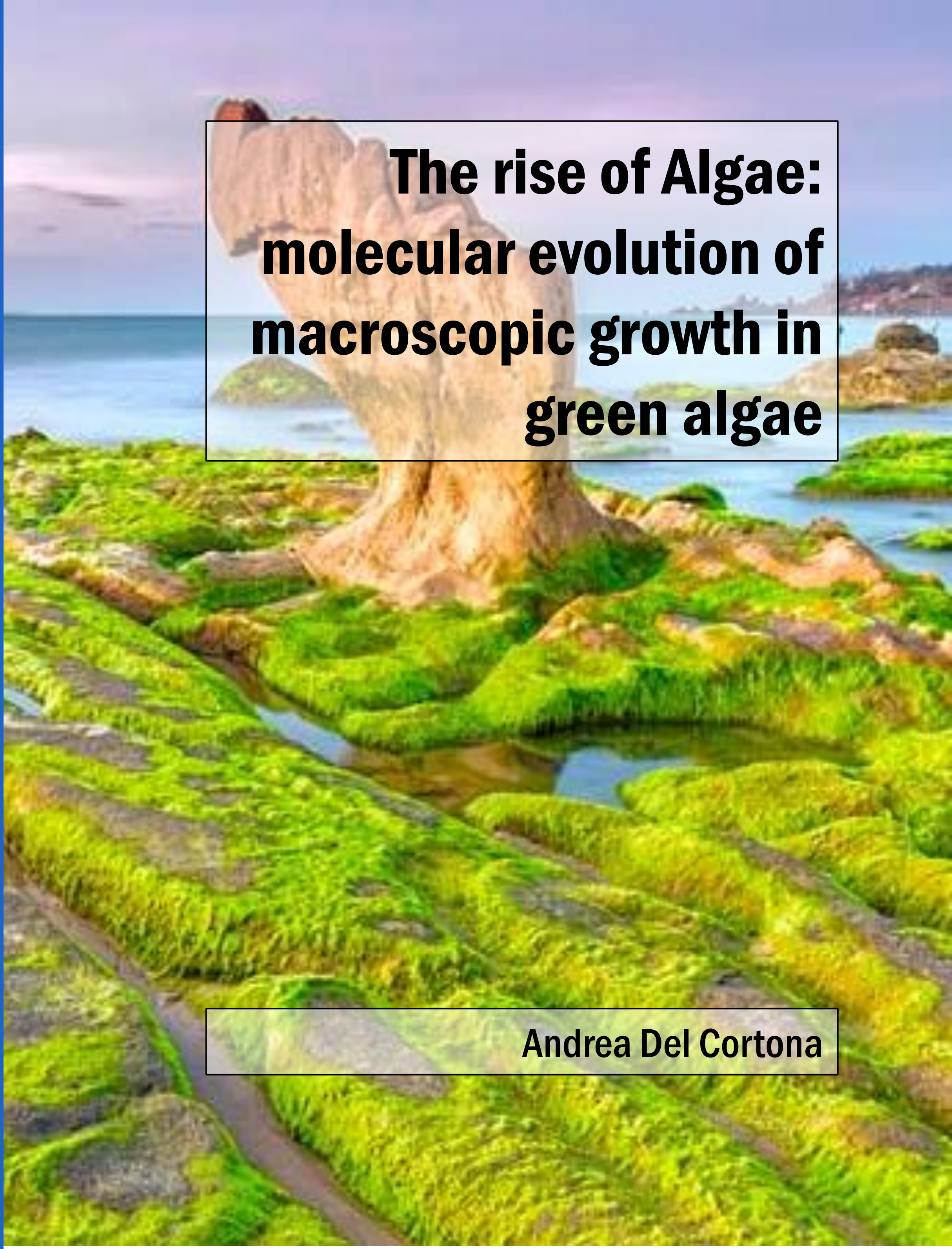


The rise of Algae: molecular evolution of macroscopic growth in green algae

Andrea Del Cortona

GHENT
UNIVERSITY

FACULTY
OF SCIENCES

VIB-UGENT
CENTER FOR PLANT
SYSTEMS BIOLOGY

The rise of Algae: molecular evolution of macroscopic growth in green algae

Andrea Del Cortona

2018

The rise of Algae: molecular evolution of macroscopic growth in green algae

Andrea Del Cortona

Ghent University

Faculty of Sciences, Department of Biology
Phycology Research Group
Department of Plant Biotechnology and Bioinformatics
Comparative Network Biology Group

The rise of Algae: molecular evolution of macroscopic growth in green algae

by

Andrea Del Cortona

Supervisors

Prof. Dr. Olivier De Clerck
Prof. Dr. Frederik Leliaert
Prof. Dr. Klaas Vandepoele

Dissertation submitted in partial fulfilment of the requirements for the degree of Doctor (Ph.D.) in Science, Biology (Ghent University)

Academic Year 2017-2018

Source of cover page picture:

https://labiotech.eu/algae-review-industry-biotech-greentech-biofuels-nutrition-scrubbing/

Members of the examination committee

Prof. Dr. Koen Sabbe, Ghent University (Chair)

Dr. Eske De Crop, Ghent University (Secretary)

Prof. Dr. Denis Baurain, University of Liège

Dr. Rolf Lohaus, Ghent University

Prof. Dr. Pavel Škaloud, Charles University

Prof. Dr. Olivier De Clerck, Ghent University*

Prof. Dr. Frederik Leliaert, Meise Botanic Garden and Ghent University*

Prof. Dr. Klaas Vandepoele, Ghent University*

*Non-voting members

This study was supported by: Ghent University (BOF/01j04813)

Contents

List of Figures

List of Tables

List of Abbreviations

BIR	Break-Induced Replication
Bp	base pairs
BUSCO	Benchmarking Universal Single-Copy Orthologs
Bya	Billion years ago
COG	Clusters of Orthologous Groups
coreGF	conserved Gene Family
cpDNA	chloroplast DNA
DNA	Deoxyribonucleic Acid
dsDNA	double-stranded Deoxyribonucleic Acid
Gb	Gigabase pairs
HMW	High Molecular Weigth (Deoxyribonucleic Acid)
HOM	Homologous family
IR	Inverted Repeats
Kb	Kilobase pairs
KO	KEGG Orthology
LMW	Low Molecular Weight (Deoxyribonucleic Acid)
LSC	Large single copy region
Mb	Megabase pairs
ML	Maximum likelihood
MLBS	Multilocus bootstrap support
mtDNA	mitochondrial genome
Mya	Million years ago
NGS	Next-Generation Sequencing
NOG	Non-supervised orthologous groups
rDNA	Ribosomal Deoxyribonucleic Acid
RNA-seq	Ribonucleic Acid sequencing
RT-LTRs	Long Terminal Repeat Retrotransposons
SMRT	Single-Molecule Real Time (sequencing)
SSC	Small single copy region
ssDNA	single-stranded Deoxyribonucleic Acid
TBCD	Trentepohliales, Bryopsidales, Cladophorales, Dasycladales
TCD	Trentepohliales, Cladophorales, Dasycladales
TPM	Transcripts per million
UTC	Ulvophyceae, Trebouxiophyceae, Chlorophyceae
UU	Ulvales, Ulotrichales

Acknowledgements

Innanzi tutto vorrei ringraziare i miei genitori che per piú di dieci anni mi hanno sostenuto moralmente ed economicamente in questa lunga maratona universitaria corsa in tre stati differenti. Una delle cose piú importanti che mi avete transmesso e insegnato é la volontá di sapere e la passione dello scoprire e dell'imparare, con la consapevolezza che non si é mai arrivati e che c'é sempre qualcosa di nuovo da imparare al prossimo incrocio. Queste lezioni lasciano un'attitudine alla vita che nessuna universitá e nessuna scuola dá. Grazie.

Grazie Gabriele, tutti i computer con cui ho lavorato (e giocato) me li hai regalati te. I miei primi ricordi sono legati all'Amiga 500 e alle navicelle spaziali. Ancora le sogno la notte. Penso che é solo grazie a questo allenamento e a tutte le cose stilosissime che abbiamo condiviso in tutti questi anni che ce l'ho fatta a sopravvivere a questa tesi passata dietro a un computer. Soprattutto mi hai insegnato a non arrendermi mai e a costruire e aggiustare, e non mi sono arreso. Grazie.

Anita, grazie per avermi sostenuto durante le parti piú difficili di questa tesi ("Tunneeeeeeel!!!"), grazie di essermi venuta a prendere tutti i giorni nel quartiere disagio, grazie di aver sopportato le alghe e di rendere ogni giorno una scoperta bellissima. Grazie per tutti i momenti meravigliosi che mi regali ogni giorno. Grazie di avermi preso per mano. Grazie.

Mathias ("il miglior amico del mi' fratello" & "il fratello del mi' migliore amico"), abbiamo vissuto il Belgio come una famiglia e abbiamo visto cose che voi umani non potete neanche immaginare: lo zighettatore, Eurolines, Caparezza, JackLaFuria. É stato proprio un divertimento. Ma soprattutto, non ti dimenticare del pompelmo e ascolta sempre il Ceccherini. Lui sa. Grazie.

(Chiedi a 77 se non sai come si fa)

Come sempre, un grazie anche a i miei amici a casa (e qui diventiamo sentimentali), in ordine sparso ma tutti rigorosamente importantissimi: Luca, Lorenzo, Daniele, Gaetano, Alessio, Samuele (Thille che fa le scintille), Mario, Gianluca, Elia. Grazie. E grazie anche ai Biotechnodrugs & OstelloBraina in transferta per il mondo – sempre nel cuore: Dade, Carmelo, ZiTotó (baciamo le mani), Totó, Frabrizio. Grazie.

Thanks the new friends that I found in Flanders: Andries, Kobe, Timo, Vasilis, Abhi. Thanks for all the beers and the good time! Thanks.

To my colleagues at the Sterre and at VIB, both past and present (and future, in the end, time is just an agreement): thank you for your help and for the great time we spent together in the last 4 (well… 5) years. You made the long hours and days of this research a great and enjoyable adventure, and I am very happy to have shared this path with you all. Thanks Dimitris, Samuel, Frédérique, Kenny (\../), Christophe, Xiaojie, Eylem, Quentin, Lan Anh, Kathryn, Soria, An Liu, Jonas, Sofie D'hondt, Anja Nohe, Monica, Paolo, Viviana, Tu, Sofie Vranken, Kyung, Maria, Josefin, Renaat, Pieter, Jeroen, Maxime, Eveline, Ilse, Alexandre, Tine, Olga, Claas, Anja Hondekyn, Frederike, Willem (uiiiiiiiiii!), Darja, Javier, Katerina, Heidi, Lotte, Eli, Gust, Jihae, Caroline, Emilio, Koen, Reinhoud, Sien, Emmelien, François, Dries, Luca, Michiel, Thomas, Stephane, Shu-Min, Lieven, Sri, Elisabeth, Shubhada, Frederik, Cristina, Jan, Ken, Marcelo, Oren, Rolf, Heike. Now, try to say that all in one breath. Sam, sorry, but I win the prize for the fattening (15 kilos gain in one year - "squeak for me piggy, squeak!" "uiiiiiiiiii uiiiiiiiiii"). If I forgot someone, thank you as well, and sorry for my labile memory, I guess I have studied seaweeds (50% sea – 50% weed) for a little bit too long. Thanks to you all.

This brings me to the national and international collaborators with whom I had the immense pleasure to work together: thanks Nicolas Dauchot, thanks Heroen Verbruggen, thanks Monique Turmel, thanks Jan Janouškovec. I have learned a lot from you and it has been a real pleasure to meet you all. I hope that in the future we will have more opportunities to share further research and, perhaps, a good bottle of wine. Thanks.

Last, but not least, I would like to thanks my scientific supervisors: Frederik Leliaert, Olivier De Clerck and Klaas Vandepoele. Thank you for your guidance in these years, for the continuous inspiration and for having contributed to my growth as a rational thinker, as a scientist and as a person. I know that I am not the most communicative person and that the switch from the wet lab to bioinformatics has not been easy. I am grateful that you did not gave up on me. Thanks for all your support and thanks for everything you managed to teach me. I am happy that in the end we managed to enrich the scientific world with two incredible stories about green seaweeds. Thanks.

Thank you all very much!

Grazie a tutti, davvero.

Summary

The ulvophytes, comprising the green seaweeds, are of particular evolutionary interest because they display a wide morphological and cytological diversity with unique features among green algae. Unfortunately, the lack of a robust phylogenetic framework hampers a full appreciation of their evolutionary history. For example, up till the present day it is still not clear if green seaweeds can be traced back to a single origin or whether their diversity is the result of multiple independent transitions from unicellular, freshwater ancestors to the marine coastal environment. Elucidating this phylogenetic history is the principle aim of this thesis.

In a first introductory chapter (**Chapter 1**), we describe green algal diversity in an evolutionary perspective, discuss recent progress towards understanding the genetic underpinning of highly specialized cyto-morphologies, such as siphonous cells, and describe the recent progress made towards the understanding of the peculiar and highly diverse chloroplast genomes found among ulvophyceans. We discuss the difficulties in resolving the ancient phylogenetic relationships among ulvophyceans, and the core Chlorophyta lineages, even when applying chloroplast phylogenomic analyses. Furthermore, we highlight the advantages and opportunities in using phylotranscriptomics as an alternative tool to resolve difficult phylogenetic relationships and to unveil the evolution of major cyto-morphological innovations and molecular features associated to them.

In **Chapter 2,** we set up a phylotranscriptomic workflow and evaluated the impact of partial datasets (i.e.: transcriptomes) on comparative genomics pipelines designed to process genomic data. We evaluated the performance of the commonly used available tools in annotating transcriptomes and estimating transcriptome completeness. Green algal gene features that could be detrimental to transcriptome annotation and to the phylotranscriptomic workflow were taken into account (e.g. alternative nuclear and chloroplast genetic codes). We further evaluated the relationships between depth of sequencing and transcriptome completeness and assessed the impact of missing data on gene family circumscription. Since gene family inference was robust and did not suffer from partial transcriptomic data, phylotranscriptomics downstream analyses are expected to be reliable. However, *de*

novo assembled transcriptomes seemed to overestimate the size of inferred gene families, suggesting that precautions should be taken before using transcriptomic data in gene family expansion, gain or loss analyses.

In **Chapter 3**, we used the insights gained and the generated workflow to process transcriptome and genome data from 55 species of green algae. A dataset of 539 single-copy nuclear genes was generated, and rigorous phylogenetic reconstructions resulted in a robust, highly supported reconstruction of the phylogenetic relationships and the divergence times of the major green algal lineages. Depending on whether a concatenation or a coalescent approach was used, analyses resulted in a clade composed by Bryopsidales and Chlorophyceae sister to the remaining Ulvophyceae or a radiation comprising Bryopsidales, Chlorophyceae and Ulvophyceae, respectively. Results are interpreted in relation to global-scale (de-)glaciations during the Cryogenian period of the Neoproterozoic (750-650 mya). This study narrows down the evolutionary history of green seaweeds to two competing scenarios. We argue that it represents a robust framework to build on the understanding of the molecular key features underlying the evolution of different and unique cyto-morphological features of ulvophyceans.

In **Chapter 4,** we solved the long standing mystery over the nature of Cladophorales green seaweeds chloroplast genome. We describe a highly deviant chloroplast genome which is entirely fragmented into hairpin chromosomes. Short and long read high-throughput sequencing of DNA and RNA demonstrated that the chloroplast genes of *Boodlea composita* (Cladophorales) are encoded on 1-7 kb DNA contigs with an exceptionally high GC-content, each containing a long inverted repeat with one or two protein-coding genes and conserved non-coding regions putatively involved in replication and/or expression. These contigs correspond to linear, single-stranded DNA molecules that fold onto themselves to form hairpin chromosomes. The origin of this highly deviant chloroplast genome likely occurred before the emergence of the Cladophorales, and coincided with an elevated transfer of chloroplast genes to the nucleus. A chloroplast genome that is composed only of linear DNA molecules is unprecedented among eukaryotes and highlights unexpected variation in the plastid genome architecture.

In the **General Discussion**, our findings are integrated with the results of previous studies. We evaluate and interpret our current understanding of green algal diversification on the light of molecular evolution and ultrastructural features. Finally, future research perspectives are explored.

Samenvatting

De Ulvophyceae of groene zeewieren zijn van bijzonder evolutionair belang door hun grote morfologische en cytologische diversiteit met unieke kenmerken onder de groenwieren. Het ontbreken van een stabiel fylogenetisch kader bemoeilijkt echter de studie van hun evolutionaire geschiedenis. Tot op heden is het bijvoorbeeld nog steeds niet duidelijk of groene zeewieren eenmaal ontstaan uit zijn in de evolutie, of dat hun grote diversiteit een gevolg is van meerdere onafhankelijke transities van eencellige zoetwatervoorouders naar de meercellige zeewieren de onze kusten hun typisch aspect verlenen. Dit proefschrift heeft als eerste belangrijk doel het ophelderen van de fylogenetische relaties van groenwieren.

In een eerste inleidend hoofdstuk (**hoofdstuk 1**) beschrijven ik de diversiteit van groenwieren in een evolutionair perspectief, wordt de recente vooruitgang in het begrijpen van de genetische onderbouwing voor gespecialiseerde cyto-morfologische kenmerken zoals sifonale cellen besproken. Tevens beschrijf ik de recentste inzichten in de diversiteit en complexiteit van chloroplastgenomen in de Ulvophyceae. De moeilijkheden bij het ophelderen van de oude fylogenetische relaties binnen de Ulvophyceae en de “core Chlorophyta”, zelf op basis van volledige chloroplastgenomen, worden bediscussieerd. Verder bespreek ik potentieel van fylotranscriptomische analyses om deze moeilijke fylogenetische relaties op te helderen, en om de evolutie te ontsluieren van cyto-morfologische en geassocieerde moleculaire kenmerken.

In **hoofdstuk 2** beschrijven we een nieuw ontwikkelde fylotranscriptomische workflow en bestuderen we de impact van partiële transcriptoom datasets op “pipelines” die ontworpen zijn om genoomdata te verwerken. We evalueren de prestaties van veelgebruikte tools voor het annoteren van transcriptomen en voor het schatten van de volledigheid van transcriptomen. Specifieke genoomkenmerken van groenwieren die de annotatie van transcriptomen en fylotranscriptomische workflows mogelijk kunnen bemoeilijken, zoals alternatieve genetische codes, werden in aanmerking genomen. Verder evalueren we de relaties tussen sequencing-diepte en de volledigheid van het transcriptoom, en beoordelen we de impact van ontbrekende data op de afbakening van genfamilies. Aangezien de afbakening van genfamilies stabiel

was en niet negatief beïnvloed werd door onvolledige transcriptoomdata, wordt verwacht dat de “downstream” fylotranscriptomische analyses betrouwbaar zijn. Toch bleken de *de novo* geassembleerde transcriptomen de grootte van de afgebakende genfamilies te overschatten. Daarom is voorzichtigheid geboden is bij het gebruik van transcriptoomdata voor de evolutionaire analyse van genfamilies.

In **hoofdstuk 3** worden de verkregen inzichten en de ontwikkelde workflow uit hoofdstuk 2 gebruikt om transcriptoom- en genoomdata van 55 groenwiersoorten te verwerken. Dit resulteerde in een dataset van 539 “single-copy” nucleaire genen, die vervolgens geanalyseerd werd aan de hand van geavanceerde fylogenetische methodes. De analyses resulteerden in een sterk ondersteunde fylogenetische boom, en een tijdschaal voor de evolutie van de belangrijkste groenwiergroepen. Afhankelijk van het type van analyse (gebaseerd op een geconcateneerde dataset of een coalescentie-gebaseerde analyse), werd een relatie tussen Bryopsidales en Chlorophyceae afgeleid, of een radiatie bestaande uit Chlorophyceae, Bryopsidales en de rest van de Ulvophyceae. De fylogenetisch resultaten worden geïnterpreteerd in het licht van wereldwijde glaciaties tijdens het Cryogenium (750-650 mya). De fylogenie vormt tevens een belangrijk kader voor verdere studies naar de evolutie van genoomkenmerken die aan de grondslag liggen van de cyto-morfologische variatie binnen de Ulvophyceae.

Hoofdstuk 4 focust op de structuur van het chloroplastgenoom van groenwieren in de orde Cladophorales. We beschrijven een sterk afwijkend chloroplastgenoom dat volledig gefragmenteerd is in korte, haarspeld-vormige chromosomen. “High-throughput” sequencing van DNA (korte en lange reads) en RNA toont aan dat de chloroplastgenen van *Boodlea composita* gecodeerd zijn op 1-7 kb DNA contigs met een uitzonderlijk hoog GC-gehalte, elk met een lange omgekeerde repeat met één of twee eiwitcoderende genen en geconserveerde niet-coderende regio’s die vermoedelijk betrokken zijn bij replicatie en/of genexpressie. Deze contigs zijn lineaire, enkelstrengige DNA-moleculen die zich vouwen in een haarspeld-vormige structuur. De evolutie van deze sterk afwijkende chloroplastgenomen in de Cladophorales valt samen met een verhoogde transfer van chloroplastgenen naar de kern. Een chloroplastgenoom dat enkel bestaat uit lineaire DNA-moleculen is uniek binnen de eukaryoten, en benadrukt de grote structurele variatie in chloroplastgenomen binnen de Ulvophyceae.

In de **algemene discussie** worden de resultaten van dit proefschrift geëvalueerd en geïnterpreteerd in het licht van eerdere studies. In het bijzonder worden nieuwe inzichten in de diversificatie van groenwieren en de evolutie van moleculaire en ultrastructurele kenmerken besproken. Tot slot worden onderzoeksperspectieven verkend.

“Your matter is energy,

‘cause your energy matters”

Chapter 1 - Molecular evolution and morphological diversification of ulvophyceans (Chlorophyta)[1]

Andrea Del Cortona and Frederik Leliaert[2]

“Die Pflanzen bildet Zellen, nicht die Zellen bildet die Pflanzen”

Heinrich Anton de Bary

[1] Published as:
Andrea Del Cortona, Frederik Leliaert – Molecular evolution and morphological diversification of Ulvophytes (Chlorophyta) – Perspectives in Phycology 5-1, 27-43 June 1, 2018, https://doi.org/10.1127/pip/2017/0075

[2] Authors contribution: A.D.C., F.L.: manuscript conceptualization, drafting and writing.

Introduction

How organisms evolve different morphologies is a central question in evolutionary biology, but in many taxonomic groups, including algae, answers remain elusive. Molecular phylogenetic and evolutionary approaches provide a powerful framework for investigating the history of morphological change, and the genetic basis of morphological diversification. Because they are morphologically and cytologically so diverse, the Ulvophyceae (green seaweeds) form an interesting group to investigate the evolution of morphological change, and the evolutionary forces shaping morphological divergence.

Seaweeds are ecologically important primary producers in coastal ecosystems worldwide. Seaweeds are not a natural group, but evolved independently from the ancestors of red (Rhodophyta), green (Viridiplantae), and brown algae (Phaeophyceae). The red algae are an ancient lineage, most likely originating in the Mesoproterozoic in low-salinity environments, with a single origin of macroscopic growth in coastal habitats more than 1.1 Bya (Sánchez-Baracaldo *et al.*, 2017). The brown algae, which are entirely restricted to the coastal environment, are much more recent, diverging only in the Lower Jurassic (Berney & Pawlowski, 2006; Brown & Sorhannus, 2015). Fossil evidence for the origin of the green algal lineage points towards a Neoproterozoic origin, somewhere between 1,000 and 500 mya (Falkowski *et al.*, 2004a; Porter, 2004). Green algae comprise a wide diversity of unicellular and multicellular forms from freshwater, terrestrial, and marine environments.

There is a general consensus that an early split in the evolution of green algae gave rise to two discrete clades. One clade, the Streptophyta, contains a wide diversity of green algae from freshwater and damp terrestrial habitats (known as charophytes), from which the land plants evolved in the Ordovician (McCourt *et al.*, 2004). The second clade, the Chlorophyta, diversified as planktonic unicellular organism (prasinophytes), mainly in oceanic environments in the Neoproterozoic (Porter, 2004; Knoll *et al.*, 2006). Prasinophytes later gave rise to the core Chlorophyta, which radiated in freshwater, terrestrial, and coastal environments (Leliaert *et al.*, 2012). Traditionally, five classes are recognized in the core Chlorophyta, including the species-rich Ulvophyceae, Trebouxiophyceae and Chlorophyceae (also known as the UTC clade), and two smaller classes of unicellular algae from marine, freshwater and soil habitats, Chlorodendrophyceae and Pedinophyceae. The Chlorophyceae and Trebouxiophyceae include unicellular and multicellular algae from freshwater and terrestrial environments. Green seaweeds are nearly exclusively restricted to the Ulvophyceae, but the class also includes some freshwater and terrestrial representatives (Škaloud *et al.*, 2018).

The Ulvophyceae is one of the major classes of green algae, comprising more than 1,700 species (Guiry, 2012). Most species are macroscopic (Figure 1.1 and 1.2), but a substantial diversity of microscopic organisms are also present. Although the ulvophyceans are best known as marine species from rocky shores and coral reefs (Brodie *et al.*, 2007), an increasing number of species is being uncovered in brackish, freshwater, and moist subaerial habitats such as soil, rocks, tree bark and leaves (Škaloud *et al.*, 2018). Nine orders are currently recognized in the Ulvophyceae: Bryopsidales, Cladophorales, Dasycladales, Ignatiales, Oltmansiellopsidales, Scotinosphaerales, Trentepohliales, Ulotrichales and Ulvales.

The Ulvophyceae display a wide variety of thallus and cellular organisations (Figure 1.1 and 1.2). Four main cyto-morphological types have been distinguished by Cocquyt et al. 2010a (see Figure 1.2 caption). Type 1 includes flagellate or non-flagellate unicellular or colonial organisms with uninucleate cells. This type is present in some Ulvales, Ulotrichales, Scotinosphaerales, Oltmansiellopsidales, and the Ignatiales (Chihara *et al.*, 1986; Nakayama *et al.*, 1996; Friedl & O'Kelly, 2002; Watanabe & Nakayama, 2007; Škaloud *et al.*, 2013). Type 2 consists of multicellular filaments or blades composed of uninucleate cells. This type occurs in the Ulvales, Ulotrichales and Trentepohliales. Type 3 is the siphonocladous thallus organisation, which is characterized by multicellular thalli composed of multinucleate cells with nuclei organized in regularly spaced cytoplasmic domains (McNaughton & Goff, 1990; Motomura, 1996). This type is found in the Cladophorales and *Blastophysa* and some members of the Ulotrichales (e.g., *Urospora* and *Acrosiphonia*). Type 4 is the siphonous thallus organisation, which is characterized by thalli consisting of a single giant tubular cell. It is present in the orders Bryopsidales and Dasycladales. In most species, the siphonous cells contain thousands of nuclei, but in several species of Dasycladales, the siphonous thallus remains uninucleate throughout much of their life cycle with a giant diploid nucleus that only divides at the onset of reproduction (Berger

& Kaever, 1992). Siphonous cells typically exhibit cytoplasmic streaming, transporting organelles, nutrients, and transcripts across the thallus (Menzel, 1987; Menzel, 1994; Mine *et al.*, 2005). Some siphonous species form large seaweeds with thalli differentiated into distinct structures, including rhizoids, stolons and blades. The giant cells of siphonocladous and siphonous species are characterized by several cytological specializations, such as unique mechanisms of cell differentiation, cell division, and wounding response (Menzel, 1988; La Claire, 1992; Kim *et al.*, 2001; Mine *et al.*, 2008).

Cyto-morphological evolution

The phylogeny of Cocquyt et al. (2010a) provided a first important framework to understand the evolution and cyto-morphological diversification of ulvophyceans. Three alternative scenarios for the cyto-morphological diversification of ulvophyceans were hypothesized. In a likely scenario, ulvophyceans originated from an ancestral unicellular, uninucleate organism, and macroscopic growth emerged through different mechanisms several times independently during the evolution of the class, leading to the current variety of thallus organizations (Figure 1.2). This view is supported by the fact that several early diverging clades of ulvophyceans (Oltmannsiellopsidales, Ignatiales and Scotinosphaerales, Figure 1.1A-1.1C) include unicellular uninucleate species, in addition to colonial forms (Watanabe & Nakayama, 2007; Leliaert *et al.*, 2009; Škaloud *et al.*, 2013; Turmel *et al.*, 2017; Škaloud *et al.*, 2018).

The Ulvales and Ulotrichales both contain a multicellular species with uninucleate cells (Figure 1.1D, 1.1E), and a number of unicellular uninucleate species (e.g. *Pseudoneochloris* and *Pirula*). Multicellularity possibly evolved independently in the two clades, or alternatively, these unicellular species represent reductions from multicellular ancestral types. Multicellular thalli range from branched or unbranched filaments to more complex tubular and blade-like morphologies. An interesting aspect of multicellular growth comes from culture experiments on some multicellular Ulvales. When grown axenically, *Ulva* and *Monostroma* form callus-like colonies of undifferentiated cells. However, upon the addition of secretions from two distinct bacteria, from *Roseobacter* and *Cytophaga* for *Ulva*, and secretions from an uncharacterized member of the CFB group for *Monostroma*, the blade-like multicellular

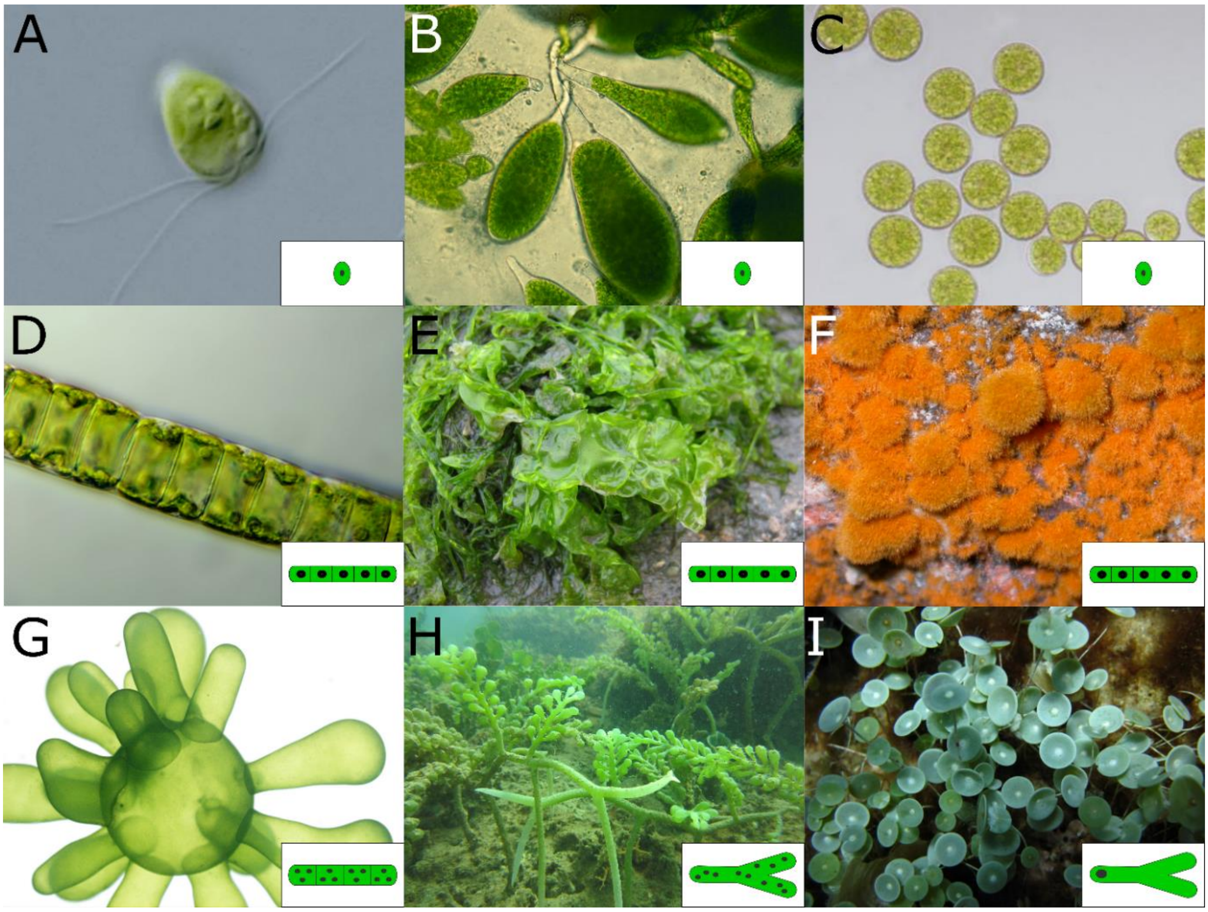


Figure 1.1: Ulvophyceae diversity.

