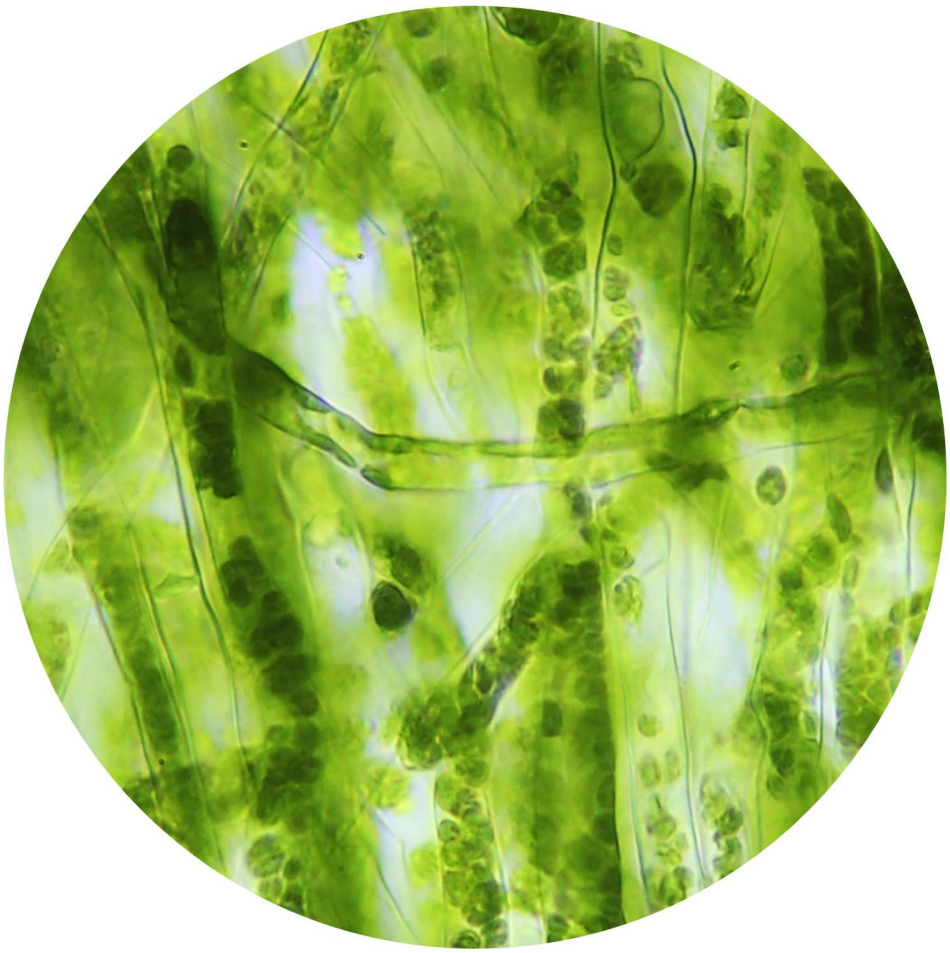
A LITTLE BIT OF GREEN:

Genome dynamics in green algae

Sonja Repetti



A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science (Honours)

Student number: 758334

Algal Biology Laboratory, School of Biosciences, University of Melbourne.

Supervisors: Dr Heroen Verbruggen, Dr Christopher J. Jackson

Email: srepetti@student.unimelb.edu.au

May 2019

Contents

DECLARATION STATEMENT

i. The work described comprises my own work except where indicated below.

ii. This work has not been submitted, either in whole or in part, for the award of a degree at The University of Melbourne or any other institution of higher education.

iii. This thesis is less than 10,000 words in length, excluding words in the abstract, acknowledgements, references, captions, tables and appendix and complies with the requirements set out for the degree of Honours of Science (BioSciences) at The University of Melbourne.

iv. Referencing is in the style of the journal New Phytologist.

v. Long read sequencing was performed by the Kathryn Holt Laboratory.

vi. Creation of the Pico-Plaza custom instance, including delineation of gene families, was performed by Dr Klaas Vandepoele and Dr Michiel Van Bel.

vii. MaSuRCa assembly was performed in collaboration with Dr Christopher J. Jackson.

viii. One draft of this thesis was read by my supervisors who provided input that was mostly general and structural (as well as identifying a couple of grammatical errors). HV also provided general feedback on a second draft of the Chapter Discussion and the General Discussion.

Sonja Repetti, 10th May 2019

ABSTRACT

The green algae (Chlorophyta) include a diverse range of organisms that differ considerably in both morphology and the structure of their genomes. Their common origin, as well as the common origins of their organelles, means that the diversity of Chlorophyta genomes reflects evolutionary forces acting differently on various lineages and, potentially, differently on the three genomes – nuclear, chloroplast and mitochondrion – within a single lineage. My project aimed to examine the evolutionary forces shaping genomes within the Chlorophyta by characterising and analysing two genomes: the nuclear genome of unidentified pedinophyte YPF701, and the mitochondrial genome of the siphonous green seaweed *Ostreobium quekettii*. Both genomes are significant due to their positions phylogenetically. YPF701 at the base of the core Chlorophyta can provide insights into gene family evolution that occurred as this group diverged, while the *O. quekettii* mitochondrial genome represents only the second mitochondrial genome sequenced in the Bryopsidales. Both projects involved combining long and short read sequencing data to assemble the genomes as well as a variety of bioinformatic tools to analyse and compare them with other Chlorophyta. The nuclear genome of pedinophyte YPF701 is a fairly small (26-34 Mb) genome that shows evidence of gene family loss along the pedinophyte lineage. My project created a more contiguous hybrid nuclear genome assembly for YPF701 that can be used to examine gene family evolution, as well as the nature of noncoding regions in this lineage. The *O. quekettii* mitochondrial genome is the largest green algal mitochondrial genome sequenced thus far (241,739 bp), and is approximately three times larger than its economical plastid genome. The genome encodes genes typical of green algal mitochondrial genomes. Most of this excess size is explained by the expansion of intergenic DNA and proliferation of introns. Several theories can explain the evolution of both genomes described in this study, which ultimately reflect an interplay of mutation, natural selection and genetic drift.

General Introduction

Phylogenetic history of Chlorophyta

Along with Streptophyta – which include the Charophytes (mostly freshwater algae) and land plants – Chlorophyta belong to the Chloroplastida lineage of eukaryotes that share a green coloured plastid, or Chloroplast, and diverged from a putative 'ancestral green flagellate' (Fig.1) (Leliaert *et al.*, 2012; Fang *et al.*, 2017). Chloroplastida, as well as the glaucophytes (Glaucophyta) and red algae (Rhodophyta), are within the Archaeplastida. Their plastids likely have a common origin from a single primary endosymbiosis event where a cyanobacterium was engulfed and retained before eventually becoming an organelle (Rodríguez-Ezpeleta *et al.*, 2005; Keeling, 2010). This primary endosymbiosis has been dated at approximately 1.5 billion years ago (Hedges *et al.*, 2004; Yoon *et al.*, 2004), however dating the subsequent divergence of the Chloroplastida has proven to be challenging (summarised in Leliaert *et al.*, 2012).

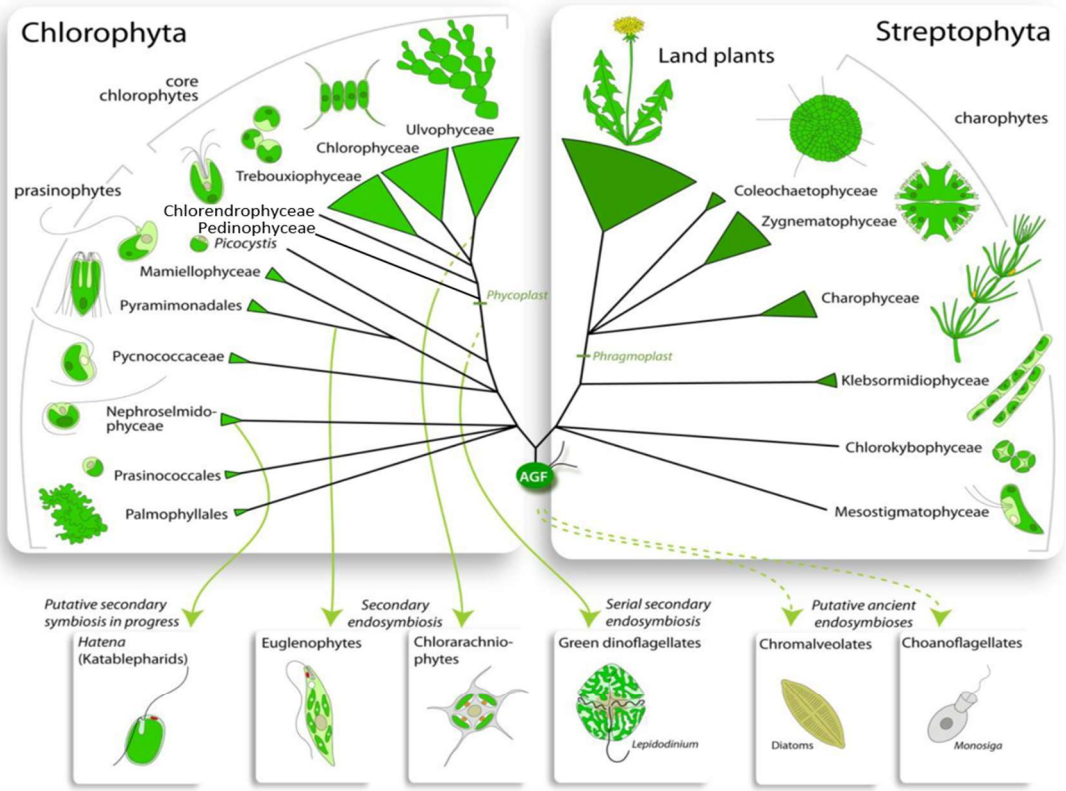


FIG. 1. – Overview phylogeny of the green lineage (Chloroplastida) and spread of green genes in other eukaryotes, adapted from Leliaert *et al.* (2012) and modified to reflect the addition of Pedinophyceae (e.g. Fang *et al.,* 2018).

The Chlorophyta are subdivided into the core Chlorophyta and the paraphyletic early-branching prasinophytes. The core Chlorophyta are a well-supported clade of classes Ulvophyceae, Chlorophyceae and Trebouxiophyceae, with the smaller and earlier diverging Chlorodendrophyceae and Pedinophyceae. The prasinophytes are mostly marine unicellular planktonic algae (Fig. 1) (Marin, 2012; Fučíková *et al.*, 2014; Fang *et al.*, 2017; Fang *et al.*, 2018).

Chlorophyta genomes

The rise of high throughput sequencing and its adoption within protist research is increasing the number of biological questions that can be explored using genome data (Oliveira *et al.*, 2018). Within the Chlorophyta there are three separate sources of such genome data: in the nucleus, as well as the mitochondrion and chloroplast.

Nuclear genomes

Despite representing a limited range of taxa, the 61 Chlorophyta whole nuclear genomes that have been submitted to Genbank vary considerably in size and properties (https://www.ncbi.nlm.nih.gov/genome/browse#!/eukaryotes/Chlorophyta). Many of these genomes have not been published with extensive maps of genes or structure, with authors instead choosing to focus on comparing lineages based on specific gene families or pathways, such as phospholipid production (Hirashima *et al.*, 2016; Hirashima *et al.*, 2018) and metabolism of sulfur (Nelson *et al.*, 2019) and starch (Deschamps *et al.*, 2008).

Comparing lineage-specific changes in gene families in response to environmental pressures has been the driving force behind the sequencing of several Chlorophyta genomes. Gene families in the genomes of polar *Coccomyxa subellipsoidea* (Blanc *et al.*, 2012), acidophilic *Chlamydomonas eustigma* (Hirooka *et al.*, 2017), halotolerant *Picochlorum* (Foflonker *et al.*, 2015), and endosymbionts *Micractinium conductrix* (Arriola *et al.*, 2018) and *Chlorella variabilis* (Blanc *et al.*, 2010) have undergone expansion that is not seen in related species who do not experience the same environmental conditions. The function of the gene families that underwent these expansions, which include lipid and polysaccharide metabolism, energy transport, ion transport, chitin synthesis, and extracellular sugar and amino acid transport, provide insight into how Chlorophyta have adapted to given environmental pressures (Blanc *et al.*, 2010; Blanc *et al.*, 2012; Foflonker *et al.*, 2015; Hirooka *et al.*, 2017; Arriola *et al.*, 2018). Reductions in gene families have also been observed, such as the loss of fermentation

pathways in the acidophile *Chlamydomonas eustigma* which would normally acidify the cytosol but are no longer required due to its acidic habitat (Hirooka *et al.*, 2017).

It is likely that interactions between organisms and their environments also play a part in shaping genome structure (Wendel *et al.*, 2016). Species in high-energy habitats with short generation times often have faster rates of molecular evolution with more mutations accumulated per unit time (Burger *et al.*, 2003; Bromham, 2011). This is proposed to have driven streamlining of the small *Picochlorum* genome (~15 Mb) (Foflonker *et al.*, 2015). We can see high rates of gene inactivation and loss particularly in parasites; their transition to an intracellular environment appears to reduce the ability of selection to retain many genes (Mira *et al.*, 2001). The genome of the obligate green-alga derived parasite *Helicosporidium* is small and compact, approximately two and a half times smaller than other free living and symbiotic Trebouxiophytes (Pombert *et al.*, 2014). This compaction in the *Helicosporidium* genome comes from contraction within gene families, particularly those linked with genome maintenance and expression, as well as reduction in the amount of noncoding DNA (Pombert *et al.*, 2014).

Genome comparisons within particular Chlorophyta lineages have also revealed differences in coding density. Prasinophytes in the genus *Ostreococcus* are some of the smallest free-living eukaryotes, and they have small genomes to match (~13 Mb) (Derelle *et al.*, 2006; Palenik *et al.*, 2007). The reduced sizes of prasinophyte genomes relative to other Chlorophyta reflect reduction in the number of gene families and individual genes, and also shortening of intergenic regions and gene fusion (Derelle *et al.*, 2006; Moreau *et al.*, 2012). The small and gene dense prasinophyte genomes may reflect the process of genome streamlining, a hypothesis that suggests selection acts to decrease the size of genomes in order to reduce the cost of replicating non-essential DNA (Giovannoni *et al.*, 2005).

Despite overall gene family size reduction in prasinophytes and *Helicosporidium*, there is evidence of lineage-specific expansion within some gene families. In *Helicosporidium* an expanded family of chitinases may digest the barriers of its insect host or remodel the parasite's cell wall (Pombert *et al.*, 2014), while the expanded gene families in *Bathycoccus* have a hypothesised role in the formation of the external scales surrounding the cell (Moreau *et al.*, 2012). *Ostreococcus* genomes have expansions within diverse gene families related to obtaining nutrients, such as iron in *O. lucimarinus*, and to photosynthesis in *O. tauri*,

reflecting optimisation of energy acquisition from the environment despite their small sizes (Derelle *et al.*, 2006; Palenik *et al.*, 2007). Other proposed sources of new coding content to Chlorophyta genomes include horizontal transfer of genes from unrelated lineages (Palenik *et al.*, 2007; Blanc *et al.*, 2010; Moreau *et al.*, 2012; Foflonker *et al.*, 2015; Hirooka *et al.*, 2017), and transfer from organelle genomes (Palenik *et al.*, 2007; Smith & Lee, 2009).

Volvox and related lineages (Volvocales) within the Chlorophyceae are an established model system for studying the transition from unicellular to multicellular life (Umen & Olson, 2012; Featherston *et al.*, 2016; Herron, 2016), as this group spans a range of morphological diversity from unicellular (e.g. *Chlamydomonas*) to differentiated multicellular forms (e.g. *Volvox*). Comparing the genome of *Chlamydomonas reinhardtii* to lineages with multiple cells indicates that the evolution of multicellularity in the Volvocales did not require large-scale genomic innovation (Hanschen *et al.*, 2016; Featherston *et al.*, 2017). A small set of gene families expanded at the evolution of colonial living in four-celled *Tetrabaena* followed by an even smaller expansion in the colonial *Gonium* (16 cells) and multicellular *Volvox* (Featherston *et al.*, 2017). These gene families are implicated in DNA repair, cell cycling, cell adhesion and extracellular functions (Hanschen *et al.*, 2016; Featherston *et al.*, 2017). Many gene family expansions are lineage-specific, such as those associated with the extracellular matrix that surrounds *Volvox* (Prochnik *et al.*, 2010; Hanschen *et al.*, 2016).

The Volvocales are also a convenient group to observe the effect of multicellularity on genome structure. The coding content of *Chlamydomonas* and *Volvox* genomes is similar (Merchant *et al.*, 2007; Prochnik *et al.*, 2010), however the *Volvox* genome is 17% larger than the Chlamydomonas genome (Prochnik *et al.*, 2010). Gene density decreases from *Chlamydomonas* to *Volvox,* while intron length increases (Hanschen *et al.*, 2016). *Volvox* has a greater amount of non-coding repetitive sequences (Prochnik *et al.*, 2010).

Organellar genomes

More than 127 chloroplast and 60 mitochondrial genomes have been published for the Chlorophyta, and these span a broader range of taxa compared with nuclear data (https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/Chlorophyta). Unlike nuclear genomes, Chlorophyta organelle genomes tend to be published with more comprehensive

genome descriptions and maps (e.g. Marcelino *et al.*, 2016; Zheng *et al.*, 2018). Alignments of multiple combined organelle genes have also been used for phylogenetic analyses to resolve lineage relationships within the Chlorophyta (e.g. Fučíková *et al.*, 2014; Cremen *et al.*, 2018). Other studies have considered organellar genome structural adaptation and unique features (Del Cortona *et al.*, 2017; Hamaji *et al.*, 2017).

Chlorophyta mitochondrial genome structure varies, particularly in the Volvocales with examples of both linear and circular forms (Hamaji *et al.*, 2017). *Yamagishiella unicocca* has a single linear mitochondrial genome with long palindromic telomeres, while the mitochondrial genome in related *Eudorina* has identical gene order but appears to form both a circular molecule *and* a linear form (Hamaji *et al.*, 2017). As such, Hamaji *et al.* (2017) hypothesised that the common ancestor of the Volvocales had a linear mitochondrial genome. Chlorophyta chloroplast genomes are usually circular, however exceptions include the fragmented hairpin plasmids seen in Cladophorales chloroplasts (Del Cortona *et al.*, 2017) and the *Acetabularia acetabulum* (Dasycladales) chloroplast genome, which is highly repetitive and may be as large as 2,000 Kb (de Vries *et al.*, 2013), though yet to be comprehensively assembled and described.

Along with their nuclear genomes, the organellar genomes of the Volvocales have been extensively sequenced and compared between species. Mirroring what is seen for nuclei, organelle DNA complexity, especially non-coding DNA, scales positively with size and cell number (Smith & Lee, 2009; Smith *et al.*, 2013; Featherston *et al.*, 2016). For chloroplast genomes, Smith *et al.* (2013) observed an increase from approximately 60% noncoding DNA in smaller lineages to greater than 80% in multicellular *Volvox carteri*. In contrast, the number of genes is only 2 greater in the largest lineages compared with the smallest (Smith *et al.*, 2013). A similar pattern was observed for mitochondrial DNA. Smith and Lee (2009) found a large proportion of noncoding DNA in both organelle genomes of *Volvox carteri*, with palindromic repeats in the mitochondrial, chloroplast, and nuclear genomes, the latter most likely via organelle to nucleus transfer. Comparative studies in Chlorophyta mitochondrial and chloroplast genomes have found that many instances of genome reduction are not due to gene loss, but rather reduction in non-coding DNA (Burger *et al.*, 2003; Smith *et al.*, 2013; Marcelino *et al.*, 2016).

Genome evolution

The common origin of the Chlorophyta, as well as that of their chloroplasts (Rodríguez-Ezpeleta *et al.*, 2005) and mitochondria (Roger *et al.*, 2017), means that the diversity of their genomes solely reflect evolutionary forces acting differently on various lineages and, potentially, differently on the three genomes within a single lineage. Although many hypotheses have been posed to explain genome evolution, at their core these theories merely describe differing contributions from the forces of mutation, natural selection, and genetic drift (stochastic changes) to the evolution of genomes.

One proposal states that excess DNA acts as a mutational target, increasing the mutation rate of associated genes. This negative impact would thereby be opposed by weak (purifying) natural selection (Lynch, 2006; Lynch *et al.*, 2006). Effective population size (N_e) is a concept used in population genetics to describe the amount of genetic drift acting on a genome; it can be defined as the population size in the Wright-Fisher model of evolution matching the level of drift observed in a more complex system (Platt *et al.*, 2018). At larger N_e, selection is expected to be more efficient and drift decreases, while at lower N_e the power of drift increases relative to selection (Lynch & Conery 2003, Lynch 2006, Lynch *et al.* 2016). The mutational hazard hypothesis (MHH) proposes that excess DNA is more likely to accumulate in genomes with a low mutation rate and small N_e (Lynch *et al.*, 2006; Smith, 2016). The MHH is supported by the streamlining of organelle genomes within various lineages including prasinophytes, red algae, and some fungi, which have high estimated mutation rates (Smith, 2016). Smith and Lee (2010) propose that in the Volvocales, the shift from unicellular *Chlamydomonas* to multicellular *V. carteri* resulted in a lower N_e that allowed non-coding DNA to persist when it would otherwise have been lost through purifying selection.

It has also been proposed that selection can influence the mutation rate of genomes. According to the drift-barrier hypothesis (Lynch *et al.*, 2016), selection acts to reduce the mutation rate with an overall limit set by genetic drift. In genomes with very high mutation rates, possessing 'antimutators' (such as DNA repair proteins) is advantageous; these antimutators mean that the mutation rate will therefore evolve downwards until the strength of selection is matched by that of genetic drift and mutation bias (Lynch *et al.*, 2016). In contrast, in genomes with a low mutation rate, having antimutators will not be sufficiently advantageous, whilst having mild mutators will not be sufficiently disadvantageous.

Therefore, mutation rate will increase until selection is strong enough to prevent the genome evolving a higher mutation rate (Lynch *et al.*, 2016). Krasovec *et al.* (2017) found support for the drift-barrier hypothesis when they combined a review of literature with further mutation rate estimates for four prasinophytes and concluded that mutation rate tends to decrease as N_e, therefore strength of selection, increases.

