW DNA fraction, a 20-kb SMRT-bell library was constructed, while for the LMW DNA fraction, a 2-kb SMRT-bell library was constructed.

Chloroplast DNA assembly and annotation. Quality of the reads from the 454 library was assessed with FastQC v.0.10.1 (http://www.bioinformatics.babraham.ac.uk, last accessed March 01, 2017). Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' end of the reads were trimmed with Fastxv.0.0.13 (https://github.com/agordon/fastx_toolkit, last accessed March 01, 2017). After trimming, reads shorter than 50 bp were discarded. *De novo* assembly of the trimmed reads was performed with MIRA v. 4.0rc5[29]. After the assembly, putative chloroplast contigs were identified by comparing their translated sequences against the NCBI non-redundant protein database using BLAST 2.2.29+[30]. Additional chloroplast contigs without similarity to protein-coding genes were identified by metagenomic binning (distribution analysis of 4-mers) with MyCC[31]. Contigs potentially coding for chloroplast tRNAs and rRNAs were identified using Infernal 1.1[32]. The chloroplast 454 contigs served as seeds for iterative contig extension with PRICE[33] (Supplementary Note 1.3), using the unpaired 454 trimmed reads as false paired-end reads. Repetitive regions in the contigs were identified with 'einverted', 'etandem' and 'palindrome' from the EMBOSS[34] package. Dotplots for all contigs were generated with YASS, using standard parameters[35]. Coverage of chloroplast contigs was evaluated by mapping the 454 reads with CLC Genomics Workbench (Qiagen, https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/, last accessed November 24, 2016). Coverage data were used to identify putative chimeric contigs (Supplementary Notes 1.1 and 1.2).

The chloroplast 454 contigs were used together with HMW DNA reads for two independent hybrid assemblies (Supplementary Note 1.4). First, we closed the gaps between the chloroplast 454 contigs with the pbahaScaffolder.py script integrated in the smrtanalysis 2.3.0 pipeline[36]. Secondly, the pre-assembled chloroplast 454 contigs were used as anchors for HMW DNA reads in a round of hybrid assembly with dbg2olc[37]. Since hybrid assemblies with uncorrected reads could not reconstruct a circular chloroplast genome, HMW DNA reads were further characterized after error correction. The high-noise HMW DNA reads were corrected using a hybrid correction

with proovread 2.12[38] using 454 reads and reads from Illumina RNA-seq libraries (see below). Corrected reads encoding chloroplast genes were identified by aligning them against a custom protein database, named Chloroprotein_db, including genes from the pico-PLAZA protein database[39] and protein-coding genes from published green algal chloroplast genomes (Chlorophyta sensu Bremer 1985, NCBI Taxonomy id: 3041).

LMW DNA reads were self-corrected with the PBcR pipeline[40] (Supplementary Note 1.5). *De novo* genome assembly of corrected reads was performed with the Celera WGS assembler version 8.3rc2[41] (Supplementary Note 1.6). Corrected and uncorrected reads as well as assembled contigs potentially encoding chloroplast genes were identified by aligning them with BLAST 2.2.29+ against Chloroprotein_db. LMW DNA harbouring chloroplast CDSs were re-assembled together with the respective chloroplast 454 contigs using Geneious v. 8.1.7 (Biomatters, http://www.geneious.com/, last accessed March 01, 2017, Supplementary Note 1.7). Contigs resulting from the orthology-guided assembly and chloroplast 454 contigs not included in the orthology-guided assembly were retained as *Boodlea* chloroplast genome contigs.

Protein-coding genes in the *Boodlea* chloroplast genome were identified with a sequence similarity search against the NCBI non-redundant protein database with BLAST 2.2.29+. Their annotation was manually refined in Geneious and Artemis 16.0.0[42] based on the BLAST search results. Repetitive elements were mapped on *Boodlea* chloroplast genome by aligning the contigs with themselves using BLAST 2.2.29+. Non-coding RNAs were identified with infernal 1.1 (cut-off value 10^{-5}). Conserved motifs were predicted with MEME suite[43], and the discovered motifs were clustered with RSAT[44]. The motifs were compared with the JASPAR-2016[45] database using TOMTOM[46] (p-value cut-off 10^{-3}).

Boodlea chloroplast genome coverage was evaluated by mapping the 454 reads with gsnap v.2016-04-04[47]. Corrected and uncorrected LMW DNA subreads and chloroplast 454 contigs resulting from the MIRA assembly were mapped against the *Boodlea* chloroplast genome with gmap v. 2014-12-06[47] using the *–nosplicing* flag. *Boodlea* chloroplast genome expression was evaluated by mapping the reads from the *Boodlea* mRNA and total-RNA libraries with gsnap and by aligning to the chloroplast genome the transcripts resulting from the *de novo* assembly of the RNA-seq libraries with gmap. Due to the high amount of repetitive sequences in LMW DNA reads and 454 contigs, the resulting annotated *Boodlea* chloroplast genome was carefully inspected in order to exclude sequencing and assembly artefacts.

Completeness of chloroplast genome was evaluated by comparing chloroplast genes in *Boodlea* to a set of 60 "core" chloroplast protein-coding genes, defined as protein-coding genes conserved among the chloroplast genomes of the following representative species of Chlorophyta (Extended Data Table 1): *Bryopsis plumosa*, *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Coccomyxa subellipsoidea*, *Gonium pectorale*, *Leptosira terrestris*, *Nephroselmis olivacea*, *Oltmannsiellopsis viridis*, *Parachlorella kessleri*, and *Pseudendoclonium akinetum*, and the streptophyte *Mesostigma viride*.