(A) Light micrograph of *Oltmannsiellopsis unicellularis* (Oltmannsiellopsidales), a flagellate unicellular uninucleate ulvophyte. Image courtesy of NCMA at Bigelow Laboratory. (B) Light micrograph of *Scotinosphaera lemnae* (Scotinosphaerales), a non-motile unicellular uninucleate ulvophyte. Image courtesy of Pavel Škaloud, Charles University, Prague. (C) Light micrograph of *Ignatius tetrasporus* (Ignatiales), a non-motile unicellular uninucleate ulvophyte. Image courtesy of UT-Austin. (D) Light micrograph of *Ulothrix* (Ulotrichales), a multicellular filamentous ulvophyte with uninucleate cells. Image courtesy of Giuseppe Vago. (E) *Ulva lactuca* (Ulvales), a multicellular blade-forming ulvophyte with uninucleate cells. Image courtesy of Kristian Peters. (F) *Trentepohlia aurea*, a multicellular filamentous ulvophyceans with uninucleate cells. Image courtesy of Alain Gerault. (G) *Valonia utricularis* (Cladophorales), a multicellular multinucleate ulvophyte. Photo by Frederik Leliaert. (H) *Caulerpa racemosa* (Bryopsidales), a siphonous, unicellular multinucleate ulvophyte. Photo by Frederik Leliaert. (I) *Acetabularia acetabulum* (Dasycladales), a siphonous, unicellular uninucleate ulvophyte. Image courtesy of StudyBlue. For each species, the corresponding cyto-morphotype has been represented as in Figure 1.2.

thallus composed of differentiated cells is restored (Matsuo *et al.*, 2005; Spoerner *et al.*, 2012). Some members of the Ulotrichales (e.g. *Acrosiphonia* and *Urospora*) have evolved a siphonocladous thallus organization, separately from the Cladophorales (see below). In these Ulotrichales, all nuclei of the multinucleate cell migrate to the future plane of cell division before mitosis, and division of the nuclei is synchronous (Hudson & Waaland, 1974; Lokhorst & Star, 1983). This contrasts with the situation in Cladophorales, where mitosis is uncoupled from cytokinesis.

The order Trentepohliales (Figure 1.1F) includes multicellular, branched filamentous or pseudoparenchymatous thalli, and likely evolved multicellularity independently from the other multicellular clades (Figure 1.2) (Cocquyt *et al.*, 2010b; Brooks *et al.*, 2015). The uninucleate cells have unique features among Chlorophyta, that are shared with land plants, such as phragmoplast-like cytokinesis and presence of plasmodesmata between vegetative cells (Chapman & Henk, 1985; Chapman *et al.*, 2001). Motile reproductive cells have multilayered structures associated with flagellar bases instead of a cruciate flagellar root system, which is typical for most core Chlorophyta (Graham & McBride, 1975). The atypical cellular characteristics of Trentepohliales and their strictly terrestrial habitat suggest intriguing evolutionary parallelisms between this clade of ulvophyceans and the streptophytes.

From a cytological perspective, the orders Cladophorales, Bryopsidales, and Dasycladales are the most peculiar clades of ulvophyceans, characterized by a siphonous or siphonocladous thallus organization, featuring highly specialized cellular characteristics.

The Cladophorales (Figure 1.1G) are mostly macroscopic or sometimes microscopic plants with a siphonocladous architecture, i.e. multicellular plants composed of multinucleate cells. Thallus morphology is very diverse, ranging from unbranched or branched filaments to blade-like or giant-celled thalli with unique cytological traits and modes of cell division (Leliaert *et al.*, 2007; Mine *et al.*, 2008). Cells range in size from a few µm to several cm, and have a large central vacuole surrounded by a thin layer of cytoplasm containing numerous nuclei and chloroplasts. Chloroplasts are often interconnected by delicate strands forming a parietal network or a more or less continuous layer. The multinucleate cells present regularly-spaced nuclei in a stationary cytoplasm and arrays of internuclear microtubules which define regular cytoplasmic domains, one for each nucleus (McNaughton & Goff, 1990). This is in contrast with the situation in Bryopsidales and Dasycladales where the cytoplasm exhibits vigorous streaming. Nuclear division is synchronous or circumscribed to discrete mitosis patches (Hori & Enomoto, 1978; Staves & La Claire, 1985; Motomura, 1996; Okuda *et al.*, 1997). Siphonocladous thallus organization is also found in the sister clade of Cladophorales, including the marine endophytic alga *Blastophysa*, where the nuclei divide in regular mitotic waves (Sears, 1967).

The Bryopsidales (Figure 1.1H) and Dasycladales (Figure 1.1I) have mostly macroscopic or sometimes microscopic thalli with a siphonous architecture. The most striking feature of Bryopsidales and Dasycladales is that a single siphon can form complex and often large plants differentiated in root-like, stem-like and blade-like structures that present metabolic and transcript partitioning (Chisholm *et al.*, 1996; Ranjan *et al.*, 2015). Like in the Cladophorales, the cell consists of a large central vacuole, surrounded by a thin parietal layer of cytoplasm. However, both Bryopsidales and Dasycladales exhibit cytoplasmic streaming, which enables transportation of organelles and nutrients, as well as RNA transcripts, throughout the siphonous thallus. In Bryopsidales and some species of Dasycladales, the siphon contains thousands to millions of nuclei that divide by asynchronous mitosis. In most species of Dasycladales, however, the siphonous thallus contains a single giant diploid nucleus that only divides at the onset of reproduction (Berger & Kaever, 1992). This is for example the case in *Acetabularia* (mermaid's wineglass), the best studied genus of Dasycladales, where the relationship between complex cytoskeleton organization, discrete transcripts distribution and complex thallus morphology has been well characterized (Menzel, 1994; Serikawa *et al.*, 2001; Vogel *et al.*, 2002; Mine *et al.*, 2005; Mine *et al.*, 2008).

Although Cladophorales, Bryopsidales, and Dasycladales share large multinucleate cells, Cocquyt et al. (2010a) suggested that their ancestor may have been an uninucleate organism that gained macroscopic growth through cell enlargement in two distinct evolutionary events, possibly as a result of selective pressures for macroscopic growth in marine benthic environments, eventually leading to siphonocladous or siphonous thalli. A multinucleate siphonous/siphonocladous thallus may have presented several advantages: multiple genome copies in a single cell grant a buffer against deleterious mutations and a higher metabolic rate due to additional copies of ribosomal DNA (rDNA) (Niklas, 2014).

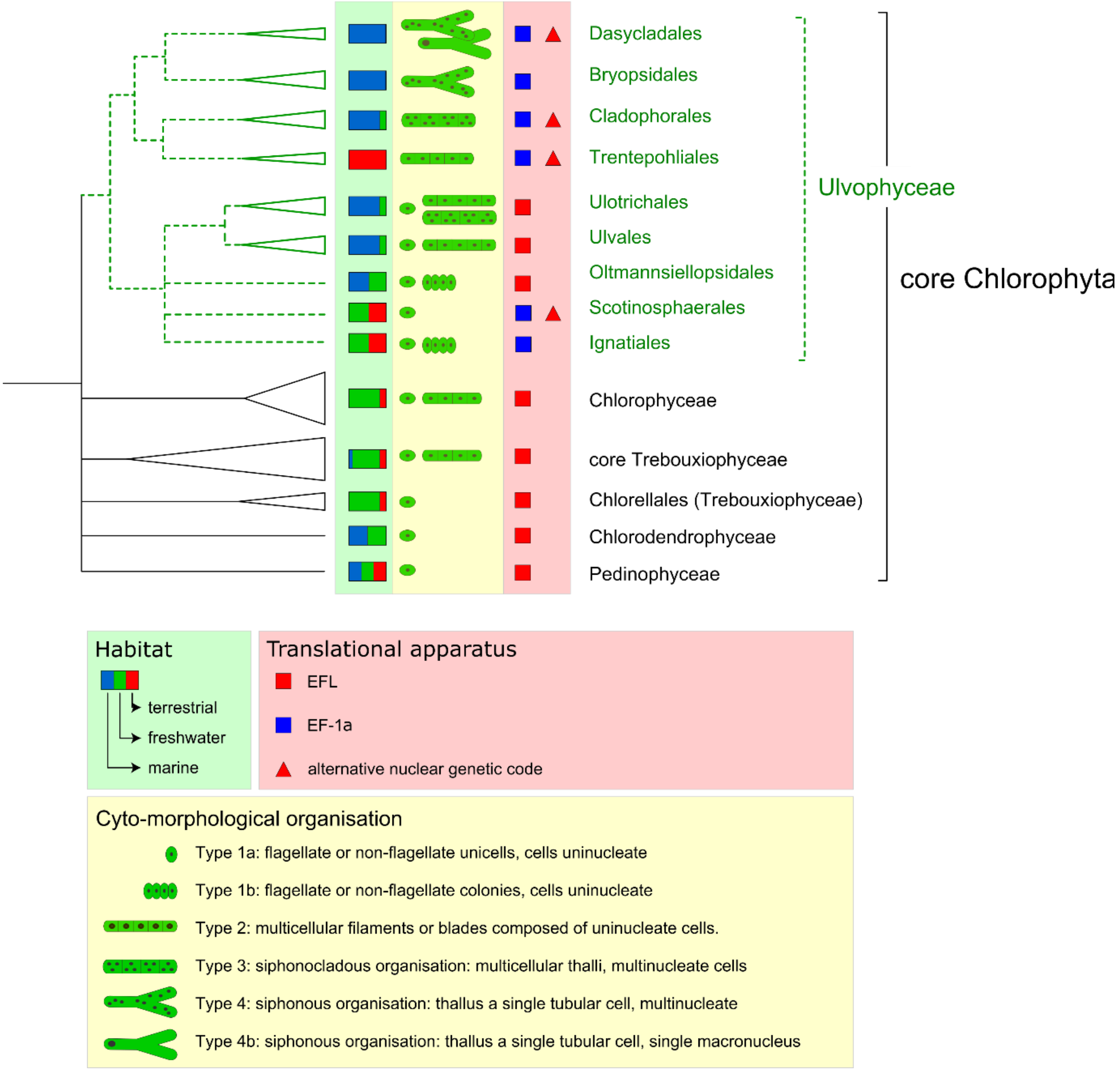

Dasycladales
Bryopsidales
Cladophorales
Trentepohliales
Ulotrichales
Ulvales
Oltmannsiellopsidales
Scotinosphaerales
Ignatiales
Ulvophyceae
core Chlorophyta
Chlorophyceae
core Trebouxiophyceae
Chlorellales (Trebouxiophyceae)
Chlorodendrophyceae
Pedinophyceae
Habitat
terrestrial
freshwater
marine
Translational apparatus
EFL
EF-1a
alternative nuclear genetic code
Cyto-morphological organisation
Type 1a: flagellate or non-flagellate unicells, cells uninucleate
Type 1b: flagellate or non-flagellate colonies, cells uninucleate
Type 2: multicellular filaments or blades composed of uninucleate cells.
Type 3: siphonocladous organisation: multicellular thalli, multinucleate cells
Type 4: siphonous organisation: thallus a single tubular cell, multinucleate
Type 4b: siphonous organisation: thallus a single tubular cell, single macronucleus

Figure 1.2: Overview of core Chlorophyta molecular and cytomorphological features.

Overview phylogeny of the core Chlorophyta with a focus on the Ulvophyceae. For each clade, the habitat, cyto-morphology and distribution of elongation factors and non-canonical nuclear genetic code is reported. Ulvophyte orders are reported in green, dashed lines between the orders indicate our current lack of knowledge on their phylogenetic relationships (see as well Figure 1.3).

The ancestor of Bryopsidales and Dasycladales was possibly a uninucleate unicellular enlarged cell, as supported by the presence of a single macronucleus in Dasycladales and in the zygote of certain Bryopsidales (Burr & West, 1971; Liddle *et al.*, 1976). In the scenario proposed by Cocquyt et al. (2010b), the evolution of siphonous thallus organization required two discrete steps. In the first step, the ancestral cell underwent elongation and development of the macronucleus and cytoplasmic streaming. Beyond a certain cell size, however, the time required for macromolecules and especially messenger RNAs to travel from one side of the cell to the other increases drastically, which would have resulted in a switch to a multinucleate status (Niklas, 2014). Multinucleate siphons probably evolved in Bryopsidales and Dasycladales independently (Cocquyt *et al.*, 2010b).

Make-over of the translational apparatus

Translation is the first stage of protein biosynthesis and an essential process in all living systems. It involves a complex interaction of several macromolecules, including messenger, ribosomal and transfer RNAs, ribosomal proteins and several associated protein factors. Although, the translation machinery is highly conserved across eukaryotes, lineage specific deviations exist, including alterations of the genetic code, lateral transfers and elevated evolutionary rates of essential genes. One of these lineages are the ulvophyceans, which features several atypical translation-related features in some representatives of the class.

First, the ulvophyceans present a complex distribution of an alternative nuclear genetic code (Figure 1.2), according to the phylogeny recovered by Cocquyt et al. where Bryopsidales, Cladophorales, Dasycladales and Trentepohliales formed a distinct clade (Cocquyt *et al.*, 2010a). While Cladophorales, Dasycladales and Trentepohliales evolved a non-canonical nuclear genetic code, where the stop codons TAG and TAA are reassigned to glutamine, Ignatiales and Bryopsidales conserve a standard nuclear genetic code (Gile *et al.*, 2009; Cocquyt *et al.*, 2010a). A stepwise acquisition model, involving ambiguous intermediates with a dual function of TAG and TAA as both coding and terminating codon (Santos *et al.*, 2004), was suggested as an explanation for the observed pattern.

Second, two distinct and mutually exclusive elongation factors, EF-1α and EFL, are scattered over the different lineages of ulvophyceans (Noble *et al.*, 2007; Cocquyt *et al.*, 2009; Gile *et al.*, 2009; Cocquyt *et al.*, 2010a). EF-1α and EFL function as translation initiation, elongation and termination by recruiting aminoacyl tRNAs to the ribosomes (Negrutskii & El'skaya, 1998; Keeling & Inagaki, 2004). While Ulvales have the elongation factor EFL like the rest of chlorophytes; Bryopsidales, Cladophorales, Dasycladales and Ignatiales present instead the elongation factor EF-1α (Cocquyt *et al.*, 2009). Although the model of Cocquyt et al. could not generate an unequivocal prediction for EFL acquisition in Chlorophyceae, the suggested step-wise transformation of the translational apparatus during the evolution of ulvophyceans is a likely scenario (Cocquyt *et al.*, 2009; Cocquyt *et al.*, 2010a). It has to be noted, however, that a complex EF-1α and EFL distribution is typical of most of the eukaryotic lineages (Keeling & Inagaki, 2004; Mikhailov *et al.*, 2014), and that both elongation factors have been found to co-occur in distantly related eukaryotic clades (Kamikawa *et al.*, 2013). A conservative model to describe the distribution of EF-1α and EFL among eukaryotes has not been formulated to date.

At last, phylogenies inferred from nuclear encoded rDNA sequences show that several ulvophyte clades are preceded by extremely long branches, indicating lineage-specific rate acceleration of the rDNA genes (Leliaert *et al.*, 2009). These factors combined provide clues that the diversification of ulvophyceans, and in particular the siphonous and siphonocladous lineages, coincided with profound changes in the translational machinery. A better understanding of the evolution of translation in the Ulvophyceae translation apparatus will require further analysis of the genetic code, codon usage, and characterisation of specific genes, including elongation factors, release factors, ribosomal proteins, and tRNAs in a phylogenetic framework.

A shaky phylogeny of ulvophyceans

A solid phylogeny of the Ulvophyceae is an important first step to understand the evolution of cyto-morphological types in the class. Unfortunately, as will be discussed below, the relationships among the main clades of ulvophyceans are still uncertain. Even monophyly of the class is under debate, as is the relationship of the Ulvophyceae with other classes of core Chlorophyta.

The original circumscription of the class was based on a set of ultrastructural characteristics, including a counter-clockwise orientation of the flagellar root system, cytokinesis by furrowing, a closed persistent mitotic spindle and the absence of a phycoplast (Mattox & Stewart, 1984; O'Kelly & Floyd, 1984; Sluiman, 1989a; Floyd &

O'Kelly, 1990). Some species have flagellate reproductive cells with cell walls or flagella covered by organic body-scales (Sluiman, 1989a). However, because none of these characters are unique to the Ulvophyceae, the monophyly of the Ulvophyceae has been questioned. For example, a counter-clockwise orientation of the flagellar root system is also found in species of Trebouxiophyceae, cytokinesis by furrowing occurs in most green algae, a closed mitosis occurs also occur in the Chlorophyceae, a persistent mitotic spindle is also characteristic for many charophyte green algae, and a phycoplast is absent in prasinophytes (Mattox & Stewart, 1984; O'Kelly & Floyd, 1984; Leliaert *et al.*, 2012). Organic body-scales occur in a various green algae, and are generally regarded as an ancestral character of the green algae (Melkonian, 1990). The order Trentepohliales, which has been affiliated with the Ulvophyceae based on nuclear rDNA data (Zechman *et al.*, 1990), displays atypical ultrastructural features, such as the presence of a phragmoplast and multilayered structures associated with flagellar bases in motile cells, instead of a cruciate flagellar root system (Graham & McBride, 1975). The order is also highly atypical from an ecological point of view as they are entirely restricted to terrestrial habitats.

Molecular systematics brought new hope to resolve ulvophyte relationships (Figure 1.3). Phylogenetic analyses of nuclear ribosomal rDNA datasets (mainly 18S) have supported the circumscription of traditional orders but were not able to fully resolve the relationships among them (Chappell *et al.*, 1991; Lopez-Bautista & Chapman, 2003; Watanabe & Nakayama, 2007; Leliaert *et al.*, 2009). In addition, 18S phylogenetic analyses were also not able to solve the question of monophyly of the class: most studies recovered the Ulvophyceae as a monophyletic group, although never with strong phylogenetic support. Instead these studies consistently recovered two distinct clades: a first clade comprising the Ulvales and Ulotrichales along with the Oltmannsiellopsidales, Scotinosphaerales and Ignatiales, and a second clade consisting of Trentepohliales, Cladophorales, Bryopsidales and Dasycladales (Zechman *et al.*, 1990; Watanabe *et al.*, 2001; Lopez-Bautista & Chapman, 2003; Watanabe & Nakayama, 2007; Cocquyt *et al.*, 2009; Škaloud *et al.*, 2013) (Figure 1.3). The first study to support monophyly of ulvophyceans with high support was a phylogenetic analysis based on 10 genes (eight nuclear and two plastid) (Cocquyt *et al.*, 2010b). This study also confirmed the divergence of the two main ulvophycean clades (Ulvales-Ulotrichales versus Bryopsidales-Cladophorales-Dasycladales-

Ignatiales-Trentepohliales), which was also supported by other genetic features, such as presence of the elongation factor-1 alpha or elongation factor-like gene, and the presence of a non-canonical genetic code (Cocquyt *et al.*, 2009; Gile *et al.*, 2009; Cocquyt *et al.*, 2010a).

More recently, phylogenetic analysis based on chloroplast multi-gene data have again altered our view on ulvophyte relationships, and have indicated a polyphyletic Ulvophyceae, consisting of two or more separate lineages that are interspersed with other core chlorophytan clades (Leliaert *et al.*, 2012; Fang *et al.*, 2017). An early chloroplast phylogenetic analysis of 23 chloroplast (cp) genes recovered *Caulerpa* (Bryopsidales) as more closely related to *Chlorella* (Trebouxiophyceae) than to the other two ulvophycean taxa in the phylogeny (*Oltmannsiellopsis* and *Tupiella*) (Zuccarello *et al.*, 2009), and a phylogeny inferred from 42 cp genes indicated a relationship between *Bryopsis* and Chlorophyceae (Lü *et al.*, 2011). Phylogenomic analyses with increased taxon sampling (53 taxa, 7 cp genes + 18S; and 38 taxa, 53 cp genes) suggested a relationship between *Oltmannsiellopsis*+ and *Tetraselmis* (Chlorodendrophyceae), and similarly to the two previous studies suggested polyphyly of the Ulvophyceae (Fučíková *et al.*, 2014). More recent, chloroplast phylogenomic analyses were also unable to confirm monophyly of the ulvophyceans, but more importantly could not resolve relationships among the main core chlorophytan lineages (Leliaert & Lopez-Bautista, 2015; Melton *et al.*, 2015; Sun *et al.*, 2016; Turmel *et al.*, 2016a; Fang *et al.*, 2018). However, increased taxon sampling (100 Chlorophyta, including 15 ulvophyceans) and increased chloroplast gene sampling (79 protein coding genes, 3 rRNA and 26 tRNA genes) recovered a monophyletic Ulvophyceae, although with relatively low support (Turmel *et al.*, 2017). In this study relationships among some of the main ulvophyte clades (Ulvales, Ulotrichales, Oltmannsiellopsidales, Ignatiales, and Bryopsidales) were relatively well resolved. Some important ulvophyte clades, however, are currently missing from chloroplast phylogenomic analyses, including the Cladophorales and Scotinosphaerales, and for Trentepohliales and Dasycladales, only partial gene content is available.

What is obvious is that it is extremely difficult to resolve the early divergences of the core Chlorophyta and Ulvophyceae because of the antiquity of these green algae, with divergences likely in the Proterozoic (Verbruggen *et al.*, 2009), and the rapidity of the

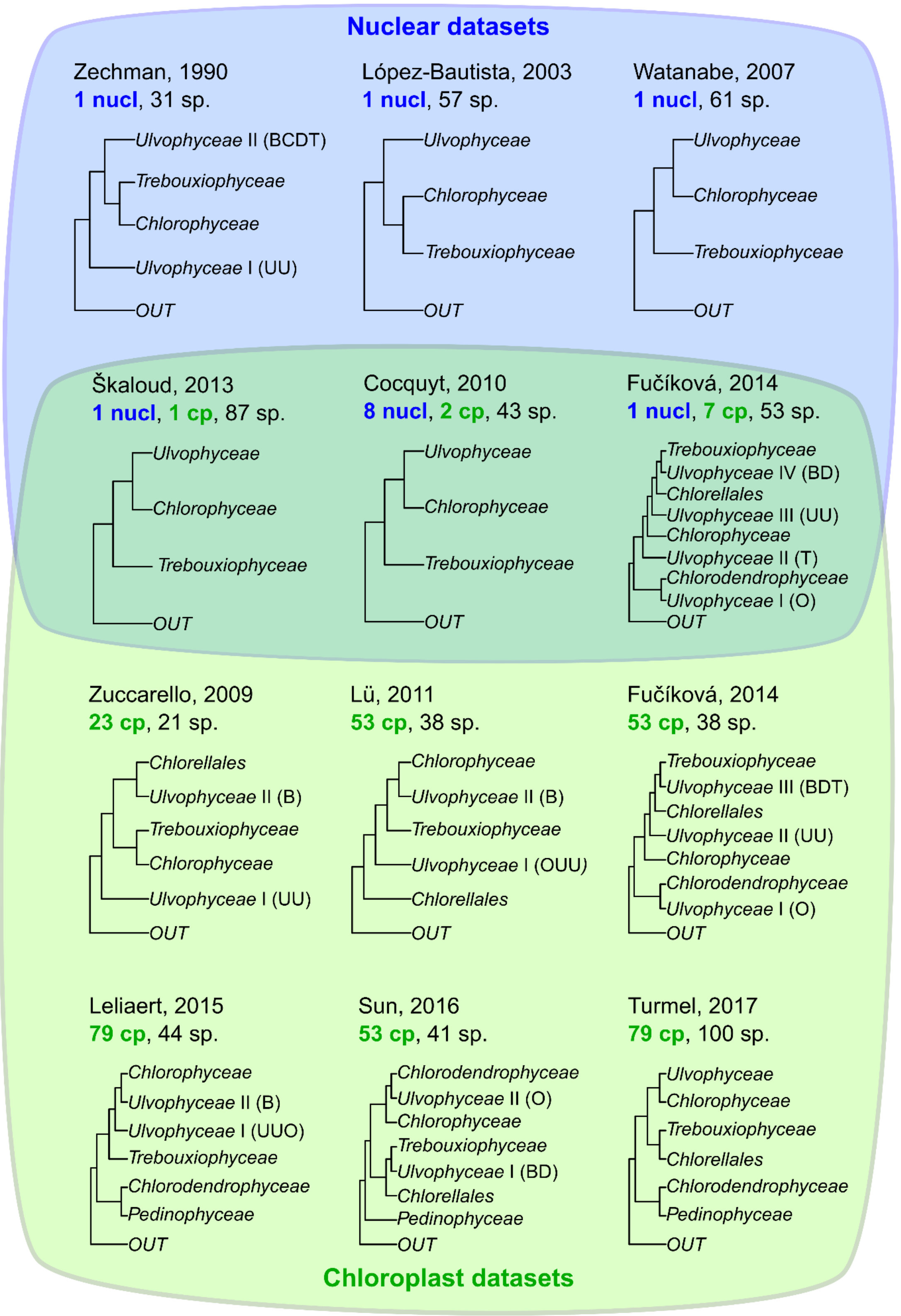


Figure 1.3: Core Chlorophyta relationships inferred in different studies.

Overview of the phylogenetic relationships among core Chlorophyta inferred from different studies. cp: Chloroplast genes; nucl: nuclear genes; sp.: species. B: Bryopsidales; C: Cladophorales; D: Dasycladales; O: Oltmannsiellopsidales; T: Trentepohliales; UU: Ulvales-Ulotrichales.

early evolutionary radiations, as evidenced by the very short branches leading to the main clades (Lemieux *et al.*, 2014a; Leliaert & Lopez-Bautista, 2015; Turmel *et al.*, 2017).

Evolution of complex chloroplast genome architectures in ulvophyceans

Although chloroplast genomes of ulvophyceans are relatively poorly represented compared to other classes of Chlorophyta, the ulvophyte chloroplast genomes available to date hint to a complex evolutionary history unprecedented in the other clades of green plants. While several species of ulvophyceans have chloroplast genomes with a canonical gene content and structure, the chloroplast genomes of Cladophorales, Trentepohliales, and Dasycladales are still to be fully described and understood, and show unique and unprecedented features.

The chloroplast genomes fully sequenced to date span between 81,997 bp (*Ostreobium queketii*) and 262,888 bp (*Pleurastrum sarcinoideum*) in length, coding for 96-110 genes commonly found in the chloroplast genomes of other green algae (Table 1.1). Since the ulvophyte chloroplast genomes sequenced to date share 95 genes, the difference in genome size is mostly due to variation in the number of introns and length of intergenic regions (Turmel *et al.*, 2017). The circular chloroplast genome of most green algae and land plants has a conserved quadripartite structure where a large inverted repeat sequence (IR, typically containing the rDNA operon and various other genes) divide the genome into two single-copy (SC) regions (Wicke *et al.*, 2011; Jansen & Ruhlman, 2012; Lang & Nedelcu, 2012). IRs have been lost independently several times in the Viridiplantae, and at least three times during the evolution of ulvophyceans. While Ignatiales and Oltmannsiellopsidales have retained the quadripartite structure (Pombert *et al.*, 2006b; Turmel *et al.*, 2017), Bryopsidales have lost the IRs (Leliaert & Lopez-Bautista, 2015; Marcelino *et al.*, 2016; Cremen *et al.*, 2018). Moreover, within the Ulvales, the genus *Ulva* has lost the IRs, while *Pseudoneochloris* has retained them (Melton *et al.*, 2015; Turmel *et al.*, 2017). Within the Ulotrichales, the genera *Pleurastrum* and *Rhexinema* have lost the IRs (Turmel *et al.*, 2016a). While other Ulotrichales for which the chloroplast genome sequence is available seems to have retained both copies of IRs, *Chamaetrichon capsulatum* chloroplast genome shows three IRs copies, an event unprecedented in green plants (Pombert *et al.*, 2005; Turmel *et al.*, 2017). Furthermore, Ignatiales, *Pseudoneochloris marina* and *Chamaetrichon capsulatum* (Ulotrichales) showed divergent IR copies, an additional unique feature which has never been reported before for green plants (Turmel *et al.*, 2017).

The trend of chloroplast genome expansion due to increase of non-coding sequences is most apparent in the Dasycladales. Complete chloroplast genome data are still unavailable for this order, but for *Acetabularia*, the chloroplast genome has been relatively well characterized. Classical molecular studies reported a substantial fraction (up to 80%) of chloroplasts lacking DNA (Woodcock & Bogorad, 1970; Luttke, 1988); in DNA containing chloroplasts, evidence based on electron microscopy was provided for an extremely large chloroplast genome. The *Acetabularia* chloroplast genome has been estimated to be about 10 times larger than the typical chloroplast genome of Viridiplantae based on electron microscopy, restriction enzymes patterns and renaturation kinetics, making it more similar in size to a small bacterial genome (Burton & Hugh, 1970; Padmanabhan & Green, 1978; Herrmann & Possingham, 1980; Tymms & Schweiger, 1985). The chloroplast genome appears to be bloated by long (10 kb) repetitive sequences tandemly arranged, with no similarity to rDNA/IRs of green algae, based on hybridization of restricted chloroplast DNA (Tymms & Schweiger, 1985). In addition, several minicircles with sequence similarity to chloroplast DNA have been isolated from chloroplasts of *Acetabularia cliftonii* (Green, 1976; Ebert *et al.*, 1985) and *Acetabularia acetabulum* (Mazza *et al.*, 1980). Only recently, by a Whole Genome Shotgun sequencing approach, 63 contigs for a total of almost 300 kb were assembled from *Acetabularia acetabulum* chloroplast DNA, coding for 51 chloroplast genes (de Vries *et al.*, 2013). The assembled contigs showed exceptionally long intergenic regions (several kb long) and long Open Reading Frames, up to 7,785 bp in length, with no similarity to known protein-coding genes (de Vries *et al.*, 2013). An even more exceptional chloroplast genome structure has been recently described in the Cladophorales. For years the structure and gene content of the chloroplast genome for this order has been elusive: universal primers for ulvophyceans chloroplast genes failed to amplify any product and virtually no

Table 1.1: Ulvophyte chloroplast genome sequences published.

Adapted from Turmel et al. 2017.

