bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

One Sentence Summary: Chloroplast genome in Cladophorales green algae is reduced and fragmented into multiple linear single-stranded DNA molecules.

Authors: 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]

Affiliations:

[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.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[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.

*Correspondence to: andrea.delcortona@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract:

Chloroplast genomes, relics of an endosymbiotic cyanobacterial genome, are circular double-stranded DNA molecules. While fragmented mitochondrial genomes evolved several times during the evolution of eukaryotes, fragmented plastid genomes are only known in dinoflagellates. Here we show that the chloroplast genome of the green alga *Boodlea composita* (Cladophorales) is reduced and fragmented into hairpin plasmids. Extensive sequencing of DNA and RNA demonstrated that the chloroplast genome is fragmented into 1-7 kb, GC-rich DNA contigs, each containing a long inverted repeat with protein-coding genes and conserved non-coding region. These contigs correspond to linear single-stranded DNA molecules that fold onto themselves to form hairpin plasmids. An elevated transfer of chloroplast genes to the nucleus coincided to *Boodlea* chloroplast genome reduction. The genes retained in the chloroplast are highly divergent from their corresponding orthologs. A chloroplast genome that is composed only of linear DNA molecules is unprecedented among eukaryotes.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Main Text:

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 (Fig. 1, A to C). Currently, and in stark contrast to other algae (*1-4*), 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 (*5, 6*). Cladophorales contain abundant plasmids within chloroplasts (*7, 8*), representing a Low Molecular Weight (LMW) DNA fraction (Fig. S1). 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 (*7-10*). Some of these hairpin plasmids contain putatively transcribed sequences with similarity to chloroplast genes (*psaB*, *psbB*, *psbC* and *psbF*) (*9*).

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 (Fig. S2 to S4 and Tables S1 and S2). 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 (Fig. S5E and S6) and, despite high coverage by sequence reads, they could not be extended by iterative contig extension (Fig. S7 and S10 to S15). The inverted repeats were also found on contigs with

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

no sequence similarity to known proteins, raising the number of contigs of chloroplast origin to 136 (Table S2). 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 (Fig. 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 (Tables S3 and S4). 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 (Fig. 1E and S8). 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*; Fig. S9), next to still acting as stop codon. Dual meaning of UGA as both stop and sense codons has recently been reported from a number of unrelated protists (*11-13*). Interestingly, a different non-canonical genetic code has been described for Cladophorales nuclear genes, where UAG and UAA codons are reassigned to glutamine (*14*), 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 (Tables S5 to S5). 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 (Fig. S6 and S10 to S15). 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 (Fig. S6, S7 and S10 to S15). Moreover, complete congruence between RNA and DNA sequences excluded the presence of RNA editing (Fig. S10 to S15).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Additional transcripts of 66 genes that have been located in the chloroplast in other Archaeplastida were identified (Table S8). 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 (*15*). 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 (Fig. S2 and S3). 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. 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 (Table S8). 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 (Fig. S15 and Table S8).

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 (Table S8). An orthology-guided assembly of chloroplast 454 contigs and the LMW DNA reads resulted in 34 contigs between 1,179 and 6,925 bp in length, henceforth referred to as “chloroplast genome” (Table S2 and S9 and Fig. S17). Several of these contigs are long near-identical palindromic sequences, including full-length coding sequences (CDSs), and a somewhat less conserved tail region (Fig. 1, F to H and S5A). The remaining contigs appear incomplete but are often indicative of the presence of similar palindromic structures (Fig. S5). Such palindromes allow regions of the

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (*7*). These hairpin plasmids are directly evidenced in several genes by long LMW DNA reads and are consistent with the structure of all chloroplast contigs (Table S9). The 16S rRNA gene was split across two distinct hairpin plasmids and much reduced compared to algal and bacterial homologs (Fig. S18), similar to what is found in the chloroplast genomes of peridinin-pigmented dinoflagellates and nematode mitochondrial genomes (*16, 17*). 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 (Fig. S19), 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 (*7, 9*) 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 (*18*) (Fig. S5).

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, Fig. S20). Some RT-LTRs were also abundant in the 454 contigs (Fig. S21). These

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 (*19*, *20*).

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 (*18*). Indeed, the two distant 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 (Table S8). Other non-canonical chloroplast genome architectures have recently been observed, such as a monomeric linear chromosome in the alveolate microalga *Chromera velia* (*21*) and three circular chromosomes in the green alga *Koshicola spirodelophila* (*22*), but these

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References and Notes:

1. M. Turmel, C. Otis, C. Lemieux, Dynamic Evolution of the Chloroplast Genome in the Green Algal Classes Pedinophyceae and Trebouxiophyceae. *Genome Biol. Evol.* **7**, 2062-2082 (2015).