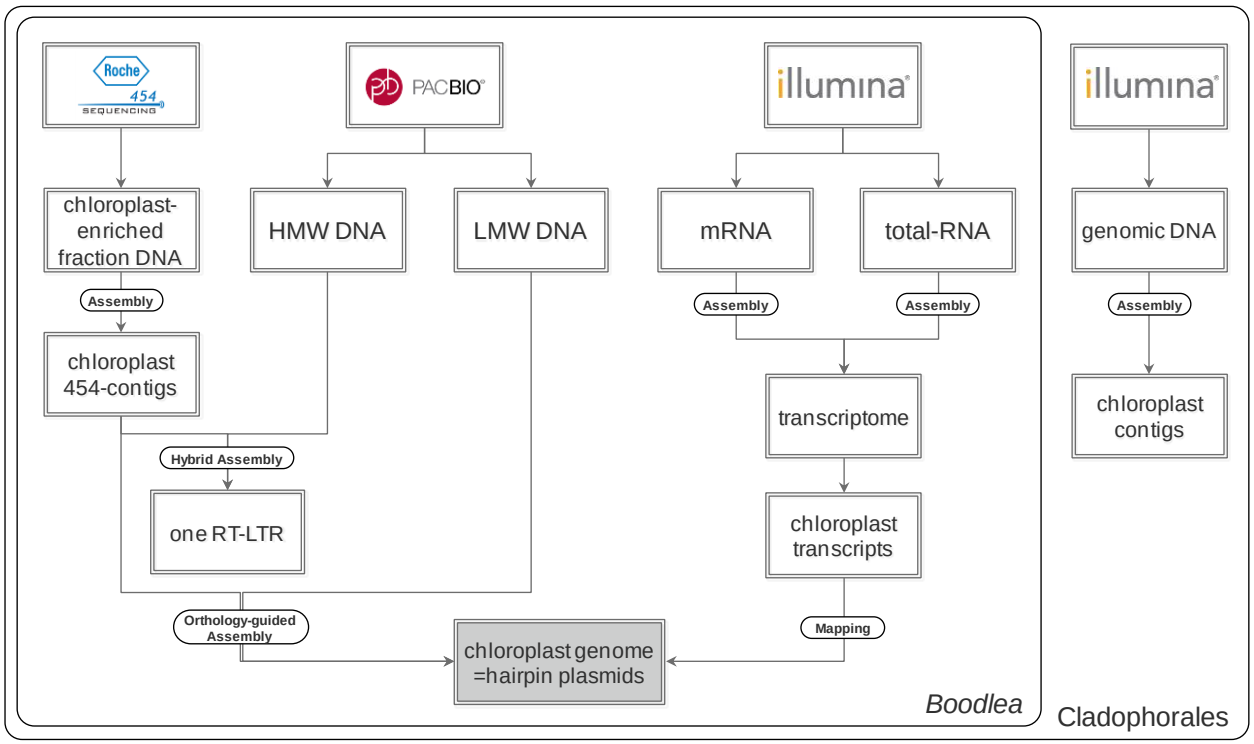
RNA extraction, quantification and sequencing. Total RNA was isolated by using a modified CTAB extraction protocol[48]. RNA quality and quantity were assessed with Qubit and Nanodrop spectrophotomete, and RNA integrity was assessed with a Bioanalyzer 2100. Two cDNA libraries for NextSeq sequencing were generated using TruSeq™ Stranded RNA sample preparation kit (Illumina, USA): one library enriched in poly(A) mRNA due to oligo-(dT) retrotranscription and one total RNA library depleted in ribosomal RNA with Ribo-Zero Plant kit (Epicentre, USA). The two libraries were sequenced on one lane of Illumina NextSeq 500 Medium platform at 2x76 bp by VIB Nucleomics Core facilities (Leuven, Belgium, Table S4.1).

Transcriptome assembly and annotation. Quality of the reads from the two RNA-seq libraries was assessed with FastQC. Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' end of the reads were trimmed with Fastx. After trimming, reads shorter than 30 bp were discarded. Read normalization and *de novo* assembly of the libraries were performed with Trinity 2.0.4[49]. The resulting contigs (hereafter, transcripts) were compared using sequence similarity searches against the NCBI non-redundant protein database using Tera-BLAST™ DeCypher (Active Motif, USA). Taxonomic profiling of the transcripts was performed using the following protocol (Table S4.2): for each transcript, sequence similarity searches were combined with the NCBI Taxonomy information of the top ten BLAST Hits in order to discriminate between eukaryotic and bacterial transcripts or transcripts lacking similarity to known protein-coding genes. Transcripts classified as "eukaryotic" were further examined to assess transcriptome completeness and to identify chloroplast transcripts. These transcripts were analysed using Tera-BLAST™ DeCypher against Chloroprotein_db. Transcriptome completeness was evaluated with a custom Perl script that compared gene families identified in the *Boodlea* transcriptome to a set of 1816 "core" gene families shared between Chlorophyta genomes present in pico-PLAZA 2.0[39], following Veeckman et al. guidelines to estimate the completeness of the annotated gene space[50] (Table S4.3).

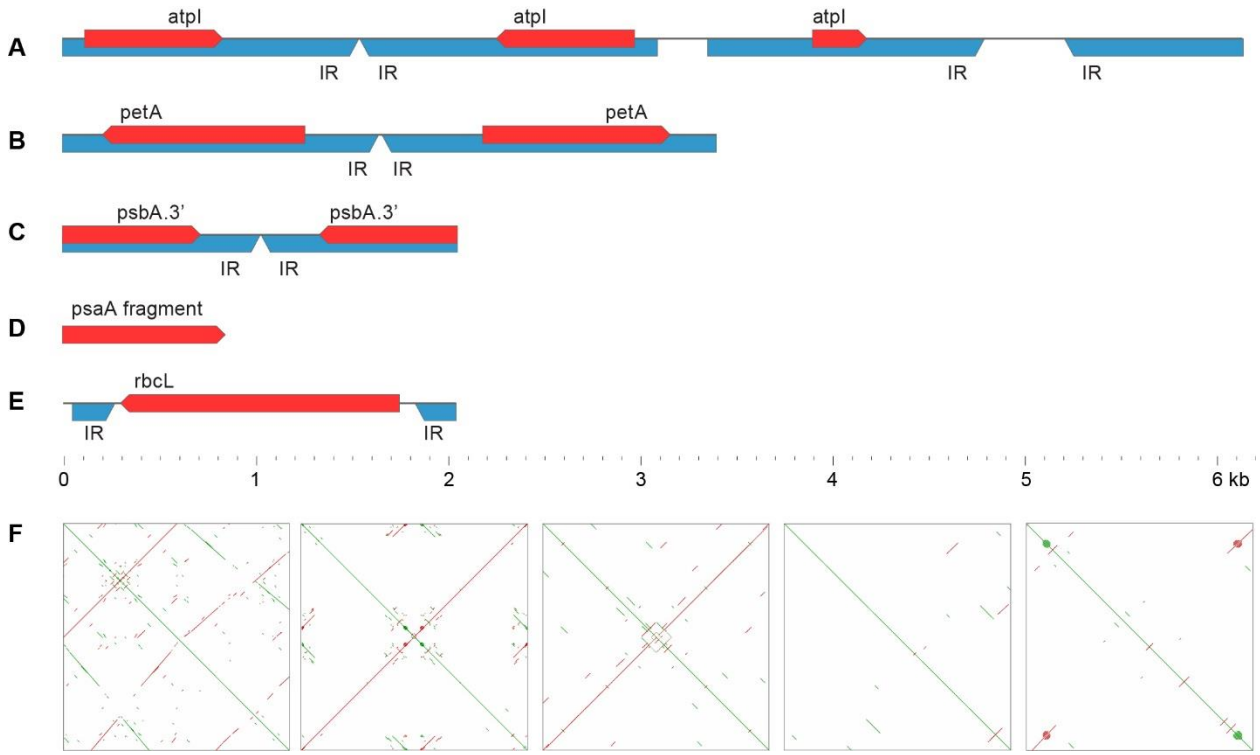
DNA sequencing of additional Cladophorales species and phylogenetic analysis. Genomic DNA from 10 additional Cladophorales species was sequenced using Illumina HiSeq 2000 technology and assembled (Supplementary Note 7.1). Although comparable sequencing approaches resulted in the assembly of circular chloroplast genomes for other green seaweeds[51,52], only short chloroplast contigs (200-8,000 bp) were assembled from these libraries. In contrast to the genes found in the *Boodlea* hairpin plasmids, most of the chloroplast genes identified in the additional Cladophorales libraries were fragmented (Supplementary Table S7.2).

