bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder.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 plastid genome in Cladophorales green algae is encoded by hairpin plasmids

Andrea Del Cortona[1,2,3,4,a], Frederik Leliaert[1,5,a], 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, 1860 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

[a] These authors contributed equally to this work

[*] Correspondence and requests for materials should be addressed to O.D.C. (olivier.declerck@ugent.be).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder.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

Virtually all plastid (chloroplast) genomes are circular double-stranded DNA molecules, typically between 100-200 kb in size and encoding circa 80-250 genes. Exceptions to this universal plastid genome architecture are very few and include the dinoflagellates where genes are located on DNA minicircles. Here we report on the highly deviant chloroplast genome of Cladophorales green algae, which is entirely fragmented into hairpin plasmids. Short and long read high-throughput sequencing of DNA and RNA demonstrated that the chloroplast genes of *Boodlea composita* are encoded on 1-7 kb DNA contigs with an exceptionally high GC-content, 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. The origin of this highly deviant chloroplast genome likely occurred before the emergence of the Cladophorales, and coincided with an elevated transfer of chloroplast genes to the nucleus. A chloroplast genome that is composed only of linear DNA molecules is unprecedented among eukaryotes and highlights unexpected variation in the plastid genome architecture.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

Introduction

Photosynthetic eukaryotes originated more than 1.4 billion years ago following an endosymbiotic event in which a heterotrophic ancestor of the Archaeplastida engulfed a cyanobacterium that became stably integrated and evolved into a membrane-bound organelle, the plastid[1, 2]. Following this primary endosymbiosis, an intricate history of plastid acquisition via eukaryote-eukaryote endosymbioses resulted in the spread of plastids to distantly related eukaryotic lineages[3, 4, 5].

Plastids have retained a reduced version of the genome inherited from their cyanobacterial ancestor. A core set of genes involved in the light reactions of photosynthesis, ATP generation, and functions related to transcription and translation is typically retained [6, 7]. Many genes have been lost or transferred to the nuclear genome and, as a result, plastids are dependent on nuclear-encoded, plastid-targeted proteins for the maintenance of essential biochemical pathways and other functions such as genome replication, gene expression, and DNA repair[8]. Despite the wide diversity in size, gene content, density and organisation of plastid genomes among different eukaryotic lineages[9, 10, 11], nearly all plastid genomes consist of a single circular-mapping chromosome, typically between 100-200 kb and encoding circa 80-250 genes[7, 12].

While fragmented mitochondrial genomes evolved several times independently during the evolution of eukaryotes[11, 13], fragmented plastid genomes are only known in dinoflagellates[14] and a single green algal species[15]. In peridinin-containing dinoflagellates, the chloroplast genome is fragmented into plasmid-like DNA minicircles of 2-3 kb, most of which carry one gene only[14, 16]. Larger minicircles of up to 12 kb have also been described[17], as well as minicircles containing two genes[18], and 'empty' minicircles without genes[19, 20]. The genes located on these minicircles mostly encode key components of the major photosynthetic complexes, including subunits of photosystems I and II, the cytochrome b6f complex, and ATP synthase, as well as rRNAs and a few tRNAs[14]. The only other alga with a fragmented chloroplast genome is the green alga *Koshicola spirodelophila*, but here the level of fragmentation is minor: the plastid genome is divided into three large circular chromosomes totalling 385 kb, with a gene content comparable to other green algae[15]. In addition, plastid minicircles that coexist with a conventional plastid genome have been observed in a few algae, including dinoflagellates with haptophyte-derived plastids[21] and the green alga *Acetabularia*[22, 23].

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

Although plastid genomes generally assemble as circular-mapping DNAs, they can take multiple complex conformations *in vivo*, including multigenomic, linear-branched structures with discrete termini[24, 25]. The alveolate *Chromera velia* is the only known alga with a linear-mapping plastid genome with telomeric arrangement[26], and is also atypical in that several core photosynthesis genes are fragmented. Linear plastid genomes, however, may be more widespread as several plastid genomes currently do not map as a circle[27].

Currently, and in stark contrast to other green algae[9, 28, 29, 30, 31], little is known about the gene content and structure of the chloroplast genome in the Cladophorales (Ulvophyceae), 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. 1a-c). Most attempts to amplify common chloroplast genes have failed[32], with only one highly divergent *rbcL* sequence published thus far for *Chaetomorpha valida*[33]. An atypical plastid genome in the Cladophorales is suggested by the presence of abundant plasmids that have been observed in the chloroplasts of several species[34, 35]. These plasmids represent a Low Molecular Weight (LMW) DNA fraction, visible on agarose gels of total DNA extracts (Fig. 1d). Pioneering work revealed that these plasmids are single-stranded DNA (ssDNA) molecules of 1.5-3.0 kb that fold in a hairpin configuration and lack sequence similarity to the nuclear DNA[36, 37]. Some of the hairpin plasmids contain putatively transcribed sequences with similarity to chloroplast genes encoding subunits of Photosystems I and II (*psaB*, *psbB*, *psbC* and *psbF*)[36].

Here, we describe intriguing features of the plastid genome of Cladophorales, focusing on *Boodlea composita*. Through the integration of different DNA sequencing methods, combined with RNA sequencing, we found that chloroplast protein-coding genes are highly expressed and are encoded on 1-7 kb linear single-stranded DNA molecules. Due to the wide-spread presence of inverted repeats, these molecules fold into a hairpin configuration. A chloroplast genome that is composed only of linear DNA molecules is unprecedented among eukaryotes and highlights unexpected variation in plastid genome architecture.

Results and Discussion

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

DNA and RNAseq data

Our reconstruction of the chloroplast genome of *Boodlea composita* is based on different high-throughput DNA sequencing methods (Supplementary Fig. 1, Materials and Methods). The choice of short read DNA sequencing of a chloroplast-enriched fraction using Roche 454 technology was based on comparable sequencing approaches in other plants and algae that successfully resulted in assembly of chloroplast genomes [30, 38]. To overcome possible assembly artefacts in a hypothetical scenario of an inflated chloroplast genome bloated by repetitive elements, long-read sequencing of the High Molecular Weight (HMW) DNA fraction using Pacific Biosciences Single-Molecule Real-Time (SMRT) method was applied, while long read sequencing of the LMW DNA fraction allowed characterization of the previously observed plasmids in the chloroplast[34, 35]. To allow comparison of the results of *Boodlea* with other species of Cladophorales, we generated additional DNA sequence data from nine other species using Illumina HiSeq 2000 technology. Finally, two deep-coverage RNA-seq libraries, a total-RNA library and a mRNA library enriched for nuclear transcripts, were generated to confirm the transcription of genes, and to inform whether genes are nuclear versus plastid encoded.

A prodigious chloroplast genome with reduced gene set

Assembly of the chloroplast-enriched DNA reads generated using Roche 454 technology did not result in a typical circular chloroplast genome. Instead, 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 code for components of the major thylakoid transmembrane protein complexes (ATP synthase, cytochrome b6f, and photosystems I and II). The contigs contained inverted repeats at their termini (Fig. 2, group E contig) and, despite high coverage by sequence reads, they could not be extended by iterative contig extension. Sequence similarity searches and a metagenomic binning approach (distribution analysis of 4-mers) demonstrated that 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. These contigs are further referred to as “chloroplast 454 contigs”. The length distribution of the chloroplast 454 contigs was consistent with the size of the LMW DNA fraction as estimated by agarose gel electrophoresis of *Boodlea* genomic DNA (Fig. 1d, Supplementary Fig. 14).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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 failure to assemble a circular chloroplast genome might be due to repetitive elements that impair the performance of short-read assemblers[39]. Inflated chloroplast genome bloated by repetitive elements have been documented in several green algae [40, 41, 42]. 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. Only 22 HMW DNA reads (ca. 0.044 %) harboured protein-coding genes commonly present in chloroplast genomes of Archaeplastida (Fig. 3). All, but three of these genes (which likely correspond to carry-over LMW DNA), contained introns, were absent in the chloroplast 454 contigs, and revealed a high ratio between mapped mRNA and total-RNA reads, altogether suggesting that they are encoded in the nucleus (Supplementary Fig. 2). 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 (Fig. 3). An orthology-guided assembly, where the chloroplast 454 contigs harbouring protein-coding genes guided the assembly of LMW DNA reads with sequence similarity to chloroplast genes, resulted in 34 contigs between 1,179 and 6,925 bp in length, henceforth referred to as "chloroplast genome" (Fig. 4, Supplementary Table 1).