However, increased genome sequencing has revealed that mutation rates can vary widely among lineages and even between genome compartments of the same lineage (Smith, 2016). Green alga *Dunaliella salina* contains inflated organelle genomes, both chloroplast and mitochondrial, but there are order-of-magnitude differences in mutation rates between the two compartments, with substitution rates between two strains of *D. salina* 2-13 times greater in mtDNA than ptDNA (Del Vasto *et al.*, 2015). Such findings make it difficult to draw direct connections between mutation rates and genome architecture (Smith, 2016).

Sequencing and comparing Chlorophyta genomes

While published land plant genomes have been compared extensively to examine their evolutionary history (summarised in Wendel *et al.*, 2016), a similar in-depth evolutionary study has not yet been performed for the Chlorophyta. This study would be enhanced by the sequencing of a greater range of Chlorophyta genomes (Pombert *et al.*, 2014).

The vast majority of Chlorophyta taxa have not yet had their nuclear or organellar genomes sequenced. Published Chlorophyta nuclear genomes vary in the completeness of their assemblies, ranging between many smaller contigs (short continuous DNA sequences), fewer larger scaffolds (built from overlapping contigs), whole chromosomes, and only two complete genomes that include all chromosomes, are gapless and lack long runs of ambiguous bases (https://www.ncbi.nlm.nih.gov/genome/browse#!/eukaryotes/Chlorophyta). It can be difficult to assemble some nuclear and organellar genomes, particularly larger genomes with complex structure and repetitive regions. Long-read sequencing can help resolve repetitive regions and discern large scale genome structure (Goodwin *et al.*, 2015; Koren & Phillippy, 2015; Oliveira *et al.*, 2018). By combining these long reads with short reads – which tend to contain less errors – into hybrid assembles, one can take advantage of the complementary strengths of both to overcome the problem of sequence complexity and successfully characterise genomes, producing more complete high quality assemblies (Rhoads & Au, 2015; Wendel *et al.*, 2016).

The studies outlined in my two chapters utilised a combination of long and short reads in order to assemble two new genomes to contribute this active field of research: the nuclear genome of the unicellular green flagellate pedinophyte YPF701 and the mitochondrial genome of the siphonous green seaweed *Ostreobium quekettii*. Due to their sizes and phylogenetic position relative to already sequenced lineages, these two genomes are well positioned to investigate the evolution of green algal nuclear genomes and the evolution of mitochondrial genomes in the Bryopsidales respectively. Background, specific aims, and future directions for each case study are explained in detail in their respective chapters. In the General Discussion, I consider potential limitations and what my results contribute to my aim of understanding how evolutionary forces have shaped Chlorophyta genomes.

CHAPTER 1
Microalga in the middle: the nuclear genome of Pedinophyte YPF701

Introduction

Although the number of sequenced genomes is increasing steadily, there has not yet been a comprehensive comparative study examining evolution at the scale of the whole Chlorophyta. Numerous studies (e.g. Derelle *et al.*, 2006; Palenik *et al.*, 2007) have analysed both the coding and noncoding content of the nuclear genomes of the basal pedinophyte lineages. Most core Chlorophyta studies, however, have focussed on the coding content of specific lineages (see General Introduction), apart from in the Volvocales where comparisons of unicellular, colonial and multicellular lineages have looked at both coding and noncoding content (e.g. Prochnik *et al.*, 2010; Hanschen *et al.*, 2016; Featherston *et al.*, 2017). Nuclear genomes have now been published for all classes of core Chlorophyta except the Chlorodendrophyceae, and Pedinophyceae (pedinophytes).

Pedinophytes are small (2.5-7 microns), usually naked, unicellular green flagellates found in water or soil habitats and sometimes in symbioses (Karpov & Tanichev, 1992; Marin, 2012), including within the dinoflagellate *Noctiluca miliaris* (Sweeney, 1976) and the radiolarian *Thalassolampe margarodes* (Cachon & Caram, 1979). Taking such symbiosis to the extreme, the secondary green chloroplast found in the dinoflagellate *Lepidodinium* appears to have originated from a pedinophyte lineage (Kamikawa *et al.*, 2015; Jackson *et al.*, 2018). Cells swim with their single emergent flagellum trailing backwards, curving around the cell (Karpov & Tanichev, 1992; Jones *et al.*, 1994; Marin, 2012). Pedinophyte morphology varies greatly, and they have been described in a variety of environments from freshwater, to marine, to hyperhaline (Karpov & Tanichev, 1992; Jones *et al.*, 1994). The class Pedinophyceae was originally erected by Moestrup (1991). Marin (2012) resolved the Pedinophyceae in phylogenies of nuclear and chloroplast-encoded rRNA operons as an independent class, sister to the core Chlorophyta.

The position of pedinophytes as sister to the rest of the core Chlorophyta means that they are well placed to examine the evolution of the core Chlorophyta, including the gene family evolution that occurred as this group diverged. A draft nuclear genome of unidentified pedinophyte YPF701 had already been assembled in the Verbruggen lab, from short-read

data that was generated as part of work exploring the evolutionary origins of secondary chloroplasts (Jackson *et al.*, 2018). My project aimed to improve this genome by creating a hybrid assembly, with the addition of long-read sequencing data, in order to create a more contiguous and complete assembly which can be used to examine gene family evolution as well as the nature of noncoding regions in this lineage.

Methods

Sequence data available at start of project

Short Illumina sequencing reads for YPF701 as well as an assembly of these short reads made using the program SPAdes were already available in the Verbruggen Lab.

Culturing, DNA extraction, sequencing and hybrid assembly of pedinophyte YPF-701

I cultured the pedinophyte strain YPF701 (NIES Microbial Culture Collection strain NIES-2566) in K- enriched seawater medium (Keller *et al.*, 1987) at 20 °C on a 10:14 hour light:dark cycle. To reduce bacterial load, cultures were treated one week prior to extraction with antibiotics (cefotaxime 0.72mg/mL, carbenicillin 0.72mg/mL, kanamycin 0.03mg/mL and amoxicillin 0.03mg/mL). Cells were harvested by centrifugation (10 min, 3,000g). Total genomic DNA was extracted using a modified CTAB protocol, in which the CTAB extraction buffer was added directly to the cell pellets (see Supplementary Methods in Appendix) (Cremen *et al.*, 2016).

DNA was quantified with a Qubit Flurometer (Invitrogen, Waltham, MA, USA) as 108.528μg, and contamination was assessed with a NanoDrop Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) (260/280: 1.92, 260/230: 1.50).

Long read sequencing data generation and initial handling was performed by members of Dr Kathryn Holt's Lab, formerly at Bio21 Institute. Library preparation and long read sequencing on a minION sequencer (Oxford Nanopore Technologies, Oxford, UK) was performed by Dr Louise Judd, while Ryan Wick performed initial adaptor removal and quality and size filtering.

Nanopore reads were assembled with the previously sequenced Illumina short reads into a hybrid assembly with the program MaSuRCa 3.2.8 (Zimin *et al.*, 2013), using default settings and setting JF_SIZE = 200000000.

Comparison of hybrid and short-read assemblies

I compared our new hybrid genome assembly with the existing SPAdes assembly of purely short-read data. Genome completeness was assessed using BUSCO with the Eukaryota dataset, which uses a set of single-copy genes found in 90% of included species to estimate completeness of genomes for expected gene content, with the assumption that these genes are present (Waterhouse *et al.*, 2017). Comparison of the two assemblies was performed in QUAST 5.0.2 (Mikheenko *et al.*, 2018), a tool which estimates various metrics including N50, with lower threshold for contig length set at 1000 bp and the eukaryotic genome flag. The length of scaffolds was obtained by loading the genomes into Geneious version 11.1.2 (Kearse *et al.*, 2012) and examining summary statistics.

Comparative genome analysis in Pico-Plaza

During the culturing, DNA extraction, sequencing and assembly of the hybrid nuclear genome, comparative genome analyses were performed using the short read assembly on a custom version of Pico-Plaza (Vandepoele *et al.*, 2013), an online genome database and integrative evolutionary sequence analysis tool, which was built containing genomes and annotations of 23 Chloroplastida species (Fig. 3).

Highly conserved single gene families from TribeMCL (Enright *et al.*, 2002), present in all 23 species, were used for phylogenetic analysis. An unedited concatenated alignment of these 47 single copy genes (41,020 amino acid positions), created in Geneious using MAFFT, was used to construct a phylogenetic tree of the inferred species topology with RAxML version 8.2.10 (Stamatakis, 2014) (model PROTGAMMAWAG, 100 bootstraps).

The phylogenetic profile of TribeMCL gene families (excluding orphans, gene families with only one copy in one species) was retrieved from Pico-Plaza and converted into phylip format using Mesquite version 3.6 (Maddison & Maddison, 2018). This file and the inferred species tree topology were used to reconstruct the most parsimonious gain and loss scenario for every gene family using the Dollop program from PHYLIP version 3.695 (Felsenstein, 2005), with the Dollo parsimony method and printing of states at all nodes of the tree. Further processing

with Orthomcl Tools (DOI 10.5281/zendo.51349) allowed this output to be mapped onto the inferred species tree in R version 3.5.3 (R core Team, 2019) using the packages 'ape' (Paradis & Schliep, 2018), 'RColorBrewer' (Brewer, 2019) and 'ggtree' (Yu *et al.*, 2017). Genome size estimates were mapped onto a subset of the phylogenetic tree with only the Chlorophyta in R version 3.5.3 with the package 'phytools' (Revell, 2012).

Results

The new hybrid genome assembly for pedinophyte YPF701 comprises 1877 scaffolds with a total length of 34,071,101 bp. This genome has not yet been annotated or filtered to remove contamination and organellar genomes. I compared the new hybrid assembly with the existing short-read assembly and found that GC content differs slightly between the two assemblies, with the hybrid assembly containing some regions of lower %GC (Table 1, Fig. S1). When examining only scaffolds greater than 1000 bp, in order to compare the hybrid assembly with the short-read assembly that was set to retain only scaffolds greater than this length, the hybrid assembly contains fewer, longer scaffolds (Table 1, S1). The mean and maximum scaffold lengths are greater in the hybrid assembly, as is the N50 (Table 1, Fig. S2, Table S1).

Table 1 Comparison of the hybrid and short-read assemblies of the pedinophyte YPF701 nuclear genome.

	Hybrid	**Short-read**
Total length (bp)	34,071,101	26,770,386
GC (%)	66.91	69.90
N50	1,083,765	35,582
Number of scaffolds (>1000 bp)	1,877 (257)	1,597 (1,597)
Mean scaffold length (bp)	1,8151.9	16,762.9
Minimum scaffold length (bp)	304	1,001
Maximum scaffold length (bp)	2,736,781	216,425

Analysis of genome completeness indicates that the hybrid assembly captures at least 86% of the eukaryotic BUSCO dataset. The short-read assembly is similar, and contains fewer duplicated and more fragmented BUSCOs (Fig. 2).

The predicted size of the pedinophyte YPF701 genome is relatively small, larger than the prasinophyte lineages but smaller than most of the core Chlorophyta (Fig. 4).

A concatenated alignment of highly conserved single gene families resolved the position of YPF701 as a member of the core Chlorophyta with high confidence (high bootstrap values). However, it failed to resolve the relationships between the core Chlorophyta classes, as well as the placement of *Ulva mutabilis* (Fig. 3). The number of genes and gene families in YPF701, predicted using the short-read assembly, was less than many of the core Chlorophyta except for members of the Trebouxiophyceae, but greater than the sequenced prasinophytes apart from the genus *Micromonas*. Considerable gene family loss was predicted along the pedinophyte branch (Fig. 3).

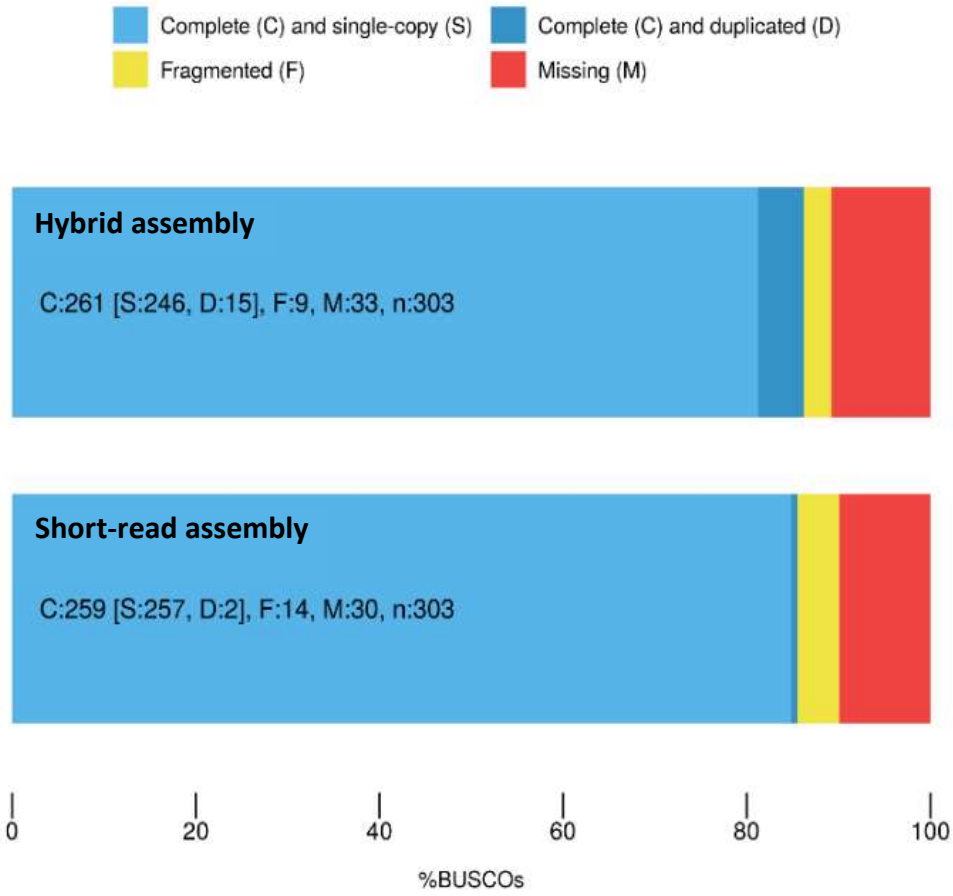
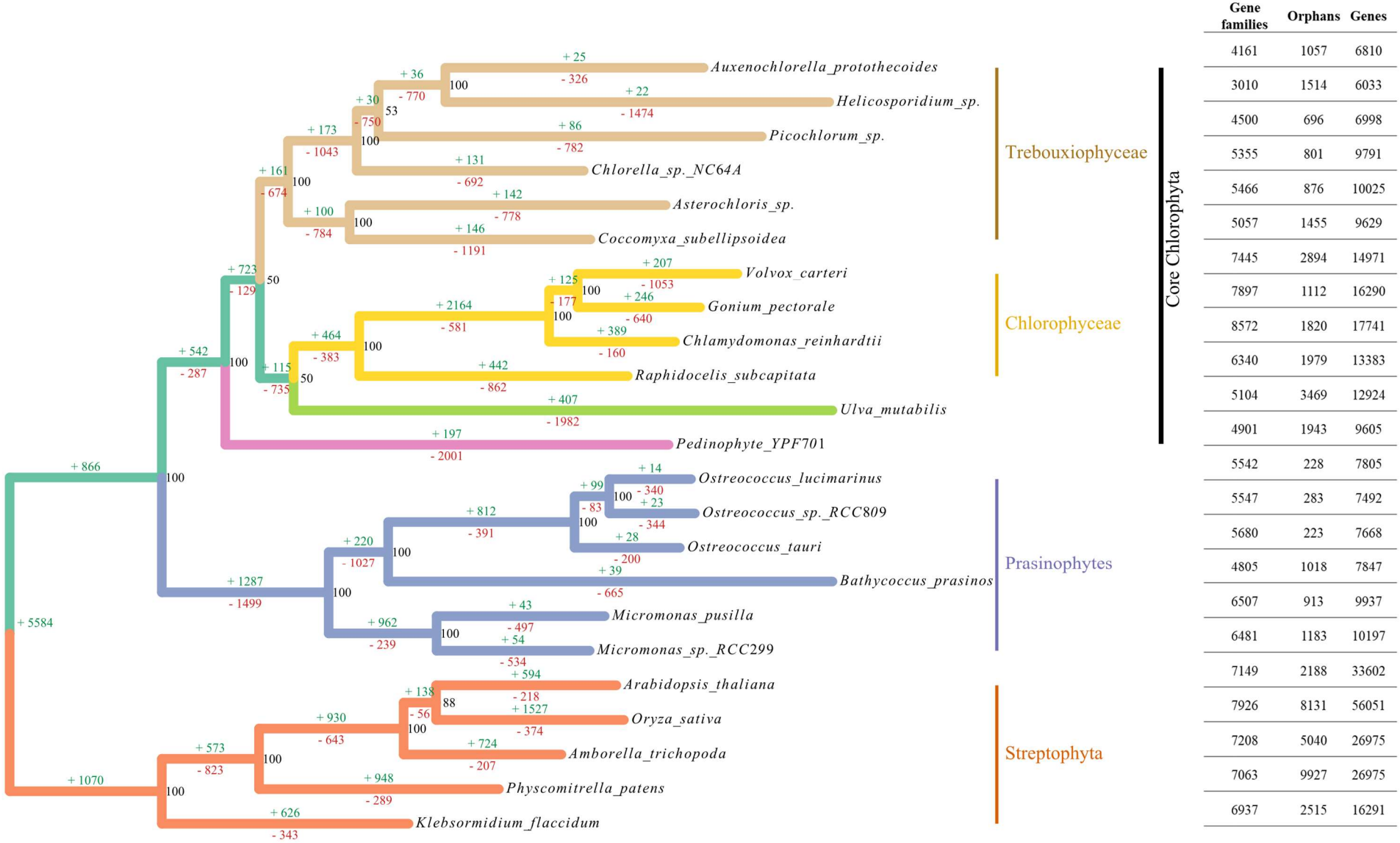


Fig. 2. – BUSCO assessment results for the hybrid and short-read genome assemblies, representing the number of sequences in the BUSCO Eukaryotic dataset (total 303) identified in the assemblies. The hybrid assembly contains more duplicated and fewer fragmented BUSCOs.



Gene families	Orphans	Genes
4161	1057	6810
3010	1514	6033
4500	696	6998
5355	801	9791
5466	876	10025
5057	1455	9629
7445	2894	14971
7897	1112	16290
8572	1820	17741
6340	1979	13383
5104	3469	12924
4901	1943	9605
5542	228	7805
5547	283	7492
5680	223	7668
4805	1018	7847
6507	913	9937
6481	1183	10197
7149	2188	33602
7926	8131	56051
7208	5040	26975
7063	9927	26975
6937	2515	16291

Fig. 3. – Phylogenetic tree of green algae and land plants including predicted pattern of gain and loss of gene families during evolution. Maximum likelihood bootstrap values are indicated in black at each node. The number of gene families acquired (in green) or lost (in red), indicated along each branch in the tree were estimated using the Dollo parsimony principle. The number of gene families, orphans (single-copy gene families in a single species) and predicted genes are indicated for each species.

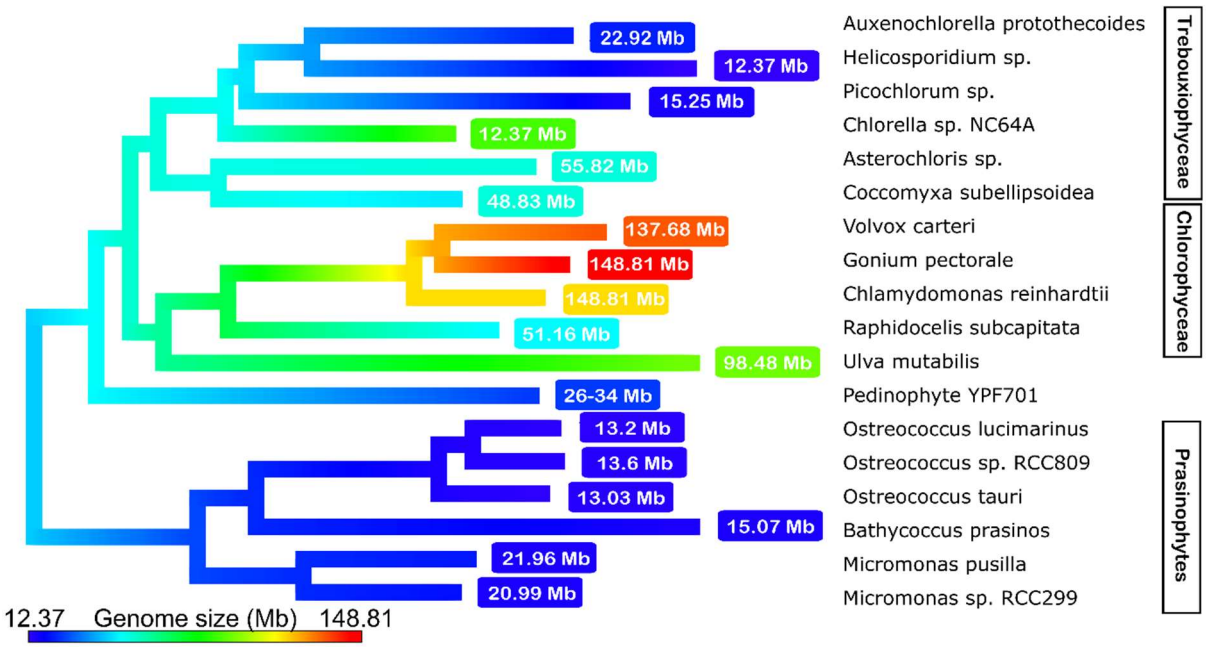


Fig. 4. – Genome sizes mapped onto a phylogenetic tree of sequenced Chlorophyta. The Pedinophyte genome size is small compared with most core Chlorophyta.

Discussion

The YPF701 nuclear genome assemblies capture at least 86% of the eukaryotic BUSCO dataset. This is similar to other recently sequenced Chlorophyta nuclear genomes, such as *C. lentillifera* (86.4%) (Arimoto *et al.*, 2019) and *U. mutabilis* (92%) (De Clerck *et al.*, 2018), and suggests that they are both reasonably complete assemblies capturing most of the coding content of the pedinophyte nuclear genome. Compared with the short-read assembly, the hybrid assembly includes longer contiguous scaffolds with fewer fragmented predicted BUSCOs. This is a benefit of the long-read sequences which can help resolve repetitive regions and complex genomic features such as transposable elements, high copy genes and duplications (Treangen & Salzberg, 2012; Goodwin *et al.*, 2015; Koren & Phillippy, 2015).