Taxon	Accession	Genome Size (bp)	IR	A + T (%)	# Genes	Introns
Bryopsidales						
Tydemania expeditionis FL1151	NC_026796	105,200	—	67.2	109	11
Bryopsis hypnoides	NC_013359	153,426[a]	—	66.9	108[a]	12
Bryopsis plumosa West4718	NC_026795	106,859	—	69.2	108	13
Caulerpa racemosa UNA00072801	NC_032042	176,522	—	66.4	106	18
Caulerpa cliftonii	KX808498	131,135	—	62.4	105	11
Ostreobium quekettii	LT593849	81,997	—	68.1	110	6
Derbesia sp.	KX808497	115,765	—	70.3	107	12
Halimeda discoidea	KX808496	122,075[b]	—	67.8	104	14
Cladophorales						
Boodlea composita FL1110		?[c]	—	43.5	22	?
Dasycladlaes						
Acetabularia acetabulum	HG18425-HG18474, HG794360	?[d]	?[e]	68.4	51	8[f]
Ignatiales						
Ignatius tetrasporus UTEX 2012	KY407659	239,387	2	63	107	9
Pseudocharacium americanum UTEX 2112	KY407658	239,448	2	63	107	9
Oltmannsiellopsidales						
Oltmannsiellopsis viridis NIES 360	NC_008099	151,933	2	59.5	104	5
Dangemannia microcystis SAG 2022	KY407660	166,355	2	66.3	106	8
Ulvales						
Pseudoneochloris marina UTEX 1445	KY407657	134,753	2	70.7	102	15
Ulva sp. UNA00071828	KP720616	99,983	—	74.7	100	5
Ulva fasciata	NC_029040	96,005	—	75.1	100	5
Ulva linza QD08	NC_030312	86,726	—	78.5	96	5
Ulotrichales						
Chamaetrichon capsulatum UTEX 1918	KY407661	189,599	3	69.2	104	16
Tupiella akineta UTEX 1912	NC_008114	195,867	2	68.5	105	27
Trichosarcina mucosa SAG 4.90	KY407656	227,181	2	62.8	103	14
Rhexinema paucicellulare SAG 29.93[g]	KX306824	221,431	—	68.5	104	31
Pleurastrum sarcinoideum UTEX 1710[g]	KX306821	262,888	—	68.5	104	27

a Based on the reannotated version of Leliaert and Lopez-Bautista, 2015.
b Halimeda has one scaffold with an unknown number of repeats annotated with 100 Ns.
c 91,391 bp assembled in 34 contigs.
d 295,664 bp assembled. The size of the chloroplast genome is estimated around 2 Mb.
e No information available.
f introns identified in the currently available sequence. The actual number may be larger.
g *Gloeotilopsis planctonica* and *Gloeotilopsis sarcinoidea* were reassigned to *Rhexinema paucicellulare* and *Pleurastrum sarcinoideum* respectively, according to (Škaloud et al., 2017)

sequence for chloroplast genes was available (Cocquyt *et al.*, 2010b; Deng *et al.*, 2013; Fučíková *et al.*, 2014). The Cladophorales have been characterized by the presence of abundant plasmid-like molecules in the chloroplasts, referred as Low Molecular Weight DNA. These plasmid-like molecules were characterized as single-stranded DNA with extensive repetitive sequences (La Claire *et al.*, 1997). The plasmids were expressed, had poor sequence similarity to some chloroplast genes and appeared to be associated with the chloroplast pyrenoid (La Claire *et al.*, 1997; La Claire *et al.*, 1998; La Claire & Wang, 2000; La Claire & Wang, 2004). While circular and linear plasmids can naturally occur in the chloroplast of green algae in addition to the canonical chloroplast genome (Green, 1976; Mazza *et al.*, 1980; Ebert *et al.*, 1985; Turmel *et al.*, 1986), a canonical circular chloroplast genome appears to be lost in Cladophorales. Whole Genome Shotgun sequencing of a chloroplast-enriched fraction showed that only plasmid-like molecules are present in the chloroplast of *Boodlea composita*, a representative of Cladophorales (Del Cortona *et al.*, 2017), Chapter 4. These plasmid-like molecules are long palindromic single-stranded molecules, 1-7 kb long, that fold intramolecularly to form hairpins. 34 hairpins were assembled, coding for 22 chloroplast genes. The genes presented atypical features, such as high divergence from orthologous algal genes, an alternative genetic code and exceptionally high GC content. In addition to the plasmid-like molecules, the sequenced chloroplast DNA abounded with retrotransposons. Del Cortona et al. (2017) suggested that hairpins were generated by an ancient retrotransposon invasion, which first led to an expansion of an ancestral circular chloroplast genome, followed by its reduction and fragmentation in hairpin plasmids, next to a massive transfer of chloroplast genes to the nucleus. Intriguingly, the gene set retained in the fragmented chloroplast genome of *Boodlea composita* is similar to the gene set found in the reduced and fragmented chloroplast genome of peridinian dinoflagellates (Howe *et al.*, 2008), suggesting a minimal set of required genes for a functional photosynthetic chloroplast.

For the remaining orders of ulvophyceans, little information is available regarding the chloroplast genome structure and content. Chloroplast genes in the terrestrial order Trentepohliales are rather divergent in sequence compared to corresponding orthologous genes in other ulvophyceans. For this order, only sequences from seven chloroplast genes are available (Rindi *et al.*, 2009; Fučíková *et al.*, 2014). For the order of Scotinosphaerales, only a small number of partial chloroplast genes are available, and no data is present on chloroplast genome structure (Škaloud *et al.*, 2013).

Phylotranscriptomics: a new and powerful tool to resolve difficult phylogenetic relationships

The phylogenetic studies to date, whether they were based on single gene, or on chloroplast or nuclear multi-gene data, failed to reach an overall consensus on the phylogenetic relationships among the main clades of ulvophyceans and core chlorophytes. This indicates that much larger amounts of data will be needed to resolve phylogenetic relationships in these green algae. New tools, such as phylotranscriptomics, are rapidly developing and may help to shed light on these obscure relationships.

The advent of Next Generation Sequencing (NGS) has lowered the resources and the investments needed to sequence DNA and RNA, resulting in an accumulation of publicly available sequence data (van Dijk *et al.*, 2014; Muir *et al.*, 2016). The ability to sequence massive amount of data for a reasonable price in a timely manner allowed the rise of initiatives and international consortia that aim to generate sequences for thousands of samples and/or species, such as 1,000 Plants project (1KP) (Matasci *et al.*, 2014), 5,000 Insect Genomes (i5k) (i5k-Consortium, 2013), the vertebrates Genome 10K (Koepfli *et al.*, 2015), the 1,000 fungal genome project 1KFG (Grigoriev *et al.*, 2014), the Tara Oceans (Pesant *et al.*, 2015), and the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling *et al.*, 2014). Transcriptome sequencing provides a cheaper and faster alternative to genome sequencing, allowing an immediate overview of a wide array of expressed genes, without the burden of gene-prediction and other difficulties of de novo genome assembly and annotation. In fact, whole-genome sequencing projects often suffer from intrinsic obstacles: extracting high quality DNA may be complicated, repetitive elements and polymorphism associated with large diploid or polyploid eukaryotic genomes may fragment or prevent the assembly, large gene families may be collapsed into chimeric sequences, gene prediction itself may be difficult in absence of sequence data from close relative species (Baker, 2012; Schatz *et al.*, 2012). In addition, transcriptomes allow to overcome the sparse taxonomic sampling of whole-genome projects (Keeling *et al.*, 2014). Moreover, user-friendly platforms for the analysis of non-model species transcriptomes have been developed (Van Bel *et al.*, 2013). As a result, an increasing community of scientists is applying RNA-seq technology to investigate a wide variety of scientific hypotheses and to answer phylogenetic and evolutionary questions.

Phylotranscriptomic analysis has proven to be a powerful tool to resolve ancient and difficult phylogenetic relationships, such as the early branching fungi and metazoans, Lepidoptera, Ostracoda, and Pancrustacea (Torruella *et al.*, 2012; Oakley *et al.*, 2013; Kawahara & Breinholt, 2014). In the Viridiplantae, a phylotranscriptomic approach has been applied to resolve phylogenetic relationships among its major lineages, mainly focussing on charophyte green algae and the origin and early diversification of land plants (Finet *et al.*, 2010; Wodniok *et al.*, 2011; Finet *et al.*, 2012; Laurin-Lemay *et al.*, 2012; Timme *et al.*, 2012; Wickett *et al.*, 2014). Phylotranscriptomics can also help to trace the evolution of major innovations and the associated molecular features in a broad phylogenetic context, as shown by Janouškovec et al. (2017) in a study on dinoflagellates evolution. Analysis of a phylogeny inferred from 101 nuclear protein coding genes revealed the genetic underpinning of major molecular innovations, such as the coincidence of the origin of the theca and the radiation of cellulase (Janouškovec *et al.*, 2017). Other examples of phylotranscriptomics as key tool to resolve the phylogeny and trace molecular innovations in non-model species are the studies of female sex organ evolution in pleurocarpous mosses (Johnson *et al.*, 2016), independent gain and losses of flagella and chitin synthase in opisthokonts (Torruella *et al.*, 2015), and parallel losses of complex multiphase life cycle in amoebae (Kang *et al.*, 2017).

Even though analysis and handling of large datasets is not trivial, a semi-automated in house phylotranscriptomics pipeline can be set up. Modular and customizable code for each step has been made publically available for the scientific community to use (Grant & Katz, 2014). The workflow of a phylotranscriptomics pipeline could be divided into three major steps: collection and generation of the sequences, orthologous groups identification and species tree inference (Figure 1.4). In order to address and resolve the outstanding phylogenetic questions, an adequate taxon sampling is necessary.

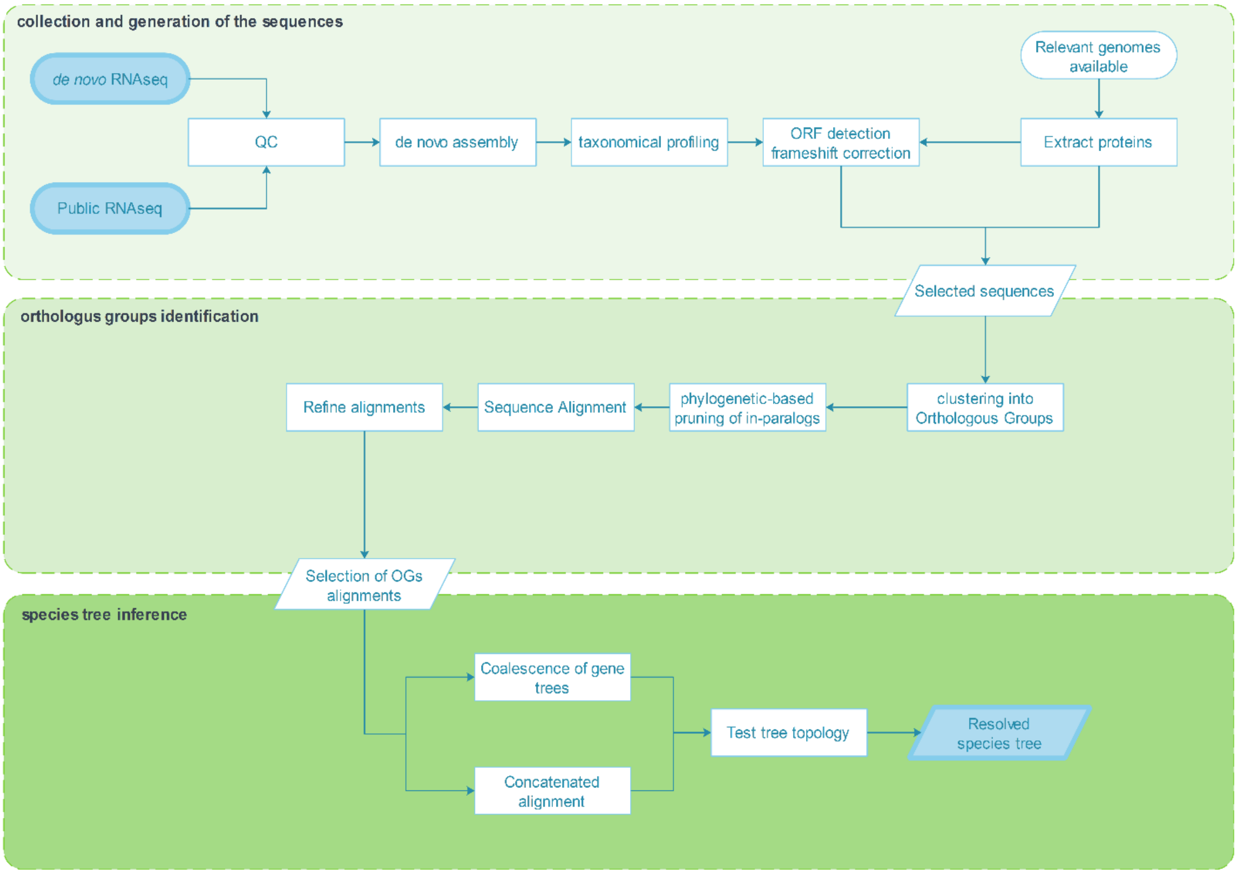


Figure 1.4: Phylotranscriptomic pipeline.

Common workflow for phylotranscriptomics analyses reposting the steps necessary to infer a species tree starting from RNA-seq reads.

Therefore, the first step of a phylotranscriptomics study is to retrieve genomic and transcriptomic data from publicly available repositories and to generate RNA-seq data for the species of interest that are still missing. Usually, it is possible to retrieve the nucleotide sequences and the corresponding amino acid translations for annotated genes in a published genome. For most publicly available transcriptomic studies, however, only the raw sequenced reads are disclosed. As a consequence, a quality control step is required on both retrieved and generated RNA-seq datasets: removal of sequencing adapter, trimming of reads and filtering based on quality scores (Andrews, 2010; Wysoker *et al.*, 2013; Bolger *et al.*, 2014). Trimmed high quality reads are then assembled into contigs, where each contig ideally corresponds to a transcribed gene (Grabherr *et al.*, 2011). Transcripts with high sequence similarity are clustered together, and usually only the longest one or the highest expressed isoform is retained as a representative for downstream analysis (Li & Godzik, 2006; Schvartzman *et al.*, 2018). Given the fragmented nature of transcriptomic data and the noise deriving sequencing and assembling errors, predicting the Open Reading Frame (ORF) of each transcript is a challenging task. Frame correction of the transcripts based on orthologous sequences is recommended (Gouzy *et al.*, 2009). Since simple ORF prediction based on presence of an in-frame start and a stop codon would generate misleading results, an orthology-guided approach is preferred (Van Bel *et al.*, 2013). Unless the RNA-seq data were generated from axenic monocultures, a taxonomical profiling of the transcripts is necessary to eliminate sequences from contaminants. Each selected sequence should have an unique identifier which possibly reports as well a unique species-code to keep the traceability and simplify the interpretation of the results (Grant & Katz, 2014).

In the second step, the collection of the selected coding sequences for all the species of interest are clustered together into Orthologous Groups (OGs). This is a crucial step since it is fundamental to distinguish between orthologs, inparalogs and outparalogs sequences in order to infer the correct phylogenetic relationships between species (Moreira & Philippe, 2000; Gabaldón, 2008). Despite all being genes descending from a common ancestor, they refer to distinct relationships: orthologs genes in the compared species derived from a single ancestral gene in the common ancestor; inparalogs genes result from species-specific duplications; while in outparalogs genes the duplication occurred before the speciation event (Koonin, 2005).

For orthology inference, the selected sequences are compared to a set of pre-defined orthologous (Chen *et al.*, 2006; Powell *et al.*, 2014; Simão *et al.*, 2015). Alternatively, *de novo* clustering of the sequences is obtained after all-against-all comparison and Markov graph-based clustering (Li *et al.*, 2003; Emms & Kelly, 2015). These methods have high sensitivity but are prone to the inclusions of outparalogous in the OGs, therefore several approaches for phylogenetic-guided pruning of the paralogous have been implemented (Boussau *et al.*, 2013; Kocot *et al.*, 2013; Yang & Smith, 2014; Ballesteros & Hormiga, 2016). The sequences in each OG are then aligned and the alignment refined to trim poorly-align regions (Talavera & Castresana, 2007; Katoh & Standley, 2013). After sequence alignment, relevant OG are selected for the downstream analysis, generally based on the number of species represented in the OG.

In the third step, multiple phylogenetic analyses are performed on the curated sequence alignment for each selected OG to decipher their phylogenetic signal and build a resolved species tree. A plethora of strategies are usually adopted in order to solve outstanding phylogenetic questions. The analyses can be ran on the nucleotide and/or on the corresponding amino acid sequences. When resolving ancient relationships, the nucleotide analysis is often restricted to the first and second codon position due to possible high among-lineage variation of the GC content in the third codon position and saturation of the phylogenetic signal (Breinholt & Kawahara, 2013; Wickett *et al.*, 2014), although other methods of site-stripping have been proposed to more selectively remove fast-evolving sites, e.g (Verbruggen & Theriot, 2008). Analyses can be performed with or without partitioning of the genes and codon positions into model parameters classes. Partitioned analyses better handle rate heterogeneity across genes and across fast-evolving positions within each gene (Lanfear *et al.*, 2012). The alignments of the selected OGs can either be concatenated for supermatrix analyses (but see Philippe *et al.* 2017 and Shen *et al.* 2017) or each OG tree can be inferred independently in a coalescence-based analysis for co-estimation of species and gene trees (Boussau *et al.*, 2013; Mirarab *et al.*, 2014). Despite different phylogenetic analyses not always converge to a unified answer, results are often largely consistent, as shown by the 52 distinct analyses performed to resolve the relationships between land plants (Wickett *et al.*, 2014; Shen *et al.*, 2017).

Transcriptomics insights into development of siphonous ulvophyceans

The first insights in the morphogenesis of siphonous ulvophyceans were gained even before the formalization of the “central dogma in molecular biology” (Crick, 1958), when the role of RNA in development was postulated but not unequivocally demonstrated (Caspersson & Schultz, 1939). In early graft experiments, the role of the macronucleus contained in the rhizoids in controlling the morphogenesis of grafted-caps in different *Acetabularia* species was investigated (Hämmerling, 1953).

Hämmerling found that the macronucleus continuously creates an apical-basal gradient of "nucleus-dependent morphogenetic substances" that controls the development and determines the morphology of the cap. These substances were furthermore described as "*products of gene action, which stand between the gene and character*" (Hämmerling, 1953).

More recently, the role of cytoskeleton organization and discrete transcripts distribution was molecularly characterized, confirming the hypothesis that messenger RNAs (mRNAs) are the "nucleus-dependent morphogenetic substances" described by Hämmerling and collaborators (Mine *et al.*, 2008). Four distinct classes of *Acetabularia* mRNAs were identified based on their differential distribution. The first class of transcripts is uniformly distributed, the second and third class have an apical/basal or a basal/apical gradient matching the asymmetrical distribution of specific metabolic activities, and the last class has a developmental-specific pattern of localization (Serikawa *et al.*, 2001; Vogel *et al.*, 2002). Actin filaments of the cytoskeleton are involved in the transport and compartmentalization of the transcripts, since actin-1 inhibitor cytochalasin D would disrupt these differential distribution (Menzel, 1994; Vogel *et al.*, 2002; Mine *et al.*, 2005). Interestingly, mRNAs and rRNAs are transported to the apex of *Acetabularia* at two different speeds, with mRNAs moving much faster than the rRNAs. Moreover, the number of ribosomes would not be sufficient to bind all mRNAs synthetized in the macronucleus. This observation suggests that in *Acetabularia* mRNAs are not transported as ribosome-mRNA complexes, but rather as an alternative form of messenger-ribonucleoprotein complexes (Kloppstech & Schweiger, 1975b; Kloppstech & Schweiger, 1975a).

Similarly, the contribution of differential transcripts accumulation in the metabolic and morphological partitioning of the multinucleate siphonous algae of the Bryopsidales was demonstrated by transcriptomic analysis of *Caulerpa taxifolia* (Chisholm *et al.*, 1996; Coneva & Chitwood, 2015; Ranjan *et al.*, 2015). Comparative transcriptomics between different subcellular structures of *Caulerpa* evidenced a strong apical-basal transcript distribution, with transcripts of genes responsible for different function accumulated in different subcellular locations. In the basal region of *Caulerpa* (holdfast, stolon, base) there is an enrichment for genes responsible for housekeeping functions and DNA replication, repair and expression. The apical region (apex, pinnules, rachis) is instead enriched for genes responsible for RNA translation, protein metabolism and vesicle movements. Interestingly, there is a gradient of accumulation of small interfering RNAs toward the apical region of *Caulerpa* (Ranjan *et al.*, 2015). Unlike in *Acetabularia*, the most probable mean of transcript transport in *Caulerpa* is cytoplasmic streaming, which is generated by bundles of cortical microtubules (Sabnis & Jacobs, 1967; Kuroa & Manabe, 1983).

An interesting lesson that can be learnt from siphonous ulvophyceans is that transcript patterning is uncoupled from cellular patterning, and that the association between cytoskeleton and transcript patterning generates the protoplast morphology. These observations represent an intriguing parallelism to vascular plants, where morphology and cell fate are determined by positional clues and constrained by physical boundaries rather than determined by the cell germlines, like it happens in metazoans (Sulston *et al.*, 1983; Hamant *et al.*, 2008; Salazar-Ciudad, 2010). Certainly the symplastic connectivity between cells in vascular plants remembers the ulvophyceans siphonous organization. Independence of morphology determination from cell division was shown as well in tobacco and wheat, where irradiated seedlings would develop proper first leaf foliage even when cell division was hampered (Haber, 1962; Haber & Foard, 1963). These observations are in agreement with the independence of morphology from multicellularity, as described by the organismal-theory (Sharp, 1926; Kaplan & Hagemann, 1991). According to this theory, multicellularity is a consequence of the unified protoplast compartmentalization into discrete portions by the cell membranes/walls deposition, rather than the result of single cells aggregation and subsequent specialization, like proposed by the cell-theory (Kaplan & Hagemann, 1991). Although being very controversial, a theoretical reconstruction of the geometric constrains regulating the development and morphology of the embryo was recently proposed for vertebrates as well, where the embryo geometry regulates patterns of gene expression and vice versa (Edelman *et al.*, 2016).

Outstanding questions

The path to solve the intriguing mysteries of ulvophyceans is disseminated with opportunities and challenges. Due to the scarce availability of green algal nuclear genome sequences compared to land plants, difficult and ancient relationships within

Viridiplantae have been resolved through sequencing of chloroplast genomes of relevant clades and phylogenomics analysis of their genes (Lemieux *et al.*, 2014b; Lemieux *et al.*, 2014a; Lemieux *et al.*, 2015; Turmel *et al.*, 2015; Leliaert *et al.*, 2016; Turmel *et al.*, 2016a; Fang *et al.*, 2018). However, as discussed previously, ulvophyceans are poorly represented in these studies, and nuclear and chloroplast data fail to converge to the same topology. Even ulvophyceans monophyly is under debate, and a polyphyletic Ulvophyceae would scramble the current interpretation of cyto-morphological evolution in this group. While the amount of available sequence data is indeed growing at a rapid rate, their interpretation and analysis in non-model organisms is not as easy and straightforward as in model organisms, where genetic and molecular basis are available.

Despite being a very promising approach to solve outstanding problems regarding phylogeny and molecular evolution, phylotranscriptomics analyses have intrinsic challenges. It is inherently difficult to avoid systematic data errors using high-throughput sequencing technologies during the construction and automated analysis of phylogenomics supermatrices (Philippe *et al.*, 2017). In addition to that, a handful of genes drive uncertainties in phylogenomic studies. This implies careful analysis, moreover, it was suggested to label as unresolved those phylogenetic relationships that are determined by a handful of positions or genes (Shen *et al.*, 2017). Furthermore, while phylogenomics and phylotranscriptomics can capture innovations in gene content, they fail to account for innovations arising from gene regulatory complexes and from epigenetic mechanisms (Sebé-Pedrós *et al.*, 2016).

Despite all the challenges and the fact that that RNA-seq data available is currently sparse in the core Chlorophyta, and ulvophyceans in particular (Xu *et al.*, 2012; Zhang *et al.*, 2012; Li *et al.*, 2014; Ranjan *et al.*, 2015; Del Cortona *et al.*, 2017), a phylotranscriptomic approach will likely prove useful to resolve the phylogenetic relationships of ulvophyceans, the evolution of their complex morphologies and of the make-over of their translational apparatus. Comparative transcriptomic data will also provide clues toward the genetic underpinning of mechanisms of nuclear and cell division, cell differentiation, polarity and growth, delimitation of nuclear-cytoplasmic domains in Cladophorales, and cytoplasmic streaming in siphonous ulvophyceans. In order to solve these outstanding questions, RNA-seq data from representatives of ulvophyceans major lineages and of each cyto-morphological type should be generated. Apart from representatives from the main clades of ulvophyceans (Ulvales, Ulotrichales, Trentepohliales, Bryopsidales, Dasycladales and Cladophorales), crucial taxa to be analysed include Oltmannsiellopsidales, Scotinosphaerales and Ignatiales, as these taxa represent uninucleate unicellular and colonial organisms that probably diverged early from the rest of the ulvophyceans before any changes in the translational apparatus and before the evolution of complex morphologies (Pombert *et al.*, 2006b; Cocquyt *et al.*, 2010a; Škaloud *et al.*, 2013). Additional species expected to provide useful information on the evolution of ulvophyceans complex morphologies are representatives of the genus *Ostreobium*, which is the earliest diverging lineage of Bryopsidales (Verbruggen *et al.*, 2009; Verbruggen *et al.*, 2017) and *Blastophysa rhizopus*, an endophytic algae that appears to be cytologically related to the Cladophorales (Sears, 1967; Chappell *et al.*, 1991; Cocquyt *et al.*, 2010b). As a fundamental support for a phylotranscriptomic analysis of ulvophyceans, *Ulva mutabilis* genome, a rising model system to study morphogenesis, was sequenced (De Clerck *et al.*, 2018).

Objectives and outline of this thesis

Two major hypothesis have been addressed in this investigation:

1. Did green seaweeds have a single origin, with macroscopic growth evolving from simpler organisms to more complex cytological organisations?

2. Has the chloroplast genome of Cladophorales been completely transferred to the nucleus?

The first hypothesis of this thesis implies resolving the phylogenetic relationships among the main lineages of the core Chlorophyta, with special care for the phylogenetic placement of the Ulvophyceae and relationships within this class. In addition, the dynamics and the timing of green seaweeds diversification need to be elucidated. In order to achieve this goal, we combined deep genome and transcriptome sequencing with publicly available datasets to obtain a balanced and representative taxon sampling of the major clades of Chlorophyta. By using a careful selection of the markers, complementary phylogenetic analyses, models of evolution, and partition strategies we hope to obtain a resolved phylogeny of green algae, and green seaweeds in particular.

The second hypothesis involves sheding light on the chloroplast genome architecture of Cladophorales, which has remained a mystery up till now. Chloroplast isolation coupled with DNA and RNA sequencing with innovative technologies will provide an exhaustive overview of chloroplast and nuclear encoded genes, and of the genome architecture. Furthermore, the abundant LMW (plasmid-like) fraction in the chloroplasts will be sequenced as well, to provide a profile of these pervasive molecules that characterize Cladophorales green seaweeds.

Because just like any other technique, phylotranscriptomics comes with its own potential problems, **Chapter 2** provides a critical overview of transcriptome assembly and annotation in the absence of a reference genome. Different approaches to evaluate transcripts and gene space completeness were tested. Moreover, potential pitfalls, such as the impact of partial and missing data on gene family inference and gene family sizes estimation are assessed and discussed.

In **Chapter 3**, a green algal phylogenetic reconstruction based on 539 nuclear markers mined from genomes and transcriptomes of 55 species is inferred to unravel the evolutionary history of the Chlorophyta, and green seaweeds in particular. The topologies inferred with complementary phylogenetic analyses are evaluated with rigorous statistical testing to present a robust and highly supported phylogenetic reconstruction.

In **Chapter 4**, the deviant chloroplast genome of Cladophorales is presented. DNA and RNA libraries from 10 Cladophorales species are analysed to describe the nature of this highly aberrant organellar genome.

In the general discussion (**Chapter 5**), I interpret and describe the significance of the phylogenetic analyses in light of what was already known about the biology and evolution of the core Chlorophyta, and expand on aspects of chloroplast genome evolution in green algae. Finally, I discuss future research avenues.

Chapter 2 - A pipeline for gene family inference in green algal transcriptomes

Andrea Del Cortona, François Bucchini and Klaas Vandepoele[3]

"Produci, consuma, crepa

Produci, consuma, crepa

Produci, consuma, crepa

Sbattiti, fatti, crepa

Sbattiti, fatti, crepa

Sbattiti, fatti, crepa

Cotonati i capelli, riempiti di borchie, rompiti le palle, rasati i capelli

Crepa, crepa, crepa"

CCCP - Morire

[3] Authors contribution: A.D.C., K.V.: study design; A.D.C.: data analysis; F.B.: TRAPID plugin and PLAZA4.0 build; A.D.C.: manuscript conceptualization, drafting and writing.

Abstract

In this work we build a semi-automated pipeline for filtering and annotating *de novo* assembled transcriptomic data, using algal datasets as a test case. We further evaluated the relationships between depth of sequencing and transcriptome completeness and also assessed the impact of different partial (transcriptomic) data in the inference of gene families by state-of-the-art phylogenomic pipelines. Since gene family inference was robust and did not suffer from partial transcriptomic data, phylotranscriptomics downstream analyses are expected to be reliable. However, d*e novo* assembled transcriptomes seemed to overestimate the size of inferred gene families, suggesting that precautions should be taken before using transcriptomic data in gene family expansion, gain or loss analyses. Our pipeline had been created for green algal transcriptomic data (e.g.: green algal genomes were used to populate reference databases; green algae alternative nuclear genetic code and chloroplast translation tables were taken into account), but it can easily be adapted for analyses of other lineages. This pipeline represents a contribution to a wider, more general purpose: a flexible and user-friendly workflow to address transcriptomic data, annotation and downstream analyses in the absence of reference genomes. Such tool is fundamental to make phylotranscriptomic analyses accessible to those researchers who lack bioinformatics knowledge or access to large computing infrastructures.

Introduction

To date, the available green algal genome sequences suffer from sparse taxon sampling, and many major lineages of core Chlorophyta are not represented at all. A remarkable example is the absence of publicly available genome sequences for ulvophyceans, except for the recent publication of *Ulva mutabilis* genome (De Clerck *et al.*, 2018). The clades for which genomic sequences are available often center on economically relevant species, e.g. production of biofuel and second metabolites by Chlorellales and Trebouxiophyceae (Blanc *et al.*, 2012; Nelson *et al.*, 2017; Roth *et al.*, 2017), or on specific molecular features, e.g. dynamics of genome reduction in prasinophytes (Derelle *et al.*, 2006; Palenik *et al.*, 2007; Worden *et al.*, 2009) and evolution of multicellularity in volvocine green algae (Chlorophyceae) (Merchant *et al.*, 2007; Prochnik *et al.*, 2010; Hanschen *et al.*, 2016; Featherston *et al.*, 2018). This selectivity is not optimal for broad phylogenomic and comparative analyses. Extensive and taxon-rich transcriptome sequencing initiatives, such as the MMESTP and the 1-KP projects (Keeling *et al.*, 2014; Matasci *et al.*, 2014), populated more neglected clades with transcriptomes, however a rigorous estimate of transcriptome completeness for many sequenced species is missing. Moreover, green seaweeds display a huge range of genome sizes, from 25 Mbp up to more than 2 Gbp (Kapraun, 2007).

To explore an organism's gene space, genome sequencing is not always the most obvious choice, despite being inclusive, due to intrinsic challenges (e.g. availability of high-quality high-molecular weight DNA and haploid cells, the amount of repetitive elements and sequence duplication) and species-specific variability in genome size and complexity (Schatz *et al.*, 2012). Transcriptome sequencing provides a faster and cheaper alternative to genome sequencing, but is only able to capture a portion of the total gene space of an organism. Additional challenges are inherent to *de novo* assembly and annotation of transcriptomes, when a reference genome is not available. The initial RNA purified from cells or tissues is a mixture of RNA species with variable stoichiometry: i.e. mature and immature messenger RNAs, ribosomal RNAs, non-coding RNAs and residual genomic DNA contaminants. The experimental design, the selection and enrichment methods for specific RNA species therefore have a considerable impact on the relative abundance of RNA species. This variability influences the quality and homogeneity of downstream analysis when the RNA-seq libraries are retrieved from different sources. The ideal scenario would require consistency across all samples for all these parameters (Griffith *et al.*, 2015).

Another layer of complexity is given by currently applied sequencing technologies, which rely either on short Next-generation Sequencing (NGS) reads or on long, noisy Single Molecules Real-Time (SMRT) reads. Despite the rise on SMRT sequencing technologies for transcriptome sequencing, their relative low output in sequenced basepairs compared to NGS sequencing and the higher error rates restrict the use of SMRT to specific studies: e.g. discovery of novel isoforms, alternative splicing and polyadenylations, long non-coding RNAs, and gene fusion (Wang *et al.*, 2016; Liu *et al.*, 2017; An *et al.*, 2018). On the other hand, NGS technologies suffer from their inherently short reads, which makes *de novo* assembly of transcriptomes in the absence of a reference genome a non-trivial task. NGS assemblers face multiple challenges: uneven coverage across the transcriptome and even across transcripts due to alternative isoforms, differential levels of expression and sample heterogeneity (Grabherr *et al.*, 2011; Schulz *et al.*, 2012; Steijger *et al.*, 2013). Additional noise in the dataset arises from sequencing errors.

De novo transcriptome assemblies generally result in a higher number of contigs (reconstructed transcripts) than the actual number of genes coded by the genome (Zhao *et al.*, 2011). The redundancy is caused by the presence of allelic and splice variants, and partial assembly or misassembly of sequences due to sequencing and assembly errors. Depending on the quality of the starting RNA, library prep, depth of sequencing and read length, a considerable amount of assembled transcripts covers only fractions of a gene (Wall *et al.*, 2009). Moreover, a *de novo* assembled transcriptome usually does not cover the whole gene space of an organism, due to differential gene expression in different tissues at different time points and life stages. It is therefore important to have reliable metrics to evaluate the completeness of a transcriptome: i.e. the percentage of gene space represented in the transcriptome and the corresponding gene completeness coverage.