A maximum likelihood phylogenetic tree was inferred from the amino acid sequences of 19 chloroplast genes from *Boodlea*, nine other Cladophorales species, 41 additional species of Archaeplastida, and 14 Cyanobacteria species (Supplementary Note 7.2).

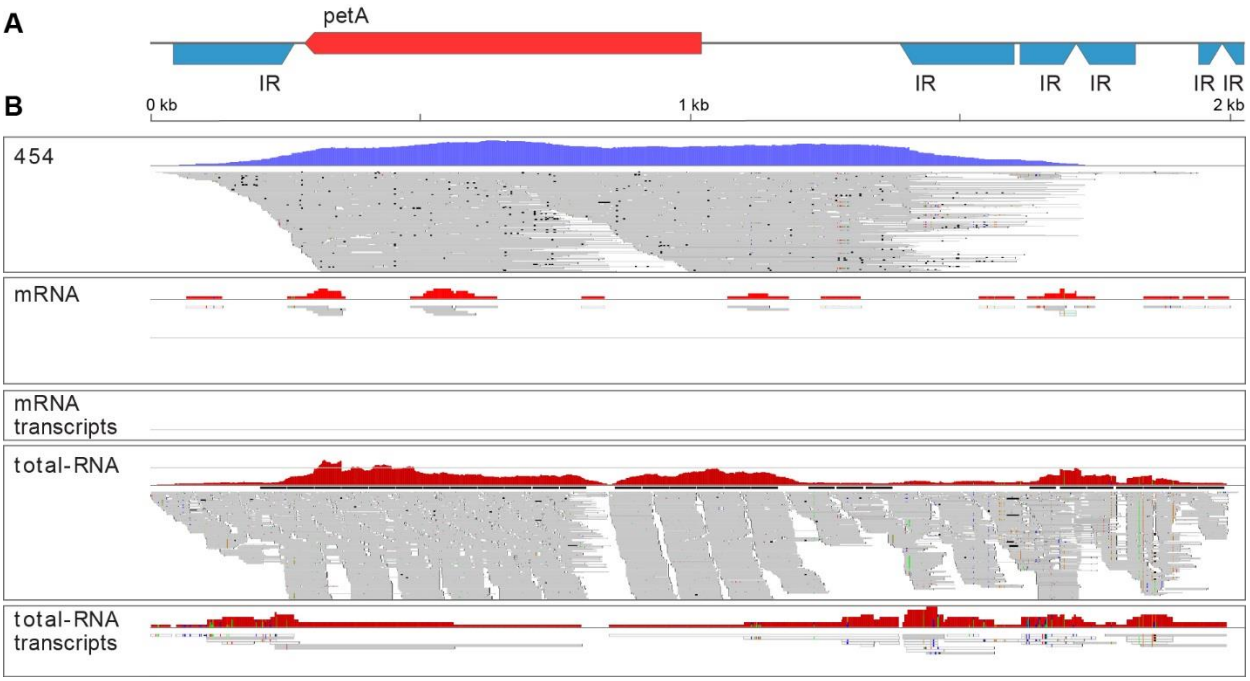
Extended Data display items



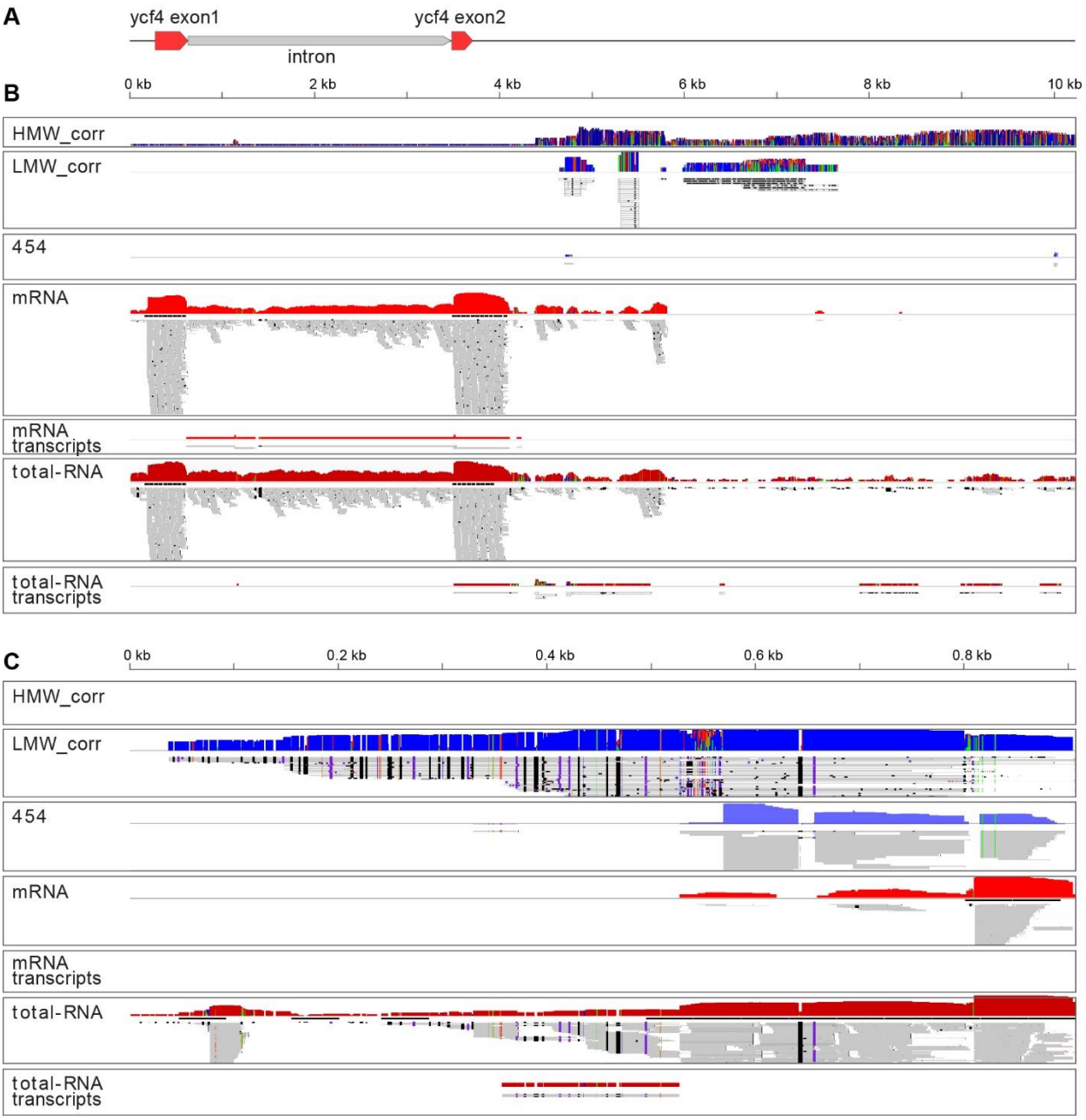
Extended Data Figure 1 | Sequence datasets generated in this study and their main analyses. The diagram in the inner box "*Boodlea*" describes the workflow used to characterize the chloroplast genome structure and organization. The datasets represented in the outer box "Cladophorales" were used for phylogenetic inference and confirmation that the chloroplast genome of Cladophorales is distributed over hairpin plasmids. HMW DNA: High Molecular Weight DNA; LMW DNA: Low Molecular Weight DNA, RT-LTR: Long Terminal Repeat Retrotransposon.