The GC content between the two assemblies differs. This might be due to the presence of bacterial contamination in the unfiltered hybrid assembly, which contains some low %GC regions not inconsistent with levels found in bacterial genomes (Hildebrand *et al.*, 2010).

Comparison of the short-read assembly with those of other sequenced Chlorophyta shows that the genome is relatively small. It is larger than the prasinophytes but smaller than most of the core Chlorophyta except some members of the Trebouxiophyceae that appear have undergone genome compaction following their divergence from other members of the core Chlorophyta (Pombert *et al.*, 2014; Foflonker *et al.*, 2015) (see General Introduction). As

seen in many prasinophytes, small genome size in YPF701 accompanies a small cell size that likely reflects specialisation to a nanoplanktonic lifestyle as a way to reduce competition (Marin, 2012). The pedinophyte lineage appears to have undergone considerable gene family loss. The small nuclear genome of YPF701 likely also reflects reduction in noncoding DNA, as is seen in the prasinophytes (Derelle *et al.*, 2006; Moreau *et al.*, 2012). This is difficult to discern from assembly statistics alone as the often-repetitive noncoding content tends to be more difficult to assemble, although it is hoped that the use of long-read sequencing will overcome some of these issues (Goodwin *et al.*, 2015). YPF701 may have undergone positive selection for genome streamlining (Giovannoni *et al.*, 2005). As it is unicellular and many of the sequenced core Chlorophyta are colonial and multicellular, YPF701 may have a larger N_e, also increasing the power of purifying selection against noncoding DNA relative to genetic drift (Lynch *et al.*, 2006; Smith, 2016). The sequenced chloroplast genomes of numerous pedinophytes, including YPF701, are relatively small and entirely lack introns (Marin, 2012; Jackson *et al.*, 2018). The mitochondrial genome of *Pedinomonas* contains only a single intron (Turmel *et al.*, 1999). These compact organellar genomes, although representing different genomic compartments, support the idea that stronger selection is acting in this class, potentially on all three genomes. Once it is annotated, the more contiguous hybrid nuclear genome assembly for YPF701 might help us to examine the noncoding regions in the genome and estimate coding density to determine the extent to which streamlining has occurred.

The hybrid genome assembly for YPF701 can also be used in a more thorough comparative genomics analysis. Increasing the Ulvophyceae genomes included, with the recently sequenced *Caulerpa lentillifera* genome (Arimoto *et al.*, 2019) and the genome for *Ostreobium quekettii* currently under preparation in the Verbruggen lab, will provide additional genomic information enabling a more comprehensive exploration of the evolution of Chlorophyta nuclear genomes and hopefully better phylogenetic resolution of the relationships between the core Chlorophyta classes. After prediction of gene family gains and losses, Gene Ontology terms and InterPro domains can be used to examine the functions of novel genes and gene families, such as the 542 gene families predicted in this study to have been gained at the base of the core Chlorophyta, as well as for gene families that have experienced expansion or contraction of gene numbers in select lineages of the Chlorophyta. Comparative genome analysis of the draft genome of the streptophyte alga *Chara braunii* and

related lineages revealed features important for the colonisation of land that evolved prior to the diversification of land plants (Nishiyama *et al.*, 2018). A similar comparative genomics study of the Chlorophyta investigating the shared and unique features of the core Chlorophyta and prasinophytes would be a pertinent use of our pedinophyte genome data.

CHAPTER 2
An uneconomical genome: The complete mitochondrial genome sequence of *Ostreobium quekettii* SAG6.99

Introduction

The Ulvophyceae have garnered interest for evolutionary studies due to the high morphological diversity within the class, including a range of cell types (Cocquyt *et al.*, 2010; Fang *et al.*, 2017). This includes the order Bryopsidales, a lineage of siphonous seaweeds with thali comprised of a single giant tubular cell containing cytoplasm, with many nuclei and other organelles, free to move around the entire plant (Fig. 5) (Vroom & Smith, 2001; Vroom & Smith, 2003; Verbruggen *et al.*, 2009; Mine *et al.*, 2015).

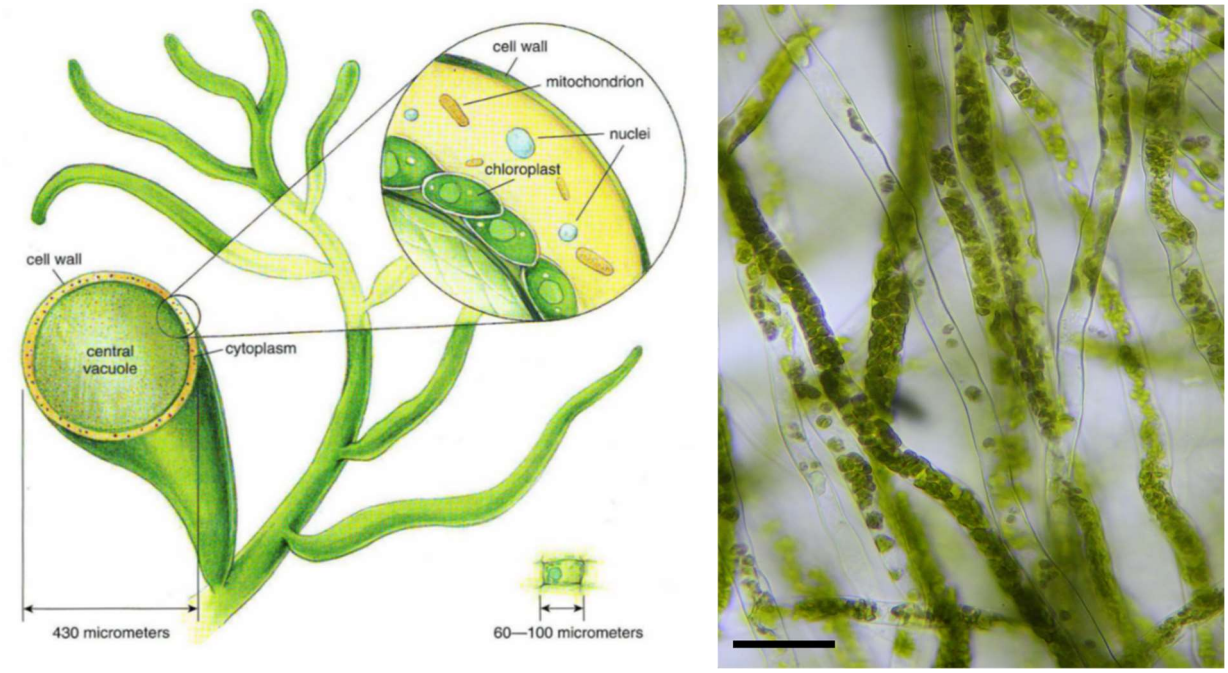


Fig. 5. – Siphonous green algae. Left: Cross section of a typical siphonous green alga, compared with a spinach leaf (from Vroom & Smith, 2001). Right: The siphons of *Ostreobium*, featuring cytoplasmic contraction of damaged siphons (green spheres). Scale bar = 25µm.

Greater effort has been made to sequence and characterise chloroplast genomes rather than mitochondrial in the Bryopsidales, which is the case for many plastid-bearing taxa (Smith & Keeling, 2015; Fang *et al.*, 2017). Within the Bryopsidales, the size and gene arrangement of chloroplast genomes varies considerably, with coding content remaining relatively consistent yet differing amounts of noncoding content including introns and intergenic DNA (Cremen *et al.*, 2018), similar to what is seen in the Volvocales (see General Introduction).

It was only last year that the first mitochondrial genome was published for a member of the Bryopsidales. Zheng *et al.* (2018) sequenced the circular mtDNA of the sea grape *Caulerpa lentillifera*, the largest green algal mitochondrial genome sequenced thus far at 209,034 bp. This is an order of magnitude larger than the shortest sequenced Chlorophyta mitochondrial genome: the 13 Kb linear genome of *Polytomella capuana,* a colourless green alga in the Volvocales (Smith & Lee, 2007). This genome expansion was mostly from an increase in non-coding DNA: intergenic sequences and introns (Zheng *et al.*, 2018).The *C. lentillifera* mitochondrial genome is considerably larger than the chloroplast genome previously reported for this lineage (Genbank accession MG753774.1).

Another member of the Bryopsidales currently of considerable ecological interest is the genus *Ostreobium*, an endolithic, or boring, alga in the suborder Ostreobineae (Fig. 5) (Verbruggen *et al.*, 2017). *Ostreobium* is present in a diverse range of calcium carbonate environments and is one of the most common genera of boring autotrophs in coral reefs (Tribollet, 2008). Its endolithic lifestyle means that *Ostreobium* inhabits environments, such as the coral skeleton, that are limited in available photosynthetically active radiation (Fig. 6) (Wilhelm & Jakob, 2006; Magnusson *et al.*, 2007). As a result, Ostreobium is optimised to absorb the low wavelengths of light that are available (Wilhelm & Jakob, 2006; Magnusson *et al.*, 2007; Tribollet, 2008).

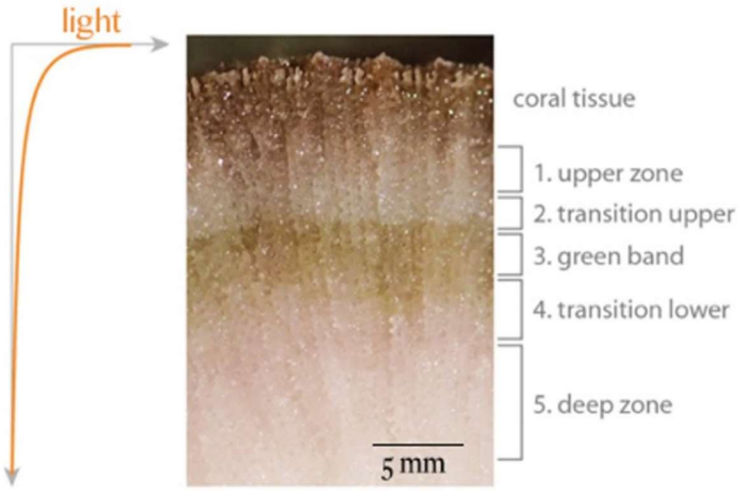


Fig. 6. – Cross section of coral showing the typical location of *Ostreobium* within the skeleton (green band). Left: schematic of light availability moving through the coral skeleton (Heroen Verbruggen).

The Ostreobineae have consistently small chloroplast genomes relative to the median for Bryopsidales of 105 Kb (Cremen *et al.*, 2018), with the chloroplast genome of *Ostreobium* sp. HV05042 the most compact found so far (80,584 bp) in the Ulvophyceae (Marcelino *et al.*, 2016; Verbruggen *et al.*, 2017). Marcelino *et al.* (2016) identified only three introns in the *O. quekettii* chloroplast genome, and there was an overall reduction in intergenic regions that resulted in its reduced size. They hypothesise that this might be due to energy limitation in

the low light environment selecting for a smaller chloroplast genome which costs less energy to be transcribed and translated (Marcelino *et al.*, 2016).

The mitochondrial genome of *Ostreobium quekettii* SAG6.99 was sequenced as part of an ongoing *Ostreobium* nuclear genome sequencing project in the Verbruggen Lab. The aim of my project was to combine long and short read sequencing data to assemble the genome, annotate and then compare it with other mitochondrial genomes published for the Chlorophyta including *Caulerpa lentillifera.* I aimed to look for evidence of selection also acting upon the mitochondrial genome, as proposed by Marcelino *et al.* (2016) to explain the reduced size of the chloroplast genome.

Methods

Sequence data available at start of project

Short and long sequencing reads for *O. quekettii* SAG6.99 had already been generated by the Verbruggen lab and were available for use. An assembly of the long reads by the program Canu, an assembly of the short reads by the program SPAdes, and an annotated hybrid assembly generated with MaSuRCa, as well as a transcriptome from RNA sequencing data, also contributed to this project.

Identification and curation of mitochondrial genome

I used the *Caulerpa lentillifera* mitochondrial genome (Genbank accession KX761577.1) (Zheng *et al.*, 2018) as the query in a BLASTn search against the long-read assembly within Geneious version 11.1.2 (Kearse *et al.*, 2012) with default settings. Only a single contig was identified as a likely candidate for the *O. quekettii* mitochondrial genome. Sections of this candidate contig were then used as queries in BLASTx searches against the NCBI nr and nt databases to confirm that the results were mitochondrial genes in other Chlorophyta, thereby confirming it represented the mitochondrial genome.

I used this long-read contig as the query in a BLASTn search against the *O. quekettii* short-read assembly, within Geneious with default settings. I aligned top hits with the long-read contig using a combination of the Geneious aligner, MAFFT and consensus align (with MAFFT), as well as manual curation. Scaffolds from the short-read assembly were used as a reference to manually correct the long-read contig within Geneious. In order to verify

circularity of the genome, I searched for a short-read scaffold which spanned the two overlapping ends of the long-read contig when aligned.

Genome annotation

I annotated the genome using MFannot (Beck & Lang, 2010) and DOGMA (Wyman *et al.*, 2004) with very relaxed settings (protein identity cut off 25%, RNA identity cut off 30%). I confirmed annotations of predicted protein coding genes through extraction of open reading frames and BLAST searches of these against the NCBI nr and nt databases, as well as alignment with transcripts that were recovered as hits from BLASTn searches against the *O. quekettii* transcriptome within Geneious, using default settings.

I identified tRNAs using tRNAscan-SE (Lowe & Chan, 2016), tRNAfinder (Kinouchi & Kurokawa, 2006), ARAGORN (Laslett & Canback, 2004) and tRNADB-CE's BLASTN/Pattern Search (Abe *et al.*, 2010). rRNAs were identified with RNAmmer (Lagesen *et al.*, 2007) and RNAweasel (Lang *et al.*, 2007).

I created a map of the genome with Circos (Krzywinski *et al.*, 2009), and annotation in Inkscape 0.92 (www.inkscape.org).

Open reading frames

Free standing open reading frames (ORFs) were predicted using ORF finder in Geneious with a minimum length of 300 bp. These were used as queries in BLASTx searches against the NCBI nr and nt databases (e value = e^{-1}), and the translated ORFs were used as queries in a batch sequence search against the Pfam database (Finn *et al.*, 2016). Only ORFs that had valid BLAST results and identified Pfam domains were retained in the final genome annotation. These ORFs, as well as ORFs from the chloroplast genome (Marcelino *et al.*, 2016) of *O. quekettii* and mitochondrial genome of *C. lentillifera,* were clustered based on all-against-all BLAST+ similarities using CLANs (Frickey & Lupas, 2004) in order to determine relationships indicative of common origins. This was performed through the MPI Bioinformatics toolkit (Zimmermann *et al.*, 2018), with BLOSSUM62 scoring matrix and extraction of BLAST HSPs up to e-values of 1e-4. The output from CLANs was annotated in Inkscape.

Introns

Transcriptome sequences are expected to have introns spliced out in most cases, so alignment of predicted genes with transcripts – identified from BLAST searches against the transcriptome – as well as intron-lacking homologous green algal mitochondrial genes and proteins, allowed me to infer the presence of introns. Intron class was predicted using RNAweasel and Rfam sequence search (Griffiths-Jones *et al.*, 2003).

In order to identify if there were introns that showed a common origin with other introns in the *O. quekettii* mitochondrial genome and/or the *O. quekettii* plastid genome and the mitochondrial genome of *C. lentillifera,* a distance matrix was constructed by comparing the introns, that were not disrupted by ORFs, from these genomes using Clustal Omega (Sievers *et al.*, 2011) with the '--distmat-out' and '--full' flags. This distance matrix was used as the input to construct a neighbour joining tree using Neighbor within the PHYLIP package (Felsenstein, 2005). I also constructed a distance matrix using only the ORF-lacking introns in the *O. quekettii* chloroplast and mitochondrial genomes and this was used to construct a neighbour joining tree. This neighbour joining tree was further visualised and annotated in MEGA (Kumar *et al.*, 2018). Clusters of introns identified from the *O. quekettii* intron neighbour joining tree were aligned in Geneious using MAFFT.

Repeats

I searched for repeats in the genome using the tandem repeats database (Gelfand *et al.*, 2006), the RepeatFinder package in Geneious with a minimum repeat length of 50 bp or more (as per Pombert *et al.*, 2004), and REPuter (Kurtz *et al.*, 2001) with minimal repeat size setting of 12 bp (as per Smith & Lee, 2009). Forward, reverse, complement, and reverse complement repeats were all considered under REPuter.

Rates of evolution

To obtain an estimate of the relative rates of evolution in the *O. quekettii* mitochondrial and chloroplast genomes, I aligned all the protein coding and rRNA genes common between the *O. quekettii* and *C. lentillifera* mitochondrial and chloroplast genomes in Geneious using the default aligner. I generated estimates of base substitutions per site between sequences using the Jukes-Cantor model (Jukes & Cantor, 1969) in MEGA. Using PAL2NAL (Suyama *et al.*, 2006), I also attempted to obtain d_N/d_S ratios, a ratio of distances equal to substitution rates

multiplied by divergence time or phylogeny branch length that can be used as an estimate of selection efficiency, with lower values indicating stronger purifying selection (Neiman & Taylor, 2009).

Mitochondrion-targeted genes

I searched in *O. quekettii* for homologues of nuclear-encoded genes that are targeted to the mitochondrion in plants, encoding DNA repair machinery (MSH1, RECA proteins and OSB1), and the transporter TatB. *Arabidopsis* sequences were used as queries in BLAST searches with default settings against the draft nuclear genome of *O. quekettii* in Pico-Plaza as well as the *O. quekettii* transcriptome in Geneious. I searched for putative targeting signals to the mitochondrion with DeepLoc-1.0 using default settings (Almagro Armenteros *et al.*, 2017).

Results

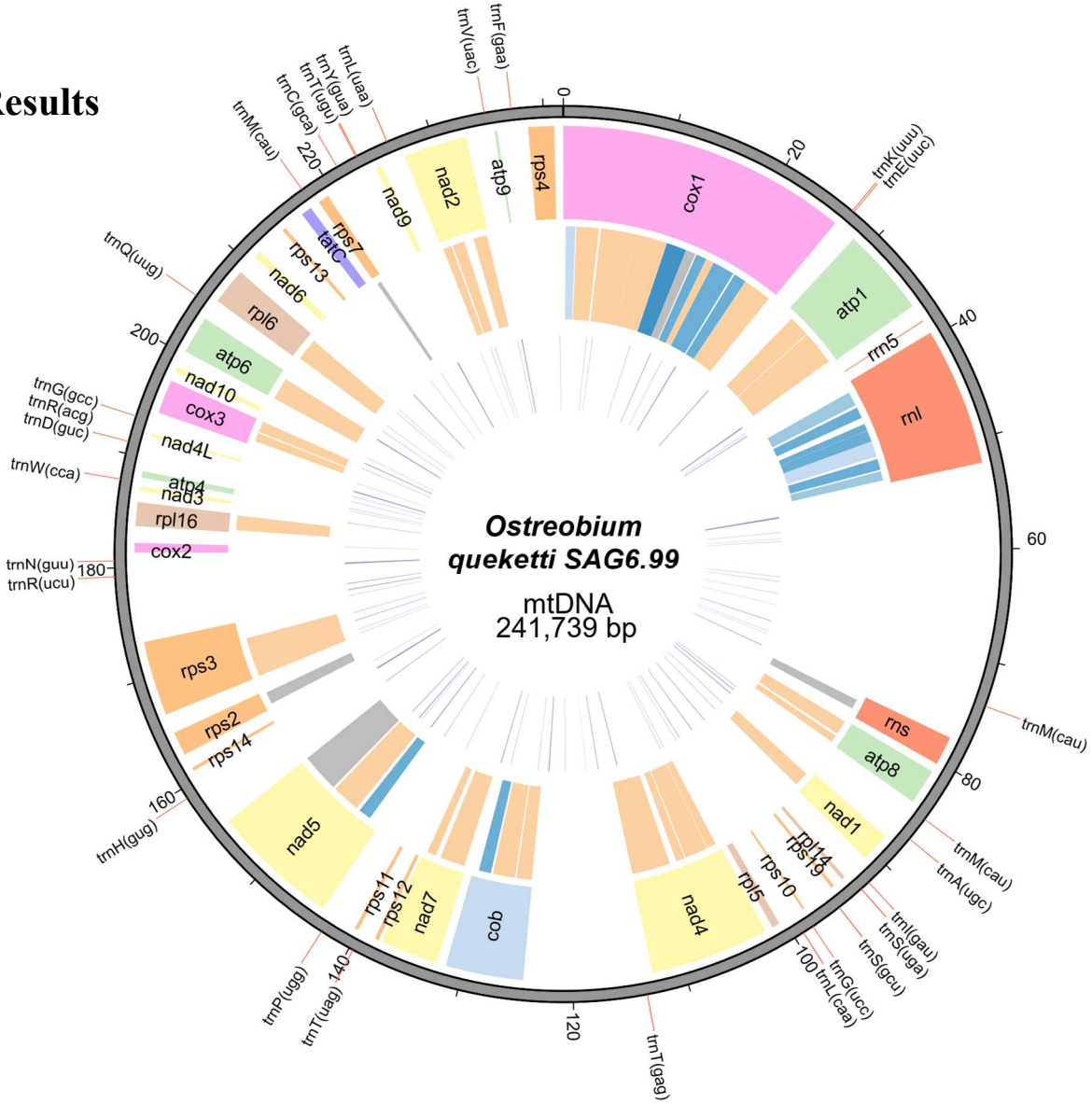


Fig. 7. – Mitochondrial genome map of *Ostreobium quekettii* SAG6.99. The position of tRNAs are shown on the outer track (red lines). The first inner circle represents the position, size and the names of the protein-coding and rRNA genes. The introns are shown in the second inner circle and are colour coded according to intron types/subtypes: group I derived (very light blue), group IA (light blue), group IB (blue), group ID (dark blue), group II (orange), unknown (grey). The third inner circle represents the position and length of repeats.