Several contigs of the *Boodlea* chloroplast genome display long imperfect palindromic sequences that include full-length coding sequences (CDSs), and a less conserved tail region (Fig. 2). The remaining contigs have similar palindromic structures but appear to be not completely assembled. Such palindromes allow regions of the single-stranded LMW DNA molecules to fold into hairpin-like secondary structures. Additional smaller inverted repeats were identified in many of the contigs, which may result in more complex secondary structures (Fig. 4).

Chloroplast 454 contigs could not be scaffolded with long HMW DNA reads, nor did an hybrid assembly between chloroplast 454 contigs and long HMW DNA reads generate a circular chloroplast genome (Supplementary Figs. S1, Materials and Methods). The LMW DNA reads are concordant and consistent with the imperfect palindromic sequences of the assembled chloroplast genome, indicating that the palindromes are not a result of assembly artefacts (Fig. 4, Supplementary Table 1). As a consequence, we conclude that the chloroplast genome is not a single large molecule but it is instead fragmented in several molecules in the LMW DNA.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

A chloroplast genome that is entirely fragmented into plasmids is in line with earlier observations of abundant LMW DNA in chloroplasts of several species of Cladophorales[34]. The hairpin configuration of the plasmids derived from our sequence data corresponds with earlier data based on electron microscopy, endo- and exonuclease digestion experiments, acridine orange staining, and denaturing gel electrophoresis[34, 37]. Fluorescence *in situ* hybridization, and Southern blot hybridisation indicated that these plasmids were present within the chloroplast only[35], supporting the congruence between chloroplast 454 contigs and sequences from the LMW fraction (Fig. 3).

The chloroplast genome contigs of *Boodlea* feature an exceptionally high GC-content, ranging from 54 to 60 % in the gene-containing contigs (average 57%) (Supplementary Table 1). These values are concordant with the high density of the LMW fraction observed in CsCl/ bisbenzimide gradients[34], and also with sequence data from cloned plasmids of *Ernodesmis* (51-59% GC)[36]. Plastid genomes are generally AT-rich, and in green algal species GC-content typically ranges between 26 and 43%[12, 43]. GC-rich plastid genomes are very rare, but higher values have been reported for the trebouxiophycean green algae *Coccomyxa subellipsoidea*, *Paradoxia multiseta* (both 51% GC), and Trebouxiophyceae sp. MX-AZ01 (58%)[29]. These species, however, feature standard plastid genome architectures.

The size of the *Boodlea* chloroplast genome could not be estimated by inspection of k-mer frequency distributions of the reads in the 454 library nor from those of the uncorrected and corrected LMW DNA reads[44]. Histograms of k-mer frequency distributions revealed several small peaks, indicating a heterogeneous population of molecules present in different stoichiometries, and the signal to noise ratio was too small to make a comfortable estimation of the sizes (Supplementary Fig. 3). The cumulative length of the 34 *Boodlea* chloroplast genome contigs is 91 kb (Supplementary table 3). However, if we would consider the large and heterogeneous population of LMW DNA reads bearing no similarity to protein-coding genes ("empty" plasmids, see below) as part of the chloroplast genome, its size could be regarded as much larger.

The largest known circular-mapping chloroplast genomes have been documented for two clades of green algae, the Chlorophyceae and the Ulvophyceae. Inflation of the 521-kb chloroplast genome of *Floydiella terrestris* (Chlorophyceae) resulted mainly from the proliferation of dispersed, heterogeneous repeats (>30 bp) in intergenic regions, representing

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

more than half of the genome length[42]. Intergenic regions of the *Volvox carteri* (Chlorophyceae) chloroplast genome, instead, are populated with short palindromic repeats (average size of 50 bp) that constitute ca. 64% of the predicted 525-kb genome[40]. The mechanisms by which such palindromic selfish DNA spread throughout the *Volvox* chloroplast genome is not clear, but the presence of a reverse transcriptase and endonuclease may point toward retrotranscription[40, 45]. For *Acetabularia acetabulum* (Ulvophyceae), the chloroplast genome was sequenced only partially and its size was estimated to exceed 1 Mb; it has exceptionally long intergenic regions and features long repetitive elements (>10 kb) arranged in tandem[41, 46]. The *Boodlea* chloroplast genome is rich as well in non-coding DNA, constituting 92.2% of the 136 chloroplast 454 contigs and 72.8% of the assembled chloroplast genome, comparable to that in inflated chloroplast genomes of other green algae (*Floydiella terrestris,* 82.1%; *Volvox carteri,* ca. 80%; *Acetabularia acetabulum,* ca. 87% of the sequenced chloroplast genome)[40, 41].

The non-coding DNA regions (ncDNA) of the hairpins showed high sequence similarity among one another. Within the ncDNA, we identified six conserved motifs, 20 to 35 bp in length, which lack similarity to known regulatory elements (Fig. 6). 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. These motifs were also present in 1,966 (ca. 1.8 %) LMW DNA reads lacking genes. This observation supports earlier findings of abundant non-coding LMW DNA molecules in the Cladophorales[34, 36]. In contrast, a very small fraction of the HMW DNA reads (15 corrected reads) displayed the same ncDNA motifs and these were present exclusively on long terminal repeat retrotransposons (RT-LTRs) (Supplementary Fig. 4, Supplementary Note 1). RT-LTRs were also abundant in the 454 contigs (Supplementary Fig. S16). The abundance of RT-LTRs in the 454 contigs and the presence of ncDNA motifs in both the *Boodlea* chloroplast genome and nuclear RT-LTRs is suggestive of DNA transfer between the nucleus and chloroplast and may allude to the origin of the hairpin plasmids. Hypothetically, an invasion of nuclear RT-LTRs in the chloroplast genome may have resulted in an expansion of the chloroplast genome and its subsequent fragmentation into hairpin plasmids during replication. Chloroplast genome fragmentation could be caused by recombination between repetitive elements and displacement of the palindromic sequences from the lagging strand during the chloroplast genome replication[47, 48], and it is consistent with the

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

expectation that recombination and cleavage of repetitive DNA will produce a heterogeneous population of molecules, as observed in dinoflagellates plastid genomes[14], and in the *Boodlea* LMW DNA.

A fragmented chloroplast genome is a common feature of Cladophorales

DNA sequence data were obtained from 9 additional Cladophorales species, representing the main lineages of the order: *Chaetomorpha aerea*, *Cladophora albida*, *C. socialis*, *C. vadorum*, *Dictyosphaeria cavernosa*, *Pithophora* sp., *Siphonocladus tropicus*, *Struvea elegans*, and *Valonia utricularis* (Supplementary Table 7). Although comparable sequencing approaches resulted in the assembly of circular chloroplast genomes for other algae, including green seaweeds[49, 50, 51], only short chloroplast contigs (200-8,000 bp) were assembled from these libraries, similar to *Boodlea composita*. Interestingly, a similar set of chloroplast genes was identified in all sequenced Cladophorales species. In contrast to the genes found in the *Boodlea* hairpin plasmids, however, most of the chloroplast genes identified in the additional Cladophorales libraries were fragmented, probably due to sequencing of the shorter Illumina reads (Supplementary Table 9). These findings support the idea that fragmentation of the chloroplast genome occurred before or early in the evolution of the Cladophorales.

Highly divergent chloroplast genes are transcribed

The 21 chloroplast protein-coding genes of *Boodlea* and the other species of Cladophorales display extremely high sequence divergence compared to orthologous genes in other photosynthetic organisms (Fig. 7b). A maximum likelihood phylogenetic tree based on a concatenated alignment of 19 chloroplast genes from Archaeplastida and Cyanobacteria species is presented in Fig. 7a. Despite their high divergence, the Cladophorales sequences formed a monophyletic group within the core Chlorophyta, sister to the Trentepohliales (Supplementary Fig. 5), a position that is supported by phylogenetic analyses of nuclear genes[52, 53]. The high sequence divergence of chloroplast genes in the Cladophorales supports the notion that organellar genomes with extremely derived architectures, including those of peridin-containing dinoflagellates, also tend to fall at the extreme ends of the range observed at the mutation rate (or gene sequence divergence) level[11, 54].

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

For some *Boodlea* chloroplast genes, the identification of start and stop codons was uncertain and a non-canonical genetic code was identified (Supplementary Fig. 6). The canonical stop codon UGA was found 11 times internally in six genes (*petA*, *psaA*, *psaB*, *psaC*, *psbC* and *rbcL*), but was also present as a genuine termination codon in several genes, *petA* and *psaA* included. At seven of these 11 positions, the corresponding amino acid residue in orthologous genes was conserved (i.e. present in more than 75% of the taxa in the alignment), but different amino acids were observed at these positions: V, Q, I and L (Supplementary Fig. 6b). For the remaining positions, the amino acid in the alignment was not conserved, and therefore the amino acid coded by the UGA codon could not be determined with certainty.