Extended Data Figure 2 | Schematic representation of assembled *Boodlea* chloroplast DNA contigs. Red arrows indicate CDSs, blue arrows indicate repetitive elements. **a.** Group A (e.g. *atpI*, 6,139 bp): Palindromic sequence with a less conserved tail. One single LMW DNA read belongs to this group. Long palindromic sequences characterized by two full-length CDSs in opposite orientations. The first half of the read is a perfect palindrome itself, containing the two inverted CDSs. The second half of the read is less conserved, but still similar to the palindrome in the first half of the read. **b.** Group B (e.g. *petA*, 3,398 bp): Palindromic sequences with full-length CDSs. 57 LMW DNA reads belong to this group. Perfect palindromes with full-length CDSs in opposite orientations, resembling the perfect palindromic stretch in the first half of group A reads. **c.** Group C (e.g. *psbA*, 2,043 bp): palindromic sequences with fragments of CDS. 252 LMW DNA reads belong to this group. **d.** Group D (e.g. *psaA*, 856 bp): short reads with fragments of CDS that lack extensive repetitive elements. 2,099 LMW DNA reads belong to this group. **e.** Group E (e.g. *rbcL*, 2,036 bp): Full-length CDSs delimited by inverted repeats. 23 chloroplast 454 contigs and 120 LMW DNA reads belong to this group. Similar to group E contigs and reads, the remaining 113 chloroplast 454 contigs lacking full-length CDSs are delimited by inverted repeats. **f.** Dotplots showing the abundance of repetitive elements, from left to right, for group A, group B, group C, group D and group E reads/contigs, respectively. Each dotplot was generated by aligning the contig/read with itself. Green lines indicate similar sequences, while red lines indicate sequences similar to the respective reverse complements.



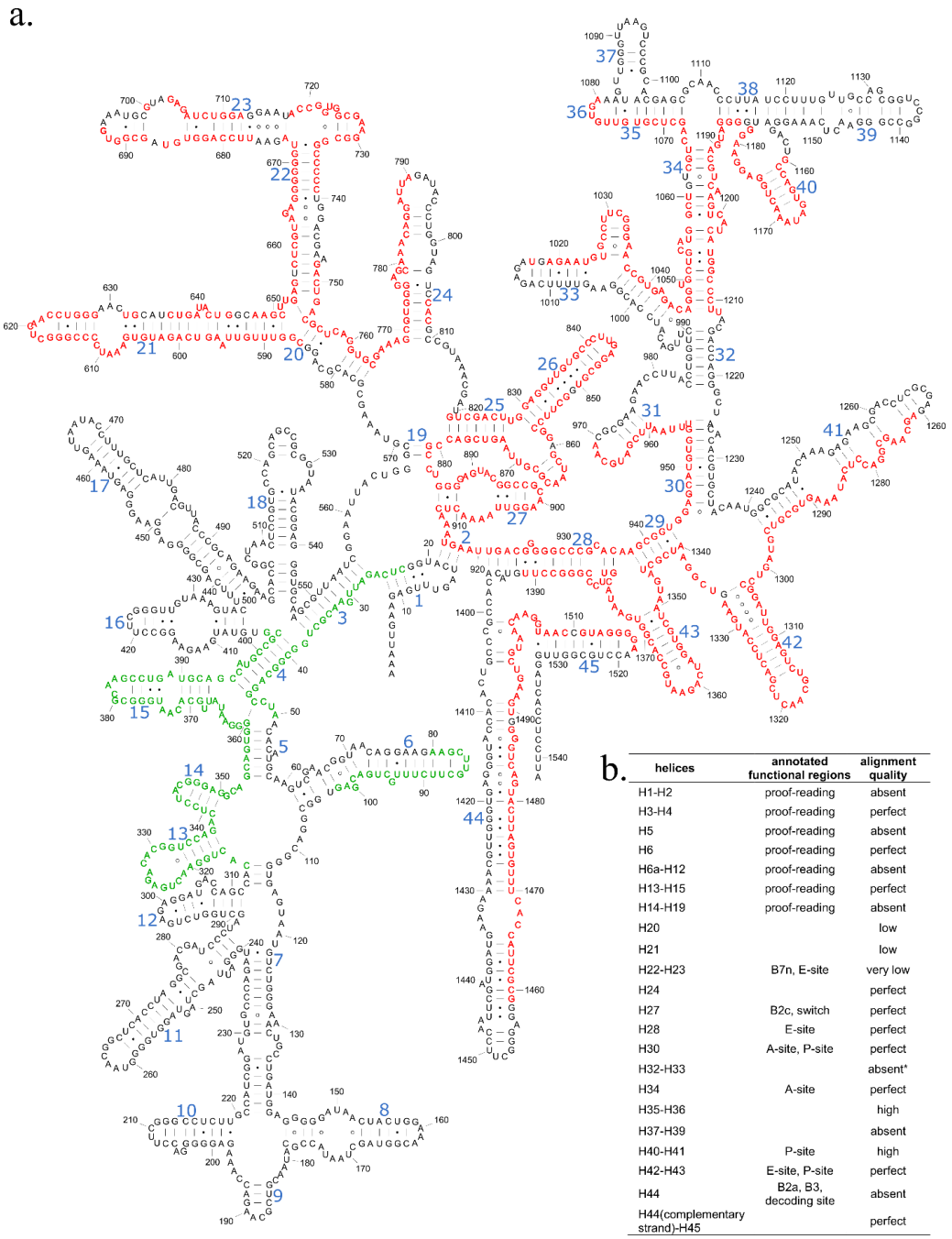
Extended Data Figure 3 | Chloroplast 454 contigs. a. Representation of the *petA*-containing contig (2,012 bp), a representative of the chloroplast 454 contigs. The red arrow indicates the CDS, the blue arrows indicate inverted repeats. **b.** Corresponding Genome Browser track, from top to bottom: 454 read coverage [range 0-235], mRNA library read coverage [range 0-3], assembled mRNA transcripts mapped [0], total-RNA library read coverage [range 0-876], assembled total-RNA transcripts mapped [range 0-7].