Genome

The mitochondrial genome of *O. quekettii* SAG6.99 assembled into a 241,739 bp circular-mapping molecule (Fig. 7). Most of the genome is noncoding DNA (Table 2). The overall GC content of the genome is 48.3% (Table 2), which is higher than the average for eukaryotic mitochondrial genomes (38%) (Smith & Lee, 2007), as well as that of the *O. quekettii* chloroplast genome (31.9%) (Marcelino *et al.*, 2016). However, it is lower than in *Caulerpa lentillifera* mtDNA (50.9%) (Zheng *et al.*, 2018), and not extreme for green algae (Del Vasto *et al.*, 2015). All 64 codons are used (Table S2) and the 28 tRNAs encoded by the *O. quekettii* mtDNA (Table S3) appear to be sufficient to recognise all of these codons assuming the standard genetic code and maximum use of wobbling and superwobbling (Alkatib *et al.*, 2012).

Gene Content

The genome encodes 3 rRNAs and 28 tRNAs (Table 3, 4), resembling other green algal mitochondrial genomes (Table S4). It also encodes 34 protein coding genes commonly found in green algae, including uncommon *nad10* and *tatC* (Table 3, 4, S4). I identified a putative nuclear-encoded *tatB* gene in *O. quekettii*, similar to a sequence identified in *Arabidopsis*, that was predicted to be targeted to the mitochondrion (DeepLoc-1.0: Mitochondrion 0.5094, Membrane 0.8249)

Repeats

The *O. quekettii* mitochondrial genome contains 373 repeats that represent 5% of the total genome (Fig. 7, Table 2), with a minimum length of 31 and maximum of 299 bp (mean 107.3±61.7SD).

Introns

18 of the 34 protein-genes contain one or multiple intron(s) (Table 3, S5), with as many as 11 in cox1 which comprises 1,578 bp of coding content spread over 26,493 bp of the genome (Fig. 7). Introns include both type I and type II introns, as well as five whose class could not be confidently determined (Table 3, S5). There does not appear to be any strong similarity between introns in the *O. quekettii* and *C. lentillifera* mitochondrial genomes, with introns from the two species mostly forming separate clusters in neighbour joining trees (Fig. S3). Alignments of the few *C. lentillifera* and *O. quekettii* introns that did cluster together did not

show convincing homology (data not shown). Comparison of ORF-lacking introns from the *O. quekettii* mitochondrion and chloroplast genomes identified groups of similar type II introns in the mitochondrial genome, providing some evidence for intron proliferation within the lineage (Fig. S3).

Table 2 Summary of coding and noncoding content of the mtDNA of *Ostreobium quekettii* SAG6.99.

	Length (bp)	**Percent of total noncoding DNA**	**Percent of overall genome**
Genome	241,739		
GC	116,762		48.30%
Coding (rRNA, tRNA, ORFS, protein coding genes)	59,964		25%
Repeats	13,268	7%	5%
Intergenic DNA	110,890	54%	46%
introns (including ORFs)	95,011		39%
introns	70,885	39%	29%
Total noncoding DNA (excluding intron encoded ORFs)	181,775		75%
Total intronic and intergenic DNA (including intron encoded ORFs)	205,901		85%

Table 3 Genes, introns and open reading frames present in mtDNA of *Ostreobium quekettii* SAG6.99. For further information on introns and ORFs, see supplementary material.

	Number in genome
Protein coding genes	34
rRNA	3
tRNA	28
Genes containing introns	18
Introns	47
Type I	14
Type 2	28
unclear	5
Introns containing ORFs	18
ORFS	20
Intronic	20
Intergenic	0

Table 4 Protein coding and ribosomal genes present in the mtDNA of *Ostreobium quekettii* SAG6.99.

Protein genes	
Complex I (nad)	nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9, nad10
Complex III (cob)	cob
Complex IV (cox)	cox1, cox2, cox3
Complex V (atp)	atp1, atp4, atp6, atp8, atp9
SSU ribosomal proteins (rps)	rps2, rps3, rps4, rps7, rps10, rps11, rps12, rps13, rps14, rps19
LSU ribosomal proteins (rpl)	rpl5, rpl6, rpl14, rpl16
Ribosomal RNAs	5s, 16s, 23s
Putative Protein Transporter	tatC

Intron-encoded open reading frames

ORFs with identified Pfam domains were only found in introns, not intergenic DNA (Table 3). The introns contain a variety of ORFs with domains that have predicted functions related to maintenance and proliferation of introns (Fig. 8). Only one ORF identified by Zheng *et al.* (2018) in *C. lentillifera* had a positive hit to the Pfam database: ORF932 in cox1 had a putative LAGLIDADG_1 domain. Clustering analyses of ORFs revealed similarity between ORFs containing the same Pfam domains. This included ORFs with a single LAGLIDADG_1 domain, including *C. lentillifera* ORF932, and ORFS with two LAGLIDADG_1 domains (Fig. 8).

Rates of evolution

Estimates of base substitutions per site between genes from *O. quekettii* and *C. lentillifera* showed a slightly higher ($t=2.94$, $p<0.01$) number of average base substitutions per site for the mitochondrial genomes (mean 0.456±0.207SE) compared with chloroplast genomes (mean 0.370±0.016SE) (Fig. 9, Table S6). I was unable to calculate d_N/d_S ratios as genes showed signs of saturated divergence, with estimated d_S values considerably greater than one.

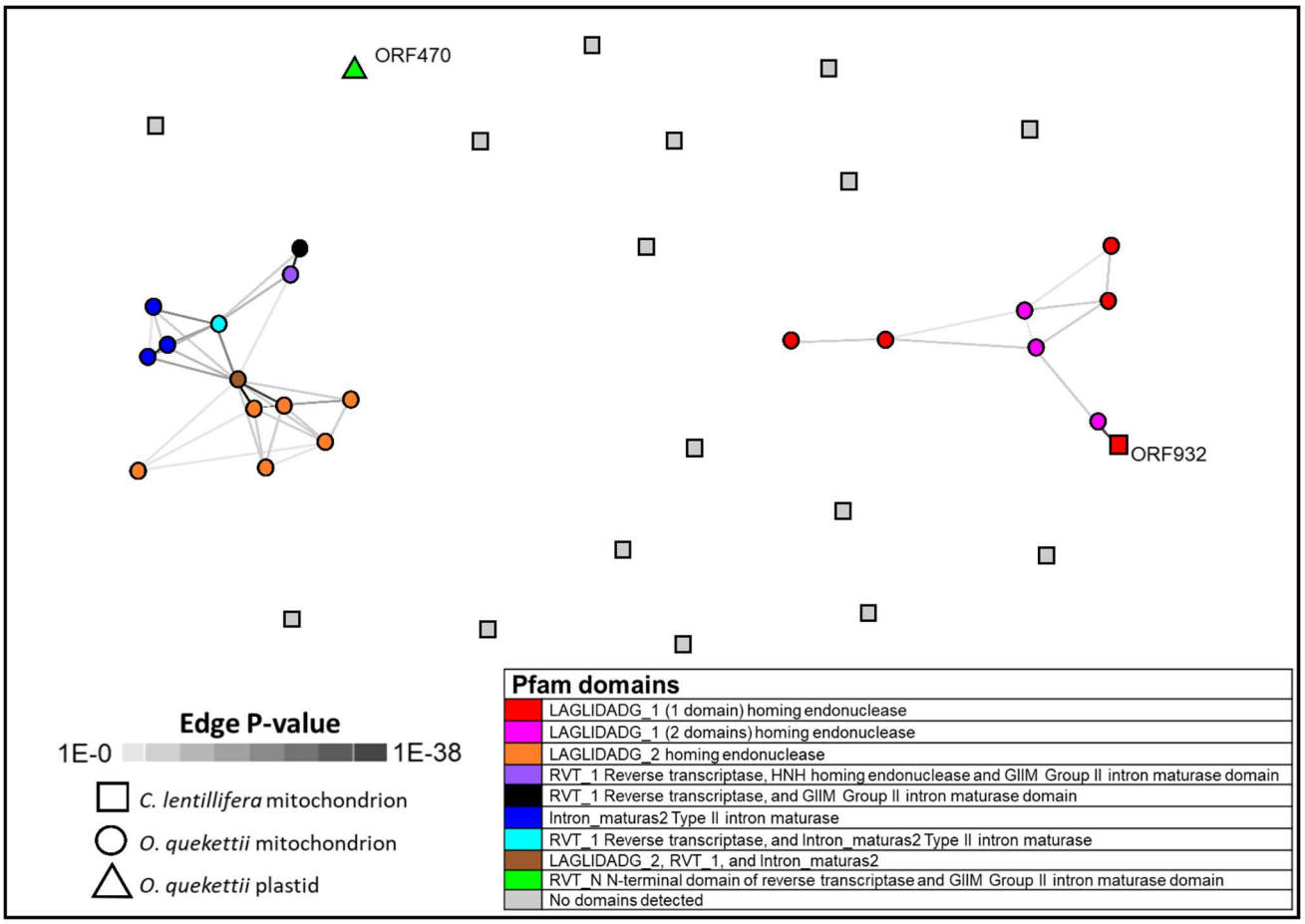


Fig. 8. Similarity network generated from all against all BLAST+ similarities of ORFs encoded in the *C. lentillifera* mitochondrion, and *O. quekettii* mitochondrion and chloroplast. Each node represents an ORF, and each edge (line) represents a significant HSP (high scoring segment pair), shaded according to p value. Generated using CLANS through the MPI Bioinformatics Toolkit (Scoring Matrix BLOSSUM62, extracting BLAST HSPs up to E-values of 1e-4, using p-values better than 1.0).

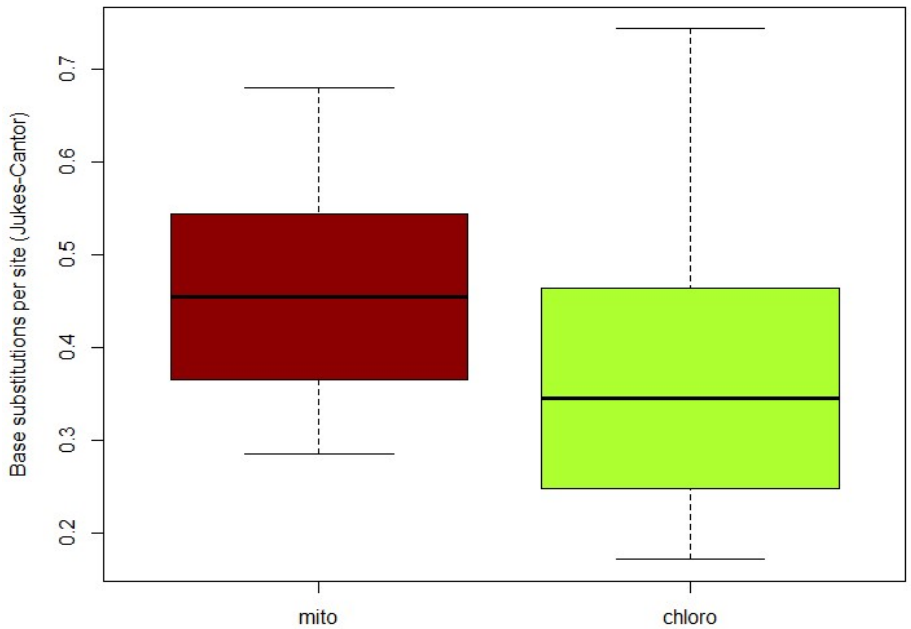


Fig. 9. Boxplots of base substitutions per site between protein coding and rRNA genes in the mitochondrion (mito) and chloroplast (chloro) genomes of *O. quekettii* and *C. lentillifera*, generated assuming the Jukes-Cantor model.

Recombination-associated repair machinery

I identified a sequence in the *O. quekettii* nuclear genome that encodes a predicted protein with sequence similarity to DNA mismatch repair protein MSH1. This sequence has a putative targeting signal to the mitochondrion (DeepLoc-1.0L: Mitochondrion 0.7151, Soluble 0.6145). Pfam and InterPro searches identified a putative specific DNA-binding GIY-YIG domain in this MSH1 homologue, as well as in potential homologues in other green algae. However, this domain lacks most of the key residues conserved among GIY-YIG family members (Garrison & Arrizabalaga, 2009). I also identified putative mitochondrion-targeted (DeepLoc-1.0L: Mitochondrion 0.9268 and 0.6334, Soluble 0.5699 and 0.6798) homologues of RECA proteins, which are also predicted to play a role in controlling mitochondrial genome maintenance in plants. Searches of the *O. quekettii* transcriptome revealed potential OSB1 homologues, predicted to be involved in homologous recombination-dependent repair, containing a central OB-fold domain but lacking a targeting signal to the mitochondrion or chloroplast (DeepLoc-1.0L: Nucleus 0.6485, Soluble 0.7227).

Discussion

The *O. quekettii* mitochondrial genome encodes all the genes commonly found in Chlorophyta mitochondrial genomes, and a majority of the ribosomal protein genes, which have been more unevenly retained across plant and algal mitochondrial genomes (Palmer *et al.*, 2000; Mower *et al.*, 2012). The genome does include some genes that are less common in Chlorophyta mitochondrial genomes. This includes *nad10* which is absent from the mitochondria of sequenced land plants and many green algae, shown to be the result of multiple independent transfers of this gene to the nucleus over evolutionary time (Mower *et al.*, 2012). The genome also retains a copy of *tatC*, a gene encoding a component of the inner membrane TAT translocase, responsible for transporting folded proteins across the membrane in bacteria but whose function in mitochondria remains unclear (Carrie *et al.*, 2016; Petrů *et al.*, 2018). *tatC* has a single alphaproteobacterial origin but has been unevenly retained across eukaryote mitochondrial genomes. It appears to have been lost at least 21 times across eukaryotes (Petrů *et al.*, 2018). The most common eukaryotic TAT is TatC encoded in the mitochondrion (Petrů *et al.*, 2018), however plant nuclear genomes also encode a TatB-like subunit, providing some evidence for a functioning Tat pathway in plant mitochondria (Carrie *et al.*, 2016). This TatB-like subunit has also been identified in some

green algae (Carrie *et al.*, 2016), and this study identified a putative *tatB* in the nuclear genome of *Ostreobium quekettii*, suggesting that there might be an active TAT pathway.

Nevertheless, most of the genome's inflated size relative to other Chlorophyta can be accounted for by expanded noncoding intergenic and intronic regions rather than extra coding material. The genome contains 47 introns, compared with only 29 in the mtDNA of *C. lentillifera* (Zheng *et al.*, 2018), including both type I and type II introns, sometimes within the same gene. Type II introns, which are found in plant mitochondrial genomes but are less common elsewhere (Lang *et al.*, 2007), are the dominant type in the *O. quekettii* mitochondrial genome. This contrasts with the *C. lentillifera* mitochondrial genome, where type I introns predominate (Zheng *et al.*, 2018). Introns from the two genomes do not show sequence similarity. Clustering analysis did show sequence similarity between ORF-lacking type II introns within the *O. quekettii* mitochondrial genome, suggesting there has been proliferation of at least type II introns within this lineage, but did not show similarity between the introns of the *O. quekettii* mitochondrial and chloroplast genomes.

Eighteen of the introns in the *O. quekettii* mitochondrial genome contain one or more ORFs. These ORFs show homology to intron-encoded proteins that act as maturases and homing endonucleases which enable splicing and promote intron mobility (Lambowitz & Belfort, 1993; Lambowitz & Zimmerly, 2011; Hausner, 2012). They contain a variety of domains, but mostly domains with the amino-acid motif LAGLIDADG, which are common in both group I and group II introns (Hausner, 2012). Three ORFs contain double LAGLIDADG domains, which in other lineages are in intron-encoded proteins with maturase activity that enable intron splicing (Lambowitz & Belfort, 1993). Four ORFs contain an RVT domain; Cremen *et al.* (2018) found evidence for the mobility of group II intron encoded ORFs containing an RVT domain within Bryopsidales chloroplast genomes.

None of the *O. quekettii* mitochondrial ORFs show similarity to the single ORF460 identified in the *O. quekettii* chloroplast which has a putative intron splicing function (Cremen *et al.*, 2018). Therefore, there is no evidence for transfer of ORFs or introns between organellar genomes in *O. quekettii*. Furthermore, there appears to be little sequence similarity between the ORFs identified in the mitochondrial genomes of *C. lentillifera* and *O. quekettii* aside from ORF932 in the *C. lentillifera*, the only ORF in the genome with a detectable Pfam

domain, which also contains a LAGLIDADG domain. Along with the fact that their introns are not readily alignable, this suggests that most if not all introns and associated proteins arose independently between *C. lentillifera* and *O. quekettii* before proliferating within their respective lineages. Alternatively, sequences have deteriorated so much that similarities are no longer recognisable, which is unsurprising given the estimated 479 million years since the diversification of the Bryopsidales into suborders during the early Paleozoic (Verbruggen *et al.*, 2009). Cremen *et al.* (2018) did find some homologous ORFs, with conserved protein domains, between Bryopsidales chloroplast genomes. Sequencing of more mitochondrial genomes within this order will help resolve if any such conservation exists within their mtDNAs.

What is particularly intriguing about the *O. quekettii* mitochondrial genome is how much larger it is (approximately 3 times) than its economical chloroplast genome (Marcelino *et al.*, 2016). This is also the case in *C. lentillifera*, although the difference between its organellar genomes is not quite as extreme (Zheng *et al.*, 2018). It is not typical of Chlorophyta, where chloroplast genomes tend to be either similar size or larger and contain more noncoding DNA than their mostly compact intron-poor mitochondrial counterparts (Leliaert *et al.*, 2012). Instead, the *O. quekettii* mitochondrial genome is more typical of land plants: bloated with introns and intergenic DNA (Leliaert *et al.*, 2012; Mower *et al.*, 2012). It appears that evolutionary forces are acting upon these two genomic compartments differently.

The mutational-hazard (MHH) (Lynch *et al.*, 2006) and drift-barrier (Lynch *et al.*, 2016) hypotheses emphasise the importance of both mutation rate and effective population size in determining genome sizes (see General Introduction). Molecular evolution rates of the *Ostreobium* chloroplast are slow compared with other Bryopsidales. Marcelino *et al.* (2016) propose that this is due to the low light habitat of *Ostreobium* which might reduce sunlight-induced DNA rearrangements and mutation. This current study represents the first attempt to estimate evolution rates in Bryopsidales mitochondrial genomes, using the mitochondrial genomes of *O. quekettii* and *C. lentillifera*. Estimates using a very simplistic model indicate a slightly higher substitution rate in the mitochondrial genomes compared with the chloroplasts, which appears to contradict the MHH that states genomes with a higher mutation rate should be smaller (Lynch *et al.*, 2006). d_S values estimated from gene alignments show signs of saturated divergence; this is unsurprising due to the considerable

time since the divergence of these lineages. d_N/d_S estimates for chloroplast and mitochondrial genes generated by sequencing and aligning genes from the organelle genomes of closely related lineages of *Ostreobium* could provide a clearer insight into the strength of selection acting on the two organellar genomes.

It has been proposed that there are two distinct strategies developed to protect organelle genomes from negative effects of non-homologous recombination. Animal mtDNAs avoid build-up of repeats and introns with a higher mutation rate due to a lack of repetitive elements meaning rearrangements are less of a concern, reducing selection pressure for efficient DNA repair (Galtier, 2011). This elevated rate means noncoding elements such as introns pose too high a mutational burden and are thus strongly selected against (Lynch *et al.*, 2006). In contrast, plants have efficient recombination-mediated DNA repair of coding DNA, which explains their low mutation rate (Odahara *et al.*, 2009; Davila *et al.*, 2011; Christensen, 2014). The nuclear-encoded RECA3 and MSH1 genes in plants are hypothesised to control mitochondrial genome maintenance, by preventing replication of short repeats while allowing recombination-dependent replication of longer repeats (Shedge *et al.*, 2007). Other putative components of the plant mitochondrial recombination machinery have been identified based on similarity to proteins functioning in other species, including DNA polymerases (Gualberto *et al.*, 2014).

I identified a putative mitochondrion-targeted MSH1 homologue in *O. quekettii.* MSH1 encodes a protein with six conserved domains (Kowalski *et al.*, 1999) including domain VI, a GIY-YIG homing endonuclease which is predicted to be responsible for specific DNA-binding and suppression of homologous recombination (Fukui *et al.*, 2018) and is specific to only the plant form of the protein (Abdelnoor *et al.*, 2006; Shedge *et al.*, 2007). Domain VI is absent from nuclear localized homologues in plants (MSH2-MSH6) and from the yeast MSH1 protein (Abdelnoor *et al.*, 2006). Although InterPro and Pfam predicted a GIY-YIG domain in the *O. quekettii* MSH1 homologue, the fact that it lacks most of the residues that are typically conserved in this domain leaves its function unresolved. *O. quekettii* also appears to encode mitochondrially-targeted RecA proteins. RecA recombinases in *Arabidopsis* are involved in strand exchange and the joining of paired DNA ends during homologous recombination (Kühn & Gualberto, 2012; Gualberto *et al.*, 2014). RECA1 is chloroplast targeted, RECA2 is dual targeted to the mitochondrion and chloroplast, and

RECA3 is targeted to the mitochondrion (Shedge *et al.*, 2007). RECA3 mutations in *Arabidopsis* result in mitochondrial rearrangements similar but not identical to MSH1 mutants (Shedge *et al.*, 2007). Odahara *et al.* (2009) propose that these RecA proteins mediate homologous recombination which is significant for suppressing short repeat-mediated genome rearrangements in plant mitochondria. They suggest that this genome stabilisation provided by RecA could allow the number of group II introns, the dominant form in the *O. quekettii* mitochondrion, to increase (Odahara *et al.*, 2009). OSB1 is another putative component of homologous recombination-dependent repair in plant mitochondria, which is also likely involved in restricting mtDNA recombination (Kühn & Gualberto, 2012; Gualberto *et al.*, 2014). Searches of the *O. quekettii* transcriptome revealed potential OSB1 homologues. However, these lack targeting signals to organelles that would provide evidence supporting their predicted function.

In contrast to the mitochondrial and chloroplast genomes of *Volvox carteri* (Smith & Lee, 2009) and the mitochondrial genomes of many land plants (Palmer *et al.*, 2000; Mower *et al.*, 2012), little of the expanded content of the *O. quekettii* mtDNA is repetitive DNA. However, most of the repeats in the *O. quekettii* mitochondrial genome are so-called 'intermediate' repeats (50-600 bp) (Kühn & Gualberto, 2012). Repeats of this length are associated with MSH1-induced recombination in *Arabidopsis* mitochondria, which can lead to accumulation of DNA as well as complex rearrangements (Gualberto *et al.*, 2014). Along with error prone repair these processes might result in low numbers of alternative genome configurations, 'mitotypes', that contribute to heteroplasmy (Gualberto *et al.*, 2014), the coexistence of different copies of an organellar genome within the same cell (Sloan & Taylor, 2012), which could eventually increase to become the dominant form of mtDNA (Kühn & Gualberto, 2012).