Deviations from the universal genetic code are widespread among mitochondrial genomes, and include loss of start and stop codons in some groups, including dinoflagellates[14, 55, 56]. In contrast, non-canonical genetic codes are much rarer in plastid genomes, and up to now have only been detected in the apicomplexans *Neospora caninum*[57], *Chromera velia*[26], and the dinoflagellate *Lepidodinium chlorophorum*[58]. In genomes of primary plastids, a non-canonical genetic code is unprecedented.

Dual meaning of UGA as both stop and sense codons has recently been reported from a number of unrelated protists[59, 60, 61, 62, 63]. While in *Saccharomyces cerevisiae* the tetranucleotide UGA-C allows increased incorporation of the near-cognate Cys-tRNA for the UGA premature termination codon[64], such preference was not observed in the *Boodlea* chloroplasts. Importantly, a non-canonical genetic code has also been described for Cladophorales nuclear genes, where UAG and UAA codons are reassigned to glutamine[65], which implies two independent departures from the standard genetic code in a single organism.

Unexpectedly, we found that the 16S rRNA gene in the *Boodlea* chloroplast genome is split across two distinct hairpin plasmids and that its size is much smaller compared to its algal and bacterial homologs and (Figs 4, 5). Fragmentation of rRNA genes has been observed in organellar genomes, including Apicomplexa, dinoflagellates, and many green algae[13]. In general, fragmentation of protein-coding and rRNA genes is more common in mitochondrial genomes than in plastid genomes[21, 26, 68, 69]. Despite considerable effort, we could not detect the 23S rRNA gene nor the 5S rRNA gene.

The transcription of the aberrant chloroplast genes was confirmed using RNAseq, and is concordant with previous results of Northern blots[36]. Transcripts of 21 chloroplast genes (that is,

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

20 protein-coding genes as well as the 16S rRNA gene) were identical to the genes encoded by the chloroplast 454 contigs (Figs. 3, 8, Supplementary Figs 7-12), providing evidence for the absence of RNA editing and corroborating the use of a non-canonical genetic code (Supplementary Figs 7-12). Lack of RNA editing was also evidenced for the 11 internal occurrences of UGA (Supplementary Figs 7-12). This observation, in combination with conservation of the sequence after the UGA codon, serves as evidence that it is not a termination codon but an alternative code. The high total-RNA to mRNA ratio observed for reads that mapped to the chloroplast 454 contigs confirmed that these genes were not transcribed in the nucleus (Supplementary Fig. 17). All coding sequences of the same protein-coding genes found on different contigs of the *Boodlea* chloroplast genome were expressed, despite minor differences in their nucleotide sequences (Supplementary Table 1).

Additional transcripts of 66 genes that have been located in the chloroplast in other Archaeplastida were identified (Fig. 3). 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.

Conclusions

We collected several lines of evidence indicating that *Boodlea composita* lacks a typical large circular chloroplast genome. The chloroplast genome is instead fragmented into multiple linear hairpin plasmids, and has a highly reduced gene repertoire compared to other chloroplast genomes. 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 (Fig. 3). This suggests that fragmentation of a conventional chloroplast genome in the Cladophorales has been accompanied with an elevated transfer of genes to the nucleus, similarly to the situation in peridinin-containing dinoflagellates[14], with plastid genomes encoding about 12 genes or less[14, 70]. Notably, the two distantly related 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 1). Other nonstandard chloroplast genome

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

architectures have recently been observed, such as a monomeric linear chromosome in the alveolate microalga *Chromera velia*[26] and three circular chromosomes in the green alga *Koshicola spirodelophila*[15], but these represent relatively small deviations from the paradigm, when compared to the chloroplast genome of the Cladophorales. The highly fragmented chloroplast genome in the Cladophorales is wholly unprecedented and will be of significance to understanding processes driving organellar genome fragmentation and gene reduction, endosymbiotic gene transfer, and the minimal functional chloroplast gene set.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

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[71] 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.[72].

Genomic DNA extraction, quantification and sequencing. Total genomic DNA from fresh *Boodlea* cultures was isolated by using a modified CTAB extraction protocol[73]. 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) (Supplementary Fig. 13, Supplementary table 2). Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' ends 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[74]. The assembly resulted in 3,735 contigs, which will be further referred to as "454 contigs" (Supplementary table 3). Length distribution of the 454 contigs is reported in Supplementary Fig. 13.

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+[75], resulting in

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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 identification of 58 contigs harbouring fragments or full-length chloroplast genes by sequence similarity search. These contigs had long stretches of conserved repetitive sequences at their 5' and 3' extremities. The conserved inverted repeats were used in a sequence similarity search with high stringency (high mismatch cost, high cost for gap opening and gap extension, long minimal word-size) against the 454 contigs to identify 18 additional contigs of putative chloroplast origin. This initial set of 76 contigs had a mean coverage of 84×, ranging between 11× and 191×. 17 of the 76 contigs had internal inverted repeats, with a sudden drop in read coverage (Supplementary Fig. 15). These contigs were regarded as chimeric contigs and were cleaved at the sites of coverage drop, raising the number of contigs of chloroplast origin to 89.

Additional chloroplast contigs without similarity to protein-coding genes were identified by metagenomic binning (distribution analysis of 4-mers) with MyCC[76], resulting in 21 clusters of 454 contigs (Supplementary Fig. 16). The initial set of 89 chloroplast 454 contigs was present in three neighbouring clusters: Cluster 14, Cluster 17 and Cluster 21. These clusters contained 122, eight and six contigs, respectively, raising the number of identified chloroplast contigs assembled from the chloroplast-enriched fraction from 89 to 136 ("chloroplast 454 contigs" in Supplementary table 3). Of these, 71 contigs had no sequence similarity to known protein-coding genes, 36 contigs harboured full-length chloroplast genes, 36 contigs harboured fragments of chloroplast genes, and 7 contigs harboured both fragments and full-length CDSs of different chloroplast genes.

Contigs potentially coding for chloroplast tRNAs and rRNAs were identified using Infernal 1.1[77]. The chloroplast 454 contigs served as seeds for iterative contig extension with PRICE[78]. Single-end 454 reads were used as false paired-end reads with expected insert size equal to the median length of the 454 reads. 141 different combinations of parameters were tested in order to optimize the contig extension. None of the selected assemblies showed a length improvement for the initial set of chloroplast 454 contigs. The length distribution of the chloroplast 454 contigs was consistent with the size of the LMW DNA fraction as estimated by agarose gel electrophoresis of *Boodlea* genomic DNA (Fig. 1d, Supplementary Fig. 14).

Repetitive regions in the contigs were identified with 'einverted', 'etandem' and 'palindrome' from the EMBOSS[79] package. Dotplots for all contigs were generated with YASS, using standard parameters[80]. Coverage of the chloroplast 454 contigs was evaluated by mapping the

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

454 reads, the mRNA and the total-RNA libraries to these contigs with CLC Genomics Workbench 7.0 (Qiagen) (Supplementary Fig. 17).

The chloroplast 454 contigs were used together with HMW DNA reads for two independent hybrid assemblies. First, we tried to close hypothetical gaps between the chloroplast 454 contigs with the pbahaScaffolder.py script integrated in the smrtanalysis 2.3.0 pipeline[81]. Secondly, the pre-assembled chloroplast 454 contigs were used as anchors for HMW DNA reads in a round of hybrid assembly with dbg2olc[82]. These analyses failed to close the hypothetical gaps between the short chloroplast 454 contigs and did not yield longer contigs. These results stand in stark contrast to the mitochondrial 454 contigs, where the same approaches yielded markedly longer contigs (see Supplementary Note 2).

Since the 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 by applying a hybrid correction with proovread 2.12[83] 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[84] 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[85]. Since the LMW DNA size is unknown and PBcR requires an estimate of the genome size for proper read correction, six different putative genome sizes were tested (100 kb, 1 Mb, 2.24 Mb, 10 Mb, 100 Mb). The best performance in terms of number of corrected reads was obtained by the combination of "10 Mb" for the estimated genome size and the *–sensitive* flag turned on; these corrected reads were used for the downstream analysis. After error correction, the number of reads was reduced from 154,852 to 106,428 (Supplementary table 2), with a similar length distribution as the uncorrected reads library (Supplementary Fig. 13).