In this work, we build an efficient, semi-automated *de novo* assembly, annotation and gene family inference pipeline for transcriptomes, able to tackle most of the shortcomings and challenges inherent to transcriptomic analyses in the absence of a reference genome. In building our pipeline, we considered several features to be fundamental for assessing the reliability of comparative and phylogenetic studies based on *de novo* assembled transcriptomes: the ability to discriminate between *bona fide* green algal sequences and contaminants in transcriptomes from uncultured organisms and environmental samples; the ability to detect and correct frameshift errors in transcriptomes generated by different sequencing technologies; the ability to assess the transcriptome completeness. In addition, we evaluated the relationships between depth of sequencing, coverage of the assembled transcripts, transcriptome completeness and number of gene families identified and the impact of transcriptomic data on state-of-the-art genome-based methods for gene family inference.

Results

Based on a dataset of genomes or transcriptomes of 55 species (15 reference genomes and 40 *de novo* assembled transcriptomes for which a reference genome was not available), representing the major clades of the green lineage. All orders of the Ulvophyceae were included (Bryopsidales, Cladophorales, Dasycladales, Ignatiales Oltmannsiellopsidales, Scotinosphaerales Trentepohliales, Ulotrichales and Ulvales, Table 2.1). Below we describe and evaluate a series of critical steps of the transcriptomic pipeline, namely:

- Taxonomic binning. The transcriptomes assembly metrics and taxonomic distribution of transcripts were described.

- Frameshift correction. The detection and correction of potential frameshift transcripts was described.

- Gene space completeness evaluation. The performances of BUSCO, coreGF and eggNOG-mapper in assessing gene space completeness of transcriptomes was evaluated.

- Depth of sequencing: transcriptome completeness and gene family size correlation. The relationship between depth of sequencing, gene space completeness and gene family size was evaluated.

- Effect of partial data on Orthology inference. The effect of using transcriptomic data in genomic-pipelines for orthology inference was evaluated.

Transcriptomes assembly metrics and taxonomic binning

The assembly metrics for the 40 transcriptomes (see Table 2.1) are reported in Table 2.2. Since the corresponding raw reads were not availble at the time of the study, *Acrosiphonia* sp., *Blastophysa rhizopus* and *Caulerpa taxifolia* transcriptome assemblies were retrieved from the respective online repositories (Table 2.1). The remaining RNA-seq libraries were assembled in house. RNA-seq data generated with 454 technology (*Botryococcus braunii*, *Chlorokybus atmophyticus* and *Ulva linza*) were

Table 2.1: Datasets used in this study.
In green, ulvophyte transcriptomes; in yellow, other green algal transcriptomes; in orange, publicly available genomes.

species	Class or Order	publication	Mitochondrial genome	Chloroplast genome	source
Caulerpa taxifolia	Bryopsidales	(Ranjan *et al.*, 2015)	N/A	N/A	SRP041084
Codium fragile[#]	Bryopsidales	N/A	N/A	N/A	TBA
Halimeda discoidea[#]	Bryopsidales	N/A	N/A	N/A	TBA
Ostreobium quekettii[#]	Bryopsidales	N/A	N/A	NC_030629.1	TBA
Blastophysa rhizopus	Chaetosiphonales	(Matasci *et al.*, 2014)	N/A	N/A	OneKP
Tetraselmis astigmatica	Chlorodendrophyceae	(Keeling *et al.*, 2014)	N/A	N/A	MMTESP0804
Tetraselmis striata	Chlorodendrophyceae	(Keeling *et al.*, 2014)	N/A	N/A	MMETSP0817
Acutodesmus acuminatus SAG 38.81	Chlorophyceae	N/A	N/A	N/A	SRR1174737
Chlamydomonas reinhardtii	Chlorophyceae	(Merchant *et al.*, 2007)	NC_001638.1	NC_005353.1	JGI4.0
Dunaliella tertiolecta	Chlorophyceae	(Keeling *et al.*, 2014)	N/A	N/A	MMTESP1127
Gonium pectorale	Chlorophyceae	(Hanschen *et al.*, 2016)	AP012493.1	NC_020438.1	NCBI
Haematococcus pluvialis	Chlorophyceae	(Gao *et al.*, 2015)	N/A	N/A	SRR2148810
Volvox carteri	Chlorophyceae	(Prochnik *et al.*, 2010)	N/A	N/A	JGI1.0
Auxenochlorella protothecoides	Chlorellales	(Gao *et al.*, 2014)	NC_026009.1	KC843975.1	KEGG
Chlorella sp NC64A	Chlorellales	(Blanc *et al.*, 2010)	NC_025413.1	KJ718922.1	JGI1.0
Picochlorum oklahomensis	Chlorellales	(Keeling *et al.*, 2014)	N/A	N/A	MMETSP1330
Boodlea composita	Cladophorales	(Del Cortona *et al.*, 2017)	MG257829 - MG257880	MG257795 - MG257828	SRR5500908
Cladophora glomerata	Cladophorales	(Matasci *et al.*, 2014)	N/A	N/A	OneKP
Acetabularia acetabulum[#]	Dasycladales	N/A	N/A	N/A	TBA
Ignatius tetrasporus	Ignatiales	(Matasci *et al.*, 2014)	N/A	NC_034712.1	OneKP
Oltmannsiellopsis unicellularis[§]	Oltmannsiellopsidales	N/A	N/A	N/A	TBA
Oltmannsiellopsis viridis[#]	Oltmannsiellopsidales	N/A	NC_008256.1	NC_008099.1	TBA
Marsupiomonas sp. [#]	Pedinophyceae	N/A	N/A	KM462870.1	TBA
Pedinomonas minor[#]	Pedinophyceae	N/A	NC_000892.1	NC_016733.1	TBA
Unknown pedinophyte YPF701[#]	Pedinophyceae	N/A	N/A	N/A	TBA

Bathycoccus prasinos	prasinophytes	(Moreau *et al.*, 2012)	FO082258	FO082259.2	ORCAE
Micromonas pusilla	prasinophytes	(Worden *et al.*, 2009)	FJ858268	FJ858269	JGI2.0
Nephroselmis pyriformis	prasinophytes	(Keeling *et al.*, 2014)	N/A	N/A	MMETSP0034
Ostreococcus tauri	prasinophytes	(Palenik *et al.*, 2007)	CR954200	CR954199	ORCAE
Picocystis salinarum	prasinophytes	(Keeling *et al.*, 2014)	N/A	NC_024828.1	MMETSP0807
Scotinosphaera lemnae[#]	Scotinosphaerales	N/A	N/A	N/A	TBA
Asterochloris sp. Cgr/DA1pho	Trebouxiophyceae	N/A	N/A	N/A	JGI2.0
Botryococcus braunii	Trebouxiophyceae	N/A	NC_027722.1	NC_025545.1	SRR069634
Coccomyxa subellipsoidea	Trebouxiophyceae	(Blanc *et al.*, 2012)	NC_015316.1	NC_015084.1	JGI1.0
Pseudochlorella pringsheimii	Trebouxiophyceae	(Zhang *et al.*, 2014)	N/A	N/A	SRR490104
Trebouxia gelatinosa	Trebouxiophyceae	(Carniel *et al.*, 2016)	N/A	N/A	SRR988248
Cephaleuros parasiticus[§]	Trentepohliales	N/A	N/A	N/A	TBA
Trentepohlia annulata[§]	Trentepohliales	N/A	N/A	N/A	TBA
Trentepohlia jolithus	Trentepohliales	(Li *et al.*, 2014)	N/A	N/A	SRR1044982
Acrosiphonia sp. SAG-127.80	Ulotrichales	(Matasci *et al.*, 2014)	N/A	N/A	OneKP
Phaeophila dendroides[§]	Ulvales	N/A	N/A	N/A	TBA
Ulva linza	Ulvales	(Zhang *et al.*, 2012)	NC_029701.1	NC_030312.1	SRR504341
Ulva mutabilis	Ulvales	(De Clerck *et al.*, 2018)	N/A	N/A	ORCAE
Chara vulgaris	Charophyceae	(Matasci *et al.*, 2014)	NC_005255.1	NC_008097.1	ERR364366
Chlorokybus atmophyticus	Chlorokybophyceae	(Timme *et al.*, 2012)	NC_009630.1	NC_008822.1	SRR064329
Chaetosphaeridium globosum	Coleochaetophyceae	N/A	AF494279.1	NC_004115.1	ERR364369
Coleochaete orbicularis	Coleochaetophyceae	(Ju *et al.*, 2015)	N/A	N/A	SRR1594679
Arabidopsis thaliana	Embryophytes	(The Arabidopsis Genome Initiative, 2000)	Y08501	AP000423	TAIR10
Oryza sativa	Embryophytes	(International Rice Genome Sequencing Project, 2005)	DQ167400	X15901	TIGR6.1
Physcomitrella patens	Embryophytes	(Rensing *et al.*, 2008)	AB251495	AP005672	JGI1.1
Selaginella moellendorfii	Embryophytes	(Banks *et al.*, 2011)	JF338143.1-JF338147.1	HM173080.1	JGI1.0
Klebsormidium flaccidum	Klebsormidiophyceae	(Ju *et al.*, 2015)	KP165386.1	NC_024167.1	SRR1594644

Mesostigma viride	Mesostigmatophyceae	(Ju *et al.*, 2015)	NC_008240.1	NC_002186.1	SRR1594255
Mesotaenium endlicherianum	Zygnematophyceae	N/A	N/A	NC_024169.1	ERR364377
Roya obtusa	Zygnematophyceae	N/A	NC_022863.1	NC_030315.1	ERR364380

#: dataset generated in this study
§: dataset available at (https://dx.doi.org/10.6084/m9.figshare.1604778)
JGI: datasets available at the DOE Joint Genome Institute (https://genome.jgi.doe.gov/portal/)
KEGG: datasets available at the Kyoto Encyclopedia of Genes and Genomes (https://www.genome.jp/kegg/genome.html)
MMETSP: datasets available at the Short Read Archive (https://www.ncbi.nlm.nih.gov/sra/)
NCBI: dataset available at the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/)
OneKP: datasets available at (http://www.onekp.com/public_read_data.html)
ORCAE: datasets available at Ghent University (http://bioinformatics.psb.ugent.be/orcae/)
SRP, SRR, ERR: datasets available at the Short Read Archive (https://www.ncbi.nlm.nih.gov/sra/)
TAIR: dataset available at The Arabidopsis Information Resource (https://www.arabidopsis.org/)
TBA: To be announced, datasets not released yet
TIGR: dataset available at the Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/)

assembled with CLC Genomic Workbench. The remaining raw reads were assembled with Trinity after filtering and trimming of low quality reads.

The number of transcripts ranged from 11,627 to 258,663 per library, with an N50 length value between 491 and 3,212 bp. All the assemblies were then clustered with CD-HIT-EST with a similarity cut-off of 97.5%. This resulted in a decrease in transcript number between 0.3 and 36.4%. The percentage of transcripts identified as eukaryotic after a sequence similarity search against the NCBI non-redundant protein database ranged from 18.2 to 75.2%. Such wide variation could be ascribed to a higher amount of genuine bacterial contaminants, the presence of non-coding RNA or lineage-specific and recently evolved genes with no sequence similarity to known proteins. The number of transcripts in the eukaryotic fraction ranged from 6,709 to 67,617, with a N50 length value ranging from 711 bp to 4,066 bp (Table 2.2).

Additional insights on the putative taxonomic distribution of the eukaryotic transcripts and on putative species-specific or potentially contaminating sequences in the transcriptomes comes from the GhostKOALA output. GhostKOALA performes similarity searches (BLASTp) between input translated CDS and an expanded KEGG GENES database that includes a non-redundant set of proteins from fully sequenced genomes and their corresponding KO terms and taxonomic affiliation. For comparison, we analyzed the protein coding fraction of the genomes (CDS) in Table 2.1 in a similar way. In addition to “Plants” sequences (i.e.: sequences assigned to Viridiplantae as taxonomic group, hereafter referred to as “green transcripts”), all the transcripts also had a variable fraction of sequences classified as “Animals”, “Protists”, “Fungi” and “Bacteria”, which could represent *bona fide* contaminants (Figure 2.1). For some Ulvophyceae species, for example *Boodlea composita* and the two *Oltmannsiellopsis* species, less than 50% of the sequences were classified as “Plants”. Interestingly, also sequences from well-characterized genomes (e.g.: *Oryza sativa*) were not classified as “Plants”.

Table 2.2: Summary of the *de novo* transcriptome assembly metrics.

	before CD-HIT		after CD-HIT		euk fraction		
Species	**# contigs**	**N50 (bp)**	**# contigs**	**N50 (bp)**	**#contigs**	**N50 (bp)**	**% euk**
Acetabularia acetabulum	258,663	759	172,881	766	33,527	1,534	19.4
Acrosiphonia sp.	41,150	1,745	41,009	1,729	14,910	2,225	36.4
Blastophysa cf. *rhizopus*	83,017	989	82,410	968	15,021	1,767	18.2
Boodlea composita	91,362	1,198	88,503	1,136	23,350	1,703	26.4
Botryococcus braunii	34,959	1,037	33,512	1,055	12,604	1,385	37.6
Caulerpa taxifolia	57,118	813	57,118	813	28,221	1,162	49.4
Cephaleuros parasiticus	63,443	1,695	55,326	1,567	13,987	2,172	25.3
Chaetosphaeridium globosum	38,126	628	26,456	586	11,770	811	44.5
Chara vulgaris	46,674	491	34,824	478	12,067	711	34.7
Chlorokybus atmophyticus	12,689	1,079	12,519	1,083	8,387	1,176	67.0
Cladophora glomerata	59,069	802	57,226	802	14,592	1,363	25.5
Codium fragile	75,407	1,066	60,707	1,090	30,111	1,387	49.6
Coleochaete orbicularis	207,738	1,708	146,497	1,600	45,236	2,328	30.9
Dunaliella tertiolectica	32,282	1,703	30,179	1,688	12,922	2,104	42.8
Haematococcus pluvialis	11,787	1,788	11,370	1,795	6,709	2,026	59.0
Halimeda discoidea	39,964	1,054	34,097	1,095	15,863	1,411	46.5
Ignatius tetrasporus	59,183	994	58,122	994	20,516	1,518	35.3
Klebsormidium flaccidum	113,738	2,098	74,342	2,146	44,330	2,380	59.6
Marsupiomonas sp.	50,280	1,333	35,846	1,469	17,957	1,773	50.1
Mesostigma viride	165,488	1,541	111,045	1,559	36,397	2,153	32.8
Mesotaenium endlicherianum	87,410	1,015	60,117	967	29,363	1,288	48.8
Nephroselmis pyriformis	73,008	1,310	65,605	1,271	31,044	1,535	47.3
Oltmannsiellopsis unicellularis	79,253	1,258	73,250	1,205	25,933	1,811	35.4
Oltmannsiellopsis viridis	129,713	1,415	102,283	1,345	40,053	1,931	39.2
Ostreobium quekettii	138,329	1,810	111,934	1,861	42,293	2,399	37.8
Pedinomonas minor	34,242	3,212	21,795	3,437	11,922	4,066	54.7
Phaeophila dendroides	135,530	1,785	127,650	1,706	45,339	2,371	35.5
Picochlorum oklahomensis	16,181	1,521	15,258	1,491	10,110	1,733	66.3
Picocystis salinarum	11,627	3,043	9,795	2,891	7,370	3,002	75.2
Pseudochlorella pringsheimii	35,760	751	31,608	669	17,468	829	55.3
Roya obtusa	55,929	950	35,927	935	19,889	1,148	55.4
Scenedesmus acuminatus	132,816	1,117	85,597	1,070	29,820	1,559	34.8
Scotinosphaera lemnae	171,583	1,733	127,230	1,603	31,153	2,328	24.5
Tetraselmis astigmatica	40,880	1,740	38,509	1,729	17,542	2,037	45.6
Tetraselmis striata	45,641	1,073	42,621	1,066	19,350	1,345	45.4
Trebouxia gelatinosa	112,450	2,343	73,026	2,396	35,354	2,875	48.4
Trentepohlia annulata	76,660	1,933	71,010	1,825	13,734	2,653	19.3
Trentepohlia jolithus	257,442	702	196,877	555	67,617	952	34.3
Ulva linza	15,866	1,124	15,253	1,140	8,298	1,342	54.4
Pedinophytes (YPF 701)	47,459	1,927	31,648	2,163	16,418	2,572	51.9

before CD-HIT: transcriptome metrics before sequence clustering with CD-HIT EST
after CD-HIT: transcriptome metrics after sequence clustering with CD-HIT EST
euk fraction: metrics of the eukaryotic fraction
% euk: percentage of transcripts (after CD-HIT) identified as eukaryotic

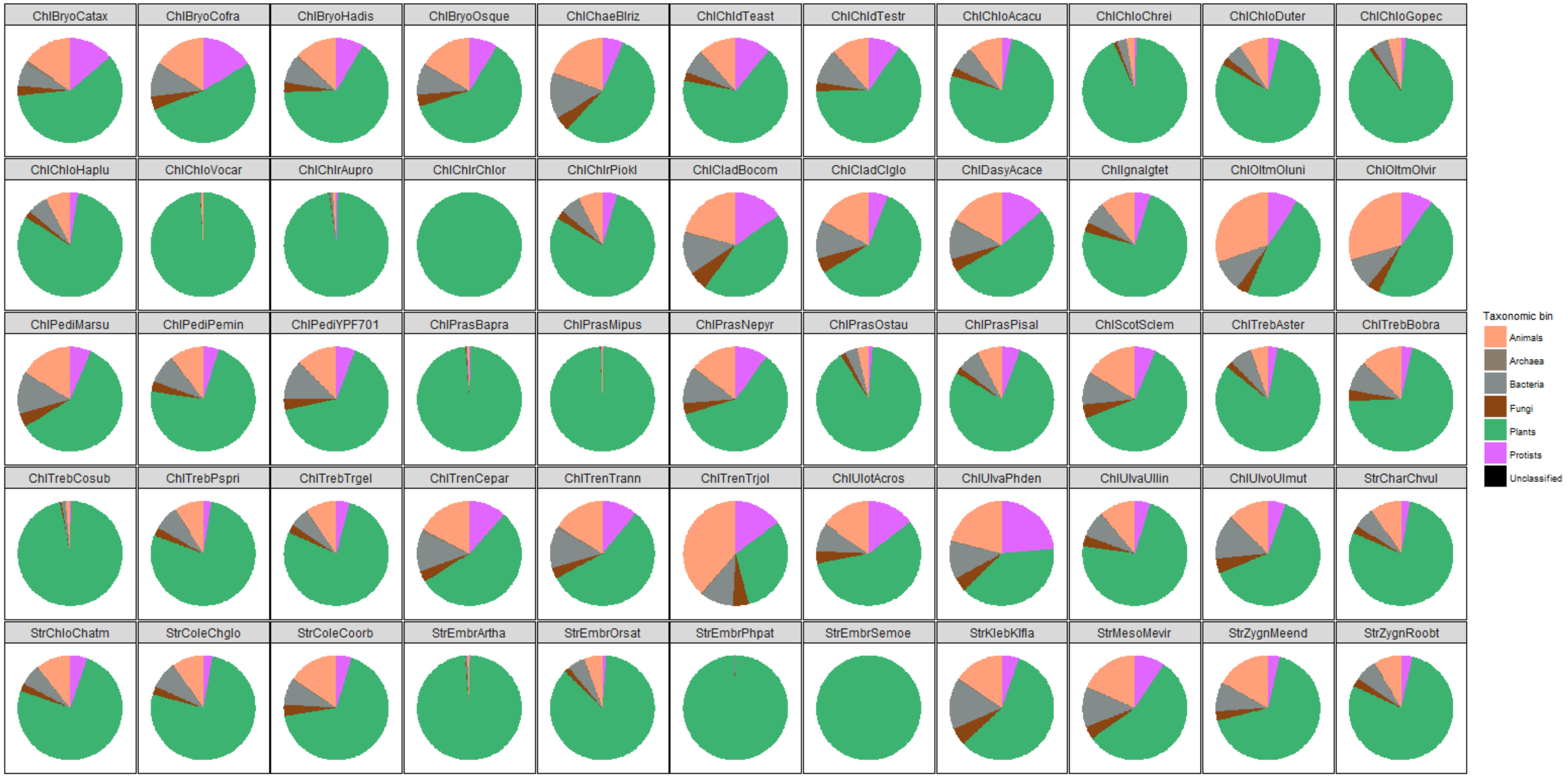


Figure 2.1: Taxonomic binning as assigned by GhostKOALA.

Pie charts report the proportion of the taxonomic bin assigned to protein-coding genes in the genomes and eukaryotic transcriptome assemblies.

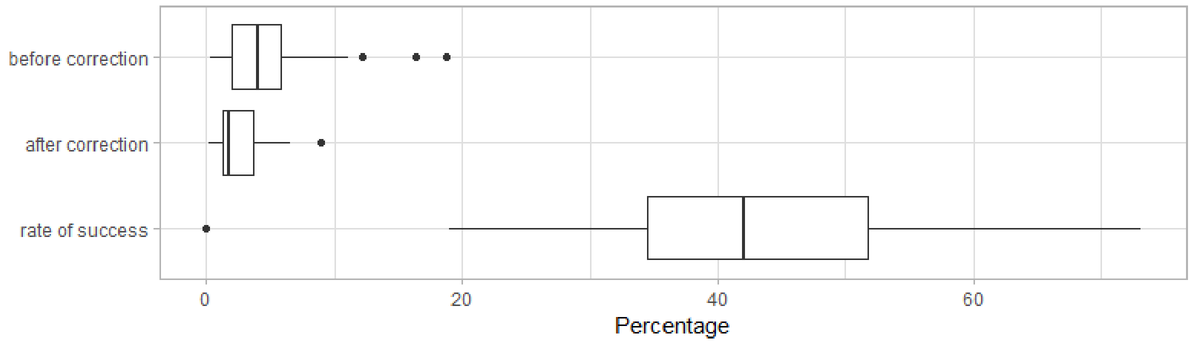


Figure 2.2: Percentage of putative frameshift transcripts and success of frameshift correction by FrameDP.

“Before correction” refers to the percentage of putative frameshift transcripts before the correction step, while “after correction” refers to the percentage of putative frameshift transcripts after the correction step. Black dots indicate outlier values.

Frameshift correction

We processed the 40 transcriptomes (Table 2.1) with the TRAPID pipeline to identify potential frameshift errors (hereafter: frameshift transcripts). Frameshift transcripts were corrected with FrameDP and the rate of success of the frameshift correction step was estimated with a subsequent TRAPID run. Depending on the sequencing technology (Table 2.3), the percentage of putative frameshift transcripts ranged from 0.3 to 18.8%, with a mean value of 4.8% (Figure 2.2). The frameshift correction step had a considerable rate of success (mean value 42.8%), measured as the percentage of the number of frameshift transcripts corrected divided by the number of initial frameshift transcripts. Surprisingly, FrameDP run did not result in any frameshift correction of the 8.9% putative frameshift transcripts of *Trentepohlia annulata.*

Gene space completeness evaluation

Gene space completeness was evaluated for the eukaryotic fraction of the transcriptome assemblies with three complementary strategies: coreGF, BUSCO and eggNOG-mapper.

The coreGF score for most of the transcriptomes analyzed was close to 1 (all the core conserved gene families identified), with the exception of the *Chara vulgaris* and *Haematococcus pluvialis* transcriptomes, which had a lower value (coreGF score0.75).

Table 2.3: sequencing technology and read depth of sequencing for each transcriptome analysis in this study.

species	**Sequencing technology**	**depth of sequencing**
Acetabularia acetabulum	Illumina NextSeq 500	103M 2x150 bp PE
Acrosiphonia sp. SAG-127.80	N/A	N/A
Acutodesmus acuminatus SAG 38.81	Illumina HiSeq 2000	21M 2x101 bp PE
Blastophysa rhizopus	N/A	N/A
Boodlea composita	Illumina NextSeq	32M 2x75 bp PE
Botryococcus braunii	454 GS FLX	1.2M
Caulerpa taxifolia	Illumina HiSeq 2000	178M 2x95 bp PE
Cephaleuros parasiticus	Illumina HiSeq 1000	47M 2x100 bp PE
Chaetosphaeridium globosum	Illumina Genome Analyzer II	8M 2x75 bp PE
Chara vulgaris	Illumina Genome Analyzer II	8M 2x75 bp PE
Chlorokybus atmophyticus	454 GS FLX Titanium	444k
Cladophora glomerata	Illumina HiSeq 2000	12M 2x90 bp PE
Codium fragile	Illumina NextSeq	11M 2x150 bp PE
Coleochaete orbicularis	Illumina HiSeq 1000	44M 2x100 bp PE
Dunaliella tertiolecta	Illumina HiSeq 2000	32M 2x50 bp PE
Haematococcus pluvialis	Illumina HiSeq 2500	12M 2x100 bp PE
Halimeda discoidea	Illumina NextSeq	7M 2x150 bp PE
Ignatius tetrasporus	Illumina HiSeq 2000	11M 2x90 bp PE
Klebsormidium flaccidum	Illumina HiSeq 1000	50M 2x100 bp PE
Marsupiomonas sp.	Illumina NextSeq	19M 2x150 bp PE
Mesostigma viride	Illumina HiSeq 1000	47M 2x100 bp PE
Mesotaenium endlicherianum	Illumina HiSeq 2000	13M 2x90 bp PE
Nephroselmis pyriformis	Illumina HiSeq 2000	23M 2x100 bp PE
Oltmannsiellopsis unicellularis	Illumina HiSeq 1000	34M 2x100 bp PE
Oltmannsiellopsis viridis	Illumina NextSeq	26M 2x150 bp PE
Ostreobium quekettii	Illumina NextSeq	24M 2x150 bp PE
Pedinomonas minor	Illumina NextSeq	20M 2x150 bp PE
Phaeophila dendroides	Illumina HiSeq 1000	60M 2x100 bp PE
Picochlorum oklahomensis	Illumina HiSeq 2500	22M 2x50 bp PE
Picocystis salinarum	Illumina HiSeq 2000	16M 2x100 bp PE
Pseudochlorella pringsheimii	Illumina Genome Analyzer IIx	35M 40bp SE
Roya obtusa	Illumina HiSeq 2000	13M 2x90 bp PE
Scotinosphaera lemnae	Illumina NextSeq 500	39M 2x150 bp PE
Tetraselmis astigmatica	Illumina HiSeq 2500	29M 2x50 bp PE
Tetraselmis striata	Illumina HiSeq 2000	20M 2x50 bp PE
Trebouxia gelatinosa	Illumina HiSeq 2000	20M 2x100 bp PE
Trentepohlia annulata	Illumina HiSeq 1000	40M 2x100 bp PE
Trentepohlia jolithus	Illumina HiSeq 2000	26M 2x90 bp PE
Ulva linza	454 GS FLX	251k
Unknown pedinophyte YPF701	Illumina NextSeq	18M 2x150 bp PE

Additional information on the quality of the transcriptome, measured as an estimate of completeness and fragmentation, comes from the percentage of reference genes covered by 50% or 90% of their length by the assembled transcripts (Figure 2.3). While it was possible to identify most of the coreGFs in most of the transcriptomes, those with a score of coreGF close to 1 could have a low coverage score (e.g.: *Caulerpa taxifolia*, *Trentepohlia jolithus*), indicating a partial assembly of the transcripts. For most of the transcriptomes analyzed in this study, at least half of the transcripts covered 50% of the reference gene or more. Furthermore, at least 25% of the transcripts covered 90% of the reference gene (representing a fraction of quasi full-length to full-length transcripts), and represent a promising initial pool of full-length transcripts for downstream phylogenetic analyses. These results were comparable to those from a similar analysis on green transcripts (Figure 2.3).

To avoid a bias toward the number BUSCO genes identified as fragmented, the eukaryotic transcripts were frameshift-corrected before the BUSCO analysis. In total, between 52.8% and 98.3% of eukaryotic BUSCO were identified in the transcriptomes, either as complete, duplicated or fragmented (Figure 2.4). Despite the high variability among the BUSCO results, for most of the transcripts 80-90% BUSCO were identified, indicating a good representation of the conserved core genes in the transcriptomes of our dataset.

The third strategy to assess the gene space completeness of the transcriptomes was based on eggNOG-mapper against three distinct non-supervised orthologous groups (NOG) clusters subsets: Chlorophyta NOGs (chloroNOG), *Viridiplantae* NOGs (virNOG), and all the NOGs present in the database (bacterial and viral NOGs included). The chloroNOG subset is virtually a subset of virNOG KO subset (2675 out of 2677 chloroNOG KO terms are included in virNOG KO terms). ChloroNOG KO terms were well represented in the majority of genomes and transcriptomes analyzed (Figure S2.1), with few notable exceptions (*Blastophysa rhizopus*, *Cladophora glomerata*, and the Trentepohliales *Cephaleuros parasiticus* and *Trentepohlia annulata* – but not *Trentepohlia jolithus*). These observations contrast with the good coreGFs and BUSCO score of these transcriptomes. Similar results were obtained for the virNOG KO terms (Figure S2.2). When compared to the ‘all KO terms collection’, most of the transcriptomes and genomes have comparable patterns of presence/absence of KO

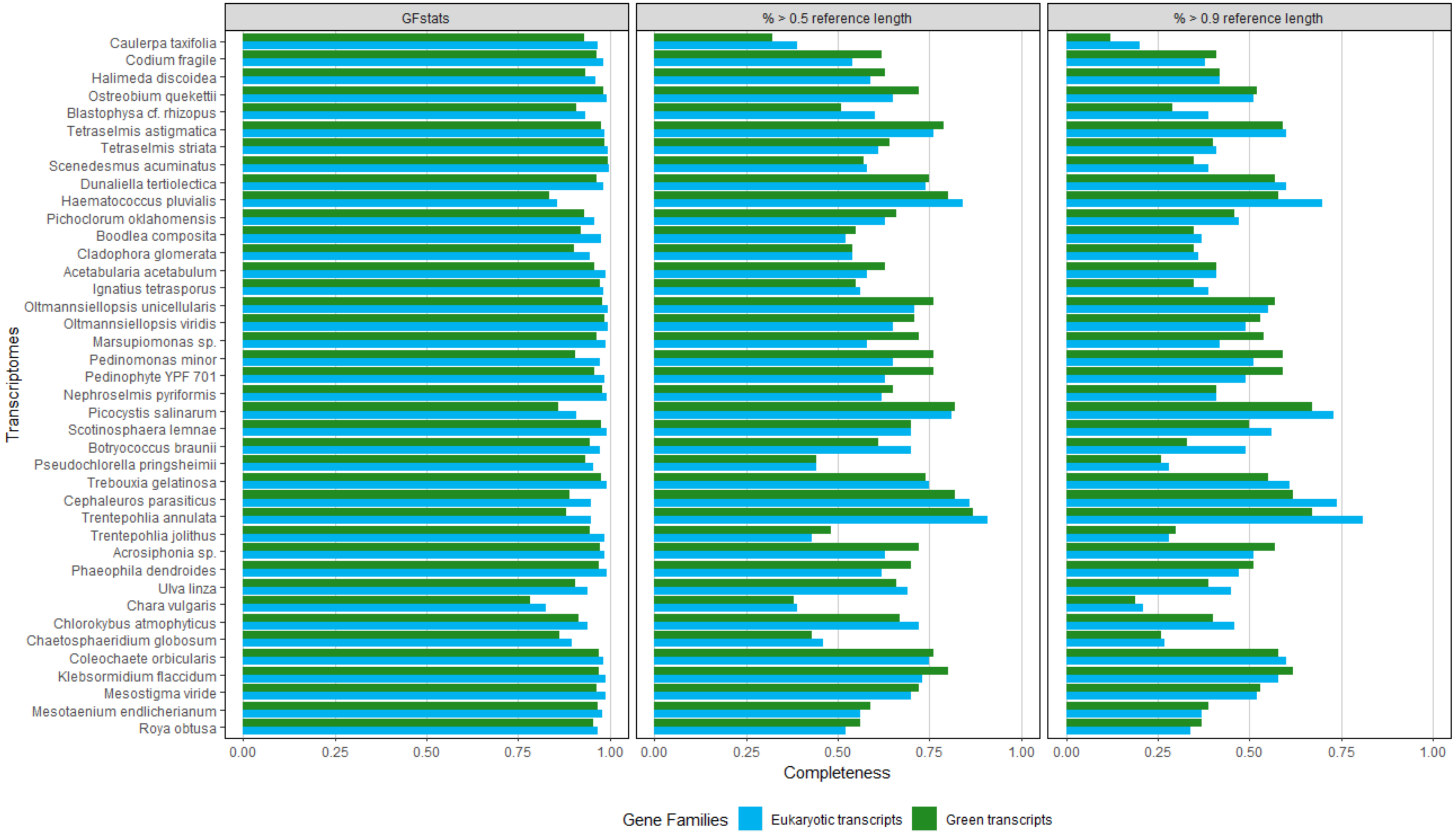


Figure 2.3: GFscore and Reference Coverage score.