While it might be tempting, based on the identification of putative recombination-associated DNA repair machinery in *O. quekettii,* to propose that a recombination-associated repair process in the *O. quekettii* mitochondrion has resulted in its inflated size, further study is required to resolve the role played by these putative mitochondrion-targeted sequences as well as to determine whether recombination is in fact occurring in the *O. quekettii* mitochondrial genome at all. In an established model system GFP fusion localization experiments might help confirm if sequences are targeted to the mitochondrion, while gene-

knockout studies could help reveal their function. For the non-model organism *Ostreobium,* aligning mitochondrial genomes of closely related lineages could at least reveal evidence of recombination or genome rearrangement if they are occurring, and we could also search for evidence of heteroplasmy, alternative forms of the mitochondrial DNA showing rearrangements, in the long-read data.

The effective population size (N_e) of genomes is also an important concept to consider as it influences the efficacy of selection (Ness *et al.*, 2015). Absent or very infrequent recombination reduces N_e, because it increases selective interference from linked loci (Neiman & Taylor, 2009). Bottlenecking of genomes, such as during the production of gametes for sexual reproduction, also leads to smaller N_e, which would reduce selection and might contribute to the higher mutational load observed in Bryopsidales mitochondrial genomes (Neiman & Taylor, 2009). The relative effective population sizes of organelle genomes in *O. quekettii* are not known. Sequencing coverage in our *Ostreobium* genome dataset was approximately seven times higher for the chloroplast compared with mitochondrial genome, and qPCR studies of *O. quekettii* organellar genes could provide a more accurate idea of relative copy numbers of mitochondrial and chloroplast genomes. However, a greater number of genomes does not necessarily correspond with greater N_e for a genome (Platt *et al.*, 2018). Sexual reproduction has not been observed in *O. quekettii,* however it has been described in a number of other Bryopsidales (e.g. Morabito *et al.*, 2010). If the sequencing of other Bryopsidales mitochondrial genomes reveals that this inflated size is a characteristic of the order, quantification of organelle and organelle genome numbers in adult plants and gametes of organisms where sexual cycles can be completed in the laboratory might reveal if there are differences in genome bottlenecks for chloroplasts and mitochondria during gamete production, which perhaps also occurred in the common ancestor of the Bryopsidales. If there is a greater reduction in the mitochondrial genome numbers compared with chloroplasts, this would reduce their N_e and thus increase the strength of genetic drift over selection.

The mutation and recombination rates in organelles are under the control of maintenance pathways that are essentially entirely nuclear-encoded (Smith & Keeling, 2015), with organelle genomes usually lacking the genes necessary for their own DNA replication and repair (Sloan & Taylor, 2012). It has been proposed that the independent evolution of similar

features in both organelle genomes arises within a species due to ‘leakage’ of nuclear-encoded proteins controlling these processes between organellar compartments, with proteins normally targeted to one organelle also becoming targeted to the other (Smith & Keeling, 2015). This does not appear to have occurred in *O. quekettii,* nor *C. lentillifera*. The effectiveness of such pathways varies considerably between organellar compartments and species (Sloan & Taylor, 2012). The great variation in organelle DNA maintenance machinery across eukaryotes might explain the broad range of organelle mutation rates observed (Smith, 2016), which tend to range more broadly and erratically in mitochondrial genomes than in chloroplasts (Smith & Keeling, 2015). With the nuclear genome for *O. quekettii* under preparation in our lab, and the nuclear genome for *C. lentillifera* recently published, there will soon be the opportunity for a more thorough study of all three genomic compartments in these two Bryopsidales, including an examination of organelle-targeted DNA maintenance machinery to further uncover the forces underpinning their divergent organelle genome sizes. Ultimately, it is likely overly simplistic to attempt to find a single explanation to cover all mitochondrial genome expansion (Smith & Keeling, 2015), with the evolution of organellar genomes in *Ostreobium* and other lineages a combination of many forces and factors.

GENERAL DISCUSSION

This study employed hybrid techniques to create the most contiguous and complete assemblies achievable within our budget for two new Chlorophyta genomes, so that we can look for evidence of the evolutionary forces shaping these genomes in both coding and noncoding DNA. These two genomes differ in source organism, genomic compartment and overall size. However, both are interesting because of their size relative to related lineages: the YPF701 nuclear genome because it is small, but intermediate between the prasinophytes and most of the core Chlorophyta, and the *O. quekettii* mtDNA because it is the largest mitochondrial genome sequenced so far in the Chlorophyta. They are both also significant due to their positions phylogenetically. YPF701 at the base of the core Chlorophyta can provide insights into the gene family evolution that occurred as this group diverged, while the *O. quekettii* mitochondrial genome represents only the second mitochondrial genome sequenced in the Bryopsidales.

My work analysing the YPF701 nuclear genome focused on the evolution of coding content. However, as for the mitochondrial genome of *Ostreobium quekettii,* the noncoding content in the YPF701 nuclear genome has also likely undergone considerable evolution. As well as the significant predicted reduction in gene family number, based on its relatively small size we can assume that there has likely also been reduction of the noncoding content in this genome. This can be examined further once the hybrid genome assembly has been quality filtered and annotated. Based on the results for the YPF701 nuclear genome as well as published pedinophyte organellar genomes, streamlining appears to have occurred in all three genomic compartments, reflecting the strong selection driving their evolution (Giovannoni *et al.*, 2005), perhaps facilitated by targeting of nuclear-encoded proteins controlling processes such as DNA replication and repair to both organelles (Smith & Keeling, 2015). An alternative testable hypothesis is that the compact YPF701 genomes might reflect mostly neutral processes due to a strong mutational bias towards deletions (Mira *et al.*, 2001).

The *O. quekettii* mitochondrial genome shows signs of greater influence by genetic drift relative to selection leading to the accumulation of noncoding content: intergenic DNA and introns. Unlike what is proposed for the pedinophytes, there appear to be different dominant forces driving the evolution of genome structure between the two organellar genome compartments in *O. quekettii*. Marcelino *et al.* (2016) suggested that selection due to a low

light environment resulted in the reduced chloroplast genome of *O. quekettii*, but it is unclear why this selection would not be able act against the expanded genome size in the mitochondrion. Perhaps effective population size, recombination and/or mutation rate, influenced by nuclear-encoded proteins, are different between the two genomes, leading to a reduction in the strength of selection to affect the evolution of the mitochondrial genome. The sequencing of more Bryopsidales will reveal if this is characteristic of the order.

Limitations

Although the time elapsed during culturing, sequencing and assembly of the nuclear genome for YPF701 gave me the opportunity to do a thorough analysis of the *O. quekettii* mitochondrial genome and a pilot study estimating gene family gains and losses using the YPF701 short-read assembly, limited time meant that I was unable to annotate the hybrid assembly and filter out any contamination. It is only once these steps have been performed that we can do a thorough comparison of the short-read and hybrid assemblies. Another limitation complicating the generation and analysis of nuclear genomes is their large size. Although it is large for a green algal mitochondrial genome, the *O. quekettii* mtDNA was sufficiently small that I was able to manually perform most of the work in error-correcting, annotating and analysing the genome. Larger nuclear genomes contain considerably more data, requiring automation that can result in some uncertainty. The noncoding content in YPF701 is unlikely to be characterised as thoroughly as the introns and intergenic DNA were in the O. *quekettii* mitochondrial genome. Hopefully, however, the use of long-read sequencing will allow it to be resolved more clearly than in the short-read assembly.

A limitation of my study of the *O. quekettii* mitochondrial genome lies in the fact that it contains solely sequencing-based analyses. Smith (2015) cautions that genome assembly data alone is a poor predictor of organelle genome structure, and calls for studies combining sequencing with traditional molecular biology techniques, as well as investigations into replication, expression and the proteome of mitochondria. Such work was not feasible within the scope of my project but, inspired by Smith's paper, I worked to move beyond merely describing the genome to explore specific hypotheses relating to the genetic forces influencing the evolution of this genome, and generated preliminary estimates of mutation as well as testable theories of evolution in this lineage that merit further study.

Conclusion

The two genomes presented in this thesis reflect evolution under differing dominant forces, with *Ostreobium* even appearing to have opposing dominant forces affecting its two organellar genomic compartments. This is unsurprising for the Chlorophyta, which show extensive morphological differences (Leliaert *et al.*, 2012) encoded in genomes that vary considerably in structure and gene content (Yurina & Odintsova, 2016), ultimately reflecting an interplay of mutation, natural selection and genetic drift.

ACKNOWLEDGEMENTS

I would like to thank Heroen Verbruggen for his guidance in this project, as well as the considerable opportunities he has granted me over the last two and a half years. Chris Jackson patiently provided his support and expertise, without which I would have achieved nothing. Richard Wetherbee, whose lecture slides changed my life and sparked my passion for protists, has been a kind and enthusiastic mentor. The Algal Biology lab members were a constant source of inspiration and new ideas, as were the members of the Marine Microbial Symbioses lab. I'd like to particularly thank Joana Costa for her help during lab work. The Simpson Lab nurtured my love of protists and gave me invaluable lab experience that boosted my confidence to take on this project. This project benefited from fruitful discussions with Joe Bielawski about evolution, as well as opportunities from Bioplatforms Australia and Melbourne Bioinformatics. I am very grateful to my friends and family, especially Quentin Bell and Aviva Green, for their support outside the lab. Easton was a sympathetic companion for my last day of revisions and a great listener as I read countless drafts aloud.

REFERENCES

Abdelnoor RV, Christensen AC, Mohammed S, Munoz-Castillo B, Moriyama H, Mackenzie SA. 2006. Mitochondrial genome dynamics in plants and animals: convergent gene fusions of a MutS homologue. *Journal of Molecular Evolution* **63**: 165-173.

Abe T, Ikemura T, Sugahara J, Kanai A, Ohara Y, Uehara H, Kinouchi M, Kanaya S, Yamada Y, Muto A et al. 2010. tRNADB-CE 2011: tRNA gene database curated manually by experts. *Nucleic Acids Research* **39**: D210-D213.

Alkatib S, Scharff LB, Rogalski M, Fleischmann TT, Matthes A, Seeger S, Schöttler MA, Ruf S, Bock R. 2012. The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS Genetics* **8**: e1003076.

Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**: 3387-3395.

Arimoto A, Nishitsuji K, Higa Y, Arakaki N, Hisata K, Shinzato C, Satoh N, Shoguchi E. 2019. A siphonous macroalgal genome suggests convergent functions of homeobox genes in algae and land plants. *DNA Research,* in press.

Arriola MB, Velmurugan N, Zhang Y, Plunkett MH, Hondzo H, Barney BM. 2018. Genome sequences of Chlorella sorokiniana UTEX 1602 and Micractinium conductrix SAG 241.80: implications to maltose excretion by a green alga. *The Plant Journal* **93**: 566-586.

Beck N, Lang B. 2010. MFannot, organelle genome annotation websever. *Canada: Université de Montréal QC*. http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl [accessed 9 May 2019].

Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, Gurnon J, Ladunga I, Lindquist E, Lucas S et al. 2012. The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biology* **13**: R39.

Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J et al. 2010. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *The Plant Cell* **22**: 2943-2955.

Brewer CA. 200x. http://www.ColorBrewer.org [accessed 9 May 2019].

Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies between mammal species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **366**: 2503-2513.

Sloan DB, Taylor DR. 2012. Evolutionary rate variation in organelle genomes: the role of mutational processes. In: Bullerwell CE, ed. *Organelle Genetics: Evolution of Organelle Genomes and Gene Expression*. Berlin, Heidelberg: Springer, 123-146.

Burger G, Gray MW, Lang BF. 2003. Mitochondrial genomes: anything goes. *Trends in Genetics* **19**: 709-716.

Cachon M, Caram B. 1979. A symbiotic green alga, Pedinomonas symbiotica sp. nov. (Prasinophyceae), in the radiolarian Thalassolampe margarodes. *Phycologia* **18**: 177-184.

Carrie C, Weißenberger S, Soll J. 2016. Plant mitochondria contain the protein translocase subunits TatB and TatC. *Journal of Cell Science* **129**: 3935-3947.

Christensen AC. 2014. Genes and junk in plant mitochondria—repair mechanisms and selection. *Genome Biology and Evolution* **6**: 1448-1453.

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

Cremen MCM, Huisman JM, Marcelino VR, Verbruggen H. 2016. Taxonomic revision of Halimeda (Bryopsidales, Chlorophyta) in south-western Australia. *Australian Systematic Botany* **29**: 41-54.

Cremen MCM, Leliaert F, Marcelino VR, Verbruggen H. 2018. Large Diversity of Nonstandard Genes and Dynamic Evolution of Chloroplast Genomes in Siphonous Green Algae (Bryopsidales, Chlorophyta). *Genome Biology and Evolution* **10**: 1048-1061.

Davila JI, Arrieta-Montiel MP, Wamboldt Y, Cao J, Hagmann J, Shedge V, Xu Y-Z, Weigel D, Mackenzie SA. 2011. Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. *BMC Biology* **9**: 64.

De Clerck O, Kao S-M, Bogaert KA, Blomme J, Foflonker F, Kwantes M, Vancaester E, Vanderstraeten L, Aydogdu E, Boesger J et al. 2018. Insights into the evolution of multicellularity from the sea lettuce genome. *Current Biology* **28**: 2921-2933.

de Vries J, Habicht J, Woehle C, Huang C, Christa G, Wägele H, Nickelsen J, Martin WF, Gould SB. 2013. Is ftsH the key to plastid longevity in sacoglossan slugs? *Genome Biology and Evolution* **5**: 2540-2548.

Del Cortona A, Leliaert F, Bogaert KA, Turmel M, Boedeker C, Janouškovec J, Lopez-Bautista JM, Verbruggen H, Vandepoele K, De Clerck O. 2017. The plastid genome in cladophorales green algae is encoded by hairpin chromosomes. *Current Biology* **27**: 3771-3782.

Del Vasto M, Figueroa-Martinez F, Featherston J, Gonzalez MA, Reyes-Prieto A, Durand PM, Smith DR. 2015. Massive and widespread organelle genomic expansion in the green algal genus Dunaliella. *Genome Biology and Evolution* **7**: 656-663.

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynié S, Cooke R et al. 2006. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences* **103**: 11647-11652.

Deschamps P, Moreau H, Worden AZ, Dauvillée D, Ball SG. 2008. Early gene duplication within chloroplastida and its correspondence with relocation of starch metabolism to chloroplasts. *Genetics* **178**: 2373-2387.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**: 1575-1584.

Fang L, Leliaert F, Novis PM, Zhang Z, Zhu H, Liu G, Penny D, Zhong B. 2018. Improving phylogenetic inference of core Chlorophyta using chloroplast sequences with strong phylogenetic signals and heterogeneous models. *Molecular Phylogenetics and Evolution* **127**: 248-255.

Fang L, Leliaert F, Zhang ZH, Penny D, Zhong BJ. 2017. Evolution of the Chlorophyta: Insights from chloroplast phylogenomic analyses. *Journal of Systematics and Evolution* **55**: 322-332.

Featherston J, Arakaki Y, Hanschen ER, Ferris PJ, Michod RE, Olson BJ, Nozaki H, Durand PM. 2017. The 4-celled Tetrabaena socialis nuclear genome reveals the essential components for genetic control of cell number at the origin of multicellularity in the volvocine lineage. *Molecular Biology and Evolution* **35**: 855-870.

Featherston J, Arakaki Y, Nozaki H, Durand PM, Smith DR. 2016. Inflated organelle genomes and a circular-mapping mtDNA probably existed at the origin of coloniality in volvocine green algae. *European Journal of Phycology* **51**: 369-377.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author.* Department of Genome Sciences, University of Washington, Seattle.

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang H-Y, Dosztányi Z, El-Gebali S, Fraser M et al. 2016. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* **45**: D190-D199.

Foflonker F, Price DC, Qiu H, Palenik B, Wang S, Bhattacharya D. 2015. Genome of the halotolerant green alga Picochlorum sp. reveals strategies for thriving under fluctuating environmental conditions. *Environmental Microbiology* **17**: 412-426.

Frickey T, Lupas A. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**: 3702-3704.

Fučíková K, Leliaert F, Cooper ED, Škaloud P, D'hondt S, De Clerck O, Gurgel CF, Lewis LA, Lewis PO, Lopez-Bautista JM et al. 2014. New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Frontiers in Ecology and Evolution* **2**: 63.

Fukui K, Harada A, Wakamatsu T, Minobe A, Ohshita K, Ashiuchi M, Yano T. 2018. The GIY-YIG endonuclease domain of Arabidopsis MutS homolog 1 specifically binds to branched DNA structures. *FEBS Letters* **592**: 4066-4077.

Galtier N. 2011. The intriguing evolutionary dynamics of plant mitochondrial DNA. *BMC Biology* **9**: 61.

Garrison EM, Arrizabalaga G. 2009. Disruption of a mitochondrial MutS DNA repair enzyme homologue confers drug resistance in the parasite Toxoplasma gondii. *Molecular Microbiology* **72**: 425-441.

Gelfand Y, Rodriguez A, Benson G. 2006. TRDB—the tandem repeats database. *Nucleic Acids Research* **35**: D80-D87.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242-1245.

Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR. 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research* **25**: 1750-1756.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. *Nucleic Acids Research* **31**: 439-441.

Gualberto JM, Mileshina D, Wallet C, Niazi AK, Weber-Lotfi F, Dietrich A. 2014. The plant mitochondrial genome: dynamics and maintenance. *Biochimie* **100**: 107-120.

Hamaji T, Kawai-Toyooka H, Toyoda A, Minakuchi Y, Suzuki M, Fujiyama A, Nozaki H, Smith DR. 2017. Multiple independent changes in mitochondrial genome conformation in chlamydomonadalean algae. *Genome Biology and Evolution* **9**: 993-999.

Hanschen ER, Marriage TN, Ferris PJ, Hamaji T, Toyoda A, Fujiyama A, Neme R, Noguchi H, Minakuchi Y, Suzuki M et al. 2016. The Gonium pectorale genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nature Communications* **7**: 11370.

Hausner G 2012. Introns, mobile elements, and plasmids. In: Bullerwell CE, ed. *Organelle Genetics: Evolution of Organelle Genomes and Gene Expression*. Berlin, Heidelberg: Springer, 329-357.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolutionary Biology* **4**: 2.
Herron MD. 2016. Origins of multicellular complexity: Volvox and the volvocine algae. *Molecular Ecology* **25**: 1213-1223.
Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics* **6**: e1001107.
Hirashima T, Tajima N, Sato N. 2016. Draft genome sequences of four species of Chlamydomonas containing phosphatidylcholine. *Genome Announcements* **4**: e01070-01016.
Hirashima T, Toyoshima M, Moriyama T, Sato N. 2018. Evolution of the phosphatidylcholine biosynthesis pathways in green algae: combinatorial diversity of methyltransferases. *Journal of Molecular Evolution* **86**: 68-76.
Hirooka S, Hirose Y, Kanesaki Y, Higuchi S, Fujiwara T, Onuma R, Era A, Ohbayashi R, Uzuka A, Nozaki H et al. 2017. Acidophilic green algal genome provides insights into adaptation to an acidic environment. *Proceedings of the National Academy of Sciences* **114**: E8304-E8313.
Jackson C, Knoll AH, Chan CX, Verbruggen H. 2018. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Scientific Reports* **8**: 1523.
Jones HL, Leadbeater B, Green J. 1994. An ultrastructural study of Marsupiomonas pelliculata gen. et sp. nov., a new member of the Pedinophyceae. *European Journal of Phycology* **29**: 171-181.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, ed. *Mammalian Protein Metabolism Volume 3*. New York, USA: Academic Press, 22-126.
Kamikawa R, Tanifuji G, Kawachi M, Miyashita H, Hashimoto T, Inagaki Y. 2015. Plastid genome-based phylogeny pinpointed the origin of the green-colored plastid in the dinoflagellate Lepidodinium chlorophorum. *Genome Biology and Evolution* **7**: 1133-1140.
Karpov S, Tanichev A. 1992. The ultrastructural study of green alga Pedinomonas tenuis Masiuk, 1970 with special reference to the flagellar apparatus. *Archiv für Protistenkunde* **141**: 315-326.
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647-1649.
Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 729-748.
Keller MD, Selvin RC, Claus W, Guillard RR. 1987. Media for the culture of oceanic ultraphytoplankton. *Journal of Phycology* **23**: 633-638.
Kinouchi M, Kurokawa K. 2006. tRNAfinder: a software system to find all tRNA genes in the DNA sequence based on the cloverleaf secondary structure. *Journal of Computer Aided Chemistry* **7**: 116-126.
Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23**: 110-120.
Kowalski JC, Belfort M, Stapleton MA, Holpert M, Dansereau JT, Pietrokovski S, Baxter SM, Derbyshire V. 1999. Configuration of the catalytic GIY-YIG domain of

intron endonuclease I-Tev I: coincidence of computational and molecular findings. *Nucleic Acids Research* **27**: 2115-2125.

Krasovec M, Eyre-Walker A, Sanchez-Ferandin S, Piganeau G. 2017. Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Molecular Biology and Evolution* **34**: 1770-1779.

Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**: 1639-1645.

Kühn K, Gualberto JM 2012. Recombination in the stability, repair and evolution of the mitochondrial genome. In: Maréchal-Drouard L, ed. *Advances in Botanical Research Volume 63*. London, UK: Academic Press, 215-252.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35**: 1547-1549.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**: 4633-4642.

Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**: 3100-3108.

Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. *Annual Review of Biochemistry* **62**: 587-622.

Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspectives in Biology* **3**: a003616.

Lang BF, Laforest M-J, Burger G. 2007. Mitochondrial introns: a critical view. *Trends in Genetics* **23**: 119-125.

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

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, De Clerck O. 2012. Phylogeny and molecular evolution of the green algae. *Critical Reviews in Plant Sciences* **31**: 1-46.

Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research* **44**: W54-W57.

Lynch M. 2006. Streamlining and simplification of microbial genome architecture. *Annual Review of Microbiology* **60**: 327-349.

Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics* **17**: 704.

Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle genomic architecture. S*cience* **311**: 1727-1730.

Maddison, WP, Maddison, DR. 2018. Mesquite: a modular system for evolutionary analysis. Version 3.6. http://www.mesquiteproject.org. [accessed 9 May 2019].