In order to estimate the *Boodlea* chloroplast genome size, k-mer frequency distributions were calculated with jellyfish 2.0[86]. K-mers ranging from 11 toto 47 were analysed for uncorrected and corrected LMW DNA reads, for the filtered 454 reads and for the 454 reads that could be mapped on the chloroplast 454 contigs (Supplementary Figure 3).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

De novo genome assembly of corrected LMW DNA reads was performed with the Celera WGS assembler version 8.3rc2[87]. The resulting assembly, hereafter called the Celera Assembly, consisted of 558 contigs (Supplementary table 3). 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. In order to identify additional short protein-coding genes, HMM profiles were generated from alignments of chloroplast genes present in Chloroprotein_db and used to search the 6-frame translations of 454 contigs and corrected and uncorrected HMW and LMW DNA reads with HMMer3[88]. To prevent assembly artefacts caused by repetitive elements and palindromic sequences, we also performed an orthology-guided assembly, in which the LMW DNA reads harbouring chloroplast CDSs were re-assembled together with the respective chloroplast 454 contigs. First, LMW DNA corrected reads and chloroplast 454 contigs were grouped according to their best BLAST hit. The corrected reads and contigs belonging to the same group were assembled using Geneious v. 8.1.7 (Biomatters, http://www.geneious.com/, last accessed March 01, 2017) with parameters "High Sensitivity/Medium", and each assembly (or lack of assembly) was visually screened to exclude potential chimeric contigs (e.g. palindromic corrected subreads should be collapsed in the same locus rather than being concatenated). Where possible, LMW DNA reads and chloroplast 454 contigs were assembled as group A molecules (Fig. 2). The orthology-guided assembly yielded 19 contigs, 2 belonging to group A, 13 to group B and 4 to group E (Supplementary Table 1). Two groups of reads could not be assembled into longer molecules, and for them, the corresponding chloroplast 454 contigs were retained. Eleven additional chloroplast 454 contigs were retained, since they were not congruent with the LMW DNA reads and could not be included in the assembly. This resulted in a total of 32 contigs containing chloroplast protein-coding genes, which together with the two later identified Group B contigs encoding the 16S rRNA gene, are regarded as the *Boodlea* chloroplast genome contigs (Fig. 4, Supplementary table 3).

Protein-coding genes in the *Boodlea* chloroplast genome contigs 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[89] based on the BLAST search results. rRNAs were identified using Infernal 1.1[77]. Repetitive elements were mapped on the *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}).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

Conserved motifs were predicted with MEME suite[90], and the discovered motifs were clustered with RSAT[91]. The motifs were compared with the JASPAR-2016[92] database using TOMTOM[93] (p-value cut-off 10^{-3}).

Boodlea chloroplast genome coverage was evaluated by mapping the 454 reads with gsnap v.2016-04-04[94]. 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[94] using the *–nosplicing* flag. Due to the high number 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 the chloroplast genome was evaluated by comparing the annotated chloroplast genes 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: *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* (Fig. 3).

RNA extraction, quantification and sequencing. Total RNA was isolated using a modified CTAB extraction protocol[95]. 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 rRNAs 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) (Supplementary Table 4).

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' ends 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[96]. The resulting contigs (hereafter, transcripts) were compared using sequence similarity searches against the NCBI non-redundant protein database

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

using Tera-BLAST™ DeCypher (Active Motif, USA). Taxonomic profiling of the transcripts was performed using the following protocol: 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 (Supplementary Table 5, Supplementary Fig. 18). 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[84], following Veeckman et al. guidelines to estimate the completeness of the annotated gene space[97] (Supplementary Table 6).

Boodlea chloroplast genome expression and presence of potential RNA editing were evaluated by mapping the reads from the mRNA and total-RNA libraries to the chloroplast genome contigs with gsnap, and by aligning the transcripts resulting from the *de novo* assembly of the RNA-seq libraries to the chloroplast genome contigs with gmap.

DNA sequencing of additional Cladophorales species and phylogenetic analysis. Sequence data were obtained from 9 additional Cladophorales species, representing the main lineages of the order (Supplementary Table 7). Total genomic DNA was extracted using a modified CTAB extraction protocol as described above, and sequenced using Illumina HiSeq 2000 technology (2×100 bp paired-end reads) on 1/5th of a lane by Cold Spring Harbor Laboratory (Cold Spring Harbor, NY, USA). Quality of the reads from the sequenced libraries was assessed with FastQC 0.10.1. Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' ends of the reads were trimmed with Fastx 0.0.13 toolkit. After trimming, reads shorter than 50 bp were discarded. Trimmed reads were assembled with CLC Genomics Workbench, MIRA and SPAdes 3.6.2[98].

The taxonomic profiling of the contigs was performed with the following protocol: for each contig, sequence similarity searches were combined with the NCBI Taxonomy ID's of the top ten BLAST hits in order to discriminate between eukaryotic and bacterial contigs and contigs with no similarity to known proteins ("NoHit"). Contigs classified as eukaryotic were further analysed to identify chloroplast contigs with a sequence similarity search using Tera-BLAST™

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

(DeCypher, www.timelogic.com) against Chloroprotein_db. After chloroplast contig identification, the assembly that allowed the reconstruction of the highest number of full-length chloroplast genes was retained. An overview of the assembly metrics is reported in Supplementary Table 8, and the length distributions of the assembled contigs are reported in Supplementary Fig. 19.

Phylogenetic analysis was based on a concatenated alignment of 19 chloroplast protein-coding genes (*atpA*, *atpB*, *atpH*, *atpI*, *petA*, *petB*, *petD*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE, psbF*, *psbJ*, *psbL*, and *rbcL*) from *Boodlea*, nine other Cladophorales species, 41 additional species of Archaeplastida, and 14 Cyanobacteria species (Supplementary Table 10). For each gene, DNA sequences were aligned separately using ClustalW codon alignment in Geneious. Gene alignments were concatenated and poorly aligned positions removed using the Gblocks server[99], preserving the DNA codons, and using the least stringent settings. The resulting alignment was translated to obtain an amino acid alignment of 6,815 positions. Phylogenetic trees were inferred from the amino acid alignment using RAxML with cpREV + Γ4 model, using default parameters[100]. Branch support was assessed by bootstrapping with 500 replicates. Phylogenetic analysis was run on the CIPRES Science Gateway v3.3[101].

Data availability. DNA 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 ***–***.

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

Acknowledgments We thank Ellen Nisbett, Christopher Howe, Bram Verhelst, Sven Gould, Joe Zuccarello, and John W. La Claire II for help and advice. This work was supported by Ghent University BOF/01J04813, the Australian Research Council (DP150100705) to H.V., and 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,

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

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.

Additional information

Competing financial interests: The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

Figures

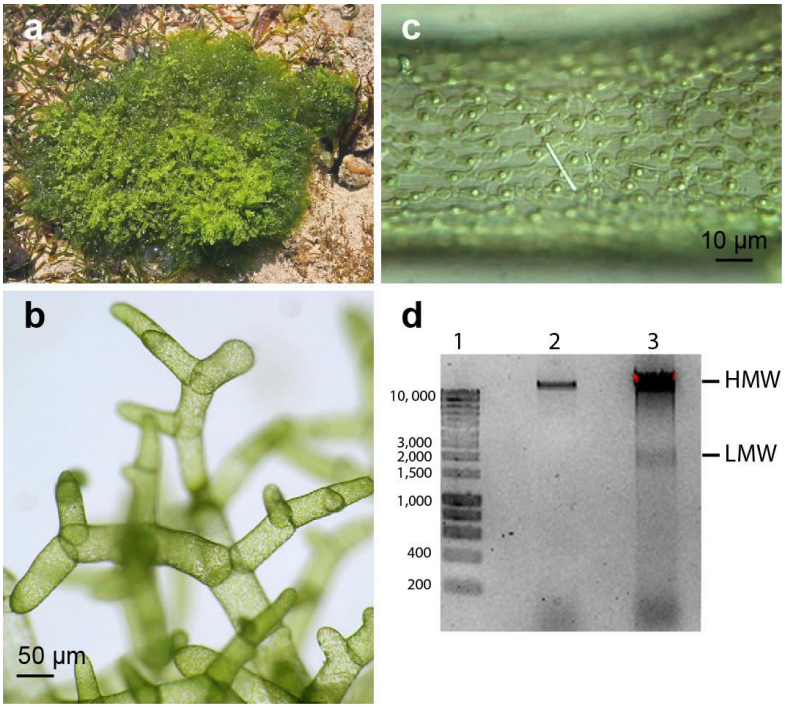


Figure 1 | *Boodlea composita*. a. Specimen in natural environment. **b.** Detail of branching cells. **c.** Detail of chloroplasts, each containing a single pyrenoid, and forming a parietal network (the white line is a calcium oxalate crystal). **d.** Native agarose gel comparing genomic DNA of *Bryopsis plumosa* (Bryopsidales) and *Boodlea composita* (Cladophorales). Lane 1: 1-kb ladder, sizes in bp; lane 2: *B. plumosa*; lane 3: *B. composita*. High molecular weight (HMW) and low molecular weight (LMW) DNA bands of *B. composita* are indicated.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