Extended Data Figure 4 | HMW DNA reads contain nuclear-encoded protein-coding genes that are normally present in the chloroplast genome. a. Representation of the HMW DNA read encoding *ycf4* (10,517 bp), one of the 22 nuclear-encoded genes identified in the HMW DNA fraction. The red arrow indicates the spliced CDS, the grey dash indicates the intron. **b.** Corresponding Genome Browser track showing from top to bottom: HMW DNA read coverage [range 0-13], LMW DNA read coverage [range 0-22], 454 read coverage [range 0-2], mRNA library read coverage [range 0-5,131], assembled mRNA transcripts mapped [range 0-2], total-RNA library read coverage [range 0-896], and assembled total-RNA transcripts mapped [range 0-3]. **c.** *psbA*, *psbB*, *psbC* HMW DNA reads are not transcribed by the nuclear machinery and they are possibly LMW DNA carry-over contaminants, rather than genuine genes transferred to the

nucleus. HMW to LMW DNA reads ratio and mRNA to total-RNA reads ratio suggest and support the first hypothesis. Genome Browser track of the *psbA* HMW DNA read (909 bp) showing from top to bottom: corrected HMW DNA read coverage [range 0-1], corrected LMW DNA read coverage [range 0-58], 454 read coverage [range 0-27], mRNA library read coverage [range 0-795], assembled mRNA transcripts mapped [0], total-RNA library read coverage [range 0-408,934], and assembled total-RNA transcripts mapped [range 0-2].

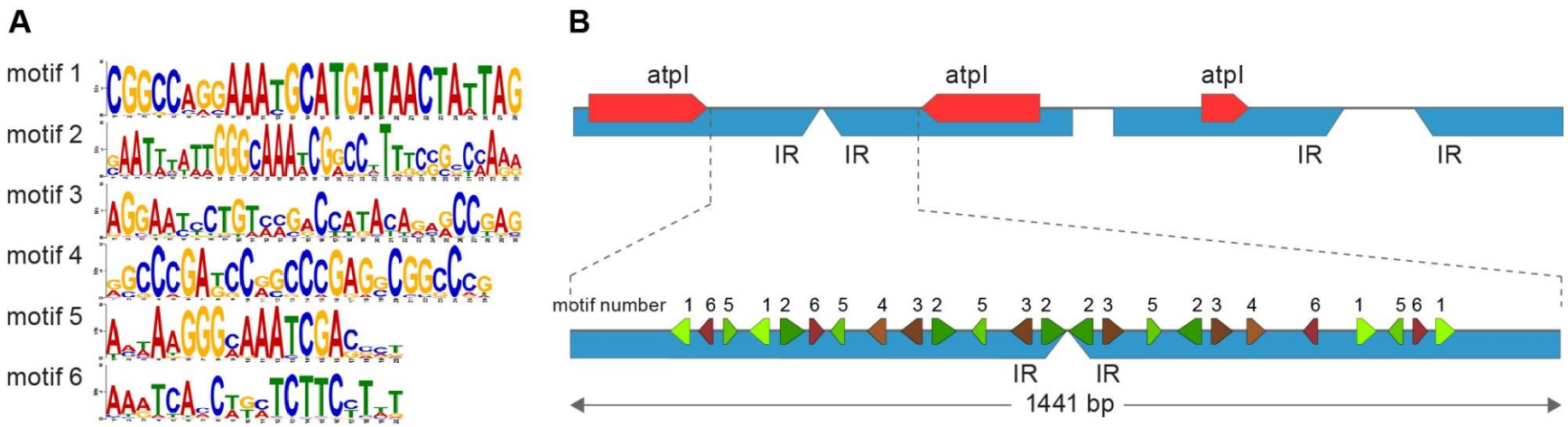
a.



b.

helices	annotated functional regions	alignment quality
H1-H2	proof-reading	absent
H3-H4	proof-reading	perfect
H5	proof-reading	absent
H6	proof-reading	perfect
H6a-H12	proof-reading	absent
H13-H15	proof-reading	perfect
H14-H19	proof-reading	absent
H20		low
H21		low
H22-H23	B7n, E-site	very low
H24		perfect
H27	B2c, switch	perfect
H28	E-site	perfect
H30	A-site, P-site	perfect
H32-H33		absent*
H34	A-site	perfect
H35-H36		high
H37-H39		absent
H40-H41	P-site	high
H42-H43	E-site, P-site	perfect
H44	B2a, B3, decoding site	absent
H44(complementary strand)-H45		perfect

Extended Data Figure 5 | *Boodlea* chloroplast 16S rRNA secondary structure. a. *Boodlea* chloroplast 16S rRNA secondary structure was compared with the *E. coli* 16S rRNA model [RF00177]. Residues shown in green and red on the *E. coli* model represent the 5' and 3' 16S rRNA regions coded by the two *Boodlea* 16S rRNA hairpin plasmids. Residues in black are absent in *Boodlea* 16S rRNA. Blue numbers indicate secondary structure helices in the 16S rRNA model[53]. **b.** Comparison between *Boodlea* and *E. coli* 16S rRNA annotated functional regions. Quality of the alignment was assessed based on the predicted posterior probability (in percentage) of each aligned region: very low < 25%; low between 25-50%; high between 50-95%; and perfect > 95%.



Extended Data Figure 6 | Conserved non-coding motifs in *Boodlea* LMW DNA. a. Sequence logos of the conserved motifs predicted in the *Boodlea* chloroplast genome. **b.** Schematic representation of the distribution of the motifs in the 1,441 bp ncDNA region from the *atpI* group A read used for the identification of additional chloroplast reads in the LMW DNA library. Motifs with conserved orientation relative to the downstream genes are represented by green arrows, while motifs without conserved orientation to the downstream genes are represented by brown arrows. CDSs are represented by red arrows, inverted repeats are represented by blue arrows.