Magnusson SH, Fine M, Kühl M. 2007. Light microclimate of endolithic phototrophs in the scleractinian corals Montipora monasteriata and Porites cylindrica. *Marine Ecology Progress Series* **332**: 119-128.

Marcelino V, Cremen MCM, Jackson CJ, Larkum AA, Verbruggen H. 2016. 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.

Marin B. 2012. Nested in the Chlorellales or independent class? Phylogeny and classification of the Pedinophyceae (Viridiplantae) revealed by molecular phylogenetic analyses of complete nuclear and plastid-encoded rRNA operons. *Protist* **163**: 778-805.

Melton III JT, Lopez-Bautista JM. 2016. De novo assembly of the mitochondrial genome of Ulva fasciata Delile (Ulvophyceae, Chlorophyta), a distromatic blade-forming green macroalga. *Mitochondrial DNA Part A* **27**: 3817-3819.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L et al. 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**: 245-250.

Michaelis G, Vahrenholz C, Pratje E. 1990. Mitochondrial DNA of Chlamydomonas reinhardtii: the gene for apocytochrome b and the complete functional map of the 15.8 kb DNA. *Molecular and General Genetics* **223**: 211-216.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142-i150.

Mine I, Sekida S, Okuda K. 2015. Cell wall and cell growth characteristics of giant-celled algae. *Phycological Research* **63**: 77-84.

Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* **17**: 589-596.

Moestrup Ø. 1991. Further studies of presumedly primitive green algae, including the description of pedinophyceae class. Nov. And resultor gen. Nov. *Journal of Phycology* **27**: 119-133.

Morabito M, Gargiulo G, Genovese G. 2010. A review of life history pathways in Bryopsis. *AAPP: Physical, Mathematical, and Natural Sciences* **88**.

Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, Van Bel M, Poulain J, Katinka M, Hohmann-Marriott MF. 2012. Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biology* **13**: R74.

Mower JP, Sloan DB, Alverson AJ 2012. Plant mitochondrial genome diversity: the genomics revolution. In: Wendel JF et al. eds. *Plant Genome Diversity Volume 1*. Berlin, Heidelberg: Springer, 123-144.

Neiman M, Taylor DR. 2009. The causes of mutation accumulation in mitochondrial genomes. *Proceedings of the Royal Society B: Biological Sciences* **276**: 1201-1209.

Nelson DR, Chaiboonchoe A, Fu W, Hazzouri KM, Huang Z, Jaiswal A, Daakour S, Mystikou A, Arnoux M, Sultana M. 2019. Potential for heightened sulfur-metabolic capacity in coastal subtropical microalgae. *iScience* **11**: 450-465.

Ness RW, Kraemer SA, Colegrave N, Keightley PD. 2015. Direct estimate of the spontaneous mutation rate uncovers the effects of drift and recombination in the Chlamydomonas reinhardtii plastid genome. *Molecular Biology and Evolution* **33**: 800-808.

Nishiyama T, Sakayama H, de Vries J, Buschmann H, Saint-Marcoux D, Ullrich KK, Haas FB, Vanderstraeten L, Becker D, Lang D. 2018. The Chara genome: secondary complexity and implications for plant terrestrialization. *Cell* **174**: 448-464.

Odahara M, Kuroiwa H, Kuroiwa T, Sekine Y. 2009. Suppression of repeat-mediated gross mitochondrial genome rearrangements by RecA in the moss Physcomitrella patens. *The Plant Cell* **21**: 1182-1194.

Oliveira MC, Repetti SI, Iha C, Jackson CJ, Díaz-Tapia P, Lubiana KMF, Cassano V, Costa JF, Cremen MCM, Marcelino VR et al. 2018. High-throughput sequencing for algal systematics. *European Journal of Phycology* **53**: 256-272.

Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S. 2007. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences* **104**: 7705-7710.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu Y-L, Song K. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proceedings of the National Academy of Sciences* **97**: 6960-6966.

Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526-528.

Petrů M, Wideman J, Moore K, Alcock F, Palmer T, Doležal P. 2018. Evolution of mitochondrial TAT translocases illustrates the loss of bacterial protein transport machines in mitochondria. *BMC Biology* **16**: 141.

Platt A, Weber CC, Liberles DA. 2018. Protein evolution depends on multiple distinct population size parameters. *BMC Evolutionary Biology* **18**: 17.

Pombert J-F, Blouin NA, Lane C, Boucias D, Keeling PJ. 2014. A lack of parasitic reduction in the obligate parasitic green alga Helicosporidium. *PLoS Genetics* **10**: e1004355.

Pombert J-F, Otis C, Lemieux C, Turmel M. 2004. The complete mitochondrial DNA sequence of the green alga Pseudendoclonium akinetum (Ulvophyceae) highlights distinctive evolutionary trends in the Chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae. *Molecular Biology and Evolution* **21**: 922-935.

Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* **329**: 223-226.

R core Team. 2019. R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria. https://www.R-project.org/.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217-223.

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics* **13**: 278-289.

Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Van de Peer Y. 2007. The complete chloroplast and mitochondrial DNA sequence of Ostreococcus tauri: organelle genomes of the smallest eukaryote are examples of compaction. *Molecular Biology and Evolution* **24**: 956-968.

Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Current Biology* **15**: 1325-1330.

Roger AJ, Munoz-Gomez SA, Kamikawa R. 2017. The origin and diversification of mitochondria. *Current Biology* **27**: R1177-R1192.

Satjarak A, Burns JA, Kim E, Graham LE. 2017. Complete mitochondrial genomes of prasinophyte algae Pyramimonas parkeae and Cymbomonas tetramitiformis. *Journal of Phycology* **53**: 601-615.

Shedge V, Arrieta-Montiel M, Christensen AC, Mackenzie SA. 2007. Plant mitochondrial recombination surveillance requires unusual RecA and MutS homologs. *The Plant Cell* **19**: 1251-1264.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**: 539.

Smith D, Lee RW. 2009. The mitochondrial and plastid genomes of Volvox carteri: bloated molecules rich in repetitive DNA. *BMC Genomics* **10**: 1.

Smith DR. 2015. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs? *Briefings in Functional Genomics* **15**: 47-54.

Smith DR. 2016. The mutational hazard hypothesis of organelle genome evolution: 10 years on. *Molecular Ecology* **25**: 3769-3775.

Smith DR, Hamaji T, Olson BJ, Durand PM, Ferris P, Michod RE, Featherston J, Nozaki H, Keeling PJ. 2013. Organelle genome complexity scales positively with organism size in volvocine green algae. *Molecular Biology and Evolution* **30**: 793-797.

Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences* **112**: 10177-10184.

Smith DR, Lee RW. 2007. Mitochondrial genome of the colorless green alga Polytomella capuana: a linear molecule with an unprecedented GC content. *Molecular Biology and Evolution* **25**: 487-496.

Smith DR, Lee RW. 2010. Low nucleotide diversity for the expanded organelle and nuclear genomes of Volvox carteri supports the mutational-hazard hypothesis. *Molecular Biology and Evolution* **27**: 2244-2256.

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

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

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**: W609-W612.

Sweeney BM. 1976. Pedinomonas noctilucae (Prasinophyceae), the flagellate symbiotic in Noctiluca (dinophyceae) in southeast asia. *Journal of Phycology* **12**: 460-464.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**: 36.

Tribollet A 2008. The boring microflora in modern coral reef ecosystems: a review of its roles. In: Wisshak M, Tapanila L, eds. *Current Developments in Bioerosion*. Berlin, Heidelberg: Springer, 67-94.

Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW. 1999. The complete mitochondrial DNA sequences of Nephroselmis olivacea and Pedinomonas minor: two radically different evolutionary patterns within green algae. *The Plant Cell* **11**: 1717-1729.

Turmel M, Otis C, Lemieux C. 2003. The mitochondrial genome of Chara vulgaris: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *The Plant Cell* **15**: 1888-1903.

Umen JG, Olson BJ 2012. Genomics of volvocine algae. In: Piganeau G, ed. *Advances in Botanical Research Volume 64*. London, UK: Academic Press, 185-243.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. *Nature Genetics* **15**: 57.

Vandepoele K, Van Bel M, Richard G, Van Landeghem S, Verhelst B, Moreau H, Van de Peer Y, Grimsley N, Piganeau G. 2013. pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environmental Microbiology* **15**: 2147-2153.

Verbruggen H, Ashworth M, LoDuca ST, Vlaeminck C, Cocquyt E, Sauvage T, Zechman FW, Littler DS, Littler MM, Leliaert F et al. 2009. A multi-locus time-calibrated phylogeny of the siphonous green algae. *Molecular Phylogenetics and Evolution* **50**: 642-653.

Verbruggen H, Marcelino VR, Guiry MD, Cremen MCM, Jackson CJ. 2017. Phylogenetic position of the coral symbiont Ostreobium (Ulvophyceae) inferred from chloroplast genome data. *Journal of Phycology* **53**: 790-803.

Vroom PS, Smith CM. 2001. The challenge of siphonous green algae. *American Scientist* **89**: 525-531.

Vroom PS, Smith CM. 2003. Life without cells. *Biologist* **50**: 222-226.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* **35**: 543-548.

Wendel JF, Jackson SA, Meyers BC, Wing RA. 2016. Evolution of plant genome architecture. *Genome Biology* **17**: 37.

Wilhelm C, Jakob T. 2006. Uphill energy transfer from long-wavelength absorbing chlorophylls to PS II in Ostreobium sp. is functional in carbon assimilation. *Photosynthesis Research* **87**: 323.

Wolff G, Plante I, Lang BF, Kück U, Burger G. 1994. Complete sequence of the mitochondrial DNA of the chlorophyte alga Prototheca wickerhamii: gene content and genome organization. *Journal of Molecular Biology* **237**: 75-86.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**: 3252-3255.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution* **21**: 809-818.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**: 28-36.

Yurina N, Odintsova M. 2016. Mitochondrial genome structure of photosynthetic eukaryotes. *Biochemistry (Moscow)* **81**: 101-113.

Zheng F, Liu H, Jiang M, Xu Z, Wang Z, Wang C, Du F, Shen Z, Wang B. 2018. The complete mitochondrial genome of the Caulerpa lentillifera (Ulvophyceae, Chlorophyta): Sequence, genome content, organization structure and phylogenetic consideration. *Gene* **673**: 225-238.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669-2677.

Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *Journal of Molecular Biology* **430**: 2237-2243.

Zou J, Bi G. 2016. Complete mitochondrial genome of a hydrocarbon-producing green alga Botryococcus braunii strain Showa. *Mitochondrial DNA Part A* **27**: 2619-2620.

APPENDIX

Chapter 1

Chapter 2

Supplementary Methods: CTAB DNA extraction protocol

A total of 10 mL of preheated (60°C) extraction buffer (2% CTAB; 5 M NaCl; 0.5 M EDTA; 1% w/v PVP; 10 mM Tris–HCl, pH 8) and 200 µL of proteinase K (20 mg/ml) were added to cell pellets. Samples were incubated at 60°C for 90 minutes, and gently inverted every 5–10 minutes. 100 µl of RNAse (10 mg/ml) was added to the mixture and incubated for a further 90 minutes. The mixture was centrifuged at full speed (9888g) at room temperature (21°C) for 10 minutes. The aqueous layer was collected and an equal volume of chloroform:isoamyl alcohol (CIA, 24:1, v/v) was added and the tubes were then inverted a few times to emulsify. The aqueous layer was collected, re-extracted with CIA and centrifuged at full speed for 5 minutes. The DNA in the aqueous phase was precipitated using an equal volume of 80% isopropanol, stored at 4°C for 90 minutes then centrifuged at 4°C for 15 minutes at full speed. Pellets were washed in 5 mL of 70% ethanol, air-dried and then re-suspended in 1 mL of 0.1 TE (TrisEDTA, ethylenediaminetetraacetic acid) buffer. Care was taken to not shear the DNA prior to long read sequencing, with only gentle inversion of tubes and the use of cut off pipette tips.

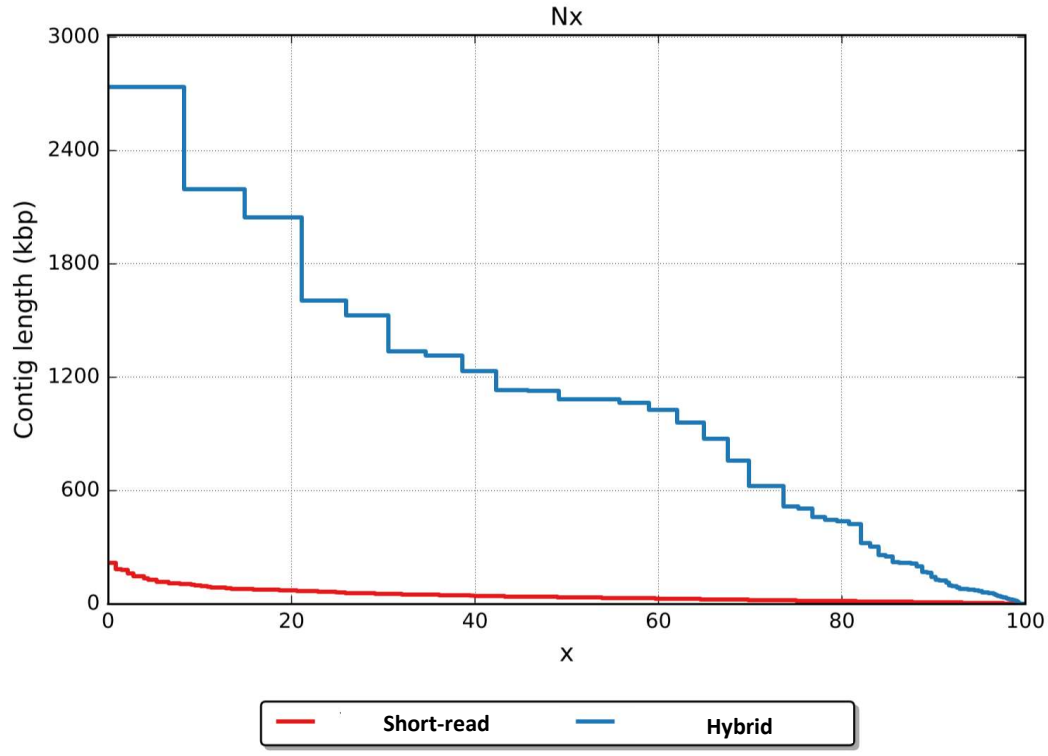


Fig. S1. Plot of contig lengths in short-read and hybrid genome assemblies for pedinophyte YPF701 from QUAST 5.0.2. The hybrid assembly is substantially more contiguous across the entire size spectrum.

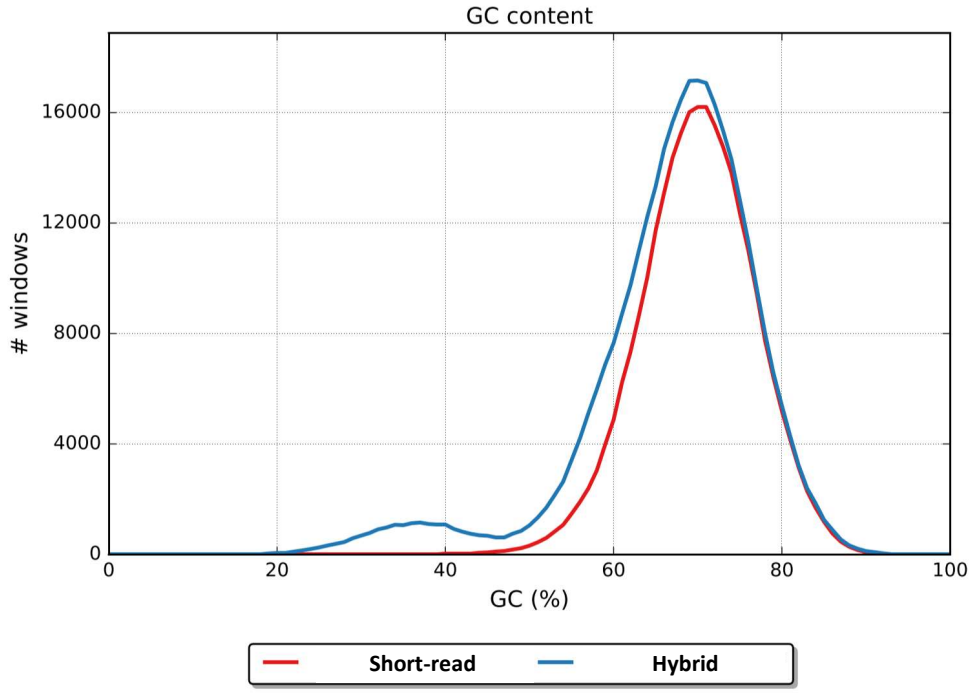


Fig. S2. Plot of GC (%) in short-read and hybrid genome assemblies for pedinophyte YPF701 from QUAST 5.0.2.

Table S1 Summary statistics for comparison of short-read and hybrid genome assemblies for pedinophyte YPF701 from QUAST 5.0.2. All statistics unless stated are based on scaffolds of size ≥ 1000 bp.

	Short-read	Hybrid
# contigs (>= 0 bp)	1597	1877
# contigs (>= 1000 bp)	1597	257
# contigs (>= 5000 bp)	999	101
# contigs (>= 10000 bp)	740	95
# contigs (>= 25000 bp)	350	78
# contigs (>= 50000 bp)	113	65
Total length (>= 0 bp)	26770386	34071101
Total length (>= 1000 bp)	26770386	33071467
Total length (>= 5000 bp)	25366432	32860715
Total length (>= 10000 bp)	23475163	32817300
Total length (>= 25000 bp)	17231994	32487712
Total length (>= 50000 bp)	8828023	32027187
# contigs	1597	257
Largest contig	216425	2736781
Total length	26770386	33071467
GC (%)	69.90	66.91
N50	35582	1083765
N75	17331	514604
L50	222	11
L75	487	20
# N's per 100 kbp	20.06	0.00

Table S2 Codon usage in the protein coding genes in mtDNA of *Ostreobium quekettii* SAG6.99.

Codon	AA	% of AA	Freq	Codon	AA	% of AA	Freq
GCA	A	26.30%	390	AAC	N	35.60%	363
GCC	A	23.50%	348	AAT	N	64.40%	656
GCG	A	17.90%	266	CCA	P	27.00%	263
GCT	A	32.30%	478	CCC	P	21.70%	211
TGC	C	43.10%	178	CCG	P	14.10%	137
TGT	C	56.90%	235	CCT	P	37.20%	362
GAC	D	34.10%	331	CAA	Q	71.90%	634
GAT	D	65.90%	641	CAG	Q	28.10%	248
GAA	E	72.20%	721	AGA	R	22.70%	384
GAG	E	27.80%	278	AGG	R	11.80%	200
TTC	F	32.50%	440	CGA	R	20.40%	345
TTT	F	67.50%	915	CGC	R	15.00%	254
GGA	G	27.20%	360	CGG	R	11.20%	189
GGC	G	20.60%	273	CGT	R	18.80%	317
GGG	G	18.30%	242	AGC	S	15.20%	280
GGT	G	33.90%	449	AGT	S	19.10%	352
CAC	H	37.10%	239	TCA	S	18.60%	344
CAT	H	62.90%	406	TCC	S	10.50%	194
ATA	I	35.10%	638	TCG	S	13.90%	257
ATC	I	20.50%	373	TCT	S	22.70%	418
ATT	I	44.50%	809	ACA	T	29.00%	340
AAA	K	69.40%	1193	ACC	T	24.00%	281
AAG	K	30.60%	525	ACG	T	13.10%	154
CTA	L	14.50%	377	ACT	T	33.90%	397
CTC	L	9.60%	249	GTA	V	27.80%	383
CTG	L	8.90%	232	GTC	V	17.70%	244
CTT	L	19.80%	515	GTG	V	23.70%	327
TTA	L	28.90%	750	GTT	V	30.70%	423
TTG	L	18.20%	473	TAC	Y	35.10%	348
ATG	M	99.00%	482	TAT	Y	64.90%	643
CTG	M	0.40%	2	TAA	*	51.50%	34
TTG	M	0.60%	3	TAG	*	28.80%	19
TGG	W	100.00%	323	TGA	*	19.70%	13

Table S3 tRNAs present in mtDNA of Ostreobium quekettii SAG6.99.

trna	start	stop	#nt	direction
tRNA-Lys(uuu)	26,865	26,937	73	forward
tRNA-Glu(uuc)	27,231	27,302	72	forward
tRNA-Met(cau)	73,803	73,875	73	forward
tRNA-Met(cau)	84,822	84,893	72	forward
tRNA-Ala(ugc)	86,936	87,008	73	forward
tRNA-Ile(gau)	92,048	92,121	74	forward
tRNA-Ser(uga)	92,893	92,978	86	forward
tRNA-Ser(gcu)	95,358	95,445	88	forward
tRNA-Gly(ucc)	98,148	98,218	71	forward
tRNA-Leu(caa)	99,103	99,188	86	forward
tRNA-Thr(gag)	113,591	113,672	82	forward
tRNA-Thr(uag)	139,600	139,679	80	forward
tRNA-Pro(ugg)	142,712	142,787	76	reverse
tRNA-His(gug)	159,029	159,100	72	reverse
tRNA-Arg(ucu)	179,218	179,291	74	forward
tRNA-Asn(guu)	180,642	180,713	72	forward
tRNA-Trp(cca)	187,710	187,781	72	forward
tRNA-Asp(guc)	191,017	191,089	73	forward
tRNA-Arg(acg)	191,095	191,168	74	forward
tRNA-Gly(gcc)	193,289	193,360	72	forward
tRNA-Gln(uug)	204,895	204,965	71	forward
tRNA-Met(cau)	217,806	217,877	72	forward
tRNA-Cys(gca)	221,376	221,447	72	forward
tRNA-Thr(ugu)	223,196	223,268	73	forward
tRNA-Tyr(gua)	223,271	223,352	82	forward
tRNA-Leu(uaa)	226,169	226,249	81	forward
tRNA-Val(uac)	234,890	234,962	73	forward
tRNA-Phe(gaa)	237,211	237,284	74	forward

Table S4 Comparison of protein coding and ribosomal RNA genes in the mitochondrial genomes of a selection of Chloroplastida including *Ostreobium quekettii* SAG6.99. (#) = the number of introns disrupting the gene, (d) = duplicated gene.