Bar plots illustrating the GFscore and the coverage score (50% and 90% of reference genes covered) of the eukaryotic and the green transcripts of each transcriptome.

terms (Figure S2.3). Noticeably, *Trentepohlia jolithus* and *Phaeophila dendroides* seems to have additional KO terms not represented by any of the other genomes/transcriptomes, neither the ones belonging to the same group. This may indicate some degree of non-*Viridiplantae* contaminants in the RNA-seq library. Alternatively, it could indicate events of gain of species-specific functions and lateral gene transfer that were absent in the ancestor and in the other members of the same clade.

Depth of sequencing: transcriptome completeness and gene family size correlation

Depth of sequencing: Transcriptome completeness evaluation

Depth of sequencing influences the amount and the length of transcripts that can be reconstructed in a *de novo* transcriptome assembly, with the underlying assumption that larger RNA-seq libraries are required for representing larger gene spaces. However, for many of the green algal species in this study, little to no info on the corresponding gene space and ploidy level is available. To obtain a good proxy for the minimum depth of sequencing required for detecting full-length transcripts, RNA-seq experiments with increasing sequencing depth were investigated in the model green algae *Chlamydomonas reinhardtii*.

Starting from the same RNA-seq library, ten subsets of reads were randomly selected, ranging from 8M to 80M reads (Table S2.1), to represent RNA-seq experiments with increasing sequencing depth. Each subset was independently *de novo* assembled with Trinity and transcripts clustered. For each assembly, eukaryotic and green transcripts transcripts were identified (Figure S2.4, Table S2.1). The completeness of the eukaryotic fraction of the transcriptome assemblies was assessed with three independent strategies: coreGF, BUSCO, eggNOG-mapper (see section 3). The coreGF score for all the assemblies was close to 1, indicating that virtually all the core conserved gene families identified. Increasing sequencing depth resulted in higher fraction of reference genes covered by 50% or 90% of their length (Figure S2.5A).

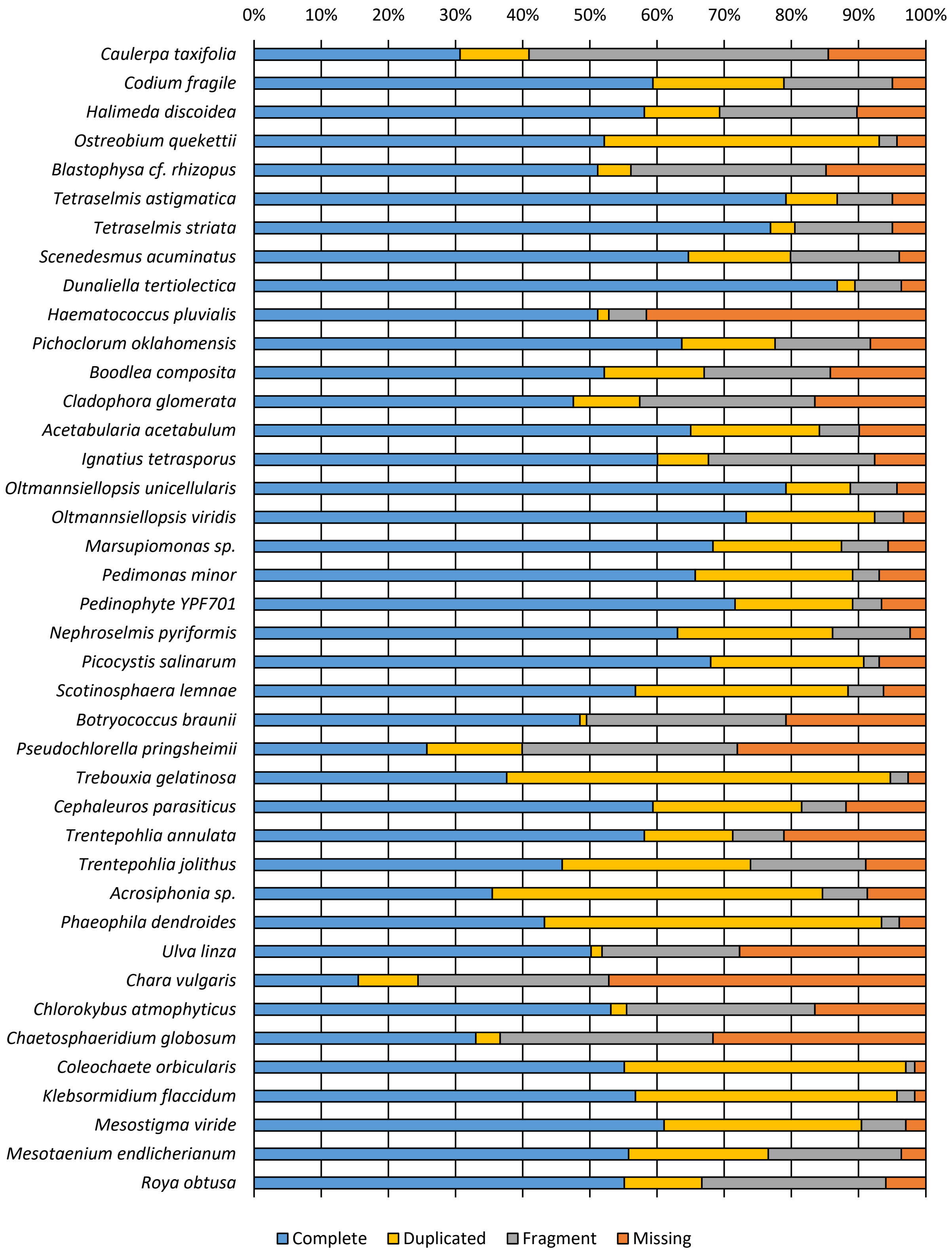


Figure 2.4: Gene space completeness predicted with the BUSCO analysis.

The bar plot reports the percentage of the BUSCO genes identified as complete, duplicated, fragmented or missing in the assemblies and in the reference genome of the total 303 BUSCO eukaryotic single-gene orthologs.

In total, between 83.9 and 97.3% of eukaryotic BUSCO were identified in the different assemblies (Figure S2.5B), while 98.3% of BUSCO genes were identified in the CDS of the *Chlamydomonas* reference genome v. 5.5. Notably and as expected, at increasing depth of sequencing, more BUSCO genes were identified as complete or duplicated and less as fragmented, indicating that higher depth of sequencing results in longer transcripts and reconstruction of multiple transcripts for each locus.

ChloroNOG KO terms were almost completely covered by the transcripts at all depths of sequencing, as well as virNOG (Figure S2.5C), except for Streptophyta-specific KO terms. When compared to all the NOG available, only a small fraction was identified. A similar analysis performed on the genome indicated discrepancies in the pattern of KO terms identified in the transcriptomes irrespective of the NOG sets analyzed. The presence of additional sequences not covered in the *Chlamydomonas* genome was confirmed by the taxonomic profile obtained with GhostKOALA (Figure S2.4).

Depth of sequencing: gene family size correlation

For each depth of sequencing, gene families were independently build following a PLAZA-like procedure from the coding sequences of 18 Viridiplantae genomes, including reference *Chlamydomonas* sequences and *Chlamydomonas de novo* assembled transcripts (see Materials and Methods for details). Then, in each gene family the number of *Chlamydomonas* reference and *de novo* assembled sequences were evaluated. The analyses of *de novo* eukaryotic transcripts and green transcripts gave comparable results (Figure 2.5, 2.6). The slope of the regression line was slightly below 1 only for the 8M reads subsets. At increasing depth of sequencing, the slope steadily increased, indicating that higher sequencing depth leads to an overestimation of gene family sizes when building gene families from *de novo* assembled transcriptomes. The intercept on the y-axis (expressed eukaryotic transcripts and green transcripts) was always higher than 0, suggesting, in general, a higher number of eukaryotic transcripts than actual genomic genes per gene families. As indicated by the correlation coefficient this trend is subject to considerable variation between gene families (Figure 2.5, 2.6).

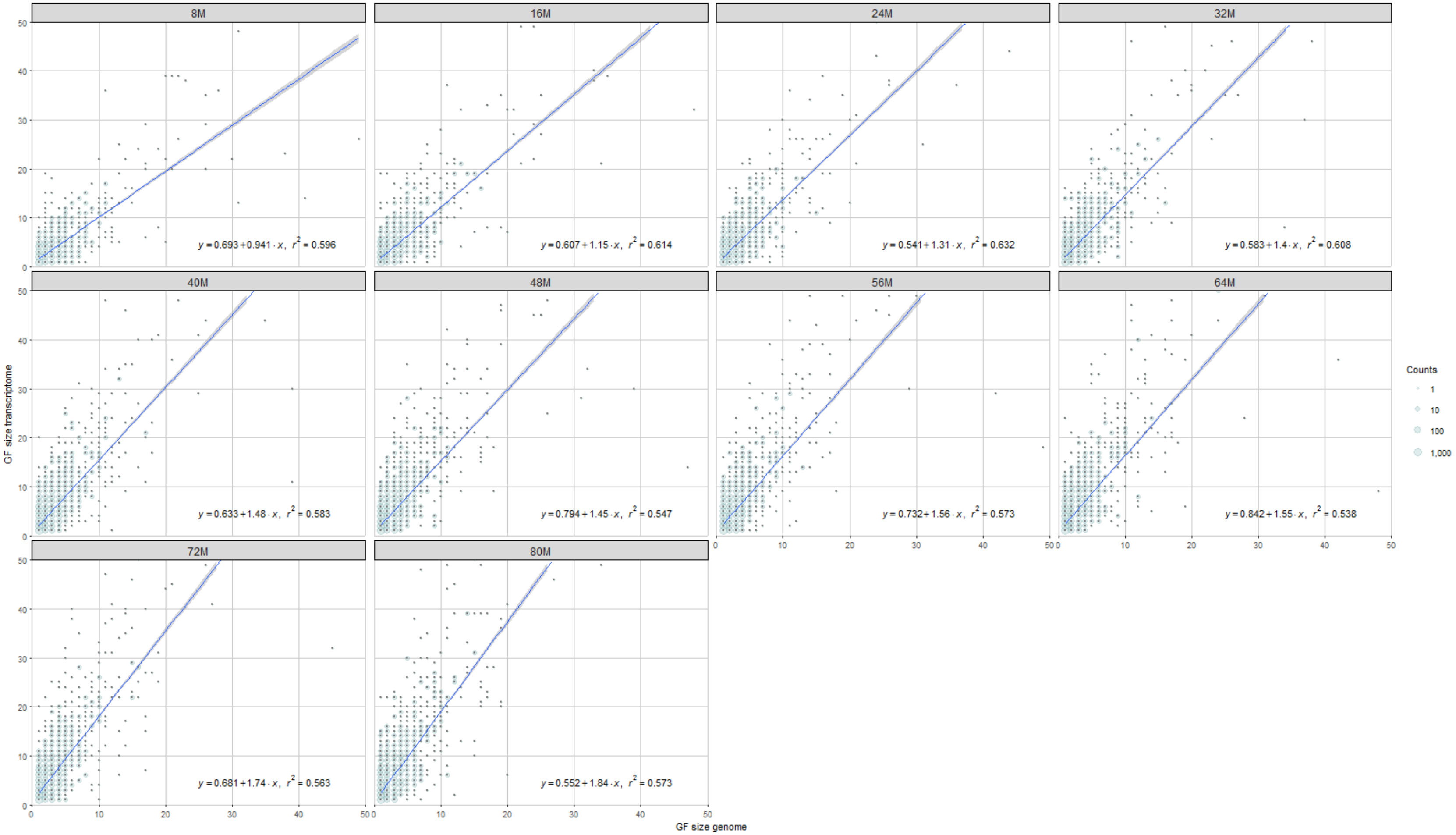


Figure 2.5: Gene family size correlation – expressed transcripts.

Each circle represents one or more gene family, the x-axis represents the number of genes from the reference *Chlamydomonas* transcriptome found expressed in that gene family, while the y-axis reports the corresponding genes found in the eukaryotic fraction of the *de novo* assembled transcriptomes. The size of the circle is proportional to the gene families with that relative number of genomic and transcriptomic genes. The blue line corresponds to the regression line, the surrounding grey area indicates the 95% confidence interval. The corresponding equation is reported, together with the coefficient of correlation r^2.

Effect of partial data on orthology inference

Despite the dataset in this study is composed by a majority of potentially fragmented data (i.e.: *de novo* assembled transcriptomes), the orthology inference was performed with PLAZA 4.0 (Van Bel *et al.*, 2018), a comparative genomic pipeline, designed to deal with a set of comprehensive full-length sequences for gene families delineation. The 40 transcriptomes and 15 genomes of Viridiplantae species present in Table 2.1 were used to build a custom PLAZA 4.0 istance (chloroPLAZA), with more than 70% of the data represented by potentially fragmented data. Therefore, we tested the impact of partial sequences on the inference of homologous gene families by the PLAZA pipeline. In order not to overestimate gene families with perfect correlation, we further filtered out HOM families composed by species-specific genes ("orphan" genes). Then, we performed pairwise comparisons between the three PLAZA builds - picoPLAZA (Vandepoele *et al.*, 2013), PLAZA 2.5, and chloroPLAZA - to assess the distribution of corresponding HOM families on different PLAZA builds and test if unitary gene families were split in different PLAZA builds (Figure 2.7). picoPLAZA against PLAZA 2.5 pairwise comparisons between gave comparable results, with less than 10% of the gene families split into 2 distinct gene families in the second PLAZA build (gene family correlation score of 0.5), while more than 90% of the gene families had a perfect a one to one correspondence (gene family correlation score of 1). chloroPLAZA comparison against picoPLAZA and PLAZA 2.5 resulted instead on a slightly higher amount of gene families split, 12.6% and 10.6% respectively.

Discussion

Taxonomic binning

Bacterial contamination is common in algal transcriptomic data obtained from field-collected material, as well as from cultures (Keeling *et al.*, 2014). Moreover, several green macroalgae engage in a mutualistic lifestyle with bacteria (Spoerner *et al.*, 2012; Wichard, 2015), or harbor intracellular bacteria (Hollants *et al.*, 2011; Hollants *et al.*, 2013; Aires *et al.*, 2015). Reliable methods to discriminate between contaminant and algal sequences are therefore fundamental.

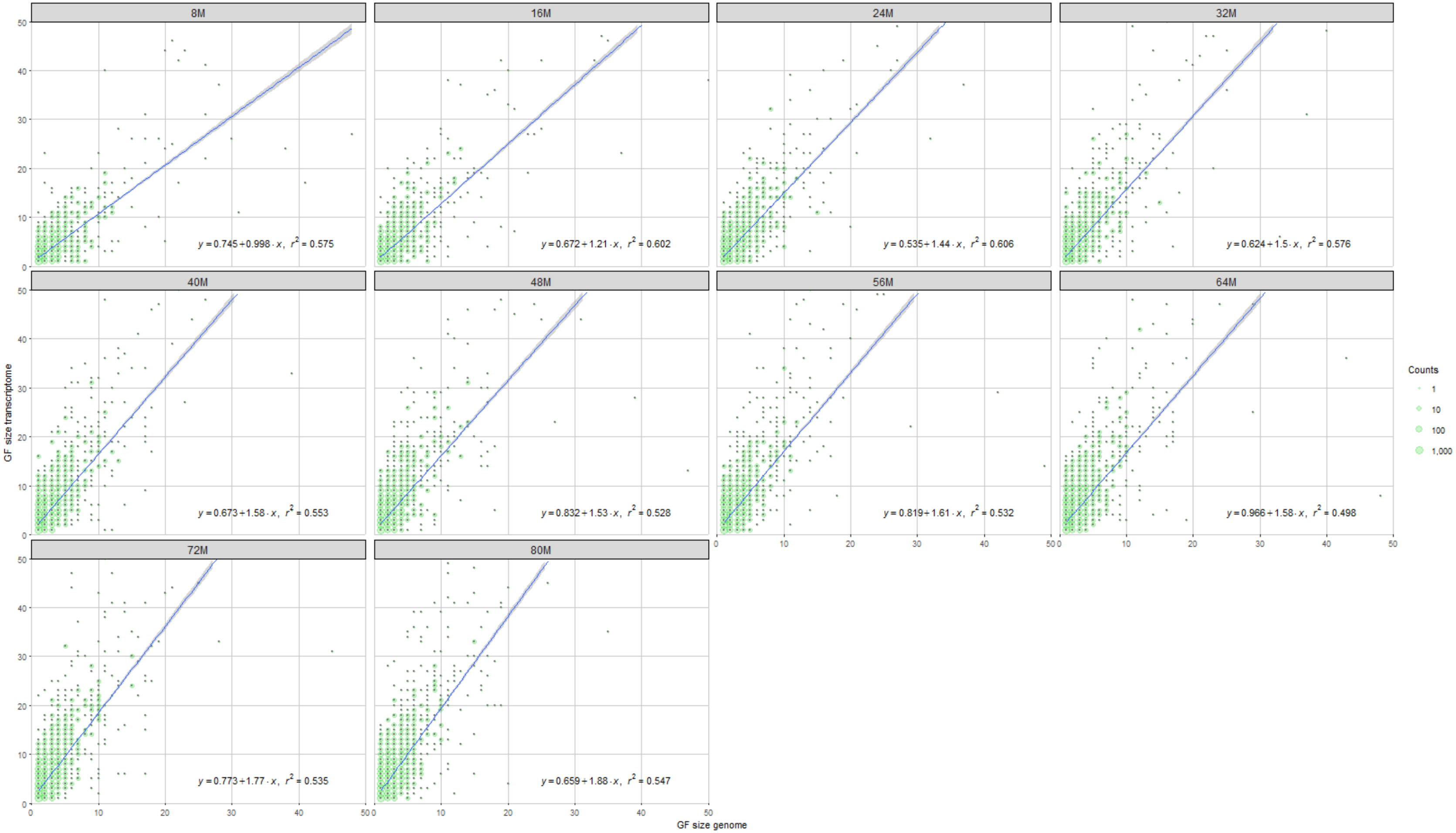


Figure 2.6: Gene family size correlation – expressed green transcripts.

Each circle represents one or more gene family, the x-axis represents the number of genes from the reference *Chlamydomonas* transcriptome found expressed in that gene family, while the y-axis reports the corresponding green genes found in the eukaryotic fraction of the *de novo* assembled transcriptomes. The size of the circle is proportional to the gene families with that relative number of genomic and transcriptomic genes. The blue line corresponds to the regression line, the surrounding grey area indicates the 95% confidence interval. The corresponding equation is reported, together with the coefficient of correlation r^2.

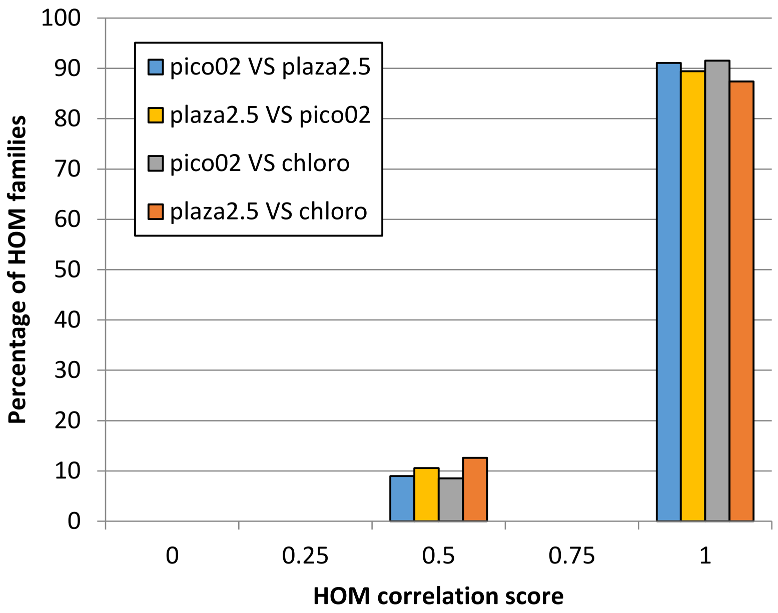


Figure 2.7: Gene family correlation between PLAZA builds.

pico02: picoPLAZA build. plaza2.5: PLAZA 2.5 build. chloro: chloroPLAZA 4.0 build. The builds were generated with sequences from Table 2.2 species.

We tested the efficiency in taxonomic profiling of an in-house method based on sequence similarity searches against the NCBI non-redundant protein database linked with taxonomic information, followed by a more detailed identification obtained with GhostKOALA. Similarity searches should handle properly the presence of the nuclear alternative genetic code, since they can perform gapped alignments. We showed that a single GhostKOALA analysis can be misleading, as is shown by the considerable amount of sequences not identified as "Plants" in well characterized green algal and land plant genomes (Figure 2.1). This is could be ascribed to the reference database GENES, which is a non-redundant collection of proteins from non-redundant pangenomes (Kanehisa *et al.*, 2014). However, the GENES database non-redundancy rule is extended up to the family level, and this should not result in mis-classification at the kingdom level. Moreover, GhostKOALA can give a useful estimate on the degree of contaminant sequences if datasets from the same species or from closely related species are analyzed. For example: difference in taxonomic distribution between *Chlamydomonas* genomic and transcriptomic sequences indicates the presence of contaminants in the transcriptomes, supported as well by different KO terms distribution in these two datasets. The fact that *Oltmannsiellopsis unicellularis* and *O.*

viridis have similar taxonomic distributions despite that their transcriptomes have been prepped and sequenced independently in different facilities, indicates that sequences classified as "Animal" could well be authentic *Oltmannsiellopsis* sequences. An additional explanation could come from the presence of undetected *Oltmannsiellopsidales* endo- or epiphytic organisms still not described nor characterized. Conversely, the taxonomic distribution of *Trentepohlia annulata* and *Cephaleuros parasiticus* sequences differs considerably from that of *Trentepohlia jolithus*, indicating that the latter transcriptome contains a considerably higher proportion of contaminant sequences, probably because the RNA was obtained from an environmental sample (Li *et al.*, 2014).

Frameshift correction

Frameshifts and premature stop codons in *de novo* assembled transcripts can represent real biological features in genes, which are corrected by RNA editing, as in mitochondrial transcripts of kinetoplastids or in chloroplast transcripts of dinoflagellates (Ochsenreiter & Hajduk, 2008; Mungpakdee *et al.*, 2014), or may represent pseudogenized genes. More often, however, frameshifts and premature stop codons result from sequencing or short read assembly errors. Frameshift errors are detrimental to the downstream comparative and phylogenetic analyses, and result in shorter and incorrectly translated peptides. In addition, due to specific drawbacks of each technology, each sequencing platform has its proper error profile which has to be taken into account (Glenn Travis, 2011; Nakamura *et al.*, 2011; Schirmer *et al.*, 2015; Schirmer *et al.*, 2016; Abnizova *et al.*, 2017). The 454 pyrosequencing technology is more prone to homopolymeric insertions, while Illumina technology is more prone to substitution: i.e. T to G transversions (Schirmer *et al.*, 2016), resulting in a theoretical higher chance to create frameshifts errors with 454 than with Illumina. Despite this, the 454 dataset did not show the highest rate of frameshifts, however, frameshift correction was most successful for the 454 datasets. Datasets generated with older Illumina chemistry did not perform worse than datasets generate with newer Illumina technologies and with longer reads. No correlation was found between depth of sequencing and percentage of frameshift transcripts, maybe because of the *in silico* read normalization step before assembling the Illumina reads with Trinity. These results may suggest that, at least for the datasets we analyzed, the putative frameshift transcripts could be ascrived to random errors during the sequencing and assembly steps, rather than dependent on the technology of sequencing. However, for a proper benchmark of the origin of frameshift errors, the same RNA library should be sequenced on different sequencing machines, which was beyond the scope of this investigation.

Our frameshift correction approach is probably not perfectly suited for clades with alternative nuclear genetic codes, because of the unpredictable behavior on this kind of datasets. In fact, at the moment, TRAPID cannot handle multiple translation tables for coding sequences detection, nor can FrameDP for the frameshift correction step. On the other hand, these tools should still be efficient in detecting frameshift transcripts. In the presence of a reassigned internal stop codon, TRAPID is more likely to flag transcripts as "partial", "quasi full-length" and "full-length", depending on transcript length and position of the internal stop codon. Only transcripts with best hits to the same reference proteins on different reading frames are flagged as "frameshift". FrameDP instead should treat internal stop codon as multiple events of frameshift that disrupt the coding sequence. This results into multiple random insertions of "N" nucleotides at the level of the internal stop codon, until the stop codon is converted to multiple ambiguous codons and the coding frame is restored. In principle, TRAPID and FrameDP should manage to correct frameshifts transcript also in the presence of alternative nuclear genetic code, although the amount of artificial insertions is unpredictable. On the other hand, there is no obvious or logic explanation for the lack of correction of *Trentepohlia annulata* frameshift transcripts. Tools designed to handle alternative nuclear genetic codes are required to confidently detect and correct frameshifts, and, at best of our knowledge, still missing.

Transcriptomes gene space completeness evaluation

Three different methods used to evaluate the completeness of transcriptome gene space, coreGF, BUSCO and eggNOG-mapper, gave complementary overviews on the gene-space completeness of the transcriptomes and of the genomes. While coreGF and BUSCO methods gave comparable results, coreGF was found to have higher sensitivity. coreGF reference database is six time larger than BUSCO (1,815 coreGF proteins against 330 BUSCO proteins) and it is populated almost exclusively with green algae genomes. This represent a reference database well suited to assess the gene space completeness of a green algal transcriptome. The BUSCO database we used is a collection of 330 quasi-single-copy orthologous genes shared by all eukaryotes. There is an Embryophyta-specific BUSCO database, but it is unbalanced towards land plants and not representative for green algae. Therefore, coreGF results as a more sensitive approach to evaluate gene space completeness in green algae. In our analysis, this difference resulted evident for some species (e.g.: *Pseudochlorella pringsheimii*), that despite having a completeness score close to 1 was missing almost 30% of BUSCO genes.

One may think that BUSCO genes are related to extremely conserved functions among eukaryotes (such as DNA replication and repair), and might not be always expressed to levels high enough for granting the detection of the transcripts. Several coreGF proteins instead, since they were determined exclusively from green algae, are probably involved in the photosynthetic processes and possibly expressed at high levels in metabolic active algae, which grant their detection during RNA sequencing. Moreover, coreGF method has the strong advantage of reporting the percentage of reference transcripts covered by 50% and 90% of their length, while BUSCO only provides a qualitative score for the transcript completeness. Recovery of full-length transcripts is a desirable outcome of RNA-seq sequencing and it is required for successful downstream analyses: the coverage score is therefore a fundamental estimate of integrity of conserved transcripts.

The eggNOG mapper is a comprehensive annotation method that can assign KO terms, gene onthology (GO) terms, KEGG pathways and cluster of orthologous groups (COG) functional categories to query sequences based on sequence similarity to precomputed clusters of orthologous groups (Huerta-Cepas *et al.*, 2017). To evaluate the completeness of a transcriptome, we detected the fraction of KO terms associated with chloroNOG and virNOG clusters identified in the transcriptomes. The KO terms identified in the different transcriptomes revealed useful insights on the completeness and on the degree of putative novel/contaminating sequences in the transcriptomes that coreGF and BUSCO methods would miss. On the other hand, many coreGF genes are not associated with a KO term: most species have coreGF score close to 1, while most of the Chlorophyta/Viridiplantae KO terms may be absent (e.g.: *Cephaleuros parasiticus* and *Trentepohlia annulata* transcriptomes). This observation is consistent with the annotation of "green cut proteins", a set of 597 nucleus-encoded proteins conserved in all photosynthetic organisms, 50% of which are not yet characterized (Karpowicz *et al.*, 2011).

Despite these three approaches rely on distinct reference databases and they are therefore not directly comparable, their concurrent use should grant a wide overview on transcripts and gene space completeness of a transcriptome. Moreover, the complementarity of the reference databases may hint toward which genes/functions are present and the possible presence of contaminant sequences

Depth of sequencing: transcriptome completeness and gene family size correlation

The analyses of transcripts and gene space completeness at increasing depth of sequencing indicated 40M reads as the depth of sequencing where most of the measured metrics reached a plateau. Despite N50 values and number of assembled transcripts increased with higher depths of sequencing, adding more reads seems not to improve drastically the recovery of conserved genes, and, instead, it results to be detrimental for certain metrics. At higher depth of sequencing, in fact, more BUSCO genes were identified as duplicated, and gene family sizes correlation between reference and *de novo* assembled sequences indicated a biased toward *de novo* sequence overestimation. The results from the gene space completeness analysis pointed towards an almost complete gene space when compared to the reference genome at all the depth of sequencing analyzed (8M-80M reads), but higher sequencing depth resulted in longer transcripts and a larger percentage of reference genes length covered.

This analysis represents a good starting point to model the relationship between depth of sequencing, gene space completeness evaluation and orthology inference. However, it is severly hampered by being restricted to *Chlamydomonas* only. It is difficult to predict how this relationship scale from *Chlamydomonas* 111 Mb genome coding for almost 18 thousand genes to the transcriptome of a green alga for which estimates on genome and gene space sizes are not available. The predictive power of this analysis would definitely benefit by being repeated on all available green algal genomes.

Effect of partial data on Orthology inference and gene family size correlation

A desirable feature of comparative transcriptomics is the possibility to compare gene family sizes between different organisms to elucidate and unveil the genetic underpinning and evolution of biological properties. Key biological innovations are often associated with gene family expansion, shrinking, gain or loss (Gardner *et al.*, 2002; Martens *et al.*, 2008; Pombert *et al.*, 2014; Dunn *et al.*, 2018). While gene family comparisons are somewhat easily executed in comparative genomic studies, it is not trivial to perform them based on transcriptomic data given that *de novo* assembled transcriptomes only partially represent the gene space of the corresponding genomes.

Our analyses on gene family reconstruction from expressed transcripts and expressed green transcripts indicate that transcriptomic data generally overestimate gene family sizes. Moreover, the corresponding correlation coefficient values were relatively low (r^2 = 0.498-0.632). This trend is observed virtually at any sequencing depth, and it could be ascribed to the redundancy of the transcriptomics datasets, i.e.: multiple transcripts are assembled for each genomic locus. Furthermore, *Chlamydomonas* is suited for this gene family size correlation assessment, since its transcription and ploidy level is much simpler if compared to other eukayrotes (e.g.: higher plants or higher mammals), and only 10% of its transcripts undergo alternative splicing (1,785 out of 17,741 protein-coding genes). Allelic variants and alternative spliced transcripts cannot account for the overestimation of gene family sizes observed in the *de novo* assembled sequences. Thus, together with the intrinsic incomplete gene space represented in the transcriptome, these results support the idea that transcriptomic data generally are not well suited for analyses of gene family expansion, and gene gain and loss.