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

3. M. Turmel, J.-C. de Cambiaire, C. Otis, C. Lemieux, Distinctive architecture of the chloroplast genome in the chlorodendrophycean green algae *Scherffelia dubia* and *Tetraselmis* sp. CCMP 881. *PLoS One* **11**, e0148934 (2016).

4. C. Lemieux, C. Otis, M. Turmel, Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front. Plant Sci.* **7**, 697 (2016).

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

6. Y. Deng, Z. Zhan, X. Tang, L. Ding, D. Duan, Molecular cloning and expression analysis of RbcL cDNA from the bloom-forming green alga *Chaetomorpha valida* (Cladophorales, Chlorophyta). *J. Appl. Phycol.* **26**, 1853-1861 (2014).

7. J. W. La Claire, G. C. Zuccarello, S. Tong, Abundant plasmid-like DNA in various members of the orders Siphonocladales and Cladophorales (Chlorophyta). *J. Phycol.* **33**, 830-837 (1997).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

8. J. W. La Claire, J. Wang, Localization of plasmidlike DNA in giant-celled marine green algae. *Protoplasma* **213**, 157-164 (2000).

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

10. J. W. La Claire, J. Wang, Structural characterization of the terminal domains fo linear plasmid-like DNA from the green alga *Ernodesmis* (Chlorophyta). *J. Phycol.* **40**, 1089-1097 (2004).

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

12. S. M. Heaphy, M. Mariotti, V. N. Gladyshev, J. F. Atkins, P. V. Baranov, 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).

13. Estienne C. Swart, V. Serra, G. Petroni, M. Nowacki, Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. *Cell* **166**, 691-702 (2016).

14. E. Cocquyt *et al.*, Complex phylogenetic distribution of a non-canonical genetic code in green algae. *BMC Evol. Biol.* **10**, 327 (2010).

15. J. R. Miller, S. Koren, G. Sutton, Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327 (2010).

16. R. Okimoto, J. L. Macfarlane, D. R. Wolstenholme, The mitochondrial ribosomal RNA genes of the nematodes *Caenorhabditis elegans* and *Ascaris suum*: consensus secondary-

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

structure models and conserved nucleotide sets for phylogenetic analysis. *J. Mol. Evol.* **39**, 598-613 (1994).

17. A. C. Barbrook, N. Santucci, L. J. Plenderleith, R. G. Hiller, C. J. Howe, Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. *BMC Genomics* **7**, 297 (2006).

18. C. J. Howe, R. E. Nisbet, A. C. Barbrook, The remarkable chloroplast genome of dinoflagellates. *J. Exp. Bot.* **59**, 1035-1045 (2008).

19. M. P. Pantou, V. N. Kouvelis, M. A. Typas, The complete mitochondrial genome of *Fusarium oxysporum*: insights into fungal mitochondrial evolution. *Gene* **419**, 7-15 (2008).

20. T. C. Walther, J. C. Kennell, Linear mitochondrial plasmids of *F. oxysporum* are novel, telomere-like retroelements. *Mol. Cell* **4**, 229-238 (1999).

21. J. Janouskovec *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).

22. S. Watanabe, K. Fučíková, L. A. Lewis, P. O. Lewis, 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).

23. R. A. Andersen, *Algal culturing techniques*. (Academic press, 2005).

24. J. D. Palmer, Physical and gene mapping of chloroplast DNA from *Atriplex triangularis* and *Cucumis sativa*. *Nucleic Acids Res.* **10**, 1593-1605 (1982).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

25. J. J. Doyle, J. L. Doyle, A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bull.* **19**, 11-15 (1987).

26. B. Chevreux *et al.*, Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, (2004).

27. G. M. Boratyn *et al.*, BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29-W33 (2013).

28. H.-H. Lin, Y.-C. Liao, Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci. Rep.* **6**, 24175 (2016).

29. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

30. J. G. Ruby, P. Bellare, J. L. Derisi, PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* **3**, 865-880 (2013).

31. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

32. L. Noé, G. Kucherov, YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **33**, W540-W543 (2005).

33. A. Bashir *et al.*, A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.* **30**, (2012).

34. C. Ye, C. M. Hill, S. Wu, J. Ruan, Z. Ma, DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* **6**, 31900 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

35. T. Hackl, R. Hedrich, J. Schultz, F. Förster, proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-3011 (2014).

36. K. Vandepoele *et al.*, pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ. Microbiol.* **15**, 2147-2153 (2013).