	Ulvophyceae			**Chlorophyceae**			**Prasinophytes**			**Pedinophytes**	**Trebouxiophyceae**		**Streptophyta**	
	Ostreobium	**Caulerpa lentilifera**	**Ulva fasciata**	**Chlamydomonas reinhardtii**	**Gonium pectorale**	**Dunaliella salina**	**Ostreococcus tauri**	**Cymbomonas tetramitiformis**	**Nephroselmis olivacea**	**Pedinomonas minor**	**Prototheca wickerhamii**	**Botryococcus braunii**	**Arabidopsis thaliana**	**Chara vulgaris**
Accession		KX761577.1	NC_028081.1	NC_001638.1	NC_020437.1	NC_012930.1	NC_008290.1	NC_036614.1	NC_008239.1	NC_000892.1	NC_001613.1	NC_027722.1	NC_037304.1	NC_005255.1
size(bp)	241739	209034	61614	15800-18900	15993	28331	44237	73520	45223	25137	55328	84583	367808	67737
tRNAs	28	20	27	3	3	3	26	23	26	9	26	23	22	26
5s	y	y	n	n	n	n	y(d)	n	y	n	y	y	y	y
16s	y(1)	y	y	y(4 segments)	y(4 fragments)	y(3 fragments, S2 has 2 introns)	y(2 segments, duplicated)	y(d)	y	y	y	y	y	y(1)
23s	y(7)	y	y	y(8 segments, L5 has 1 intron, L7 has 1 intron)	y(8 fragments)	y(6 fragments, L5 has 1 intron, L6 has 3)	y(d)	y	y(3)	y(2 fragments, L1 has 1 intron)	y(2)	y(1)	y	y(9)
rnpB	n	n	n	n	n	n	y	n	y	n	n	n	n	n
atp1	y(2)	y(2)	y	n	n	n	y	n	y	n	y	y	y	y
atp4	y	n	y	n	n	n	y(d)	y	n	n	n	y	y	y
atp6	y(1)	y(2)	y	n	n	n	y	y	y	y	y	y	y	y
atp8	y(2)	y(1)	y	n	n	n	y(d)	y	y	y	y	y	y	y
atp9	y	y(2)	y	n	n	n	y	n	y	n	y	y	y	y(2)
cob	y(3)	y(2)	y	y(1)	y	y(4)	y(d)	y	y(1)	y	y	y	y	y(3)
cox1	y(11)	y(5)	y(3)	y(2)	y	y(5)	y(d)	y(2)	y	y	y(3)	y	y	y(6)
cox2	y	y(1)	y	n	n	n	y	y	y	n	y	y	y(1)	y(1)
cox3	y(2)	y	y	n	n	n	y	y	y	n	y	y	y	y
nad1	y(1)	y(1)	y	y	y	y(1)	y	y	y	y	y	y	y(4)	y
nad2	y(3)	y(1)	y	y	y	y	y	y	y	y	y	y	y(4)	y
nad3	y	y	y(1)	n	n	n	y	y	y	y	y	y	y	y(2)
nad4	y(4)	y(4)	y	y	y	y	y	y	y	y	y	y	y(3)	y(1)
nad4L	y	y	y	n	n	n	y(d)	y	y	y	y	y	y	y
nad5	y(3)	y(3)	y	y	y(1)	y(2)	y	y	y	y	y	y	y(4)	y
nad6	y	y	y	y	y	y	y	y	y	y	y	y	y	y
nad7	y(2)	y(4)	y	n	n	n	y	y	y	n	y	y	y(4)	y
nad9	y	y(1)	y	n	n	n	y	y	y	n	y	y	y	y
nad10	y	n	n	n	n	n	y	n	y	n	n	n	n	n
rpl2	n	n	n	n	n	n	n	n	n	n	n	n	y(1)	y
rpl5	y	n	y	n	n	n	y	y	y	n	y	y	y	y
rpl6	y(1)	n	n	n	n	n	y	y	y	n	y	n	n	y
rpl14	y	n	n	n	n	n	y	y	y	n	n	n	n	y
rpl16	y(1)	n	y	n	n	n	y	y	y	n	y	y	y	y
rps1	n	n	n	n	n	n	n	n	n	n	n	n	n	y

rps2	y(1)	n	y	n	n	n	y	y	y	n	y	y	n	y
rps3	y(1)	n	y	n	n	n	y	y	y	n	y	y	y(1)	y(1)
rps4	y	n	y	n	n	n	y	y	y	n	y	y	y	y
rps7	y(1)	n	n	n	n	n	y	y	y	n	y	y	y	y
rps8	n	n	n	n	n	n	y	y	y	n	n	n	n	n
rps10	y	n	y	n	n	n	y	y	y	n	y	y	n	y
rps11	y	n	y	n	n	n	y	y	y	n	y	y	n	y
rps12	y	n	y	n	n	n	y	y	y	n	y	y	y	y
rps13	y	n	y	n	n	n	y	y	y	n	y	y	n	n
rps14	y	n	y	n	n	n	y	y	y	n	y	y	pseudogene	y
rps19	y	n	y	n	n	n	y	y	y	n	y	y	pseudogene	y
tatC/ mttB	y	ORF233	n	n	n	n	y	n	y	n	n	y	y	n
sdh3	n	n	n	n	n	n	n	n	n	n	n	n	n	y
sdh4	n	n	n	n	n	n	n	n	n	n	n	n	pseudogene	y
yejR	n	n	n	n	n	n	n	n	n	n	n	n	n	y
yejU	n	n	n	n	n	n	n	n	n	n	n	n	n	y
yejVc	n	n	n	n	n	n	n	n	n	n	n	n	n	y
ccmB	n	n	n	n	n	n	n	n	n	n	n	n	y	n
ccmC	n	n	n	n	n	n	n	n	n	n	n	n	y	n
ccmFc	n	n	n	n	n	n	n	n	n	n	n	n	y(1)	n
ccmFN	n	n	n	n	n	n	n	n	n	n	n	n	y	n
Reference	This study	(Zheng *et al.,* 2019)	(Melton III & Lopez-Bautista, 2016)	(Michaelis *et al.,* 1990; Smith *et al.,* 2010)	(Hamaji *et al.,* 2013)	(Smith *et al.,* 2010)	(Robbens *et al.,* 2007)	(Satjarak *et al.,* 2017)	(Turmel *et al.,* 1999)	(Turmel *et al.,* 1999)	(Wolff *et al.,* 1994)	(Zou & Bi, 2016)	(Unseld *et al.,* 1997)	(Turmel *et al.,* 2003)

Table S5 Introns and associated ORFs within mtDNA of *Ostreobium quekettii* SAG6.99. No = No domain detected.

GENE		INTRON 1	INTRON 2	INTRON 3	INTRON 4	INTRON 5	INTRON 6	INTRON 7	INTRON 8	INTRON 9	INTRON 10	INTRON 11
cox1	**RNAweasel**	intron I (derived, B1)	intron II (domainV)	intron II (domainV)	intron II (domainV)	intron ID	No	intron IB (complete)	intron II (domainV)	intron IB	intron IB (3', partial)	intron II (domainV)
	Rfam	No	intron II	intron II	intron II	No	No	No	intron II	No	No	No
	ORF		ORF113	ORF698	ORF576	ORF144, ORF231		ORF328			ORF257	ORF617
	ORF Pfam domains		LAGLIDADG_2	Intron_maturas2	Intron_maturas2	ORF144: LAGLIDADG_1 ORF231: LAGLIDADG_2		LAGLIDADG_1			LAGLIDADG_1 and LAGLIDADG_1 (2 domains)	Intron_maturas2
atp1	**RNAweasel**	intron II (domainV)	intron II (domainV)									
	Rfam	intron II	No									
	ORF	ORF714	ORF1168									
	ORF Pfam domains	RVT_1 and Intron_maturas2	LAGLIDADG_2, RVT_1, and Intron_maturas2									
LSU	**RNAweasel**	intron IA3	intron IB (complete)	intron IB (5', partial)	intron IB (complete)	intron I (derived, B1)	intron IB (complete)	intron IA				
	Rfam	no	no	No	No	No	No	No				
	ORF	ORF215	ORF109, ORF139		ORF300	ORF279	ORF103	ORF175				
	ORF Pfam domains	LAGLIDADG_2	ORF109: LAGLIDADG_1, ORF139: LAGLIDADG_1		LAGLIDADG_1 and LAGLIDADG_1	LAGLIDADG_1 and LAGLIDADG_1	LAGLIDADG_2	LAGLIDADG_2				
SSU	**RNAweasel**	No										
	Rfam	No										
	ORF	ORF268										
	ORF Pfam domains	LAGLIDADG_2										
cob	**RNAweasel**		intron II (domainV)	intron IB (complete)								
	Rfam	intron II	intron II	No								
	ORF		ORF626									
	ORF Pfam domains		RVT_1, HNH and GIIM									
nad7	**RNAweasel**	intron II (domainV)	intron II (domainV)									
	Rfam	No	intron II									
	ORF	ORF963										

	ORF Pfam domains	RVT 1 and GIIM										
rps2	**RNAweasel**	No										
	Rfam	No										
nad5	**RNAweasel**	No	intron II (domainV)	intron IB (3', partial)								
	Rfam	No	intron II	No								
atp8	**RNAweasel**	intron II (domainV)	intron II (domainV)									
	Rfam	intron II	intron II									
nad1	**RNAweasel**	intron II (domainV)										
	Rfam	intron II										
nad4	**RNAweasel**	intron II (domainV)	intron II (domainV)	intron II (domainV)	intron II (domainV)							
	Rfam	intron II	No	intron II	intron II							
cox3	**RNAweasel**	intron II (domainV)	intron II (domainV)									
	Rfam	intron II	intron II									
rpl6	**RNAweasel**	intron II (domainV)										
	Rfam	intron II										
nad2	**RNAweasel**	intron II (domainV)	intron II (domainV)	intron II (domainV)								
	Rfam	intron II	intron II	No								
rpl16	**RNAweasel**	intron II (domainV)										
	Rfam	intron II										
rps3	**RNAweasel**	intron II (domainV)										
	Rfam	intron II										
rps7	**RNAweasel**	No										
	Rfam	No										
atp6	**RNAweasel**	intron II (domainV)										
	Rfam	intron II										

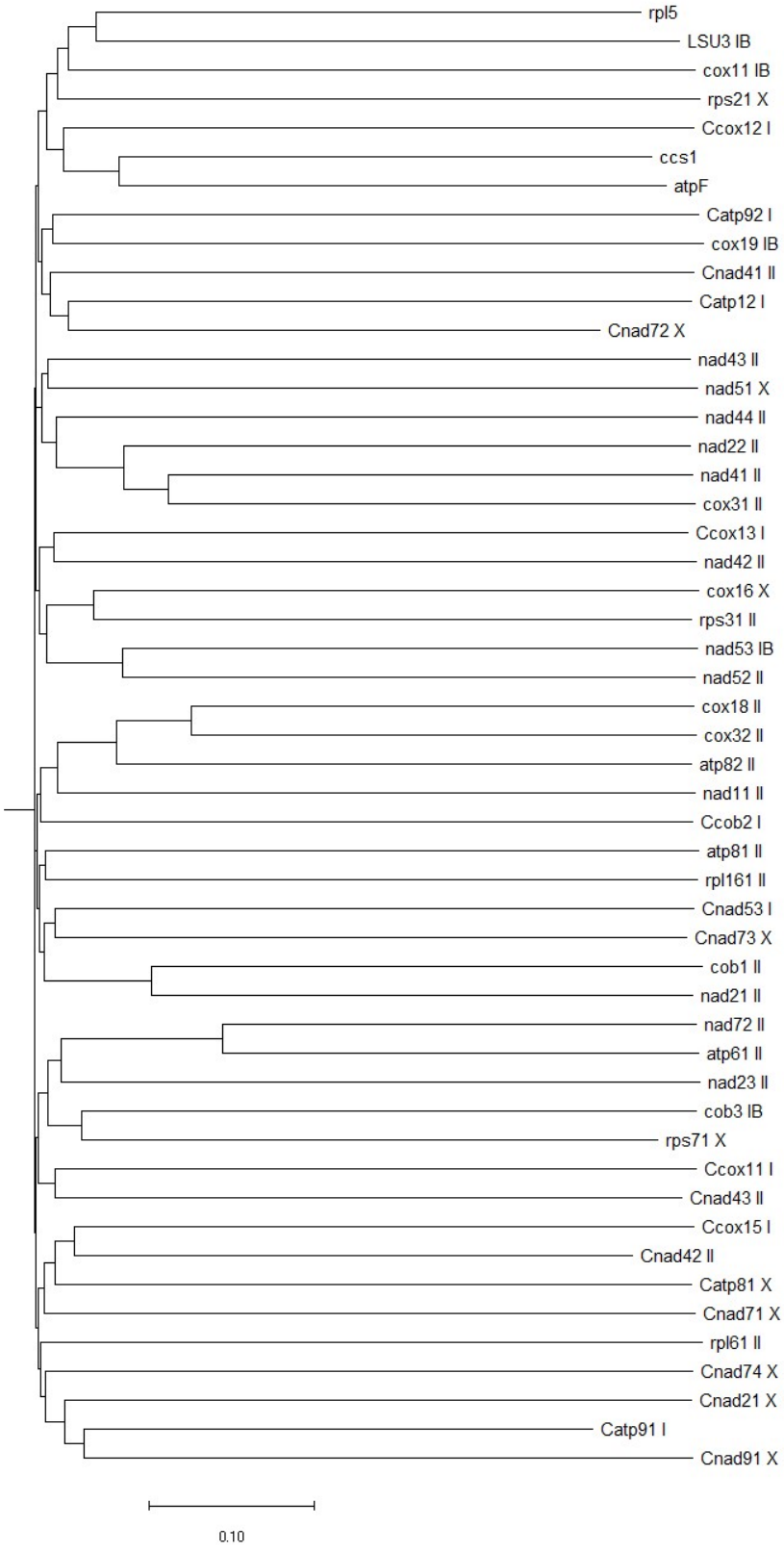


Fig. S3. Neighbour joining tree constructed from Clustal Omega distance matrix of ORF-lacking introns in *C. lentillifera* mitochondrial genome and *O. quekettii* mitochondrial and plastid genomes.

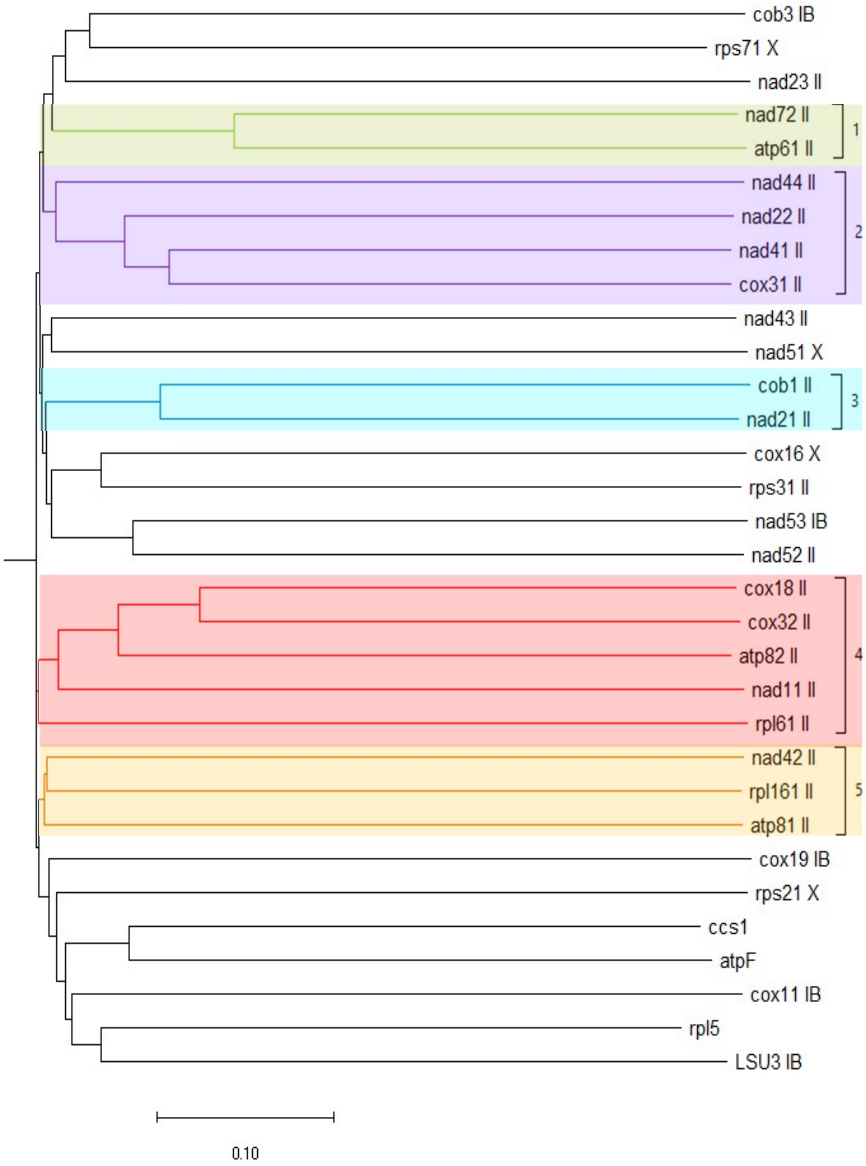
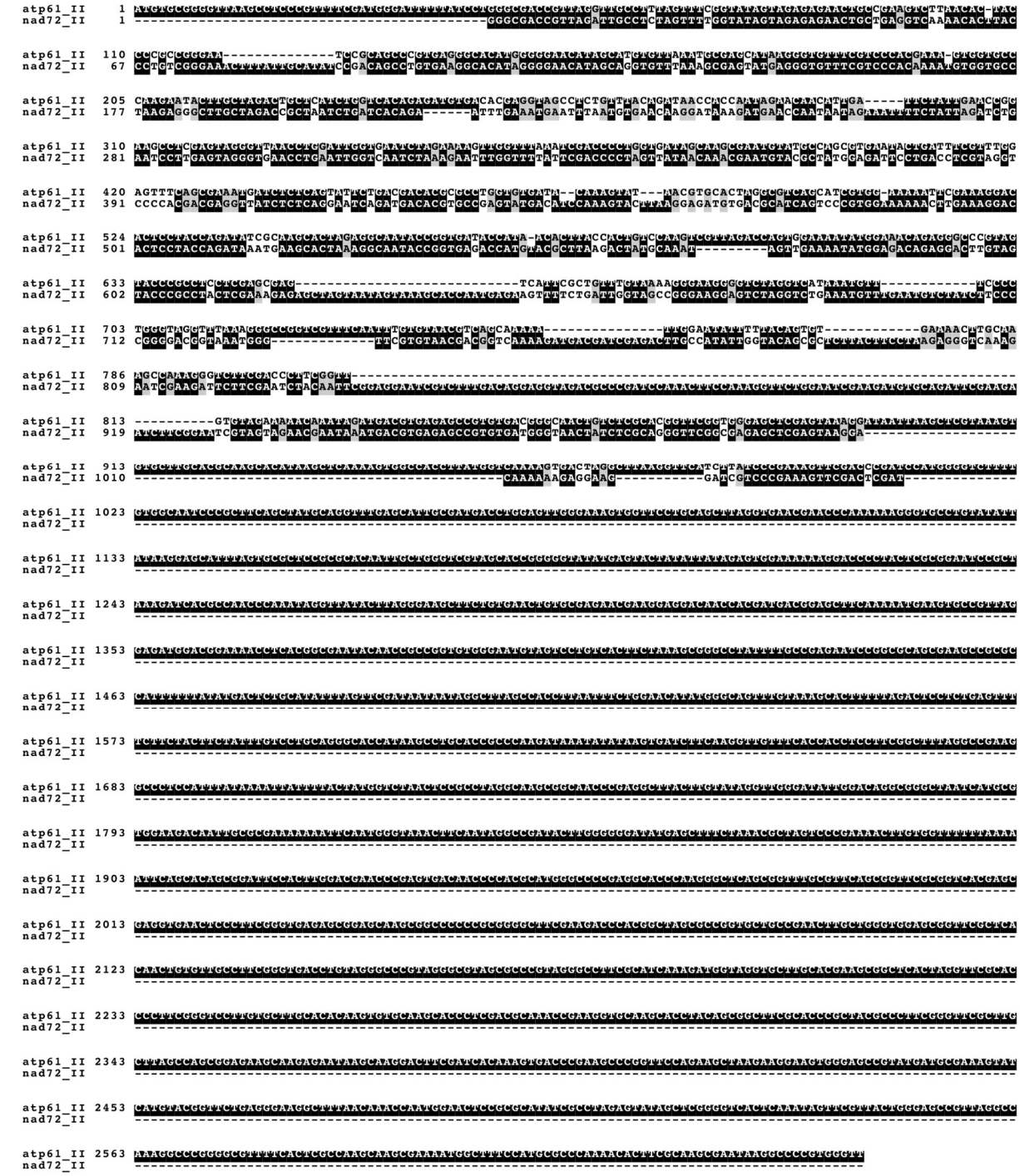
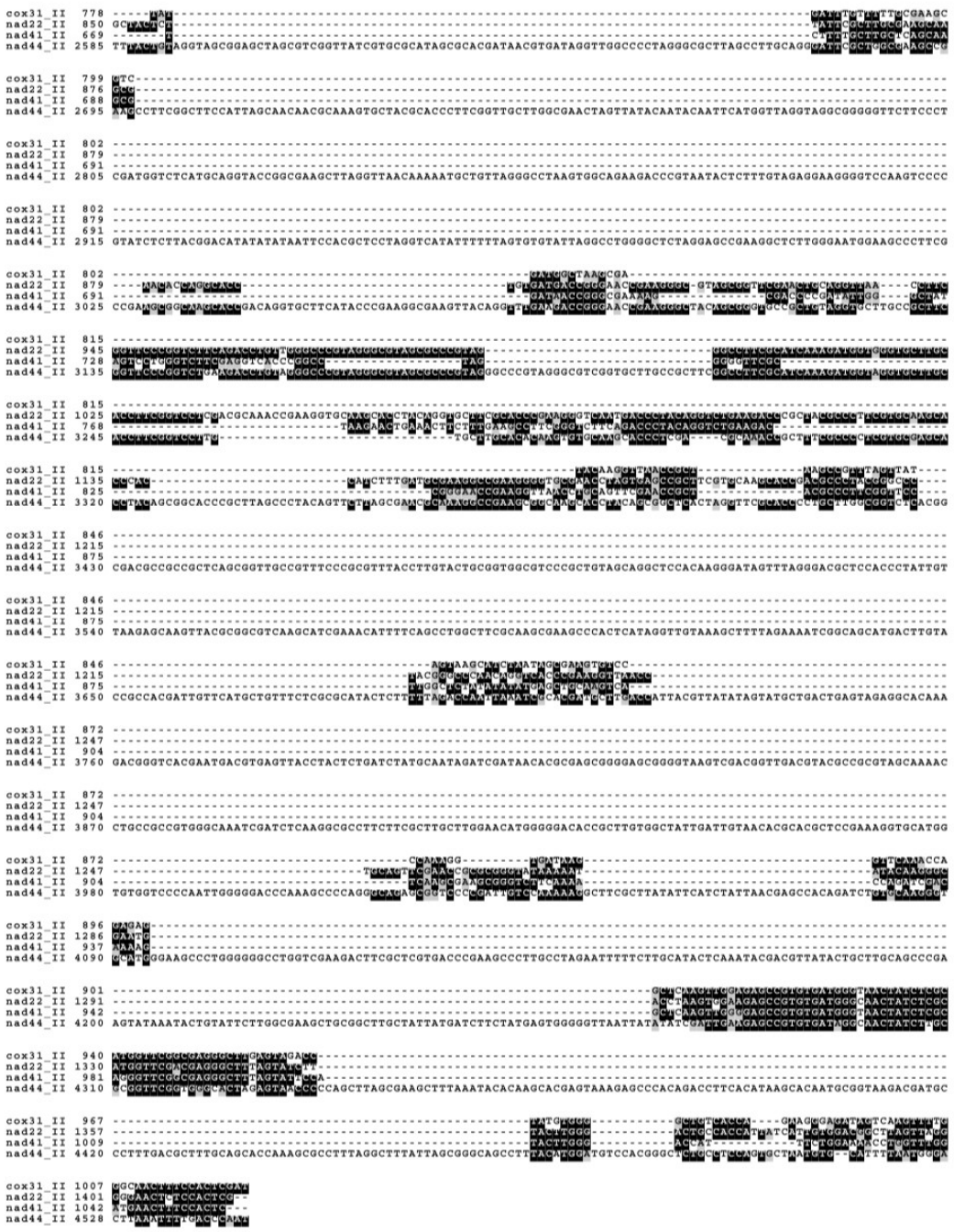


Fig. S4. Neighbour joining tree constructed from Clustal Omega distance matrix of ORF-lacking introns in *O. quekettii* mitochondrial and chloroplast genomes. Below: MAFFT alignments corresponding to clusters identified.