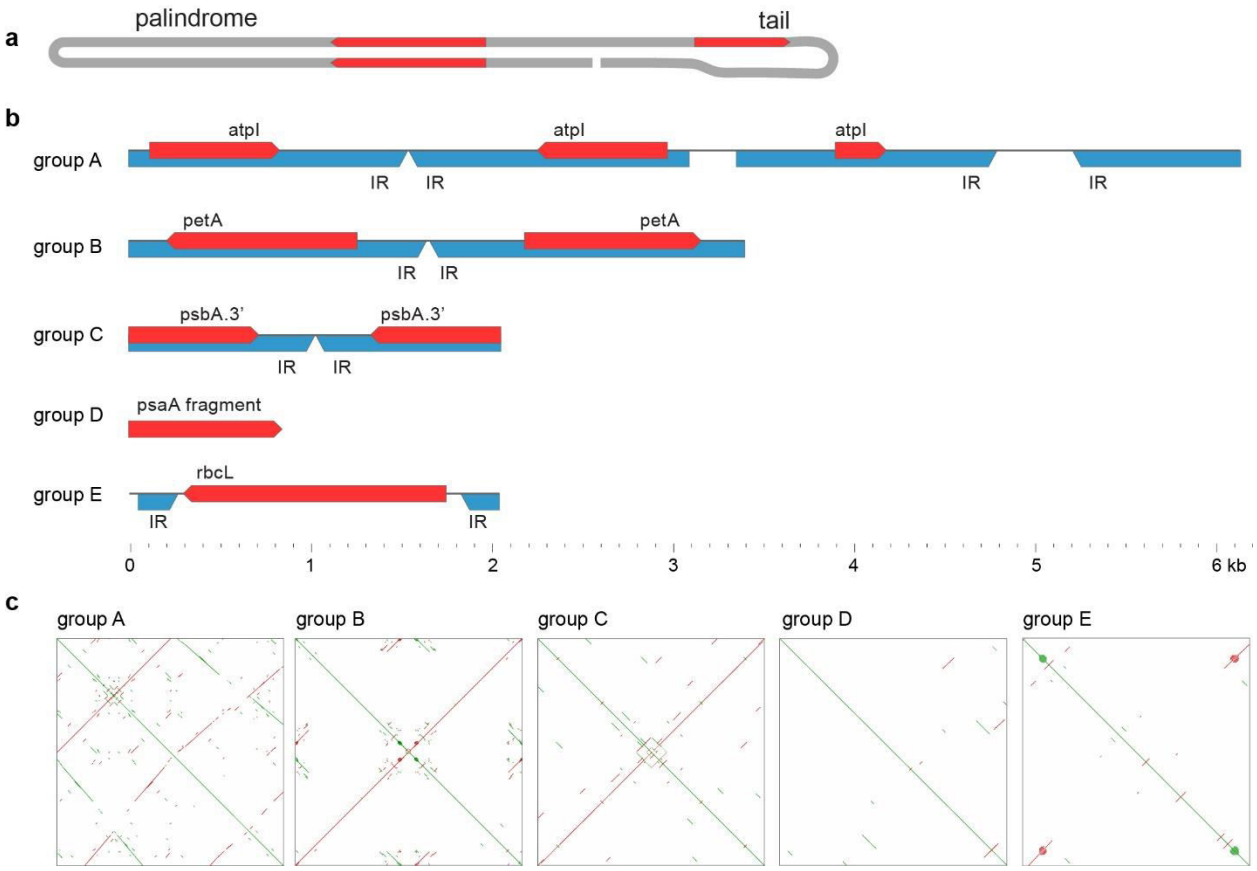


Figure 2 | Schematic representation of assembled *Boodlea* chloroplast DNA contigs. a. Schematic representation of one of the predicted native conformation of chloroplast hairpins plasmids (with a near-perfect palindromic region and a less conserved tail). Red arrows represent CDSs. **b.** Schematic representation of the assembled *Boodlea* chloroplast DNA contigs; red arrows represent CDSs, blue arrows represent major inverted repeats; scale bar at the bottom indicates length of the contigs/reads. Group A: The first half of the read is a perfect palindrome, containing the two inverted CDSs; the second half of the read (tail) is less conserved, but similar to the first half of the read. Group B: Palindromic sequences with full-length CDSs in opposite orientations, resembling the first half of group A read. Group C: palindromic sequences with fragments of CDS. Group D: short reads with fragments of CDSs that lack extensive repetitive elements. Group E: Full-length CDSs delimited by inverted repeats. Similar to group E contigs and reads, the remaining 113 chloroplast 454 contigs lacking full-length CDSs are delimited by inverted repeats. **c.** Dotplots of the five groups, showing the abundance of repetitive elements. Each dotplot was generated by aligning the contig/read with itself. Green lines indicate similar sequences; red lines indicate sequences similar to the respective reverse complements.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

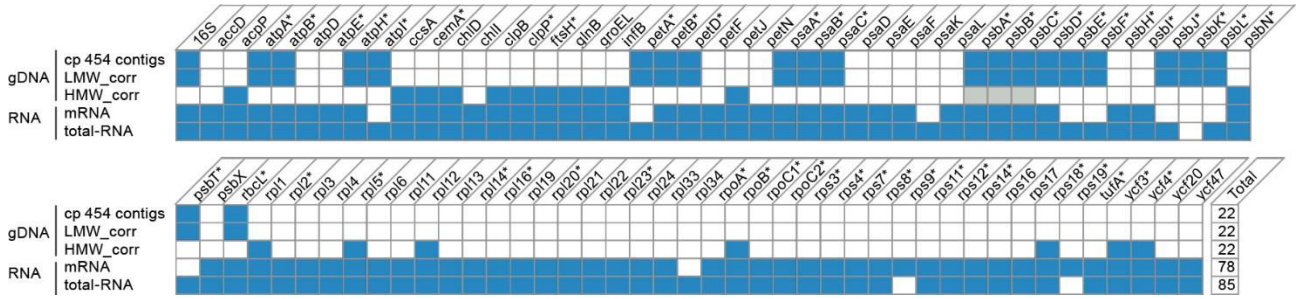


Figure 3 | *Boodlea* genes having orthologs in the chloroplast of other Archaeplastida and their distribution over the sequenced DNA/RNA. gDNA (genomic DNA): chloroplast (cp) 454 contigs, HMW and LMW corrected reads; RNA: mRNA and total-RNA assemblies. Asterisks (*) indicate "core" chloroplast genes, i.e. protein-coding genes conserved between chloroplast genomes of Chlorophyta (see Methods). 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*. Grey cells are putative carry-over contaminants: LMW DNA reads contaminants as suggested by the ratios of HMW to LMW DNA reads and mRNA to total-RNA reads (Supplementary Fig. 2).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