These results contrast with the accuracy of gene family inference of PLAZA 4.0. This method in fact is robust and does not suffer from the majority of data being fragmented sequences during the gene family reconstruction step. Perhaps, complete sequences from the 15 genomes drive the faithful inference of orthologous groups, despite representing only 30% of the total sequences. Redundant sequences from *de novo* assembled transcriptomes are likely to be grouped in the same gene family and not to

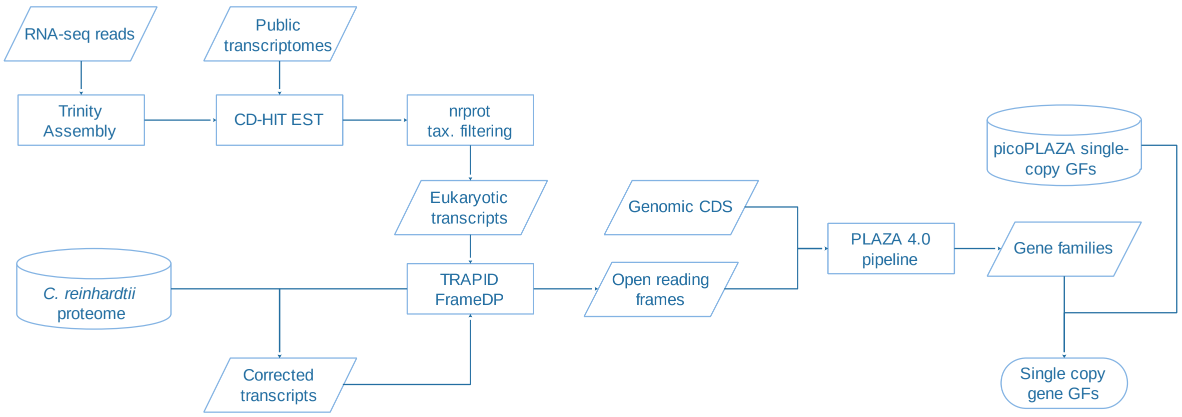


Figure 2.8: Schematic overview of the transcriptomic pipeline for green algal transcriptomes.

Detailed information on the workflow steps are reported on the text.

influence the orthology inference process. Instead, the minimal (2%) increase of split gene families observed in chloroPLAZA (Figure 2.7) are probably ascribed to partial transcripts assigned to the wrong gene families due to similar domain composition.

A transcriptomic pipeline for green algae

The information acquired during our analyses was used to build a *de novo* assembly and annotation pipeline. After quality control, RNA-seq reads are first assembled, then, the redundancy in the transcriptome is reduced using CD-HIT EST with stringent parameters. Nrprot taxonomic binning approach is used to discard bacterial sequences and sequences with no similarity with known proteins. TRAPID and FrameDP are used to correct frameshift errors and to predict open reading frames of each transcript with an orthology-guided approach based on *Chlamydomonas* proteome. A custom java plugin allows TRAPID do uses multiple translation tables during Open reading frame prediction, accounting for nuclear alternative, chloroplast and mitochondrial genetic code. Open reading frames are combined with the genomic coding sequences to build a custom PLAZA 4.0 instance, which is used to circumscribe gene families, which can be used in phylotranscriptomic analyses in Chapter 3: “Reconstruction of the early diversification of green seaweeds” (Figure 2.8).

Materials and Methods

Dataset retrieval and RNA extraction

55 species of green algae were sampled belonging to the major clades of the Chlorophyta and Streptophyta. Data consisted of 15 genomes and 40 transcriptomes; 9 transcriptomes were generated for this study, while the remaining data were retrieved from publicly available repositories (Table 2.1).

Acetabularia acetabulum, *Oltmannsiellopsis viridis* and *Scotinosphaera lemnae* cultures were grown at 20 °C and 12-h light/12-h dark cycle in Ace-25 medium (Hunt & Mandoli, 1996) and 1.5% agar-solidified Bold Basal medium (Bischoff, 1963), respectively. *Marsupiomonas* sp., *Pedinomonas minor* and the pedinophyte strain YPF-701 (NIES Microbial Culture Collection strain NIES-2566) were cultured in Guillard's F/2 medium at 20 °C and 14-h light/10-h dark cycle. For *Ostreobium* sp. HV05042, *Halimeda discoidea* and *Codium fragile* freshly collected material was used for RNA extraction. Collection details are given in Verbruggen et al. (2017). Unicellular microscopic algae cultures (*Marsupiomonas* sp., *Oltmannsiellopsis viridis, Pedinomonas minor,* pedinophyte strain YPF-701 and *Scotinosphaera lemnae*) were harvested during their exponential phase. Whole macroscopic seaweed specimens were harvested from cultures (*Acetabularia acetabulum*) or collected in their natural environment (*Codium fragile*, *Halimeda discoidea* and *Ostreobium* sp. HV05042) and ground in liquid nitrogen for RNA extraction. RNA extractions follow Palmer (1982). RNA quality and quantity were assessed with Qubit and Nanodrop spectrophotometer, and integrity was assessed with a Bioanalyzer 2100. RNA-seq libraries were sequenced as reported in Table 2.3.

Transcriptome Assembly and Taxonomic filtering

At the time of the experiment, only pre-assembled transcriptomes of *Acrosiphonia* sp., *Blastophysa rhizopus* and *Caulerpa taxifolia* transcriptomes were available: Therefore, these datasets were retrieved from the respective sources reported in Table 2.1. All the remaining assemblies were performed in house starting from the raw reads on a custom semi-automated pipeline. The pipeline consisted of the following steps. Quality of the raw reads were assessed with FastQC v.0.10.1

(http://www.bioinformatics.babraham.ac.uk, last accessed March 01, 2017). Low-quality reads (average Phreds quality score below 20) and low quality read ends were trimmed with Fastx v.0.0.13 (https://github.com/agordon/fastx_toolkit, last accessed March 01, 2017). Trimmed reads shorter than 30 bp were discarded.

Transcriptome *de novo* assembly of *Botryococcus braunii*, *Chlorokybus atmophyticus*, *Ulva linza* (Roche 454 data) were performed with CLC Genomics Workbench version 7.5.1 (http://www.clcbio.com, last accessed on June 06, 2017), using a word size of 63 and standard parameters. Transcriptome *de novo* assembly for the remaining species were performed with Trinity 2.1.1 (Grabherr *et al.*, 2011), in SE or PE mode were appropriate depending on the RNA-seq library type, after *in silico* reads normalization.

For each transcriptome, transcripts were clustered with CD-HIT-EST v. 4.6.1 (Li & Godzik, 2006) with the following parameters: -c 0.975, -d 0, -p 1 and -M 0, and the longest one was retained as representative of the cluster. The eukaryotic fraction of each transcriptome was identified with sequence similarity searches using Tera-BLAST DeCypher (Active Motif, USA) against the NCBI non-redundant protein database combined with the NCBI Taxonomy information of the top ten BLAST hits (hereafter: nrprot), where a hit to Eukaryotic sequences was sufficient to assign the transcript tot the “eukaryotic bin”. Transcripts were therefore assigned to “eukaryotic”, “bacterial”, “no hit” bins based on the outcome of search. To evaluate the amount of residual bacterial contaminants still present in nrprot and to circumscribe a fraction of green transcripts, the coding regions of eukaryotic transcripts were predicted with TRAPID and translated into the corresponding amino acid sequences with transeq algorithm from EMBOSS 6.6.0 (Rice *et al.*, 2000), using the appropriate translation table where necessary. The resulting peptides were processed with GhostKOALA (Kanehisa *et al.*, 2016). For each peptide having the best scoring hit in *Viridiplantae*, the corresponding transcript was flagged as “green”. Coding sequences from the genomes in Table 2.1 were processed as well in a similar manner as control sequences.

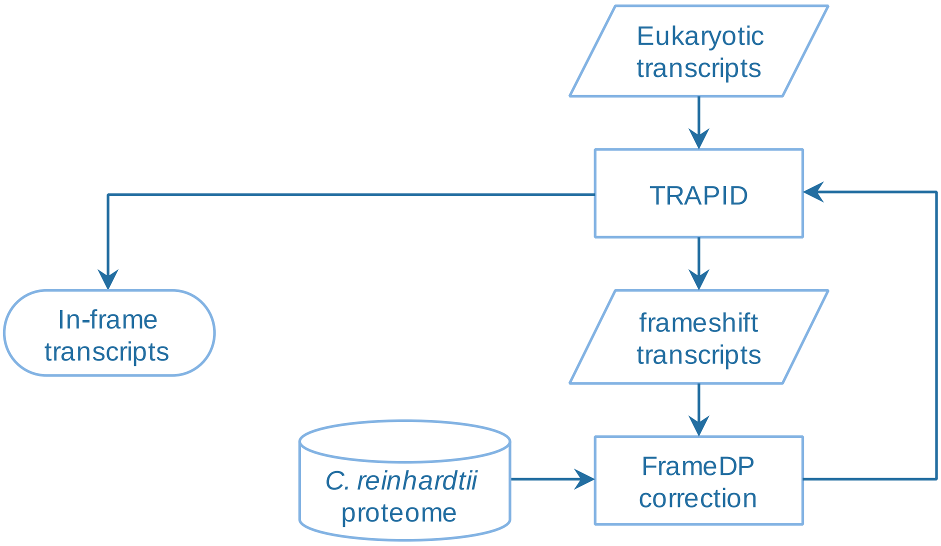


Figure 2.9:Schematic overview of the frameshift correction workflow. Detailed information on the workflow steps are reported on the text.

Frameshift correction

Frameshift correction performance (Figure 2.9) was evaluated on the eukaryotic fraction of the transcriptomes of Table 2.2. Each transcriptome was analyzed with TRAPID (Van Bel *et al.*, 2013). For each transcriptome, a subset of transcripts with potential frameshift was predicted, based on the concordance with orthologous sequences as predicted with TRAPID. This subset was corrected with a local step of frameshift correction in FrameDP 1.2.2 (Gouzy *et al.*, 2009), using the *Chlamydomonas* 4.0 proteome. The corrected subset, together with transcripts that were not classified as potentially having a frameshift, were analyzed again with the TRAPID pipeline to evaluate the efficiency of the frameshift detection and correction.

Transcriptome completeness evaluation

Evaluation of transcriptome completeness was performed by evaluating the gene space completeness of the genomes and the eukaryotic fraction of transcriptome with three different methods (Figure 2.10): coreGF score (Veeckman *et al.*, 2016); BUSCO (Simão *et al.*, 2015) and eggNOG-mapper (Huerta-Cepas *et al.*, 2017). Despite that each method relies on a different set of reference sequences, the underlying idea is conserved: estimate the portion of the reference database represented in the query sequences, thus, derive the *bona fide* completeness of the query itself. Where possible, the reference dataset should include closely related species, to increase the likelihood to represent the most complete (likely) gene space possible of the query species.

For the coreGF evaluation method, transcripts and genomic coding sequences were searched using Tera-BLAST™ DeCypher against pico-PLAZA v2.0 proteome, which is composed by genome sequences of 10 green algae, 3 land plants, one glaucophyte and 5 stramenopiles (Vandepoele *et al.*, 2013). A set of 1,815 homologous core gene families were identified with tribe-MCL (Enright *et al.*, 2002). A core gene family is defined as a gene family shared among all the Chlorophyta genomes present in pico-PLAZA 2.0. Each transcript or coding sequence was assigned to a certain gene family based on its top 5 hit, following a majority consensus rule. Finally, the GFscore was calculated as the sum of each core family identified, counted with a weight equal to one divided by the average family size.

For the BUSCO analysis, the transcripts and genomic coding sequences were searched with BUSCO 3.0.1 (Simão *et al.*, 2015) against the eukaryotic ortholog groups present on OrthoDB v9 database (Zdobnov *et al.*, 2017), using the "transcriptome" mode (flag -m tran).

For the eggNOG-mapper search (Huerta-Cepas *et al.*, 2017), three distinct sets of non-supervised orthologous groups (NOG) clusters were tested: Chlorophyta NOGs (chloroNOG), Viridiplantae NOGs (virNOG), and all the NOGs present in the emapper database v. 4.5.1 (bacterial and viral NOGs included). The coding sequences of genomes and transcriptomes were compared to the chloroNOG, virNOG and NOG clusters. Searches against chloroNOG and virNOG were run with HMMer profiles (Eddy, 2011) available at the emapper database, using optimized memory searches (flag --usemem). For the sake of speed, searches against the whole NOG clusters were run with DIAMOND 0.9.9, flag -m diamond (Buchfink *et al.*, 2014). To evaluate the gene space completeness predicted with eggNOG searches, we identified the number of KEGG Orthology (KO) terms (Kanehisa *et al.*, 2014) that were assigned to each subset by the searches. To define the total KO terms associated to chloroNOG and virNOG clusters, the protein used for populating the clusters were searched

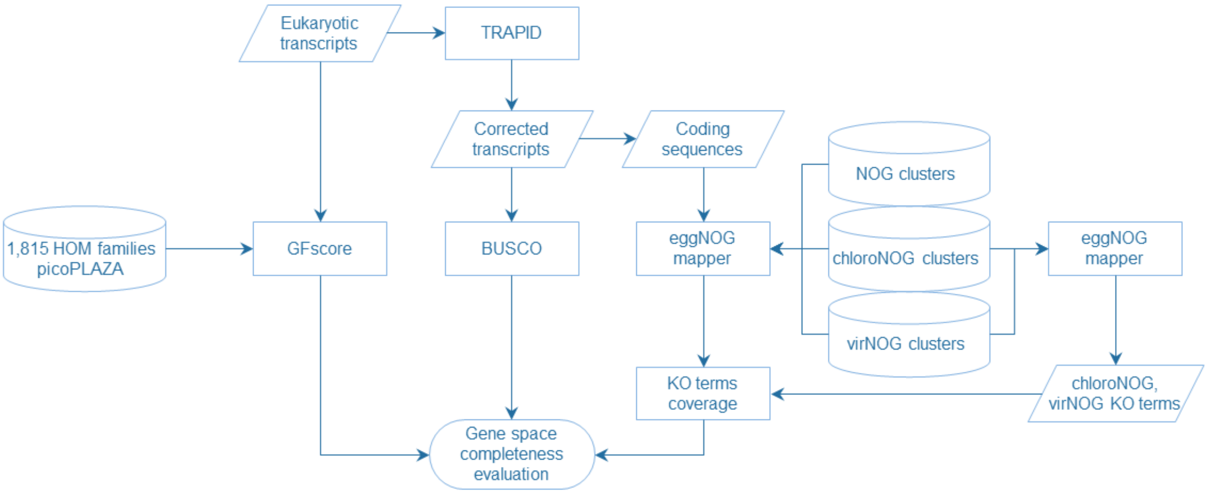


Figure 2.10: Schematic overview of the gene space completeness evaluation workflow.

Detailed information on the workflow steps are reported on the text.

against chloroNOG and virNOG clusters respectively, and the resulting KO terms identified were set as KO terms associated to the clusters.

Depth of sequencing: transcriptome completeness and gene family size correlation

Data retrieval, transcriptome assembly and frameshift correction

For the depth of sequencing evaluation, the *Chlamydomonas reinhardtii* reference genome v.5.5 and corresponding full length transcripts were retrieved from Phytozome, https://phytozome.jgi.doe.gov/pz/portal.html (Goodstein *et al.*, 2012). The SRR353973 RNA-seq library was retrieved from the Short Read Archive, https://www.ncbi.nlm.nih.gov/sra/?term=SRR353973 (Leinonen *et al.*, 2011), which represents the largest RNA-seq library available to our knowledge for *Chlamydomonas reinhardtii*.

The workflow to process this dataset was similar as described above. Briefly, quality of the 83M 2x101 bp PE raw reads were assessed with FastQC v.0.10.1. Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' ends of the reads were trimmed with Fastx v.0.0.13. After trimming, reads shorter than 30 bp were discarded. 80,660,120 PE reads were retained after trimming and filtering. Ten subsets representing 10-100% of the trimmed reads were randomly selected from the total retained reads (Table S2.1) to mimic comparable RNA-seq libraries with increasing sequencing depth using a custom python script https://github.com/brentp/bio-playground/blob/master/reads-utils/select-random-pairs.py. Each subset of PE reads was independently assembled with Trinity v. 2.1.1. Trinity assemblies were performed with standard parameters after *in silico* read normalization (flag --normalize_reads).

After assembly, the resulting transcripts were clustered with CD-HIT-EST 4.6.1 using the following parameters: -c 0.975, -d 0, -p 1 and -M 0. The longest member for each cluster was used as the representative for the cluster for downstream analyses. The eukaryotic fraction for each assembly was identified as described above. Transcripts classified as "eukaryotic" were further processed in TRAPID for identifying coding regions and correct putative frameshift errors. A fraction of green transcripts was identified with GhostKOALA. After processing with TRAPID, the transcripts and the corresponding coding regions of each assembly were examined to assess transcriptome completeness as described in the previous paragraph.

Gene family size correlation

To evaluate the members of a gene family *de novo* assembled in a transcriptome with the real number of members as identified in the corresponding fully sequenced genome, we determined homologous gene families (HOM) in a similar process to the PLAZA 4.0 build (Figure 2.11). We build a custom protein dataset by first removing *Chlamydomonas* sequences from picoPLAZA proteome and by adding the remaining picoPLAZA proteins to all predicted proteins from the sequenced genomes of *Gonium pectorale* (Hanschen *et al.*, 2016), *Auxenochlorella protothecoides* v. 1.0 (Gao *et al.*, 2014)*, Astereochloris* sp. Cgr/DA1pho v2.0*, Ulva mutabilis* v.3.0 (De Clerck *et al.*, 2018), *Selaginella moellendorffii* v. 1.0 (Banks *et al.*, 2011). Then *C. reinhardtii* complete reference transcriptome v.5.5 was retrieved from JGI, and only reference *Chlamydomonas reinhardtii* transcripts that were found expressed at a certain depth of sequencing were added to the custom protein database. For each depth of sequencing, reference *Chlamydomonas* expressed genes were identified by aligning the reads the reference transcriptome with Kallisto v. 0.43.0 (Bray *et al.*, 2016). Only transcripts with a transcript per million (TMP) value > 5 were considered as expressed. This resulted in 10 reference proteome datasets, one for each depth of sequencing (8M-80M reads), composed by complete proteomes and amino acid sequences of the reference *Chlamydomonas* transcripts found expressed at that depth of sequencing,

Then, for each depth of sequencing, *Chlamydomonas de novo* assembled eukaryotic and green transcripts that were found expressed were identified and added to the corresponding reference protein databases. The corresponding amino acid sequences of the potential coding regions were deduced with the transeq algorithm present in the EMBOSS package. For each depth of sequencing tested, the deduced peptides were added to corresponding reference proteome database. For each protein database, all-against-all sequence similarity searches were computed with DIAMOND, with 4,000 max target sequences and using the flag --more-sensitive. To define the gene families, the sequence similarity results were then clustered with Tribe-MCL v. 10-201 with the following parameters: -I 2, -scheme 4. For each gene family, the members from the *de novo* assembled transcriptome and the members from reference *Chlamydomonas* transcriptome were counted, and the counts plotted. Only gene families with members both in the transcriptome and in the genome were considered in this analysis.

Effect of partial data on orthology inference

We tested the impact of introducing partial sequences (i.e.: *de novo* assembled transcripts) in the inference of homologous gene families by the PLAZA pipeline (all-against-all sequence similarity search followed by Tribe-MCL clustering). First of all, transcriptomes and genomes for the 55 species reported in Table 2.1 were processed through the pipeline described in this chapter (eukaryotic filtering, frameshift correction). The longest coding frame was detected with a custom java script plugged into TRAPID using four different translation tables (1, 6, 11 and 16), to take into account the alternative nuclear genetic codes (Cocquyt *et al.*, 2010a), as well as the chloroplast

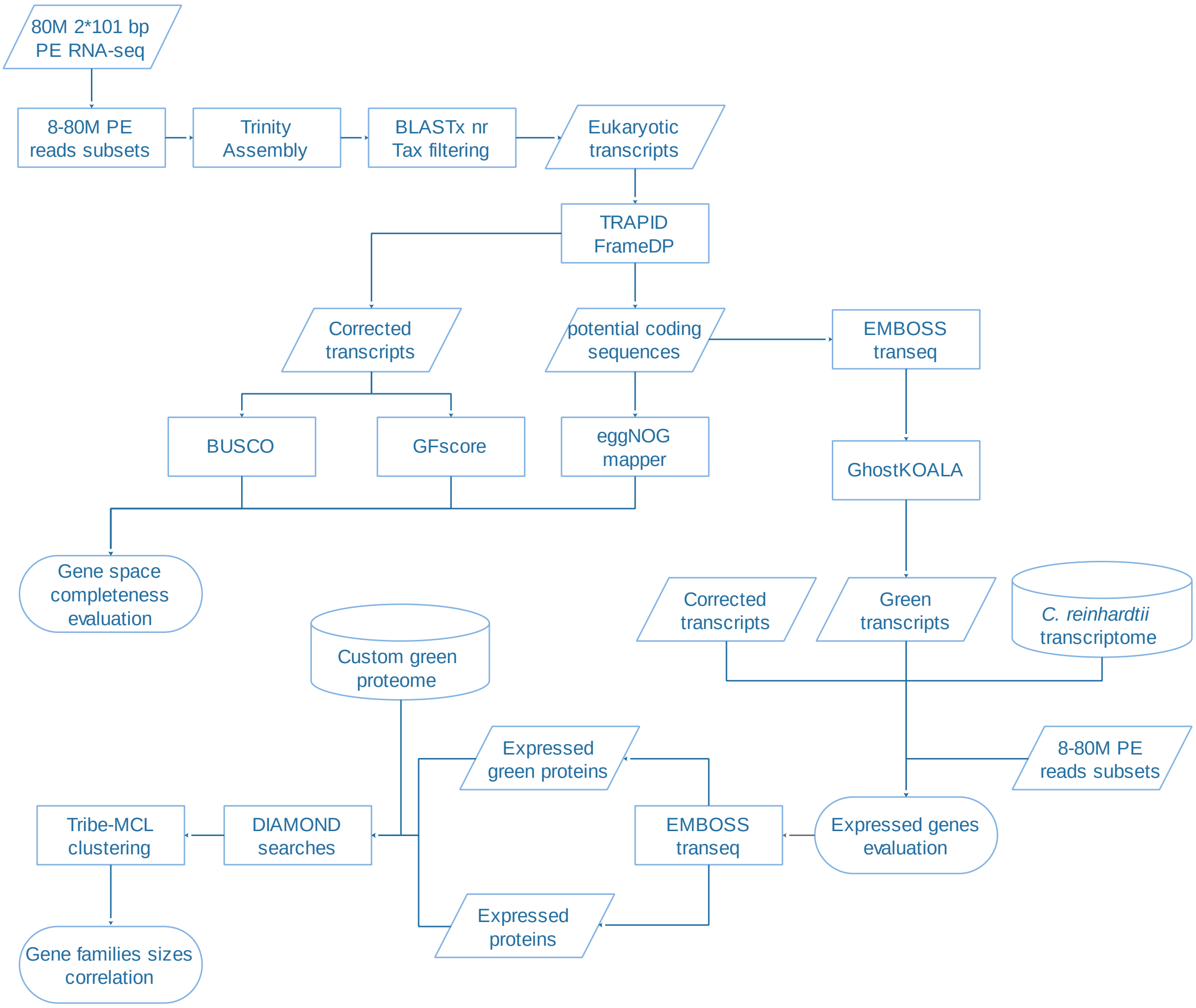


Figure 2.11: Schematic overview of the depth of sequencing evaluation workflow.

Detailed information on the workflow steps are reported on the text.

and mitochondrial translation tables. Briefly, for each transcript coding frames were predicted with each of the four translation tables. Then, for each transcript, the longest coding sequences detected was retained for the downstream analysis, and the corresponding translation table recorded. Concordance with orthologous sequences predicted by the TRAPID pipeline by sequence similarity searches was also taken into account. Each transcript was translated into the corresponding amino acid sequences with the transeq algorithm from the EMBOSS package, using the appropriate translation table. The resulting predicted peptides were processed with an in-house instance of PLAZA4.0 (Van Bel *et al.*, 2018), named ChloroPLAZA.

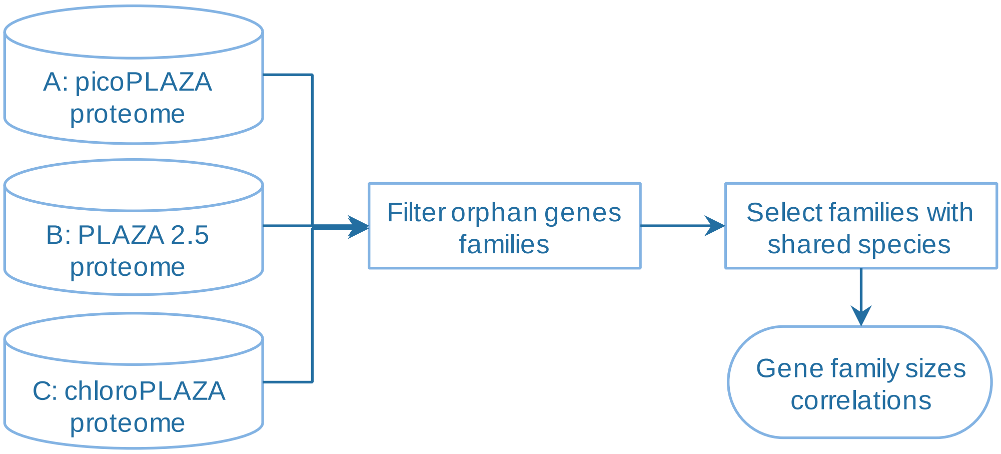

A: picoPLAZA proteome
B: PLAZA 2.5 proteome
C: chloroPLAZA proteome
Filter orphan genes families
Select families with shared species
Gene family sizes correlations

Figure 2.12: Schematic overview of the effect of partial data on orthology inference workflow.

Detailed information on the workflow steps are reported on the text.

Slim versions of picoPLAZA (build A), PLAZA2.5 (build B) and ChloroPLAZA (build C) protein databases were created, where HOM gene families containing orphan genes were removed. picoPLAZA and PLAZA2.5 (Build A and Build B, respectively) constitutes reference datasets constructed from genomic data, while ChloroPLAZA (Build C) represents the dataset composed by majority of transcriptomic data to test.

For the analysis, only HOM families containing one or more genes from *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Micromonas pusilla*, *Ostreococcus tauri*, *Physcomitrella patens* or *Volvox carteri* were selected, since for these species the genome versions to build the databases were identical and gene identifiers were conserved between builds. As a positive control, build A was tested against build B, and vice versa. To test the effect of partial data on orthology inference, build A and build B were independently tested against build C (Figure 2.12). For each of the selected HOM families, each gene from *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Micromonas pusilla*, *Ostreococcus tauri*, *Physcomitrella patens* or *Volvox carteri* of a dataset was searched and identified in the HOM families of the test dataset and the number of HOM families retrieved is recorder. The resulting rate of correlation for each HOM family was calculated as: 1/(number of Hit HOM identified), where 1 is the highest score and indicate a perfect 1 to 1 correlation, while values between 0 and 1 indicate that genes from a HOM family were found in two or more HOM families in the test dataset.

Acknowledgements

I am grateful to Shu-Min Kao for sharing the scripts to run the BUSCO analyses and for providing the *Ulva mutabilis* proteome ahead of publication; to François Bucchini for writing the TRAPID java plug-in to determine the longest reading frame based on multiple translation tables; and to Prof. Diedrik Menzel for providing *Acetabularia acetabulum* cultures.

Supplementary Information

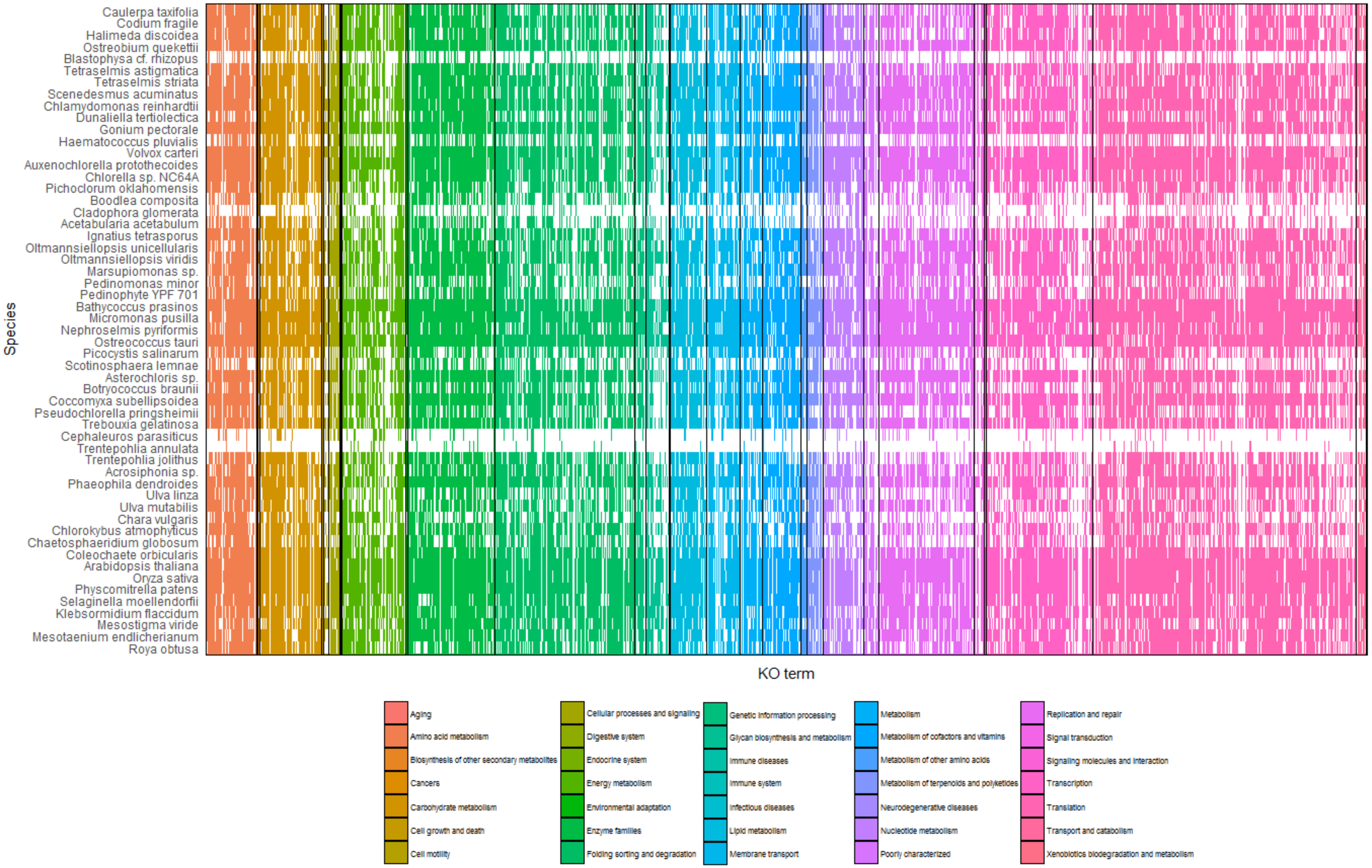


Figure S2.1: Gene space completeness evaluated with eggNOG mapper.

KO terms associated with chloroNOG clusters identified for the species Table 2.2. KO terms were grouped based on their function.