Extended Data Table 1 | *Boodlea* genes having orthologs in the chloroplast in other Archaeplastida and their distribution over the sequenced DNA/RNA

	gDNA			RNA	
	Chloroplast 454 contigs	**HMW_corr**	**LMW_corr**	**mRNA**	**total-RNA**
16S	●		●	●	●
accD				●	●
acpP		●		●	●
*atpA**	●		●	●	●
*atpB**	●		●	●	●
atpD				●	●
*atpE**				●	●
*atpH**	●		●	●	●
*atpI**	●		●		●
ccsA		●		●	●
*cemA**		●		●	●
chlD		●		●	●
chlI				●	●
clpB		●		●	●
*clpP**		●		●	●
*ftsH**		●		●	●
glnB		●		●	●
groEL		●		●	●
infB		●		●	●
*petA**	●		●		●
*petB**	●		●	●	●
*petD**	●		●	●	●
petF				●	●
petJ		●		●	●
petN				●	●
*psaA**	●		●	●	●
*psaB**	●		●	●	●
*psaC**	●		●	●	●
psaD				●	●
psaE				●	●
psaF				●	●
psaK					●
psaL				●	●
*psbA**	●	●§	●	●	●
*psbB**	●	●§	●	●	●
*psbC**	●	●§	●	●	●
*psbD**	●		●	●	●
*psbE**	●		●		●
*psbF**	●		●		●
*psbH**				●	●
*psbI**				●	●
*psbJ**	●		●		●
*psbK**	●		●		
*psbL**	●		●		●
*psbN**		●		●	●

	gDNA			RNA	
	Chloroplast 454 contigs	**HMW_corr**	**LMW_corr**	**mRNA**	**total-RNA**
*psbT**	●		●		●
psbX				●	●
*rbcL**	●		●	●	●
rpl1		●		●	●
*rpl2**				●	●
rpl3				●	●
rpl4				●	●
*rpl5**		●		●	●
rpl6				●	●
rpl11				●	●
rpl12		●		●	●
rpl13				●	●
*rpl14**				●	●
*rpl16**				●	●
rpl19				●	●
*rpl20**				●	●
rpl21				●	●
rpl22				●	●
*rpl23**				●	●
rpl24				●	●
rpl33				●	●
rpl34					●
*rpoA**				●	●
*rpoB**		●		●	●
*rpoC1**				●	●
*rpoC2**				●	●
*rps3**				●	●
*rps4**				●	●
*rps7**				●	●
*rps8**				●	●
*rps9**				●	
*rps11**				●	●
*rps12**				●	●
*rps14**				●	●
rps16				●	●
rps17				●	●
*rps18**		●		●	●
*rps19**				●	
*tufA**				●	●
*ycf3**		●		●	●
*ycf4**		●		●	●
ycf20				●	●
ycf47				●	●
TOTAL	**22**	**22**	**22**	**78**	**85**

gDNA (genomic DNA): chloroplast 454 contigs, HMW and LMW corrected reads; RNA: mRNA and total-RNA assemblies.
*** "core" chloroplast genes:** protein-coding genes conserved between the chloroplast genomes of Chlorophyta. The following 9 "core" chloroplast genes were not found in any of the *Boodlea* libraries sequenced: *atpF*, *petG*, *petL*, *psaJ*, *psbM*, *psbZ*, *rpl36*, *rps2* and *ycf1*.
§ carry-over contaminants: LMW DNA reads contaminants as suggested by HMW to LMW DNA reads and mRNA to total-RNA reads ratios (Extended Data Figure 4).

Extended Data Table 2 | Features of the 34 contigs constituting the *Boodlea* chloroplast genome and unassembled LMW DNA reads with sequence similarity to chloroplast genes (a graphical representation of the contigs is available in Supplementary Note 2)

Contig	Containing gene	# LMW DNA reads	length (bp) [total; min; max; N50]	% coding	GC% [total; CDS; nc]	PBcR full-CDS	assembled with 454	congruent with 454	Group A	Group B	Group C	Group D	Group E	start	stop	features
ctg1	**16S rRNA 5'**	64	**1,885**; 510; 3,439; 939	29.1	**54.6;** 55.3; 54.5	●	○	●	0	13	1	51	0	-	-	Group B
ctg2	**16S rRNA 3'**	5	**2,600**;1,761; 1,874; 1,836	26.7	**56.3**; 55.6; 56.4	○	●	○	0	0	5	0	0	-	-	Group B
ctg3	**atpA**	36	**3,295**; 512; 3,488; 2,697	67.5	**59.6**; 60.3;56.8	○	●	○	0	0	22	14	0	ATG	TAG	Group B
ctg4	**atpB**	346	**1,843**; 516; 10,796; 922	75.0	**58.4**; 58.8; 57.3	○	●	○	0	0	31	280	35	GTT	TAA	Group E shorter CDS
ctg5	**atpH**	96	**2,686**; 585; 2,686; 1,287	9.0	**57.4**; 57.2; 57.4	●	○	●	0	2	0	43	51	GTG	TGA	Group B
ctg6*	**atpH**	-	**1,833**	12.6	**56.9**; 57.1; 56.9	-	-	-	0	0	0	0	1	GTG	TCC	lacks 3' ?
ctg7*	**atpH**	-	**2,149**	9.4	**58.2**; 57.1; 58.3	-	-	-	0	0	0	0	1	GTG	TCT	lacks 3' ?
ctg8*	**atpH**	-	**1,873**	12.3	**57.6**; 57.1; 57.7	-	-	-	0	0	0	0	1	GTG	TCC	lacks 3' ?
ctg9*	**atpH**	-	**1,949**	10.2	**58.2**; 60.1; 58.0	-	-	-	0	0	0	0	1	?	TGA	lacks 5'
ctg10	**atpI**	132	**6,139**; 535; 6,137; 605	26.7	**55.6**; 57.1; 55.1	●	○	●	1	0	0	131	0	ATG	TAG	Group A
ctg11	**petA**	16	**3,398**; 641; 3,396; 1,833	22.4	**57.3**; 59.5; 55.2	●	○	●	0	1	4	11	0	CTG	TAG	Group B TGA → V TGA → K
ctg12*	**petA**	-	**2,012**	36.2	**57.7**; 59.3; 56.7	-	-	-	0	0	0	0	1	?	TAG	lacks 5'
ctg13*	**petA**	-	**1,721**	46.2	**57.8**; 59.5; 55.7	-	-	-	0	0	0	0	1	ATG	TAG	TGA → V
ctg14	**petB**	8	**2,695**; 1,673; 1,914; 1,841	39.9	**55.2**; 56.3; 54.4	●	○	●	0	0	7	0	1	CTG	TAA	Group B
ctg15*	**petB**	-	**1,631**	40.5	**56.2**, 55.2, 56.9	-	-	-	0	0	0	0	1	CTG	TAA	
-	**petD**	9	754; 811; 809	-	**57.2**; -; -	○	○	○	0	0	0	9	0	ATG	TAA	Group D, unassembled
ctg16*	**petD**	-	**1,395**	35.1	**55.8**; 54.8; 56.3	-	-	-	0	0	0	0	1	ATG	TAA	
ctg17	**psaA**	250	**6,925**; 572; 6,596; 1,191	39.6	**57.9**; 59.0; 57.4	○	●	○	0	0	39	211	0	ATG	TAG	Group E, TGA → Q? TGA → V?
-	**psaB**	337	-; 523; 2,493; 776	-	**57.7**; -; -	○	○	○	0	0	8	329	0	ATG	TAA	unassembled shorter (5' 30 aa)
ctg18*	**psaB**	-	**2,116**	95.9	**58.2**; 58.5; 51.1	-	-	-	0	0	0	0	1	?	TAA	lacks 5' TGA → C/V
ctg19	**psaC** **psbJ**	92	**2,971**; 503; 2,346; 726	16.6	**57.7**; 55.2; 58.2	●	○	●	0	6	0	86	0	ATG TTG	TAG TGA	Group B TGA → C
ctg20*	**psaC**	-	**1,850**	12.5	**56.1**; 56.3; 56.1	-	-	-	0	0	0	0	1	?	?	longer 5'?
ctg21	**psbA**	136	**3,599**; 551; 2,636; 893	54.5	**53.7**; 52.5; 55.5	○	●	○	0	0	37	109	0	ATG	TAG	Group A