37. S. Koren *et al.*, Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biol.* **14**, 1-16 (2013).

38. K. Berlin *et al.*, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623-630 (2015).

39. K. Rutherford *et al.*, Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944-945 (2000).

40. T. L. Bailey *et al.*, MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-W208 (2009).

41. A. Medina-Rivera *et al.*, RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res.* **43**, (2015).

42. A. Mathelier *et al.*, JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142-147 (2013).

43. S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, W. S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

44. T. D. Wu, J. Reeder, M. Lawrence, G. Becker, M. J. Brauer, in *Statistical Genomics: Methods and Protocols*, E. Mathé, S. Davis, Eds. (Springer New York, New York, NY, 2016), pp. 283-334.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. A. Le Bail *et al.*, Normalisation genes for expression analyses in the brown alga model *Ectocarpus siliculosus*. *BMC Mol. Biol.* **9**, 1-9 (2008).

46. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652 (2011).

47. E. Veeckman, T. Ruttink, K. Vandepoele, Are we there yet? Reliably estimating the completeness of plant genome sequences. *Plant Cell* **28**, 1759-1768 (2016).

48. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455-477 (2012).

49. K. Tamura *et al.*, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, (2011).

50. G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, (2007).

51. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, (2014).

52. M. A. Miller, W. Pfeiffer, T. Schwartz, paper presented at the Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery, Salt Lake City, Utah, 2011.

53. P. Maly, R. Brimacombe, Refined secondary structure models for the 16S and 23S ribosomal RNA of Escherichia coli. *Nucleic Acids Res.* **11**, 7263-7286 (1983).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments: 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 ***–***. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.D.C. (andrea.delcortona@gmail.com). 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.

Supplementary Materials:

Materials and Methods

Supplementary Text

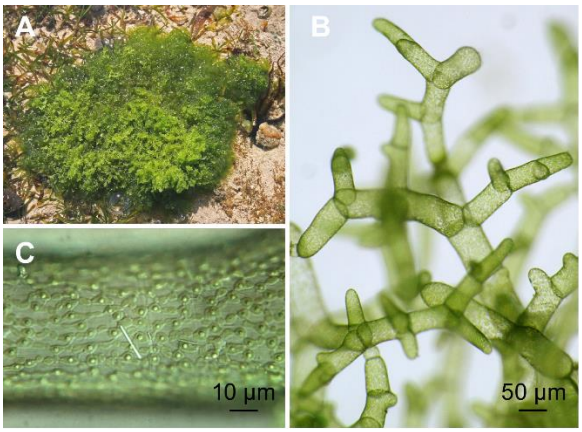
Figures S1-S21

Tables S1-S9

References (*23-53*)

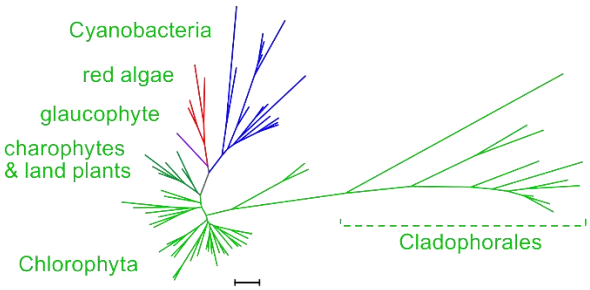
bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

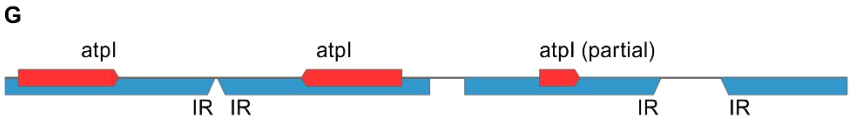
1



D

	Cladoph.	Viridipl.*	Archaepl.*	Cyanobact.
Cladophorales	0.465			
Viridiplantae*	0.506	0.238		
Archaeplastida*	0.518	0.282	0.282	
Cyanobacteria	0.547	0.357	0.359	0.376





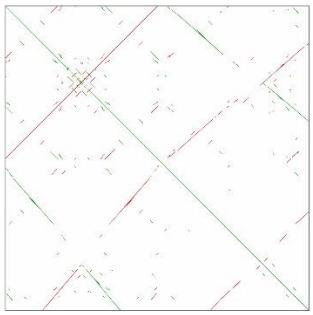


Fig. 1. *Boodlea composita* and its chloroplast hairpin plasmids. (**A**) Specimen in natural environment. (**B**) Detail of branching cells. (**C**) Detail of chloroplasts dotted with pyrenoids and forming a parietal network. (**D**) Maximum amino acid pairwise sequence distance estimated between the chloroplast genes of the taxa in the phylogenetic tree shown in Fig. 1E (* excluded Cladophorales). (**E**) Maximum likelihood phylogenetic tree inferred from a concatenated amino acid alignment of 19 chloroplast genes. Taxon names and bootstrap support values are provided in the tree shown in Fig. S8. 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 Fig. S5). Red arrows represent CDSs, blue arrows represent major inverted repeats. (**H**) Dotplot illustrating the complexity of repetitive sequences.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