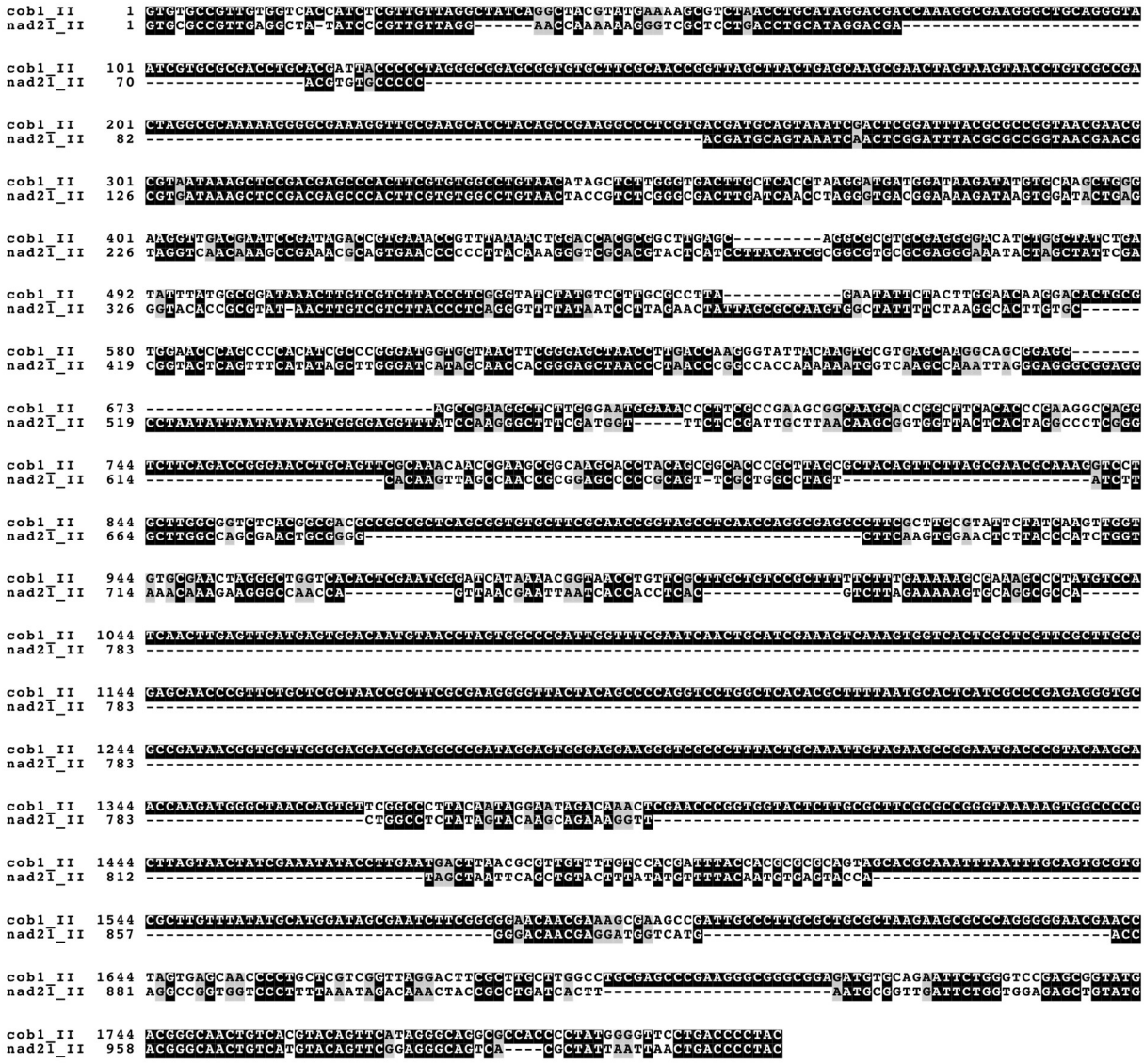
Group 1:



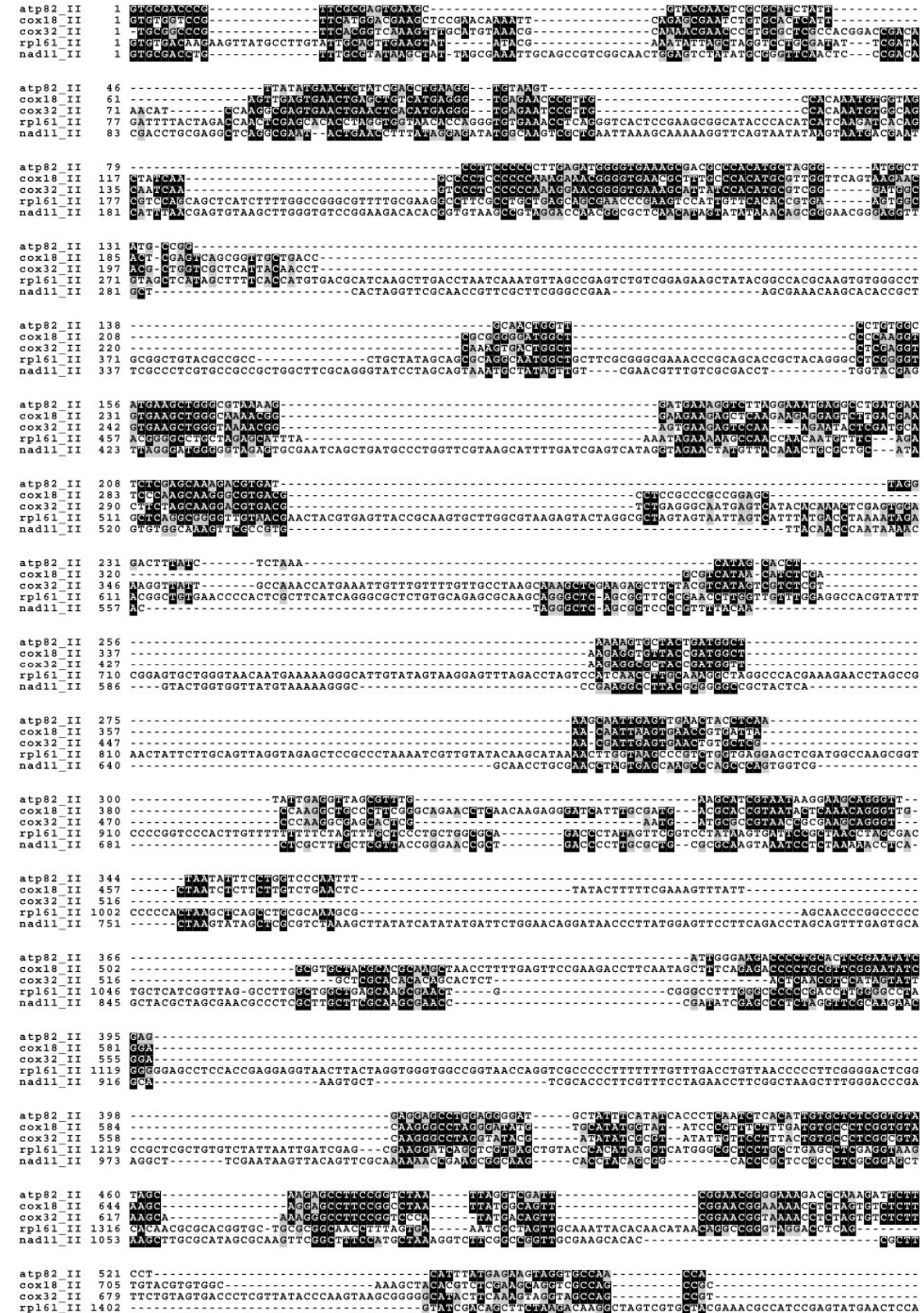
Group 2:

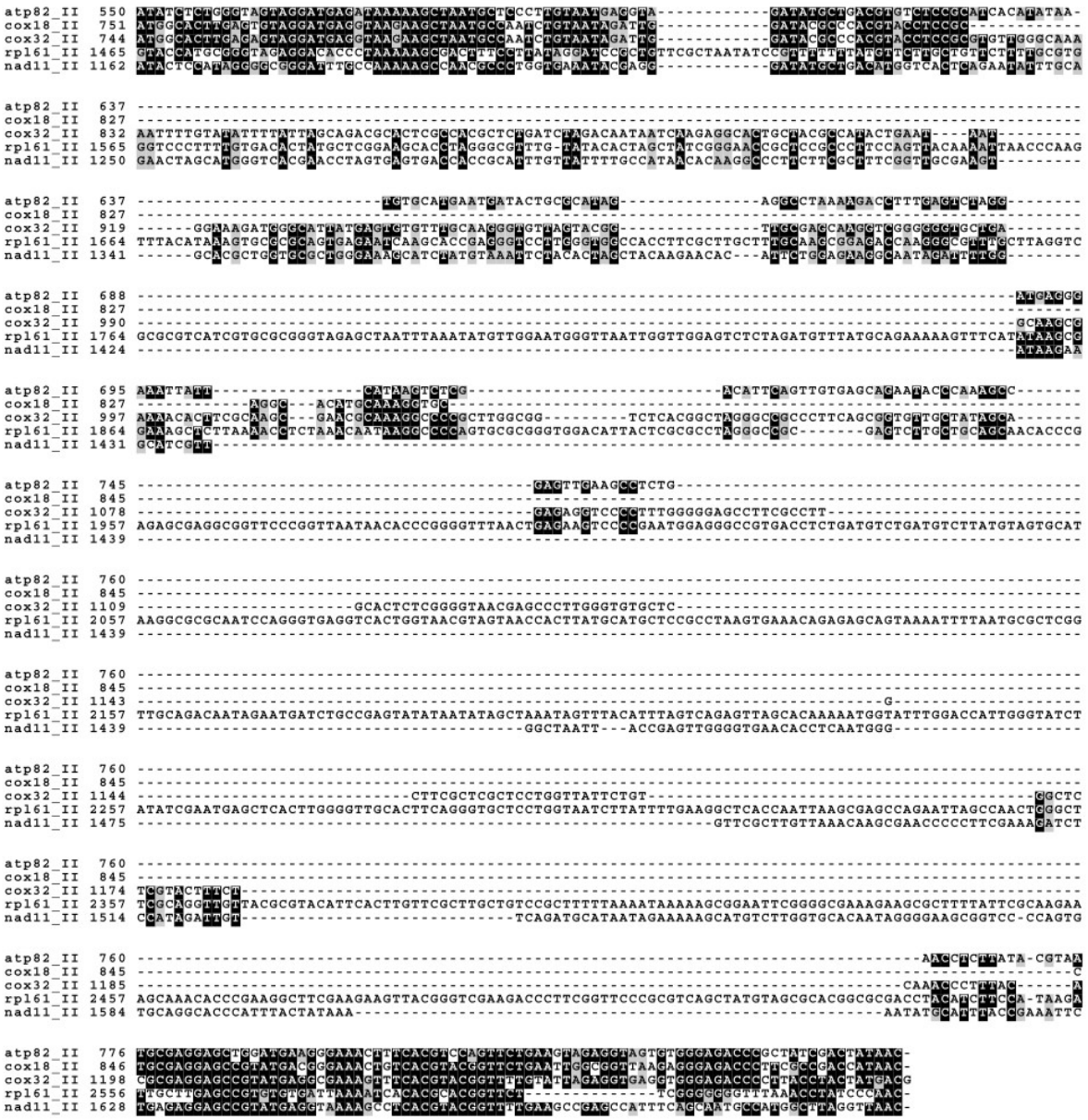


Group 3:

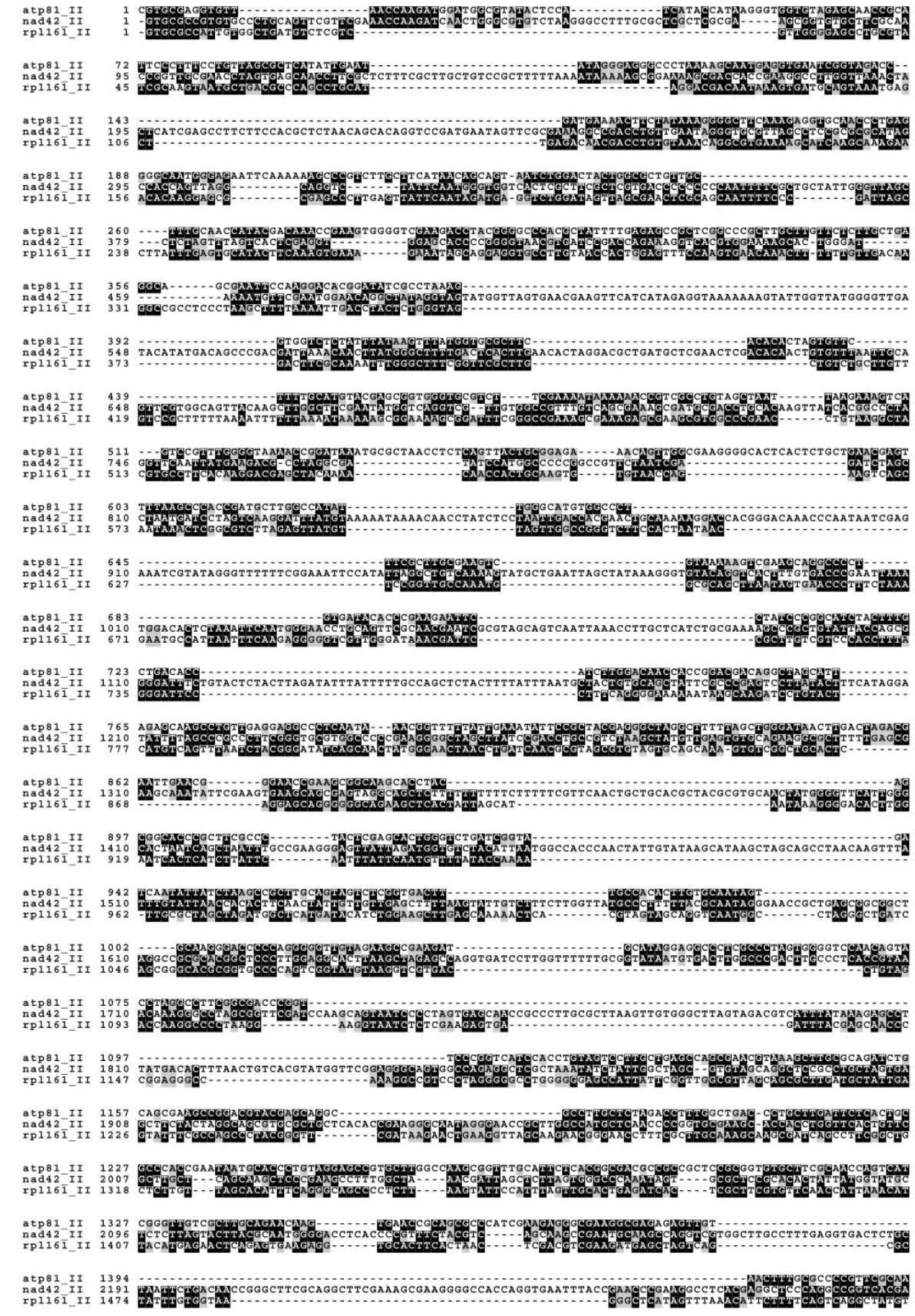

cob1_II 1
nad2I_II 1
cob1_II 101
nad2I_II 70
cob1_II 201
nad2I_II 82
cob1_II 301
nad2I_II 126
cob1_II 401
nad2I_II 226
cob1_II 492
nad2I_II 326
cob1_II 580
nad2I_II 419
cob1_II 673
nad2I_II 519
cob1_II 744
nad2I_II 614
cob1_II 844
nad2I_II 664
cob1_II 944
nad2I_II 714
cob1_II 1044
nad2I_II 783
cob1_II 1144
nad2I_II 783
cob1_II 1244
nad2I_II 783
cob1_II 1344
nad2I_II 783
cob1_II 1444
nad2I_II 812
cob1_II 1544
nad2I_II 857
cob1_II 1644
nad2I_II 881
cob1_II 1744
nad2I_II 958

Group 4:





Group 5:



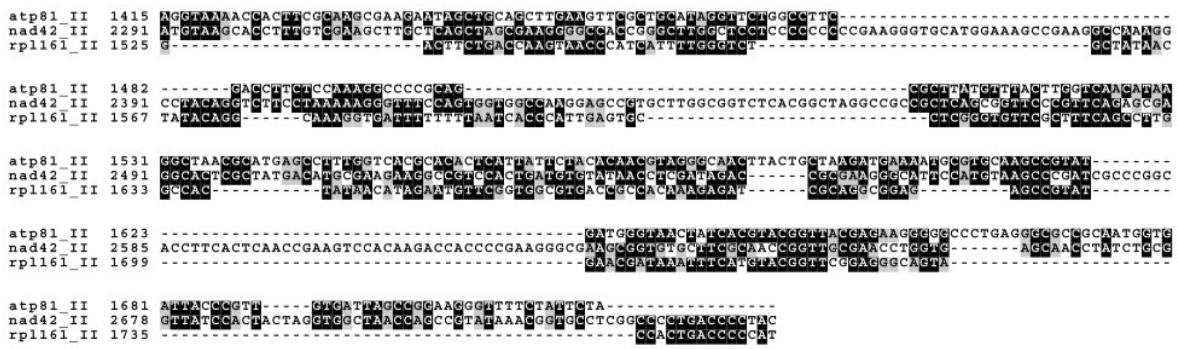

atp81_II 1415
nad42_II 2291
rpl16l_II 1525
atp81_II 1482
nad42_II 2391
rpl16l_II 1567
atp81_II 1531
nad42_II 2491
rpl16l_II 1633
atp81_II 1623
nad42_II 2585
rpl16l_II 1699
atp81_II 1681
nad42_II 2678
rpl16l_II 1735

Table S6 Estimates of base substitutions per site between genes from *O. quekettii* and *C. lentillifera* using the Jukes Cantor model conducted in MEGA. 1st, 2nd and 3rd codon positions were included. All ambiguous positions were removed for each sequence pair (pairwise deletion option).

Mitochondrion		**Chloroplast**							
atp6	0.562	atpI	0.290	petB	0.238	psbL	0.171	rps4	0.538
atp9	0.296	atpH	0.191	petD	0.214	psbM	0.254	rps7	0.463
atp1	0.425	atpA	0.297	petG	0.268	psbN	0.243	rps8	0.460
atp8	0.556	atpF	0.383	petL	0.467	psbT	0.202	rps9	0.743
cob	0.396	atpE	0.382	psaA	0.246	psbZ	0.352	rps11	0.555
cox1	0.348	atpB	0.273	psaB	0.246	rbcL	0.219	rps12	0.326
cox2	0.531	5s	0.531	psaC	0.204	rpl2	0.464	rps14	0.417
cox3	0.467	SSU	0.198	psaI	0.321	rpl5	0.502	rps18	0.629
5s	0.441	LSU	0.267	psaJ	0.310	rpl14	0.349	rps19	0.434
SSU	0.285	accD	0.537	psaM	0.339	rpl16	0.433	tufA	0.276
LSU	0.490	chlB	0.248	psbA	0.245	rpl19	0.418	ycf1	0.633
nad1	0.342	chlI	0.407	psbB	0.278	rpl20	0.479	ycf3	0.278
nad2	0.485	chlL	0.246	psbC	0.243	rpl23	0.402	ycf4	0.520
nad3	0.499	chlN	0.338	psbD	0.211	rpl32	0.503	ycf20	0.588
nad4	0.679	clpP	0.318	psbE	0.248	rpl36	0.244		
nad4L	0.364	cysA	0.397	psbF	0.241	rpoA	0.687		
nad5	0.635	cysT	0.525	psbH	0.342	rpoC1	0.707		
nad6	0.365	ftsH	0.477	psbI	0.246	rpoC2	0.426		
nad7	0.384	infA	0.38	psbJ	0.267	rps2	0.448		
nad9	0.568	petA	0.364	psbK	0.388	rps3	0.393		