ctg22	**psbB**	152	**3,090**; 552; 4,273; 1,757	61.1	**60.4**; 60.4; 60.4	○	●	○	0	0	64	88	0	ATG	TAG	Group B
ctg23	**psbC**	187	**2,041**; 500; 3,404; 646	67.0	**59.3**; 59.3; 59.3	-	-	-	0	0	5	171	11	ATG	TAA	Group E ; TGA → L
ctg24	**psbD**	90	**3,643**; 558; 2753; 751	52.0	**56.3**; 56.7;56.0	●	○	●	0	0	3	78	4	ATG	TAA	Group B, lacks 5' ?
ctg25*	**psbD**	-	**1,960**	59.5	**56.5**; 56.4; 56.7	-	-	-	0	0	0	0	1	ATG	TAA	longer 5'
ctg26	**psbE** **psbK**	10	**1,327**; 689; 2,077; 1,405	16.9	**55.8**; 53.6; 56.2	●	○	●	0	0	0	7	2	? ?	? TAG	Group E; no 5', no stop codon
ctg27	**psbF**	5	**2,617**; 1,715; 1,726; 1,718	5.2	**56.8**; 54.8; 57.0	●	○	●	0	4	0	0	0	CTC	TAG	Group B, uncorrected
ctg28	**psbL**	1	**1,179**	8.4	**55.9**; 48.5; 56.6	●	○	●	0	0	0	0	1	ATG	TAG	Group B; uncorrected
ctg29*	**psbL**	-	**1,666**	6.3	**57.8**; 49.5; 58.5	-	-	-	0	0	0	0	1	ATG	TAG	
ctg30*	**psbL**	-	**1,986**	6.2	**56.9**; 50.4; 57.3	-	-	-	0	0	0	0	1	TTG	?	arbitrary
ctg31	**psbT**	70	**4,515**	1.6	**57.9**; 47.2; 58.0	●	○	●	0	9	0	42	2	GTG	?	Group B
ctg32	**psbT**	43	**3,687**	4.3	**57.2**; 53.5; 57.4	●	○	●	0	10	0	33	0	GTG	TAG	Group B
ctg33	**psbT**	33	**3,000**	6.0	**56.9**; 50.8; 57.3	●	○	●	0	12	0	21	0	CAT	TAA	Group B
ctg34	**rbcL**	411	**4,116**; 503; 3,120; 909	65.1	**57.3**; 57.8; 56.1	○	●	○	0	0	26	385	0	ATG	TAA	Group B, TGA → C

***** chloroplast 454 contigs that could not be assembled together with LMW DNA reads
● yes
○ no
- not applicable
length: in bold is indicated the length of the contig resulting from the orthology-guided assembly (chloroplast-enriched fraction + LMW DNA); minimum, maximum and N50 lengths in bp of the LMW DNA reads. With regards to the chloroplast 454 contigs that could not be assembled together with the LMW DNA reads, their lengths are also reported in bold.
% coding: percentage of coding sequence of the orthology-guided contig or of the chloroplast 454 contigs.
GC%: GC contents of orthology-guided/chloroplast 454 contigs, of the respective CDS and of the non-coding region.
PBcRfullCDS: whether respective LMW DNA reads harbor a full-length CDS.
assembled with 454: whether a full-length CDS could be reconstructed by orthology-guided assembly of LMW DNA reads and chloroplast 454 contigs.
congruent with 454: whether LMW DNA reads harboring full-length CDSs have corresponding chloroplast 454 contigs.
groupA-E: distribution of LMW DNA reads in different Groups. All unassembled chloroplast 454 contigs belong to Group E molecules.
start: the identified start codon of the CDS. A question mark indicates that the start codon could not be univocally identified.
stop: the identified stop codon of the CDS. A question mark indicates that the stop codon could not be univocally identified.
features: in this column are reported the Groups to which the orthology-guided contigs with full-length CDSs belong, the alternative codons identified if it was possible to assemble a full-length CDS, and the eventual differences with the orthologous sequences of other green algae.

References

1 Ochoa de Alda, J. A. G., Esteban, R., Diago, M. L. & Houmard, J. The plastid ancestor originated among one of the major cyanobacterial lineages. *Nat. Commun.* **5**, 4937, doi:10.1038/ncomms5937 (2014).

2 Lang, B. F. & Nedelcu, A. in *Genomics of chloroplasts and mitochondria* (eds R. Bock & V. Knoop) (Springer Netherlands, 2012).

3 Smith, D. R. & Keeling, P. J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci.* **112**, 10177-10184, doi:10.1073/pnas.1422049112 (2015).

4 Howe, C. J., Nisbet, R. E. & Barbrook, A. C. The remarkable chloroplast genome of dinoflagellates. *J. Exp. Bot.* **59**, 1035-1045, doi:10.1093/jxb/erm292 (2008).

5 Turmel, M., Otis, C. & Lemieux, C. Dynamic Evolution of the Chloroplast Genome in the Green Algal Classes Pedinophyceae and Trebouxiophyceae. *Genome Biol. Evol.* **7**, 2062-2082, doi:10.1093/gbe/evv130 (2015).

6 Leliaert, F. *et al.* Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci. Rep.* **6**, 25367, doi:10.1038/srep25367 (2016).

7 Turmel, M., de Cambiaire, J.-C., Otis, C. & Lemieux, C. Distinctive architecture of the chloroplast genome in the chlorodendrophycean green algae *Scherffelia dubia* and *Tetraselmis* sp. CCMP 881. *PLoS One* **11**, e0148934, doi:10.1371/journal.pone.0148934 (2016).

8 Lemieux, C., Otis, C. & Turmel, M. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front. Plant Sci.* **7**, 697, doi:10.3389/fpls.2016.00697 (2016).

9 Fučíková, K. *et al.* New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Front. Ecol. Evol.* **2**, 63 (2014).

10 Deng, Y., Zhan, Z., Tang, X., Ding, L. & Duan, D. Molecular cloning and expression analysis of RbcL cDNA from the bloom-forming green alga *Chaetomorpha valida* (Cladophorales, Chlorophyta). *J. Appl. Phycol.* **26**, 1853-1861, doi:10.1007/s10811-013-0208-z (2014).

11 La Claire, J. W., Zuccarello, G. C. & Tong, S. Abundant plasmid-like DNA in various members of the orders Siphonocladales and Cladophorales (Chlorophyta). *J. Phycol.* **33**, 830-837, doi:10.1111/j.0022-3646.1997.00830.x (1997).