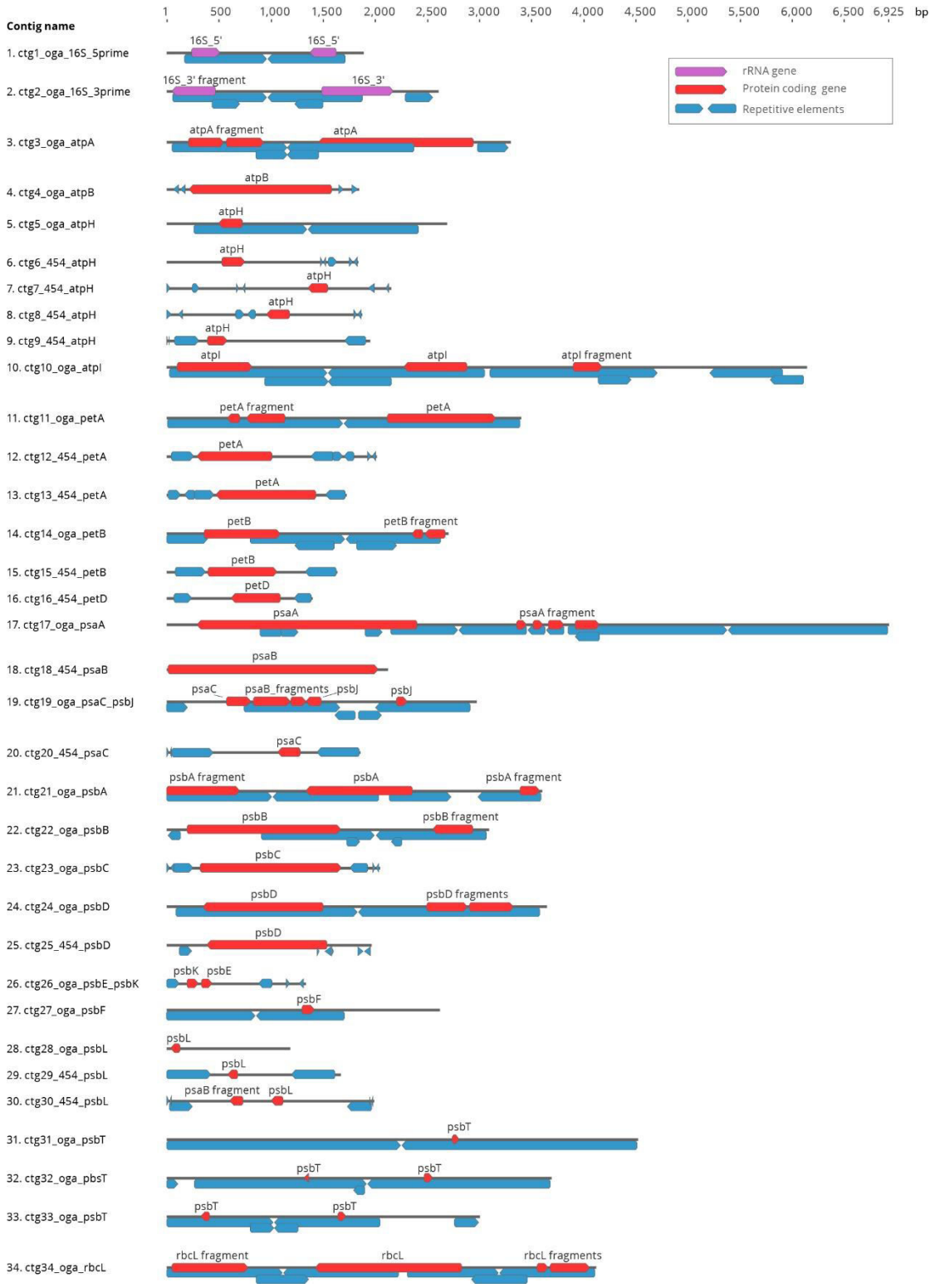
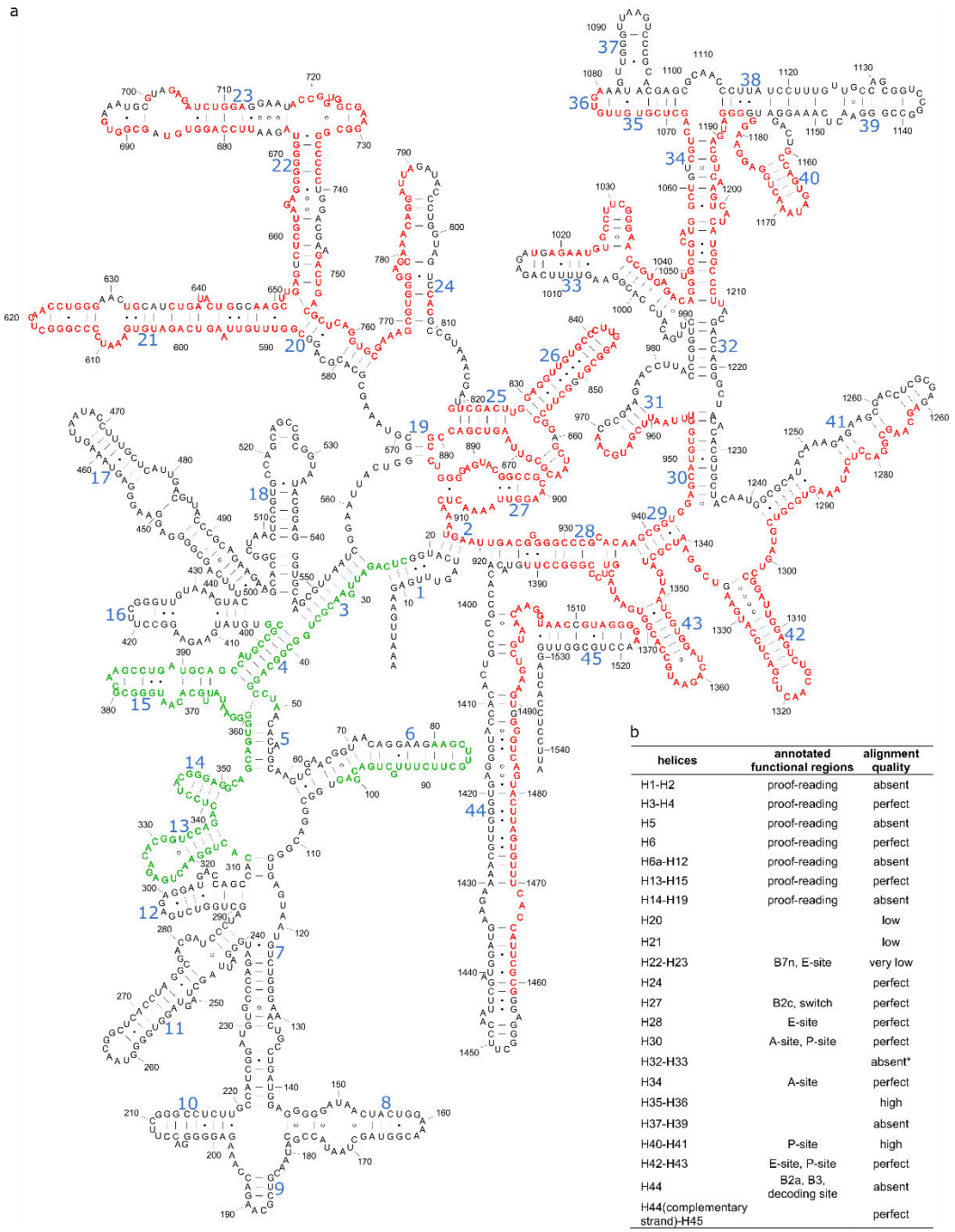


Figure 4 | Schematic representation of *Boodlea* chloroplast genome. Purple arrows indicate rRNA genes, red arrows indicate CDSs of protein-coding genes, and blue arrows indicate repetitive elements. For each contig, repetitive elements with similar length indicate similar sequences. Oga: contig obtained by orthology-guided assembly. 454: chloroplast 454 contig.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

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

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 [102]. **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%.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

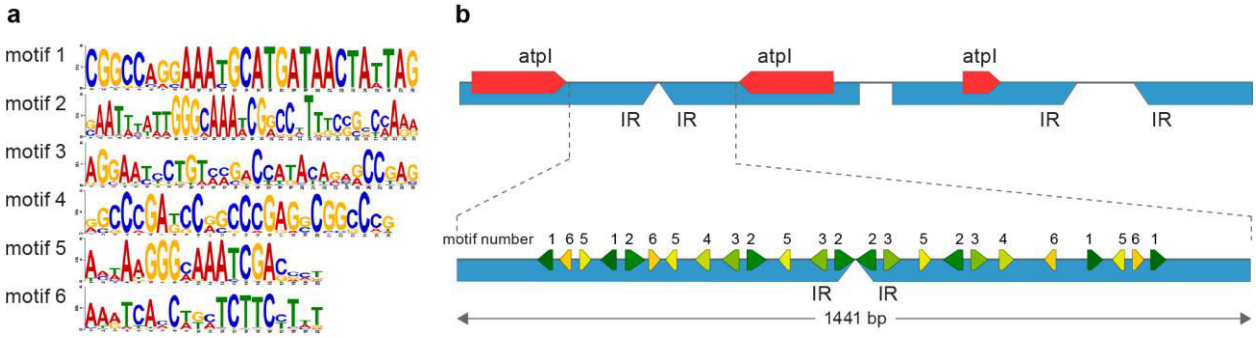
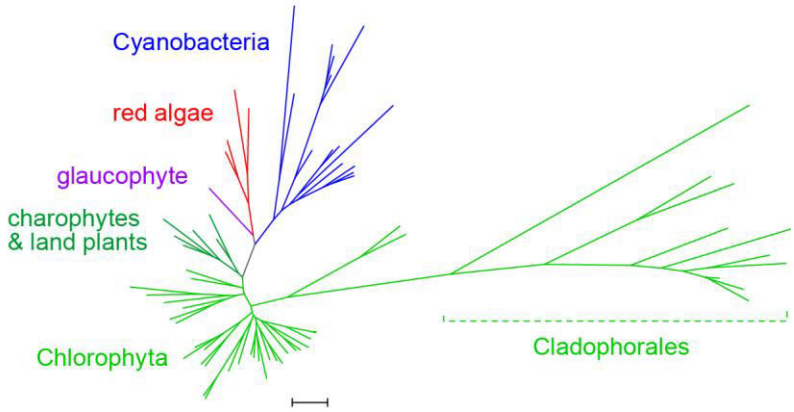