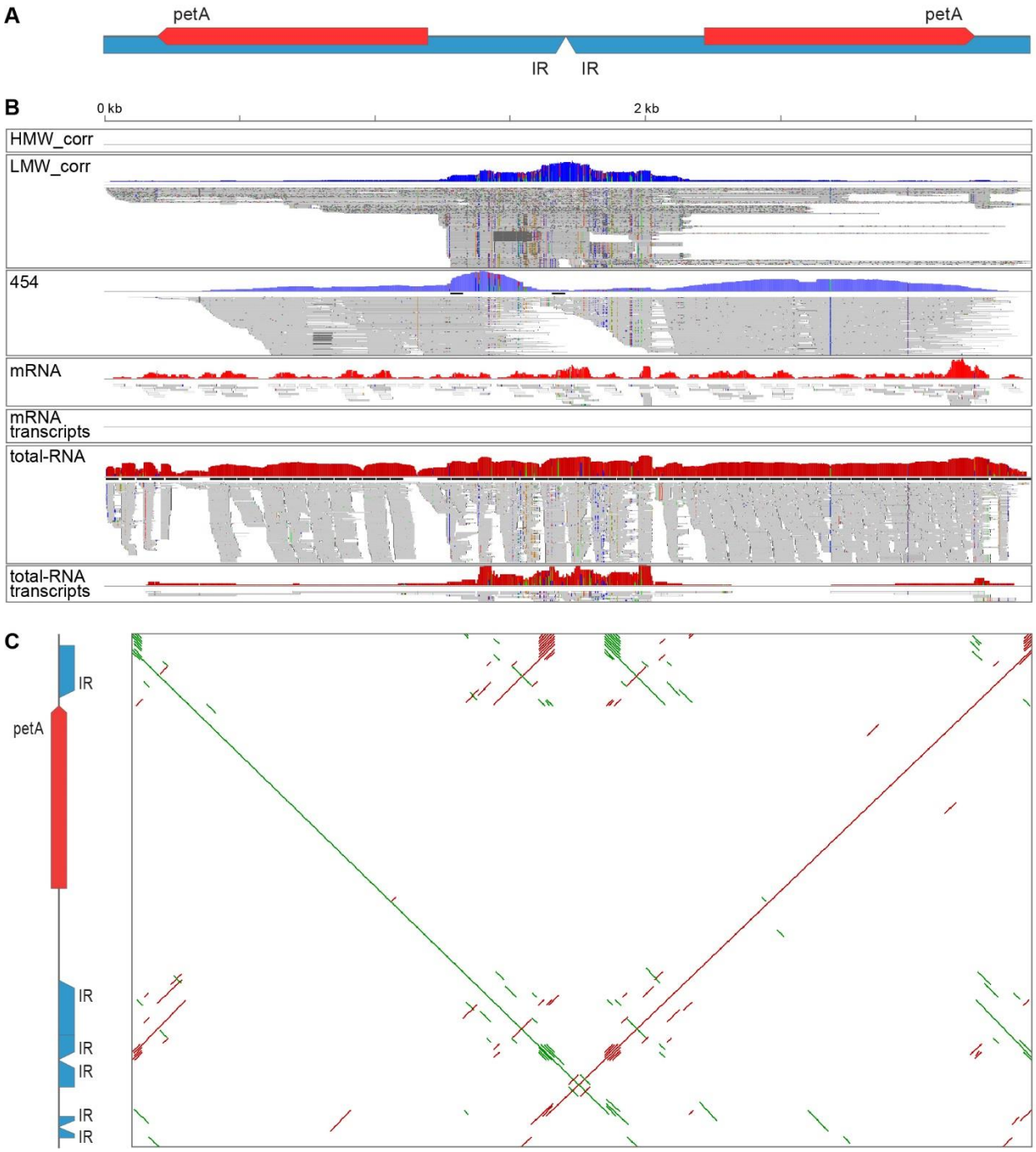


Fig. 2. LMW DNA reads containing chloroplast genes are expressed, enriched in the total-RNA fraction and congruent to the respective chloroplast 454 contigs. (**A**) Representation of *petA* LMW DNA read (3,398 bp), a representative of group B contigs and reads (see Fig.S5). The

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted June 8, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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.