12 La Claire, J. W. & Wang, J. Localization of plasmidlike DNA in giant-celled marine green algae. *Protoplasma* **213**, 157-164, doi:10.1007/BF01282153 (2000).

13 La Claire, J. W., Loudenslager, C. M. & Zuccarello, G. C. Characterization of novel extrachromosomal DNA from giant-celled marine green algae. *Curr. Genet.* **34**, 204-211 (1998).

14 La Claire, J. W. & Wang, J. Structural characterization of the terminal domains fo linear plasmid-like DNA from the green alga *Ernodesmis* (Chlorophyta). *J. Phycol.* **40**, 1089-1097, doi:10.1111/j.1529-8817.2004.04100.x (2004).

15 Zahonova, K., Kostygov, A. Y., Sevcikova, T., Yurchenko, V. & Elias, M. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr. Biol.* **26**, 1879-0445 (2016).

16 Heaphy, S. M., Mariotti, M., Gladyshev, V. N., Atkins, J. F. & Baranov, P. V. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. *Mol. Biol. Evol.* **33**, 1537-1719 (2016).

17 Swart, Estienne C., Serra, V., Petroni, G. & Nowacki, M. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. *Cell* **166**, 691-702, doi:10.1016/j.cell.2016.06.020 (2016).

18 Cocquyt, E. *et al.* Complex phylogenetic distribution of a non-canonical genetic code in green algae. *BMC Evol. Biol.* **10**, 327, doi:10.1186/1471-2148-10-327 (2010).
19 Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327, doi:10.1016/j.ygeno.2010.03.001 (2010).
20 Okimoto, R., Macfarlane, J. L. & Wolstenholme, D. R. The mitochondrial ribosomal RNA genes of the nematodes *Caenorhabditis elegans* and *Ascaris suum*: consensus secondary-structure models and conserved nucleotide sets for phylogenetic analysis. *J. Mol. Evol.* **39**, 598-613, doi:10.1007/BF00160405 (1994).
21 Barbrook, A. C., Santucci, N., Plenderleith, L. J., Hiller, R. G. & Howe, C. J. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. *BMC Genomics* **7**, 297, doi:10.1186/1471-2164-7-297 (2006).
22 Pantou, M. P., Kouvelis, V. N. & Typas, M. A. The complete mitochondrial genome of *Fusarium oxysporum*: insights into fungal mitochondrial evolution. *Gene* **419**, 7-15 (2008).
23 Walther, T. C. & Kennell, J. C. Linear mitochondrial plasmids of *F. oxysporum* are novel, telomere-like retroelements. *Mol. Cell* **4**, 229-238, doi:10.1016/S1097-2765(00)80370-6 (1999).
24 Janouskovec, J. *et al.* Split photosystem protein, linear-mapping topology, and growth of structural complexity in the plastid genome of *Chromera velia*. *Mol. Biol. Evol.* **30**, 2447-2462 (2013).
25 Watanabe, S., Fučíková, K., Lewis, L. A. & Lewis, P. O. Hiding in plain sight: *Koshicola spirodelophila* gen. et sp. nov. (Chaetopeltidales, Chlorophyceae), a novel green alga associated with the aquatic angiosperm *Spirodela polyrhiza*. *Am. J. Bot.* **103**, 865-875 (2016).

References Extended Material

26 Andersen, R. A. *Algal culturing techniques*. (Academic press, 2005).
27 Palmer, J. D. Physical and gene mapping of chloroplast DNA from *Atriplex triangularis* and *Cucumis sativa*. *Nucleic Acids Res.* **10**, 1593-1605, doi:10.1093/nar/10.5.1593 (1982).
28 Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bull.* **19**, 11-15, doi:citeulike-article-id:678648 (1987).
29 Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, doi:10.1101/gr.1917404 (2004).
30 Boratyn, G. M. *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29-W33, doi:10.1093/nar/gkt282 (2013).
31 Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 24175, doi:10.1038/srep24175 (2016).
32 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935, doi:10.1093/bioinformatics/btt509 (2013).
33 Ruby, J. G., Bellare, P. & Derisi, J. L. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* **3**, 865-880, doi:10.1534/g3.113.005967 (2013).
34 Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277, doi:http://dx.doi.org/10.1016/S0168-9525(00)02024-2 (2000).
35 Noé, L. & Kucherov, G. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **33**, W540-W543, doi:10.1093/nar/gki478 (2005).
36 Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30**, doi:10.1038/nbt.2288 (2012).

37 Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 31900, doi:10.1038/srep31900 (2016).
38 Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-3011, doi:10.1093/bioinformatics/btu392 (2014).
39 Vandepoele, K. *et al.* pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.* **15**, 2147-2153, doi:10.1111/1462-2920.12174 (2013).
40 Koren, S. *et al.* Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, 1-16, doi:10.1186/gb-2013-14-9-r101 (2013).
41 Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623-630, doi:10.1038/nbt.3238 (2015).
42 Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945, doi:10.1093/bioinformatics/16.10.944 (2000).
43 Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-W208 (2009).
44 Medina-Rivera, A. *et al.* RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.* **43** (2015).
45 Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142-147 (2013).
46 Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24, doi:10.1186/gb-2007-8-2-r24 (2007).
47 Wu, T. D., Reeder, J., Lawrence, M., Becker, G. & Brauer, M. J. in *Statistical Genomics: Methods and Protocols* (eds Ewy Mathé & Sean Davis) 283-334 (Springer New York, 2016).
48 Le Bail, A. *et al.* Normalisation genes for expression analyses in the brown alga model *Ectocarpus siliculosus*. *BMC Mol. Biol.* **9**, 1-9, doi:10.1186/1471-2199-9-75 (2008).
49 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652, doi:10.1038/nbt.1883 (2011).
50 Veeckman, E., Ruttink, T. & Vandepoele, K. Are we there yet? Reliably estimating the completeness of plant genome sequences. *The Plant Cell* **28**, 1759-1768 (2016).
51 Leliaert, F. & Lopez-Bautista, J. M. The chloroplast genomes of *Bryopsis plumosa* and *Tydemania expeditiones* (Bryopsidales, Chlorophyta): compact genomes and genes of bacterial origin. *BMC Genomics* **16**, 1-20, doi:10.1186/s12864-015-1418-3 (2015).
52 Melton, J. T., Leliaert, F., Tronholm, A. & Lopez-Bautista, J. M. The complete chloroplast and mitochondrial genomes of the green macroalga *Ulva* sp. UNA00071828 (Ulvophyceae, Chlorophyta). *PLoS One* **10**, doi:10.1371/journal.pone.0121020 (2015).
53 Maly, P. & Brimacombe, R. Refined secondary structure models for the 16S and 23S ribosomal RNA of Escherichia coli. *Nucleic Acids Res.* **11**, 7263-7286 (1983).