Figure 6 | Conserved non-coding motifs in *Boodlea* LMW DNA. a. Sequence logos of the conserved motifs predicted in the *Boodlea* chloroplast genome. **b.** Schematic representation of the distribution of the motifs in the 1,441 bp ncDNA region from the *atpI* group A read used for the identification 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 yellow arrows. CDSs are represented by red arrows, inverted repeats are represented by blue arrows.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

a



b

	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

Figure 7 | Phylogeny of Archaeplastida and Cyanobacteria based on 19 chloroplast genes, and corresponding distance matrix illustrating the large sequence divergence of Cladophorales chloroplast genes. a. Maximum likelihood phylogenetic tree inferred from a concatenated amino acid alignment of 19 chloroplast genes. The phylogeny with taxon names and bootstrap support values is shown in Supplementary Fig. 4. The scale represent 0,3 substitution per amino acid position. **b.** Maximum pairwise amino acid sequence distances of the chloroplast genes within and between clades (* excluded Cladophorales).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

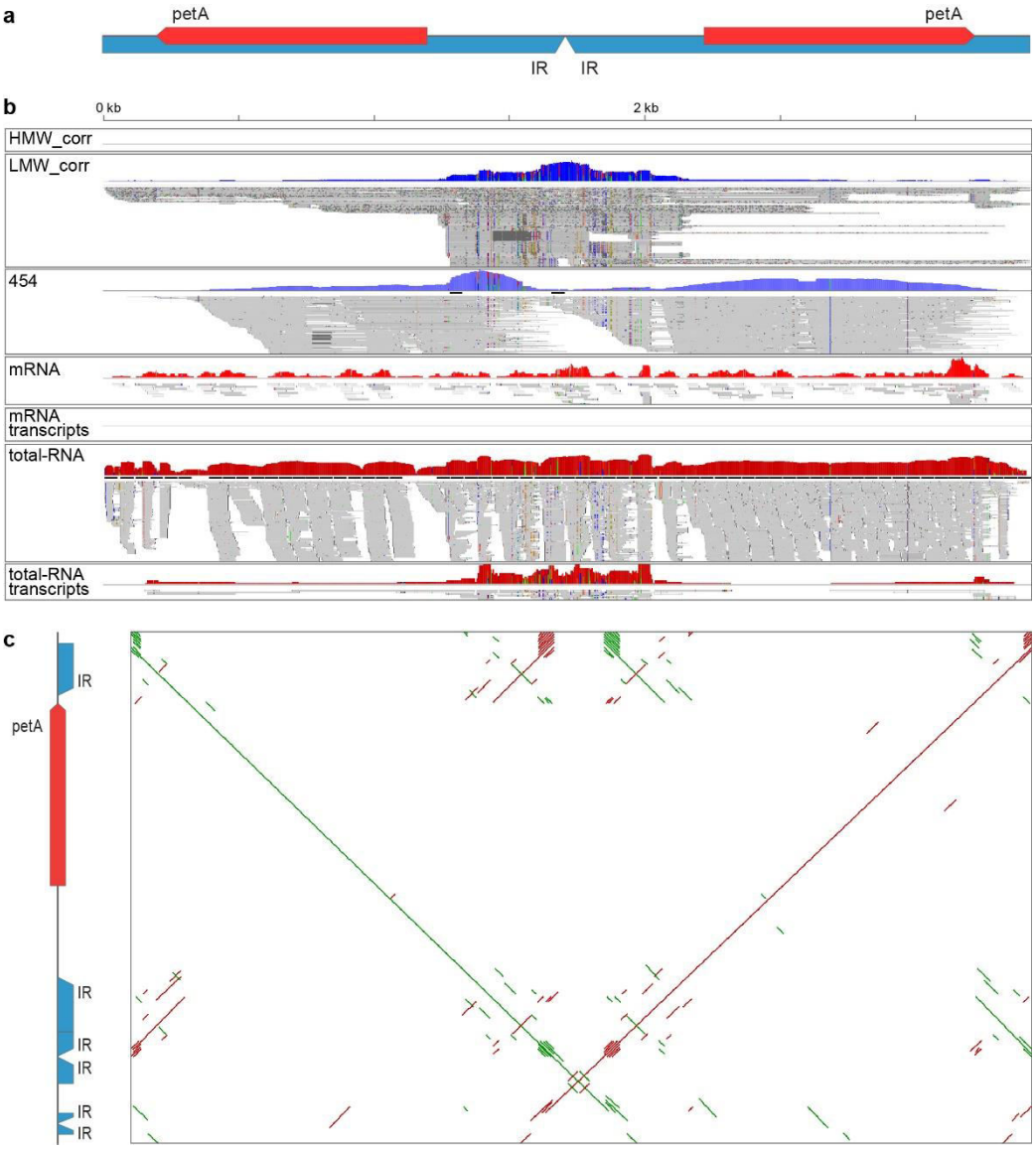


Figure 8 | LMW DNA reads containing chloroplast genes are expressed, enriched in the total-RNA fraction and congruent to the respective chloroplast 454 contigs. a. Representation of *petA* LMW DNA read (3,398 bp), a representative of group B contigs and reads (see Fig. 2). The red arrows indicate CDSs, 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; red lines indicate sequences similar to the respective reverse complements.

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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

1. Ochoa de Alda JAG, Esteban R, Diago ML, Houmard J. The plastid ancestor originated among one of the major cyanobacterial lineages. *Nature Communications* **5**, 4937 (2014).

2. Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. An early-branching freshwater cyanobacterium at the origin of plastids. *Current Biology* **27**, 386–391 (2017).

3. Gould SB, Waller RR, McFadden GI. Plastid evolution. *Annual Review of Plant Biology* **59**, 491-517 (2008).

4. Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 729-748 (2010).

5. Archibald JM. The puzzle of plastid evolution. *Current Biology* **19**, R81-R88 (2009).

6. Allen JF. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 10231-10238 (2015).

7. Green BR. Chloroplast genomes of photosynthetic eukaryotes. *The Plant Journal* **66**, 34-44 (2011).

8. Kleine T, Maier UG, Leister D. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual review of plant biology* **60**, 115-138 (2009).

9. Archibald JM, Slamovits CH. The new red algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. *Current Biology* **27**, 1-8 (2017).

10. Simpson CL, Stern DB. The treasure trove of algal chloroplast genomes. Surprises in architecture and gene content, and their functional implications. *Plant Physiology* **129**, 957-966 (2002).

11. Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences of the United States of America*, 201422049 (2015).

12. Lang BF, Nedelcu AM. Plastid genomes of algae. In: *Genomics of chloroplasts and mitochondria* (ed^(eds Bock R, Knoop V). Springer Netherlands (2012).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

13. Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. Organization and expression of organellar genomes. *Philosophical Transactions of the Royal Society B-Biological Sciences* **365**, 785-797 (2010).

14. Howe CJ, Nisbet RER, Barbrook AC. The remarkable chloroplast genome of dinoflagellates. *Journal of Experimental Botany* **59**, 1035-1045 (2008).

15. Watanabe S, Fučíková K, Lewis LA, Lewis PO. Hiding in plain sight: Koshicola spirodelophila gen. et sp. nov.(Chaetopeltidales, Chlorophyceae), a novel green alga associated with the aquatic angiosperm Spirodela polyrhiza. *American journal of botany* **103**, 865-875 (2016).

16. Zhang Z, Green B, Cavalier-Smith T. Single gene circles in dinoflagellate chloroplast genomes. *Nature* **400**, 155-159 (1999).

17. Nelson MJ, Green BR. Double hairpin elements and tandem repeats in the non-coding region of *Adenoides eludens* chloroplast gene minicircles. *Gene* **358**, 102-110 (2005).

18. Laatsch T, Zauner S, Stoebe-Maier B, Kowallik K, Maier U-G. Plastid-derived single gene minicircles of the dinoflagellate Ceratium horridum are localized in the nucleus. *Molecular Biology and Evolution* **21**, 1318-1322 (2004).

19. Hiller RG. 'Empty' minicircles and petB/atpA and psbD/psbE (cytb 559 α) genes in tandem in *Amphidinium carterae* plastid DNA. *FEBS letters* **505**, 449-452 (2001).

20. Barbrook A, Symington H, Nisbet R, Larkum A, Howe C. Organisation and expression of the plastid genome of the dinoflagellate *Amphidinium operculatum*. *Molecular Genetics and Genomics* **266**, 632-638 (2001).

21. Espelund M*, et al.* Genome fragmentation is not confined to the peridinin plastid in dinoflagellates. *PLoS One* **7**, e38809 (2012).