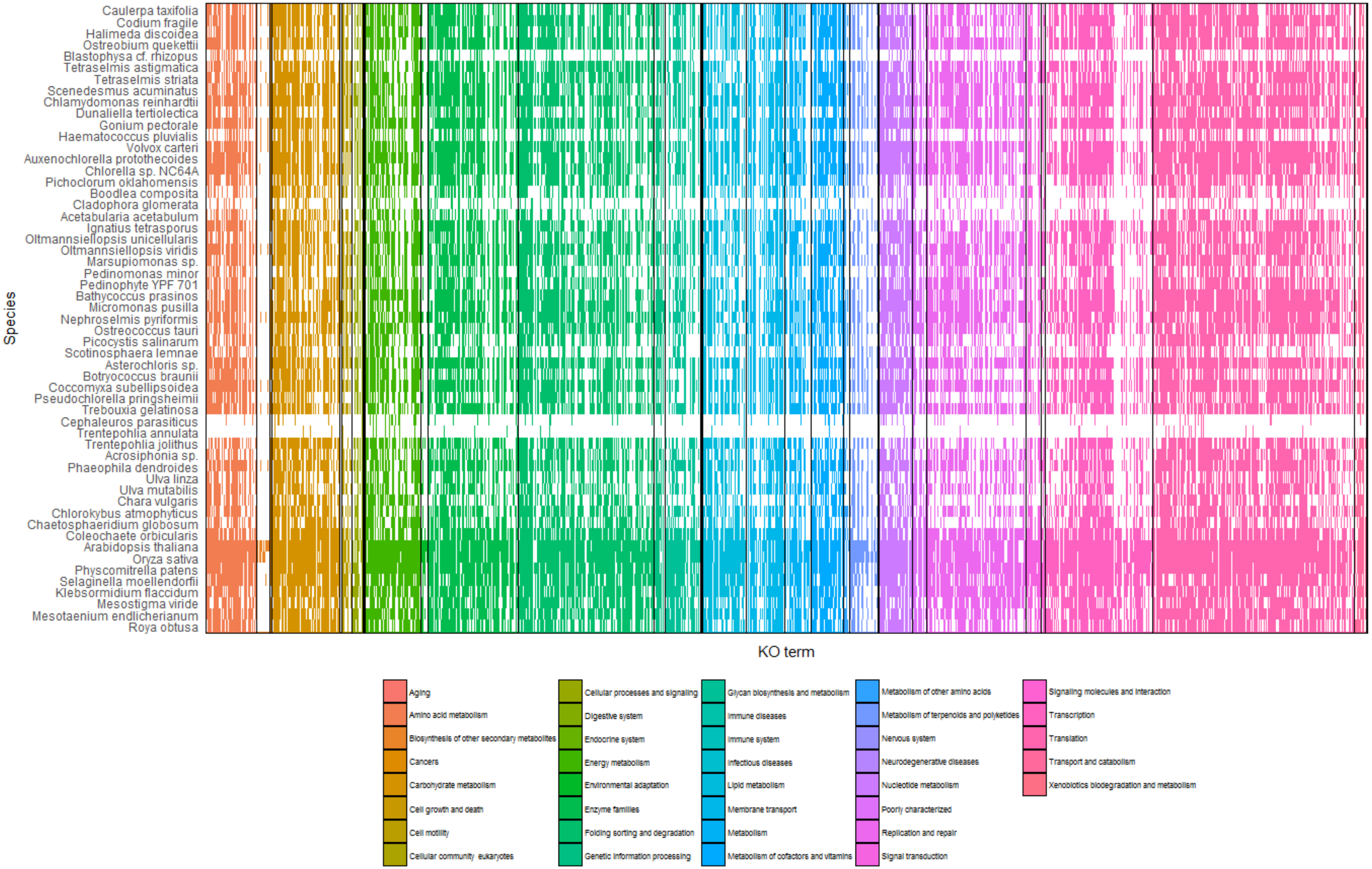


Figure S2.2: Gene space completeness evaluated with eggNOG mapper.

KO terms associated with virNOG clusters identified for the species Table 2.2. KO terms were grouped based on their function.

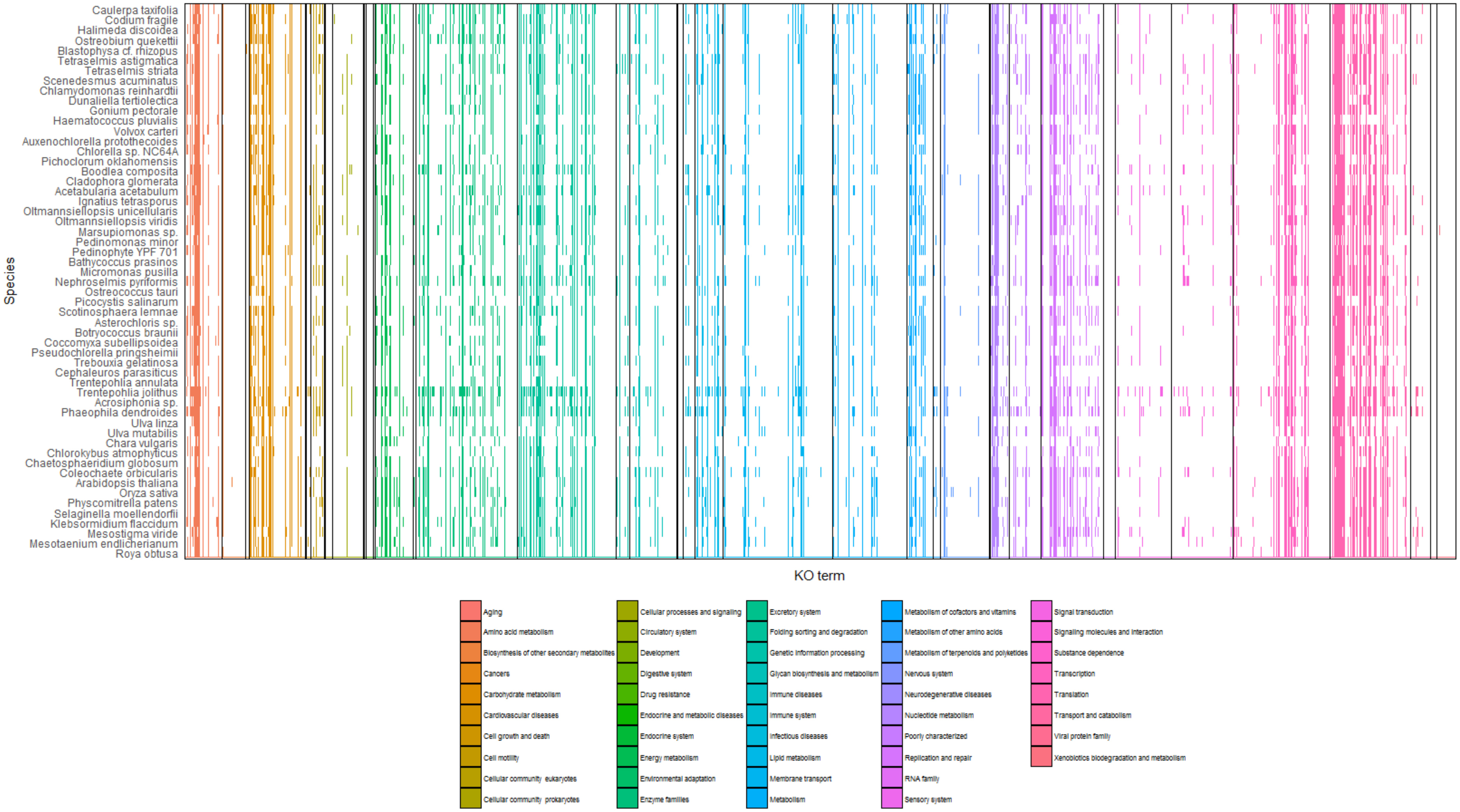


Figure S2.3: Gene space completeness evaluated with eggNOG mapper.

KO terms identified for the species Table 2.2. KO terms were grouped based on their function.

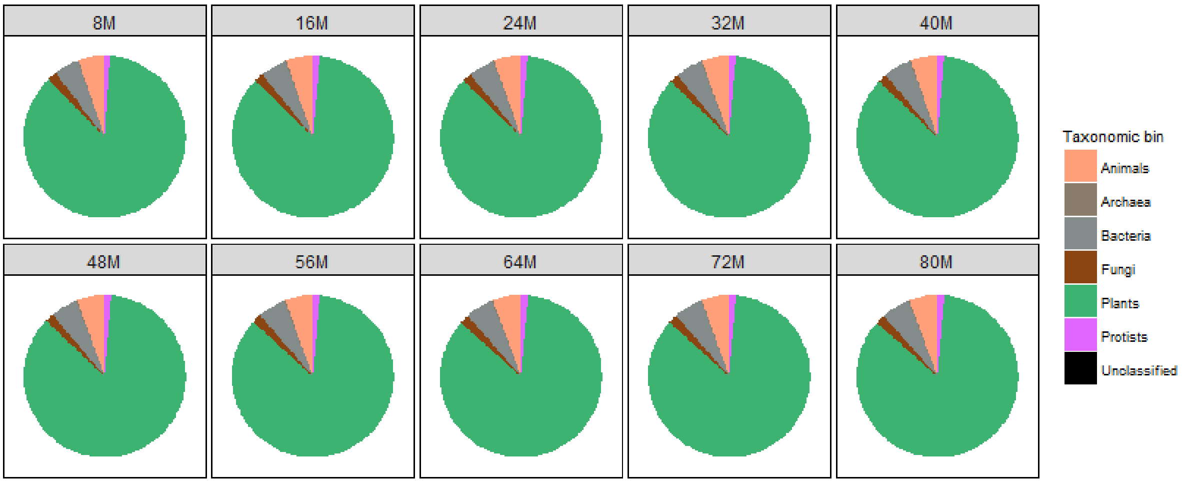


Figure S2.4: Taxonomic binning as assigned by GhostKOALA.

Pie charts report the proportion of the taxonomic bin assigned to transcriptome assembly. These results are in contrast with the GhostKOALA output for *Chlamydomonas* genomic CDSs (Figure 2.1).

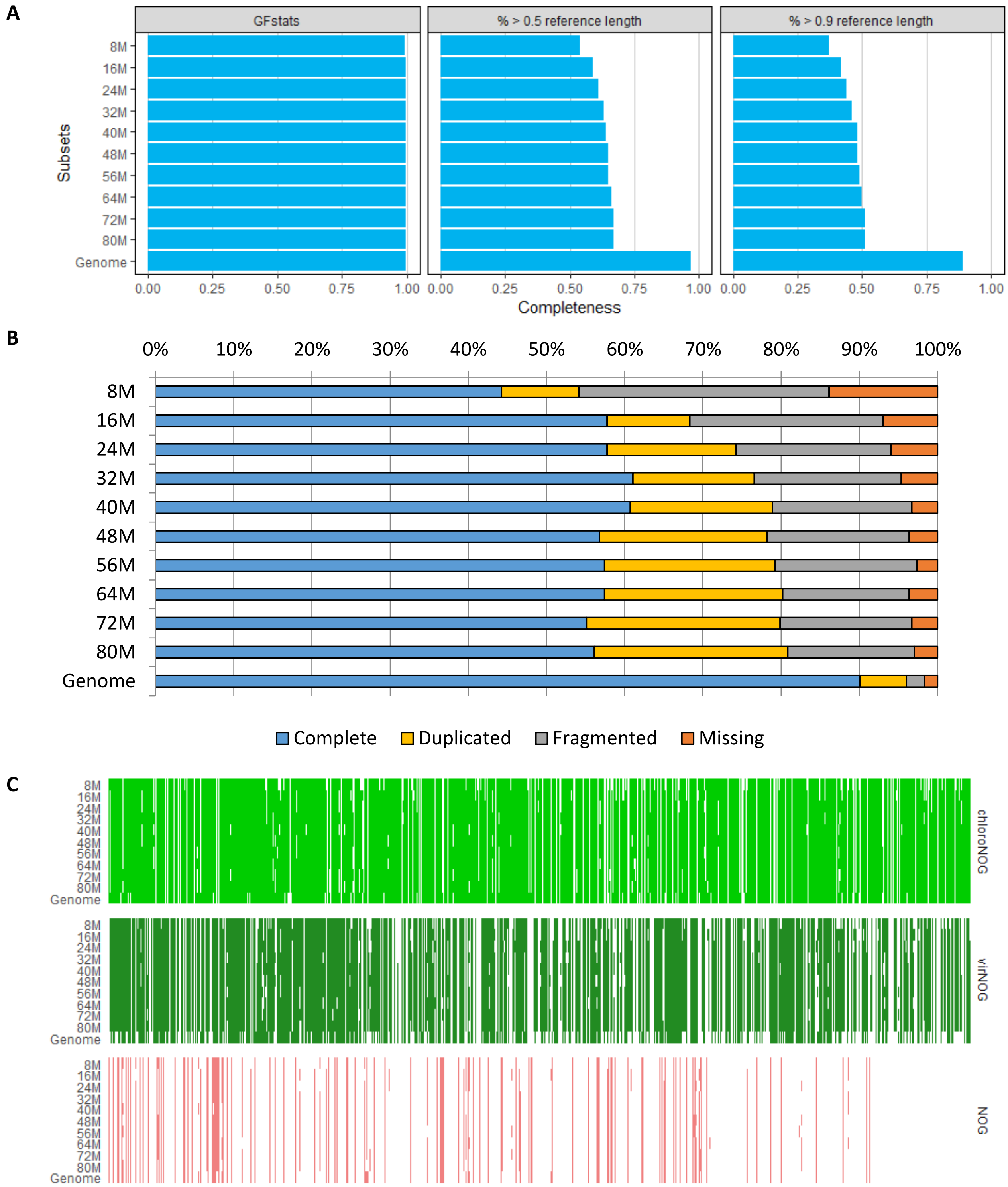


Figure S2.5: Depth of sequencing and gene space completeness evaluation.

(A) Bar plot illustrating the GFscore and the corresponding Reference coverage scores (50% and 90% of reference genes covered) respectively. (B) The percentage of the BUSCO genes identified as complete, duplicated, fragmented or missing in the assemblies and in the reference genome of the total 303 BUSCO eukaryotic single-gene orthologs is reported. (C) KO terms associated with chloroNOG, virNOG, NOG clusters identified at each depth of sequencing.

Table S2.1: Overview of the assembly metrics.

		Assembly metrics						
		before CD-HIT		after CD-HIT		euk fraction		
subset	**# PE reads**	**# seq**	**N50 (bp)**	**# seq**	**N50 (bp)**	**# euk**	**%euk**	**N50 euk (bp)**
8M	8,066,012	72,378	827	50,350	770	32,091	63.74	939
16M	16,132,024	92,565	951	62,495	909	40,593	64.95	1,112
24M	24,198,036	106,092	1,020	70,465	984	45,691	64.84	1,225
32M	32,264,048	116,322	1,066	76,399	1,041	49,328	64.57	1,287
40M	40,330,060	124,202	1,093	81,202	1,081	52,103	64.16	1,334
48M	48,396,072	131,244	1,124	85,118	1,122	54,213	63.69	1,391
56M	56,462,084	136,644	1,151	88,385	1,155	56,159	63.54	1,433
64M	64,528,096	142,075	1,180	91,436	1,189	57,764	63.17	1,471
72M	72,594,108	146,736	1,196	94,338	1,220	59,526	63.10	1,502
80M	80,660,120	150,504	1,206	96,490	1,233	60,694	62.90	1,525

before CD-HIT: transcriptome metrics before sequence clustering with CD-HIT EST
after CD-HIT: transcriptome metrics after sequence clustering with CD-HIT EST
euk fraction: metrics of the eukaryotic fraction
% euk: percentage of transcripts (after CD-HIT) identified as eukaryotic

Chapter 3 - Reconstruction of the early diversification of green seaweeds[4]

Andrea Del Cortona, François Bucchini, Chris Jackson, Michiel Van Bel, Endymion Cooper, Pavel Škaloud, Sofie D’hondt, Heroen Verbruggen, Charles Delwiche, Frederik Leliaert*, Klaas Vandepoele*, Olivier De Clerck[5]

“Neurotrasmettitori sinapsi elettrochimiche

catene sequenziali di acidi nucleici mi parlano di te”

Subsonica - Perfezione

[4] Manuscript in preparation:
Andrea Del Cortona, François Bucchini, Chris Jackson, Michiel Van Bel, Endymion Cooper, Pavel Škaloud, Sofie D’hondt, Heroen Verbruggen, Charles Delwiche, Frederik Leliaert*, Klaas Vandepoele*, Olivier De Clerck* – Reconstruction of the early diversification of green seaweeds
* equal contribution.

[5] Authors contribution: A.D.C., F.L., H.V., K.V., O.D.C.: study design; A.D.C., C.D., C.J., E.C., H.V., K.V., O.D.C., P.S., S.D.H.: data generation; A.D.C., F.L.: data analysis; F.B., M.V.B.: TRAPID plugin and PLAZA4.0 build; A.D.C., C.J.: time-calibrated phylogenetic analysis; A.D.C., F.L., K.V., O.D.C.: manuscript conceptualization, drafting and writing.

Abstract

The Neoproterozoic was marked by a change from a largely bacterial to a eukaryotic phototrophic world, thereby creating the foundation for complex benthic ecosystems which sustained the Precambrian radiation of Metazoa. Until recently, we were largely oblivious about the timing and speed of this transition. This study focusses on the green algal lineage, which have been a dominant group of photosynthetic eukaryotes in marine, freshwater and terrestrial habitats for millions of years. By applying a phylotranscriptomic approach we resolve important relationships and unveil a rapid radiation of the main green algal clades. The green seaweeds, a dominant group of mainly macroscopic primary producers in coastal environments, likely originated during the Cryogenian. We hypothesize that the ancestors of green seaweeds were benthic unicellular algae, which survived the global glaciation during the Cryogenian in isolated refugia. Broader marine photic habitats becoming available due to retreating sea ice, in combination with an increased supply of nutrients (phosphate weathering), may have favored the colonization of benthic environments. Increased cells sizes, macroscopic growth and compartmentalization of the ancestors of green seaweeds resulted as mitigation strategy to the pressure applied by grazers and to the competition with other benthic phototrophs in the novel Ediacaran environment. A rapid radiation event and parallel evolution in isolated refugia supports earlier hypotheses that the different green seaweed lineages evolved macroscopic growth independently using different mechanisms, ranging from canonical multicellularity over coenocytic cells to complex siphonous thalli.

Introduction

The diversification of green seaweeds (Ulvophyceae) in coastal environments coincided with the evolution of an astonishing diversity of thallus forms, ranging from microscopic unicellular organisms to seaweeds, macroscopic multicellular and giant-celled algae, with highly specialized cellular and physiological characteristics (Chapter 1). However, when and how many times green seaweeds have emerged has been a matter of debate.

Understanding the origin and ecological diversification of green seaweeds requires a well-resolved phylogeny, and a reliable estimate of their age. However, resolving the phylogenetic relationships of Ulvophyceae and other clades of the core Chlorophyta has been a difficult task. Early phylogenetic studies based on ultrastructural features, which have been instrumental to define higher level groupings of green algae, such as the fine structure of cytokinesis, flagellar apparatus, and mitosis, have been inconclusive. As a result, the monophyly of the Ulvophyceae has been questioned because of the absence of shared derived characters (Leliaert *et al.*, 2012). Molecular phylogenetic studies were promising initially, but nuclear and chloroplast gene data often yielded ambivalent and often contradicting results. Molecular phylogenetic analyses based on nuclear ribosomal DNA recovered a monophyletic Ulvophyceae but with low phylogenetic support (Watanabe & Nakayama, 2007; Leliaert *et al.*, 2009). These analyses recovered two distinct clades of Ulvophyceae: the Ulvales-Ulotrichales clade (UU clade) and a clade consisting of Trentepohliales, Bryopsidales, Cladophorales, and Dasycladales (TBCD clade), along with some unicellular clades of uncertain affinity, such as Oltmannsiellopsidales, Scotinosphaerales and Ignatiales (Cocquyt *et al.*, 2009; Škaloud *et al.*, 2013). A phylogenetic study based on 10 genes (eight nuclear and two plastid genes) recovered the Ulvophyceae as a well-supported monophyletic group for the first time, divided into the UU and TBCD clades (Cocquyt *et al.*, 2010a). The inferred topology suggested that within the class, macroscopic growth may have originated at least four times independently from marine unicellular ancestors. Conversely, chloroplast multigene analyses do generally not support monophyly of the Ulvophyceae, but instead indicate two or more separate ulvophyte lineages within the core Chlorophyta, indicating that green seaweeds may have evolved multiple times independently from freshwater progenitors (Fučíková *et al.*, 2014; Leliaert & Lopez-Bautista, 2015; Turmel *et al.*, 2017; Fang *et al.*, 2018). The relationships among the main core chlorophyte clades, however, were poorly supported, hampering inference of ecological and morphological diversification.

In this study, we investigated the evolutionary relationships and divergence times among green algae using a phylotranscriptomic approach. We constructed a dataset of 1,877,542 aligned sites from 539 protein-coding nuclear genes for 55 species. For 14 species, newly generated transcriptomes were analyzed. Integration with 26 publicly available transcriptomes and 15 genomes yielded a representative dataset, including the major clades of Chlorophyta and Streptophyta. We focused on the core Chlorophyta and Ulvophyceae in particular, for which we included 19 species, representing all major clades and the five distinct cyto-morphotypes. These data were analyzed using different data-filtering approaches, complementary phylogenetic methods, and rigorous statistical tests to provide the best supported phylogenetic hypotheses of the core Chlorophyta yet produced. Our results indicate an early diversification of the core Chlorophyta, and a rapid radiation of ulvophyte -lineages in the late Neoproterozoic (750-650 mya). Unexpected but coherent and highly supported relationships were recovered, such as the position of the Bryopsidales as a sister clade to the other ulvophyceans. Finally, our phylogeny offers a solid framework to understand the evolution of genomic features, such as an alternative nuclear genetic code, and unconventional chloroplast genomes (Cocquyt *et al.*, 2009; Cocquyt *et al.*, 2010a; de Vries *et al.*, 2013; Del Cortona *et al.*, 2017).

Results

Transcriptome data generation, gene family identification and phylogenetic analysis

We collected and analyzed nuclear encoded protein-coding genes from 55 species mined from 15 genomes and 40 transcriptomes. We included representatives from the major clades of Streptophyta and Chlorophyta, while keeping a focus on ulvophyceans, for which all major clades were sampled, representing the five cyto-morphotypes and three different environments (Table S2.2). This resulted in the retrieval of 1,228,821 protein-coding genes that were clustered into gene families (see Materials and Methods).

A set of 539 high-confidence single-copy gene families was identified (hereafter referred to as coreGF), based on single-copy gene families present in picoPLAZA green algal genomes (Vandepoele *et al.*, 2013). Furthermore, a subset of 355 gene families was selected, where, for each gene family, at least one species for each of the 9 main ulvophyte clades was present (i.e. Bryopsidales, Cladophorales, Dasycladales, Ignatiales, Oltmannsiellopsidales, Scotinosphaerales, Trentepohliales, Ulotrichales and Ulvales, hereafter referred to as ulvoGF). Partial sequences from the transcriptomes were either scaffolded (scaffolded dataset) or removed (unscaffolded dataset), resulting in a more comprehensive and a more conservative version of the coreGF and ulvoGF datasets, respectively (Figure S3.1, S3.2). Poorly aligned regions were removed to obtain the corresponding trimmed datasets. These operations resulted in eight single-copy gene datasets (Table S3.1): coreGF unscaffold, composed by 539 single-copy genes, with partial sequences removed; coreGF scaffold, composed by 539 single-copy genes, with partial sequences scaffolded; coreGF unscaffold TRIM, as coreGF unscaffold, but with less conserved regions filtered; coreGF scaffold TRIM: as coreGF scaffold, but with less conserved regions filtered; ulvoGF unscaffold, composed by a 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences removed; ulvoGF scaffold, composed by a 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences scaffolded; ulvoGF unscaffold TRIM, as ulvoGF unscaffold, but with less conserved regions filtered; and ulvoGF scaffold TRIM, as ulvoGF scaffold,

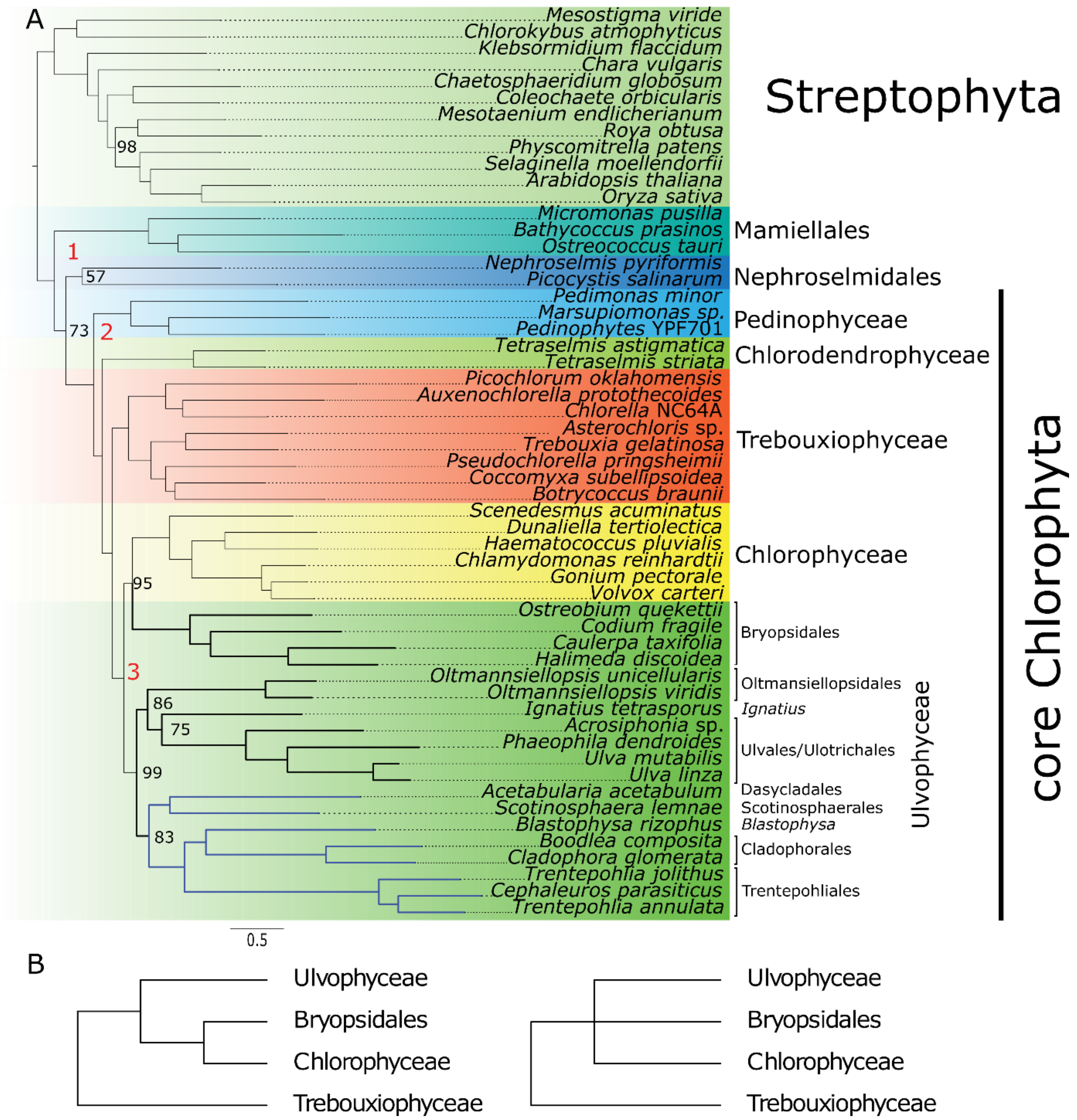


Figure 3.1: Green algal phylogenetic relationships.

(A) ML phylogenetic tree inferred from the concatenated alignment and from amino acid sequences of 539 coreGF scaffolded untrimmed gene trees. Bootstrap values are shown on the nodes. 100 BS support are omitted. The scale indicates substitutions per amino acid position. The red numbers indicate incongruences between the topology recovered by the partitioned ML analysis and the coalescence-based analysis based on the same gene dataset: 1. *Nephroselmis pyriformis* position; 2. Pedinophyceae position; 3. Bryopsidales position. The alternative topology for the Bryopsidales-Chlorophyceae-ulvophyceans relationships inferred with the coalescence-based analysis is summarised on the right hand-side scheme. Blue branches indicate presence of the alternative nuclear genetic code in the clade. (B) Alternative topologies of the Bryopsidales position as recovered by the partitioned ML analysis (on the left) and the coalescence-based analysis (on the right).

but with less conserved regions filtered (Table S3.1). We analyzed these eight datasets with complementary phylogenetic methods and tested the significance of conflicting topologies to assess the robustness of our findings.

Establishing a resolved phylogeny of Chlorophyta

Maximum likelihood (ML) supermatrix analyses of the eight datasets recovered a solid phylogenetic reconstruction of Viridiplantae (Figure 3.1). The inferred relationships among the main clades of streptophytes are in agreement with published phylogenies (Zhong *et al.*, 2013; Wickett *et al.*, 2014; Puttick *et al.*, 2018), denoting the power of our approach to confidently resolve difficult phylogenetic relationships. Several inferred relationships among the main clades of Chlorophyta were highly supported and largely congruent over different analyses and datasets, while other relationships were less stable and dependant on the dataset or analysis. We recovered Chlorodendrophyceae and Pedinophyceae as two early diverging clades at the base of the core Chlorophyta (Figure 3.1). Trebouxiophyceae and Chlorophyceae were recovered as monophyletic groups, with the Trebouxiophyceae consisting of two distinct clades, the Chlorellales and core Trebouxiophyceae.

While the supermatrix and the coalescence-based analyses confirmed monophyly of Chlorophyceae, a sister relationship between Chlorophyceae and Bryopsidales in the supermatrix analyses rendered the Ulvophyceae paraphyletic. The coalescent-based analyses, on the other hand, inferred the Ulvophyceae as sister to the Chlorophyceae, but the branch leading to the Ulvophyceae clade was extremely short. A polytomy test (Sayyari & Mirarab, 2018) could not reject the null hypothesis of a hard polytomy (Figure S3.3). Under increasing gene numbers, p-values, which indicated the ability to reject the null hypothesis of a hard polytomy, did not show any tendency to become lower. The supermatrix and coalescence-based analyses supported the same relationships among the remaining orders of Ulvophyceae. Two major clades were recovered, one clade including the Oltmannsiellopsidales, Ignatiales and Ulvales-Ulotrichales, and a second clade including the Dasycladales, Scotinosphaerales, Trentepohliales, Cladophorales and *Blastophysa*.

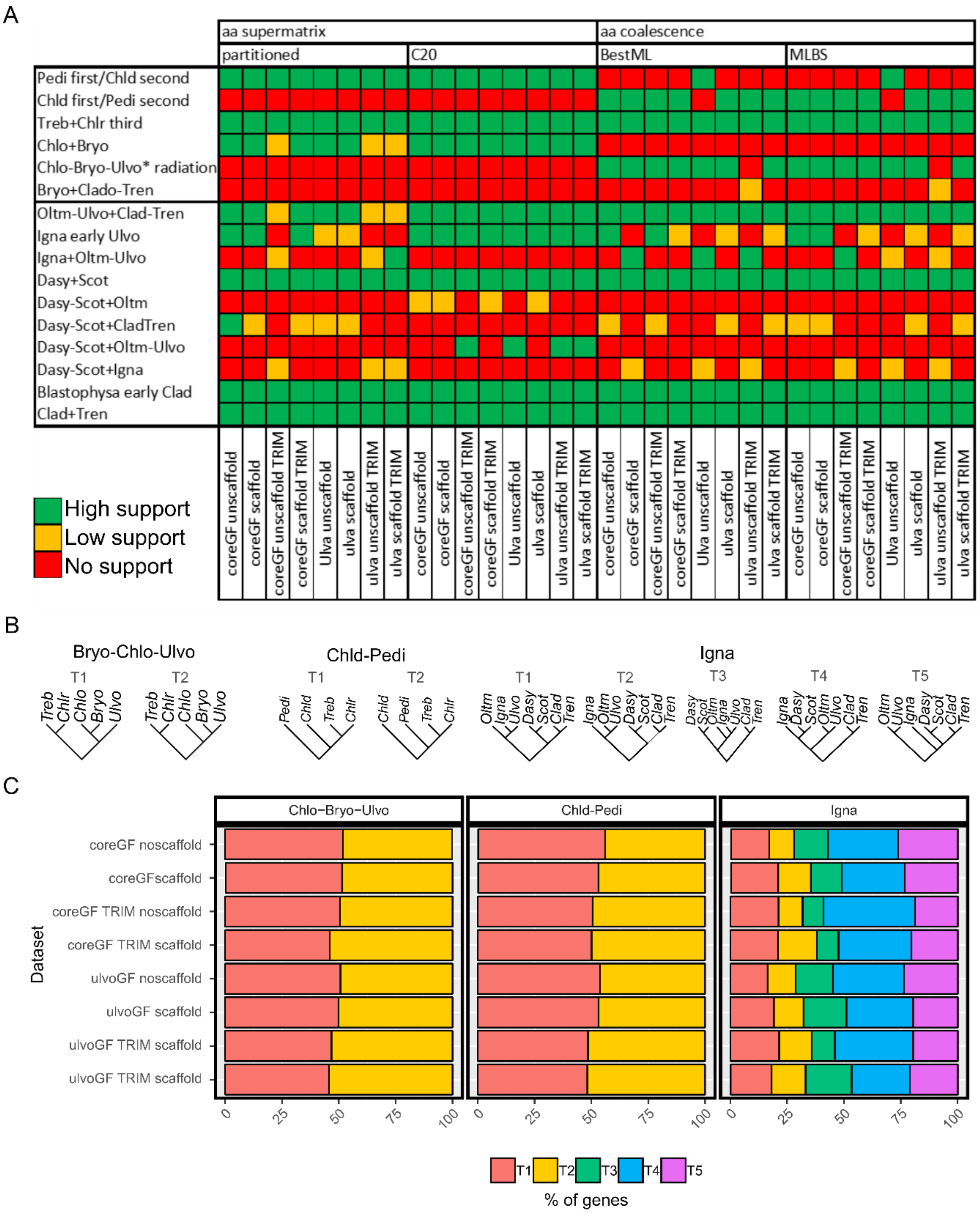


Figure 3.2: Tests of alternative topologies.