22. Green BR. Covalently closed minicircular DNA associated with *Acetabularia* chloroplasts. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* **447**, 156-166 (1976).

23. Ebert C, Tymms MJ, Schweiger H-G. Homology between 4.3 µm minicircular and plastomic DNA in chloroplasts of *Acetabularia cliftonii*. *Molecular and General Genetics* **200**, 187-192 (1985).

24. Bendich AJ. The size and form of chromosomes are constant in the nucleus, but highly variable in bacteria, mitochondria and chloroplasts. *BioEssays* **29**, 474-483 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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. Oldenburg DJ, Bendich AJ. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. *Current genetics* **62**, 431-442 (2016).

26. Janouškovec J*, et al.* Split photosystem protein, linear-mapping topology and growth of structural complexity in the plastid genome of *Chromera velia*. *Molecular Biology and Evolution* **30**, 2447-2462 (2013).

27. Gabrielsen TM*, et al.* Genome evolution of a tertiary dinoflagellate plastid. *PLoS One* **6**, e19132 (2011).

28. 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 (2016).

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* **7**, 2062-2082 (2015).

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

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

32. Fučíková K*, et al.* New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Frontiers in Ecology and Evolution* **2**, 63 (2014).

33. Deng Y, Zhan Z, Tang X, Ding L, Duan D. Molecular cloning and expression analysis of *rbc*L cDNA from the bloom-forming green alga *Chaetomorpha valida* (Cladophorales, Chlorophyta). *J Appl Phycol*, 1-9 (2013).

34. La Claire JW, Zuccarello GC, Tong S. Abundant plasmid-like DNA in various members of the orders Siphonocladales and Cladophorales (Chlorophyta). *Journal of Phycology* **33**, 830-837 (1997).

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

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

36. La Claire JW, Loudenslager CM, Zuccarello GC. Characterization of novel extrachromosomal DNA from giant celled marine green algae. *Current Genetics* **34**, 204-211 (1998).

37. La Claire JW, Wang JS. Structural characterization of the terminal domains of linear plasmid-like DNA from the green alga *Ernodesmis* (Chlorophyta). *Journal of Phycology* **40**, 1089-1097 (2004).

38. Zhang T, Zhang X, Hu S, Yu J. An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant methods* **7**, 38 (2011).

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

40. Smith DR, Lee RW. The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA. *BMC genomics* **10**, (2009).

41. de Vries J*, et al.* Is ftsH the key to plastid longevity in sacoglossan slugs? *Genome Biology and Evolution* **5**, 2540-2548 (2013).

42. 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* **2**, 240-256 (2010).

43. Leliaert F*, et al.* Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* **31**, 1-46. (2012).

44. Hozza M, Vinař T, Brejová B. How big is that genome? Estimating genome size and coverage from k-mer abundance spectra. In: *String Processing and Information Retrieval: 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings* (ed^(eds Iliopoulos C, Puglisi S, Yilmaz E). Springer International Publishing (2015).

45. Burt AT, Robert. *Genes in conflict: the biology of selfish genetic elements*. Harvard University Press (2006).

46. Tymms MJ, Schweiger H-G. Tandemly repeated nonribosomal DNA sequences in the chloroplast genome of an *Acetabularia mediterranea* strain. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 1706-1710 (1985).

47. Bikard D, Loot C, Baharoglu Z, Mazel D. Folded DNA in action: hairpin formation and biological functions in Prokaryotes. *Microbiology and Molecular Biology Reviews* **74**, 570-588 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

48. Ellis THN, Day A. A hairpin plastid genome in barley. *The EMBO Journal* **5**, 2769-2774 (1986).

49. Leliaert F, Lopez-Bautista JM. The chloroplast genomes of *Bryopsis plumosa* and *Tydemania expeditionis* (Bryopsidales, Chlorophyta): compact genomes and genes of bacterial origin. *BMC genomics* **16**, 204 (2015).

50. Melton JT, III, Leliaert F, Tronholm A, Lopez-Bautista JM. The complete chloroplast and mitochondrial genomes of the green macroalga *Ulva* sp. UNA00071828 (Ulvophyceae, Chlorophyta). *PLoS ONE* **10**, e0121020 (2015).

51. Marcelino RV, 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* **8**, 2939-2951 (2016).

52. Cocquyt E, Verbruggen H, Leliaert F, De Clerck O. Evolution and cytological diversification of the green seaweeds (Ulvophyceae). *Molecular Biology and Evolution* **27**, 2052-2061 (2010).

53. Leliaert F*, et al.* Systematics of the marine microfilamentous green algae *Uronema curvatum* and *Urospora microscopica* (Chlorophyta). *European Journal of Phycology* **44**, 487-496 (2009).

54. Zhang Z, Green BR, Cavalier-Smith T. Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: a possible common origin for sporozoan and dinoflagellate plastids. *Journal of Molecular Evolution* **51**, 26-40 (2000).

55. Waller RF, Jackson CJ. Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *Bioessays* **31**, 237-245 (2009).

56. Slamovits CH, Saldarriaga JF, Larocque A, Keeling PJ. The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *Journal of molecular biology* **372**, 356-368 (2007).

57. Lang-Unnasch N, Aiello D. Sequence evidence for an altered genetic code in the *Neospora caninum* plastid. *International Journal for Parasitology* **29**, 1557-1562 (1999).

58. Matsumoto T, Ishikawa SA, Hashimoto T, Inagaki Y. A deviant genetic code in the green alga-derived plastid in the dinoflagellate *Lepidodinium chlorophorum*. *Molecular phylogenetics and evolution* **60**, 68-72 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

59. Zahonova K, Kostygov AY, Sevcikova T, Yurchenko V, Elias M. An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Current Biology* **26**, 1879-0445 (2016).

60. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. 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).

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

62. Keeling Patrick J. Genomics: Evolution of the Genetic Code. *Current Biology* **26**, R851-R853 (2016).

63. Zinshteyn B, Green R. When stop makes sense. *Science* **354**, 1106-1106 (2016).

64. Beznoskova P, Gunisova S, Valasek LS. Rules of UGA-N decoding by near-cognate tRNAs and analysis of readthrough on short uORFs in yeast. *RNA* **22**, 456-466 (2016).

65. Cocquyt E, Gile G, Leliaert F, Verbruggen H, Keeling P, De Clerck O. Complex phylogenetic distribution of a non-canonical genetic code in green algae. *BMC Evolutionary Biology* **10**, 327 (2010).

66. Okimoto R, Macfarlane JL, Wolstenholme DR. The mitochondrial ribosomal RNA genes of the nematodes *Caenorhabditis elegans* and *Ascaris suum*: consensus secondary-structure models and conserved nucleotide sets for phylogenetic analysis. *Journal of Molecular Evolution* **39**, 598-613 (1994).

67. Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. *BMC genomics* **7**, 297 (2006).

68. Smith DR, Lee RW, Cushman JC, Magnuson JK, Tran D, Polle JEW. The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. *BMC Plant Biology* **10**, 83 (2010).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

71. Andersen RA. *Algal culturing techniques*. Elsevier (2005).

72. Palmer JD. Physical and gene mapping of chloroplast DNA from *Atriplex triangularis* and *Cucumis sativa*. *Nucleic Acids Research* **10**, 1593-1605 (1982).

73. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11-15 (1987).

74. Chevreux B*, et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* **14**, 1147-1159 (2004).

75. Boratyn GM*, et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Research* **41**, W29-W33 (2013).

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

77. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

78. Ruby JG, Bellare P, Derisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3* **3**, 865-880 (2013).

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

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

81. Bashir A*, et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology* **30**, 701-707 (2012).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

84. Vandepoele K*, et al.* pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environmental Microbiology* **15**, 2147-2153 (2013).

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

86. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770 (2011).

87. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology* **33**, 623-630 (2015).

88. Eddy SR. Accelerated profile HMM searches. *PLOS Computational Biology* **7**, (2011).

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

90. Bailey TL*, et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202-W208 (2009).

91. Medina-Rivera A*, et al.* RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Research* **43**, W50-56. (2015).

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

93. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome Biology* **8**, R24 (2007).

94. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In: *Statistical Genomics: Methods and Protocols* (ed^(eds Mathé E, Davis S). Springer New York (2016).

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

bioRxiv preprint doi: https://doi.org/10.1101/145037; this version posted August 15, 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.

96. Grabherr MG*, et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-U130 (2011).

97. 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).

98. Bankevich A*, et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455-477 (2012).

99. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56**, 564-577 (2007).

100. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

101. Miller MA, Pfeiffer W, Schwartz T. The CIPRES science gateway: a community resource for phylogenetic analyses. In: *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery* (ed^(eds). ACM (2011).

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