(A) Summary of support for hypotheses of Chlorophyta relationships, based on the amino acid alignments of 8 single-copy gene datasets from 55 Chlorophyta and Streptophyta species, across 32 distinct supermatrix or coalescence-based analyses. CoreGF unscaffold: 539 single-copy genes, with partial sequences removed. CoreGF scaffold: 539 single-copy genes, with partial sequences scaffolded. CoreGF unscaffold TRIM: as coreGF unscaffold, but with less conserved regions filtered. CoreGF scaffold TRIM: as coreGF scaffold, but with less conserved regions filtered. UlvoGF unscaffold: 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences removed. UlvoGF scaffold: 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences scaffolded. UlvoGF unscaffold TRIM: as ulvoGF unscaffold, but with less conserved regions filtered. ulvoGF scaffold TRIM: as ulvoGF scaffold, but with less conserved regions filtered. Partitioned: gene-wise partition of the supermatrix, with substitution model for each partition inferred by the gene-tree best model, allowing invariable sites and free rate of heterogeneity across sites. C20: supermatrix analyses with empirical mixture substitution model of amino acids, allowing invariable sites and free rate of heterogeneity across sites. BestML: coalescence-based analysis. MLBS: as BestML, but using the Multi-Locus Bootstrap Support approach. Green: strong support. Yellow: low support. Red: no support. Strong support refers to bootstrap values or posterior probabilities > 75% or 0.75, respectively, for the relationships depicted. (B) Summary of alternative constrained topologies tested with AU-test (Table S3.2, S3.3) and pairwise log-likelihood scores comparisons. (C) Proportion of genes supporting each of the alternative hypotheses in pairwise log-likelihood score comparisons. Bryo: Bryopsidales; Chld: Chlorodendrophyceae; Chlo: Chlorophyceae; Chlr: Chlorellales; Clad: Cladophorales+*Blastophysa*; Dasy: Dasycladales; Igna: Ignatiales; Oltm: Oltmannsiellopsidales; Pedi: Pedinophyceae; Scot: Scotinosphaerales; Tren: Trentepohliales; Ulvo: Ulvales-Ulotrichales.

For a number of specific relationships, differences were observed in the trees estimated by the different methods and datasets (Figure 3.1, 3.2A). The topologies of the supermatrix and the coalescence-based analyses consistently differed in the position of the earliest diverging core Chlorophyta (Pedinophyceae first or Chlorodendrophyceae first, respectively). Constraining the Chlorodendrophyceae as the first diverging core Chlorophyta was rejected by all but the trimmed datasets (AU-test, n = 100,000 for each dataset, Δ-likelihood = 13.766 - 66.129, p-value = 0.0971 - 0.2115). This indicates that the phylogenetic signal for the Pedinophyceae first topology mainly resides in the most variable sites of the alignments (Figure 3.2B, Table S3.2). We further characterized the proportion of genes for each dataset supporting these two alternative hypotheses (Shen *et al.*, 2017). The analysis of the gene-wise log-likelihood showed an equal number of genes supporting the two alternative topologies for all the datasets (Figure 3.2C). As for the core Chlorophyta early diversification and the position of Bryopsidales, AU-test for the trimmed datasets could not reject the topology constrained to conform by the coalescence-based analyses (Ulvophyceae monophyletic) (n = 100,000 for each dataset, Δ-likelihood = 1.132 - 49.031, p-value = 0.2526 - 0.5049, Table S3.2). Slightly more than 50% of the genes supported the monophyly of ulvophyceans for 7 of the 8 datasets (Figure 3.2C). The support for the *Ignatius* position was tested with a similar approach (Figure 3.2C, Table S3.3). The inclusion of *Ignatius* in fact, caused instability in the analyses. Removing it led to an overall higher support for the position of the Oltmannsiellopsidales-Ulvales-Ulotrichales and of Dasycladales-Scotinosphaerales clades (Figure S3.4). Alternative topologies recovered by including Ignatius in the phylogenetic analyses were evaluated with an AU-test, which failed to reject any alternative topology as significantly worse (Table S3.3). Despite that most genes (20-30%) supported a sister relationship between *Ignatius* and Dasycladales-Scotinosphaerales clade (Figure 3.2C), this topology was never recovered with high support by any of the phylogenetic analyses we performed (Figure 3.2A, Figure S3.6).

Gene-wise contribution to the phylogenetic signal was also evaluated by clustering of the gene trees based on their pairwise Jensen-Shannon distances. We performed a cluster-wise supermatrix and coalescence-based analyses, and we identified outlier clusters that had the most divergent signals and excluded or included them in the analyses (Figure S3.4). Different clusters of genes supported the different contrasting topologies for the earliest diverging core Chlorophyta clade and for the monophyly of Ulvophyceae in both the supermatrix and the coalescence-based analyses (Figure S3.5). Removing the outlier clusters increased the overall support of the topology inferred from the remaining clusters, but supermatrix and coalescence analyses did not converge to the same topology. These results altogether suggest that a robust topology of the core Chlorophyta was recovered, except for few equally likely uncertainties.

Macroscopic growth is associated with a transition to benthic marine environments 750-650 mya

To estimate a time-frame of diversification of the core Chlorophyta, we inferred a chronogram from 10 of the most clock-like genes, calculated using relaxed clock analyses with SortDate (Smith *et al.*, 2018). Calibration nodes were derived from fossil information and node age estimates from previous studies, and we also tested the inclusion or the exclusion of contentious fossils (Table S3.4). Analyses were repeated for both the ML supermatrix and the coalescence-based topologies and different molecular clock models (Table S3.5). Results indicate that the ancestor of the core Chlorophyta emerged during the Neoproterozoic, 900-800 mya (Figure 3.3, Table S3.5). The main Ulvophycean seaweeds clades (Bryopsidales, Cladophorales, Dasycladales, Ulvales) radiated during a time-window (750-650 mya) corresponding to the long-lasting global-scale Sturtian and Marinoan glaciations.

Estimated ancestral states of key ecological and cyto-morphological traits are illustrated in Figure 3.4. The core Chlorophyta likely originated in freshwater environments, with early and independent transitions to marine environments in some lineages of Chlorodendrophyceae and Pedinophyceae. The Trebouxiophyceae and Chlorophyceae diversified in freshwater environments mainly. The radiation event at the base of the diversification of Chlorophyceae, Bryopsidales and Ulvophyceae originated a major switch back to marine environments early in the evolution of Bryopsidales and Ulvophyceae, populating coastal environments with marine benthic green seaweeds.

Discussion

We present a robust, well supported phylogenetic reconstruction of the core Chlorophyta (Figure 3.1) and contextualize the relationships in light of the evolution of multicellularity and macroscopic growth, and the transition to marine benthic environments. While supermatrix analyses supported a scenario whereby Chlorophyceae are sister to the Bryopsidales, coalescence-based suggested a hard polytomy of Ulvophyceae (incl. Bryopsidales) and Chlorophyceae. Relationships within the core Chlorophyta have been notoriously difficult to resolve because of the antiquity of the main lineages, and the rapidity of the early evolutionary radiations (Cocquyt *et al.*, 2010b; Leliaert *et al.*, 2012; Marin, 2012; Turmel *et al.*, 2017; Jackson *et al.*, 2018). Pedinophyceae and Chlorodendrophyceae are the earliest diverging core Chlorophyta, followed by Trebouxiophyceae, Chlorophyceae and Ulvophyceae (UTC clade). The Trebouxiophyceae and Chlorophyceae were unambiguously recovered as monophyletic groups in all analyses, with the Trebouxiophyceae sister to a clade containing the Chlorophyceae and Ulvophyceae. Monophyly of the Ulvophyceae was less clear as different analyses yielded different results.

Incongruences between supermatrix and coalescence-based analyses, at the base of core Chlorophyta and at the radiation of Ulvophyceae, could be ascribed to the equal likelihood of the alternative topologies, the short time frame of the diversification

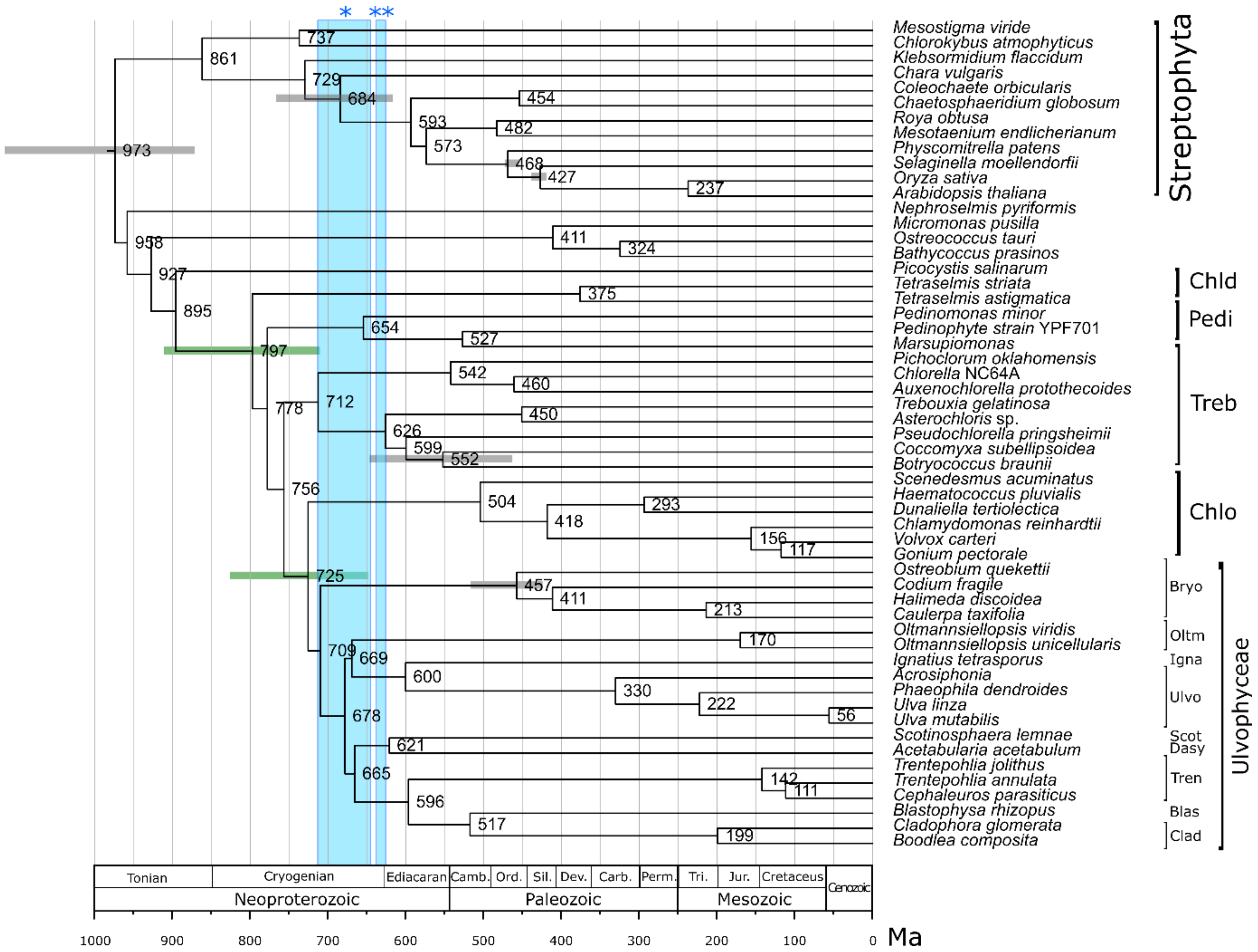

*
**
737
861
729
684
973
454
593
482
573
468
427
237
958
411
324
927
895
375
654
527
797
542
460
712
778
450
626
599
552
756
504
293
418
156
117
725
457
411
213
709
170
669
600
330
222
56
678
621
665
142
111
596
517
199
Mesostigma viride
Chlorokybus atmophyticus
Klebsormidium flaccidum
Chara vulgaris
Coleochaete orbicularis
Chaetosphaeridium globosum
Roya obtusa
Mesotaenium endlicherianum
Physcomitrella patens
Selaginella moellendorfii
Oryza sativa
Arabidopsis thaliana
Nephroselmis pyriformis
Micromonas pusilla
Ostreococcus tauri
Bathycoccus prasinos
Picocystis salinarum
Tetraselmis striata
Tetraselmis astigmatica
Pedinomonas minor
Pedinophyte strain YPF701
Marsupiomonas
Pichoclorum oklahomensis
Chlorella NC64A
Auxenochlorella protothecoides
Trebouxia gelatinosa
Asterochloris sp.
Pseudochlorella pringsheimii
Coccomyxa subellipsoidea
Botryococcus braunii
Scenedesmus acuminatus
Haematococcus pluvialis
Dunaliella tertiolectica
Chlamydomonas reinhardtii
Volvox carteri
Gonium pectorale
Ostreobium quekettii
Codium fragile
Halimeda discoidea
Caulerpa taxifolia
Oltmannsiellopsis viridis
Oltmannsiellopsis unicellularis
Ignatius tetrasporus
Acrosiphonia
Phaeophila dendroides
Ulva linza
Ulva mutabilis
Scotinosphaera lemnae
Acetabularia acetabulum
Trentepohlia jolithus
Trentepohlia annulata
Cephaleuros parasiticus
Blastophysa rhizopus
Cladophora glomerata
Boodlea composita
Streptophyta
Chld
Pedi
Treb
Chlo
Bryo
Oltm
Igna
Ulvo
Scot
Dasy
Tren
Blas
Clad
Ulvophyceae
Tonian
Cryogenian
Ediacaran
Camb.
Ord.
Sil.
Dev.
Carb.
Perm.
Tri.
Jur.
Cretaceus
Cenozoic
Neoproterozoic
Paleozoic
Mesozoic
1000
900
800
700
600
500
400
300
200
100
0
Ma

Figure 3.3: Time calibrated ultrametric tree.

Chronogram based on the 10 most clock-like genes from the scaffold trim dataset (Table S3.5). Node ages were inferred starting from the coalescence-based analysis tree using Bayesian inference assuming a relaxed molecular clock and node constrains derived from fossil records (see Supplemental Material). Node values indicate average node ages. The grey bars represent 95% confidence interval (CI) of the calibration nodes. The green bars represent 95% CI for relevant nodes for this study: the core Chlorophyta and the Chlorophyceae – Ulvophyceae diversification. The two blue bars are in correspondence of the Sturtian (*) and the Marinoan (**) glaciations. Blas: *Blastophysa*; Bryo: Bryopsidales; Chld: Chlorodendrophyceae; Chlo: Chlorophyceae; Clad: Cladophorales; Dasy: Dasycladales; Igna: Ignatiales; Oltm: Oltmannsiellopsidales; Pedi: Pedinophyceae; Scot: Scotinosphaerales; Tren: Trentepohliales; Ulvo: Ulvales-Ulotrichales.

events, and possible incomplete lineage sorting (ILS, Figure 3.2, S3.3, S3.4, S3.5). From an ultrastructural point of view, the divergence of the phycoplast-containing Chlorodendrophyceae, after the Pedinophyceae, which lack a phycoplast, would be most parsimonious. This scenario is congruent with previous analyses based on nuclear rRNA operons (Marin, 2012). Chloroplast phylogenomic data recovered contrasting positions for Pedinophyceae: either as sister to Chlorellales (Leliaert & Lopez-Bautista, 2015; Turmel *et al.*, 2016a; Turmel *et al.*, 2017; Jackson *et al.*, 2018) or, similarly to analyses based on nuclear markers, as first diverging among core Chlorophyta (Fučíková *et al.*, 2014; Melton *et al.*, 2015; Sun *et al.*, 2016; Turmel *et al.*, 2016a). The extremely short branches of the Chlorophyceae-Bryopsidales-ulvophyceans in the coalescence-based analyses support the idea of high discordance in the gene trees, rapid radiation and massive ILS among these groups. An intrinsic advantage of coalescence-based analyses is the ability to detect and handle conflicting signals in gene-trees, moreover they perform better than supermatrix analyses when ILS is high, while supermatrix analyses may converge with high confidence on the wrong species-tree (Roch & Steel, 2015; Mirarab *et al.*, 2016; Molloy & Warnow, 2018). Our tests always failed to reject the polytomy null hypothesis for all datasets (Figure S3.3). This could be a consequence of massive sequence convergence, ILS or a real hard polytomy, however, the failure to reject the polytomy hypothesis does not result in automatic acceptance of the multifucartion (Greenland *et al.*, 2016). As the addition of more genes did not help to solve the phylogeny, different types of information would be needed (e.g.: intron position, non-overlapping ultra-conserved elements in the upstream and downstream sequences of genes), and hopefully result in an unambiguous resolution of these phylogenetic relationships (Jarvis *et al.*, 2014).

An unexpected relationship that emerged from our analyses was the Bryopsidales sister to the rest of the Ulvophyceae, including a clade comprising Ignatiales, Oltmannsiellopsidales, Ulvales and Ulotrichales (UU Clade) and a clade composed of Trentepohliales, Cladophorales, Dasycladales, and Scotinosphaerales (TCD clade). The siphonous architecture of the Bryopsidales and Dasycladales would suggest a close relationship between the two clades. However, the phylogenetic position of the Bryopsidales has been difficult to resolve, with several, mainly nuclear rDNA-based studies placing it sister to the Dasycladales (Watanabe & Nakayama, 2007; Leliaert *et al.*, 2009; Cocquyt *et al.*, 2010b), while in plastid genome-based phylogenies Bryopsidales position has been very unstable, being sister to the UU clade or related to other core chlorophytan lineages (Leliaert & Lopez-Bautista, 2015; Turmel *et al.*, 2017; Fang *et al.*, 2018). The presence of a non-canonical nuclear genetic code restricted to the TCD clade (Cladophorales, Dasycladales, Scotinosphaerales, Trentepohliales), where the stop codons TAG and TAA have been reassigned to glutamine (Gile *et al.*, 2009; Cocquyt *et al.*, 2010a), supports the idea of independent evolution towards siphonous organization in Bryopsidales and Dasycladales.

The UU clade was recovered virtually by all studies so far, based on both nuclear and chloroplast markers. In our study, *Ignatius* perturbed the stability of the phylogenetic signal. The position of *Ignatius* as sister clade to Ulvales and Ulotrichales recovered in our analysis is consistent with phylogenies based on chloroplast genomic and 18S rRNA data (Watanabe & Nakayama, 2007; Turmel *et al.*, 2017), but differs from a study based on 8 nuclear and 2 chloroplast markers, where *Ignatius* was recovered as sister to the TBCD clade (Cocquyt *et al.*, 2010b). Chloroplast phylogenomic analyses provided so far only an incomplete picture for the TCD(B) clade, due to the highly deviant chloroplast genes in Cladophorales (Del Cortona *et al.*, 2017) and restricted taxon sampling for Dasycladales, Scotinosphaerales and Trentepohliales.

Despite the large confidence intervals in the time-calibrated analyses, the relationships between Chlorophyceae and Ulvophyceae are better understood when we also consider the time-frame of their divergence. Our relaxed clock analyses indicated that the Ulvophyceae emerged during the Cryogenian, a period characterized by extreme conditions that resulted in at least two global-scale glaciations (Hoffman *et al.*, 1998), the Sturtian glaciation that lasted for 50 million years, from 716 to 659 mya (Macdonald *et al.*, 2010), and the shorter Marinoan glaciation, from 645 to 635 mya (Kennedy *et al.*, 2008; Shields, 2008). The global-scale glaciations resulted in most or all the ocean surface to be frozen for millions of years ("snowball" or "slushball" earth), followed by melting and transition to a "greenhouse" world (Hoffman *et al.*, 1998; Micheels, 2008; Bechstädt *et al.*, 2018). During the global-scale glaciations, photosynthetic eukaryotes, including the ancestors of Chlorophyceae and Ulvophyceae, may have survived in isolated refugia of brackish or fresh water, similar to modern diatoms that proliferated in brine channels in the Arctic and Antarctic ice (van Leeuwe *et al.*, 2018). Despite the wide, 100 million years time window, the environmental conditions may have slowed down diversification. At the end of the Sturtian, and after the Marinoan glaciation (Marinoan Meltdown) (Shields, 2008), green seaweeds could flourish in the marine benthic environments that became available (Rise of the Algae) (Brocks *et al.*, 2017). Macroscopic growth and different cyto-morphotypes likely arose independently, possibly as a response to increased grazing pressure.

Evidences for benthic macroalgae persistence during the Marinoan-glaciation was found in black shales from the Marinoan-age Nantuo Formation in South China (Ye *et al.*, 2015). This hypothesis is supported as well by molecular fossil records, that showed a significant increase in steroid diversity and abundance during the narrow time window between the Sturtian and Marinoan glaciations, a molecular signature of Archaeplastida (Brocks *et al.*, 2017). Similarly, a switch in steroid diversity of the sediments was observed after the Permian-Triassic mass extinction that led to the dominance of algae bearing secondary plastids in the oceans (Grantham & Wakefield, 1988; Brocks *et al.*, 2017). Algal proliferation resulted in more efficient energy transfer and novel, richer food webs, allowing the evolution of larger and more complex organisms. Diversification of new organisms during the early Ediacardan period would then result in the Avalon explosion, 575 million years ago (Shen *et al.*, 2008). Despite the fossil records that indicate the presence of a putative Ulvophycean ancestor during the Neoproterozoic (i.e.: *Proterocladus*) (Butterfield *et al.*, 1994), a more exhaustive picture would suggest a problematic interpretation for *Proterocladus* and the advent of modern seaweeds diversity only after the Cryogenian global-scale glaciations (Knoll *et al.*, 2006).

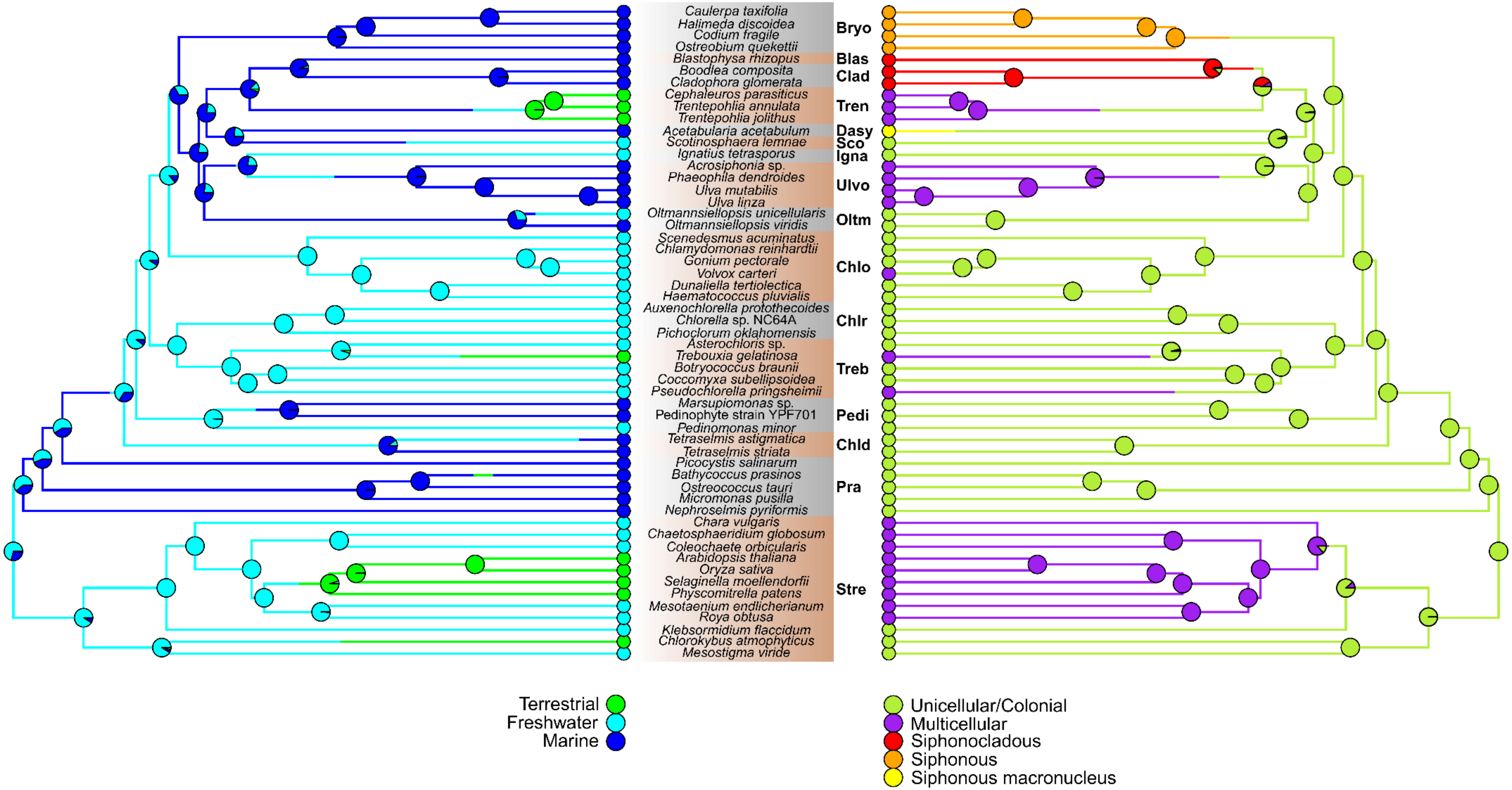


Figure 3.4: Ancestral state estimation.

Ancestral state estimation of environmental (left) and cyto-morphological (right) traits, plotted on the ultrametric tree (Figure 3.3, Table S3.5). Note that, except for Oltmannsiellopsidales, each transition to marine benthic environment coincided with evolution in macroscopic growth in Ulvophyceae.

In this perspective, the alternation of freshwater and marine ulvophyceans fits perfectly, indicating multiple independent events of transitions between marine and freshwater environments, even in a parsimonious scenario. This complex distribution of environments reflects the evolutionary history of Ulvophyceae. Despite the short timeframe for their radiation during and at the end of the Cryogenian glaciations, Ulvophyceae seized multiple opportunities for transition to new ecological niches, and each time, a novel and unique way to macroscopic growth evolved.

Conclusions

Green seaweeds' unique cyto-morphological characteristics represent distinct solutions to macroscopic growth. To gain a better understanding of green seaweed evolution, a comprehensive, large-scale, phylogeny of the core Chlorophyta was generated using the largest gene set to date, mined from whole genome and transcriptome datasets. Our results are consistent with some previous analyses based on chloroplast phylogenomics data (Turmel *et al.*, 2017; Jackson *et al.*, 2018), but shed new light on the relationships between core Chlorophyta lineages. The monophyly of Trebouxiophyceae, Chlorophyceae and Ulvophyceae was confirmed, and within the Ulvophyceae, the relationships between the green seaweeds were largely resolved. The relationships inferred are more conservative than previous hypotheses (Cocquyt *et al.*, 2010b) under several molecular features (for example, the recovery of the monophyly of the alternative nuclear genetic code). Our reconstruction does not require a complex step-wise acquisition and/or loss of shifts in the translational apparatus. Moreover, the analyses support previous hypotheses of independent evolution of macroscopic growth in the different marine benthic ulvophycean clades (Cocquyt *et al.*, 2010b). Each clade solved the quest to macroscopic growth independently, by acquiring unique cyto-morphologies.

Although the concordance between our multiple analyses is not perfect, their evolutionary and biological interpretation is not influenced by the small incongruences observed in our results. Incongruences in some relationships (e.g.: Chlorophyceae, Bryopsidales, ulvophyceans) may be due to massive amount of ILS, extinction and rapid radiation, which impact the supermatrix and the coalescence-based analyses in different ways (Roch & Steel, 2015; Molloy & Warnow, 2018). The rapid radiation and differentiation of Ulvophyceae is consistent with their large differences in fundamental cytological features, such as cellular architecture, nuclear division, ultrastructure of the flagellar base, cell wall composition, and reproduction, which ultimately led van den Hoek elevate different ulvophyte orders to the class level (Van den Hoek *et al.*, 1995). These findings represent a fundamental framework to understand the molecular players behind the different ways of evolving macroscopic growth and multicellularity, the transition from freshwater to marine environments and the makeover of the translational apparatus.

Materials and Methods

Dataset retrieval, RNA extraction and sequencing

For this study, the 15 genomes and 40 transcriptomes datasets reported in Table 2.1 were used. Dataset retrieval, RNA extraction and sequencing was performed as described in Chapter 2, Materials and Methods (Tables 2.3, Materials and Methods).

Transcriptome assembly, frameshift errors correction and ORF detection

Reads trimming, filtering and transcriptome assemblies were performed as indicated in Chapter 2, Materials and Methods. Then, for each of the 40 transcriptomes, transcripts were clustered with CD-HIT-EST v. 4.6.1 (Li & Godzik, 2006) with a similarity cut-off of 97.5%, and only the longest transcript was retained for downstream analysis as representative of the cluster. Taxonomic profiling of the transcripts was performed using the following protocol: first, the transcripts were compared to the NCBI non-redundant protein database by sequence similarity searches, using Tera-BLAST DeCypher (Active Motif, USA); then, for each transcript, sequence similarity searches were combined with the NCBI Taxonomy information of the top ten BLAST hits in order to discriminate between eukaryotic and bacterial transcripts or transcripts lacking similarity to known protein-coding genes. Only eukaryotic transcripts were retained for downstream analysis, bacterial transcripts and transcripts lacking sequence similarity to known proteins were discarded.

Transcripts with putative frameshift errors were identified after initial processing in TRAPID (Van Bel *et al.*, 2013), using the *Chlamydomonas reinhardtii* proteome as reference database for the sequence similarity searches. Transcripts carrying a putative frameshift error were corrected with a local version of FrameDP 1.2.2 (Gouzy *et al.*, 2009), using the *Chlamydomonas* proteome as reference to guide the frameshift correction step. The longest coding frame was detected with a custom java script plugged into TRAPID using four different translation tables (1, 6, 11 and 16), to take into account the alternative nuclear genetic codes (Cocquyt *et al.*, 2010a), as well as the chloroplast and mitochondrial translation tables. Briefly, for each transcript coding frames were predicted with each of the four translation tables. Then, for each transcript, the longest coding sequences detected was retained for the downstream analysis, and the corresponding translation table recorded. Concordance with orthologous sequences predicted by the TRAPID pipeline by sequence similarity searches was also taken into account. Each transcript was translated into the corresponding amino acid sequences with the transeq algorithm from the EMBOSS package, using the appropriate translation table, and added to the proteome data of the 15 genomes described in Table 2.1, resulting in 1,228,821 amino acid sequences.

Gene Family inference

Sequences were used to build a custom PLAZA 4.0 instance (Van Bel *et al.*, 2018). Briefly: all-against-all sequence similarity comparison was executed with DIAMOND v. 0.9.18 (Buchfink *et al.*, 2014), and the similarity matrix was used to infer homologous relationships among proteins using the graph-based Markov clustering method implemented in Tribe-MCL v. 10-201 (Enright *et al.*, 2002) with parameters: -scheme 4 -I 2. This resulted in the clustering of 976,181 protein sequences (79.5% of the total proteins) into 69,462 gene families, leaving 252,640 singleton proteins. Homologous relationships between sequences were further refined by building subfamilies on the same protein similarity graph with OrthoFinder v. 1.1.4 (Emms & Kelly, 2015), which resulted in the identification of 158,039 subfamilies. A procedure was applied to identify and flag outlier proteins from gene families and subfamilies if they showed similarity only to a minority of all family members (Proost *et al.*, 2009). Single-copy families were selected by identifying the 620 picoPLAZA single–copy gene families (Vandepoele *et al.*, 2013). At this point, to remove potential contaminants in the transcriptomic data from the single-copy gene families, only sequences that were classified as “Viridiplantae” after an additional sequence similarity search with the GhostKOALA (Kanehisa *et al.*, 2016) webserver were retained for downstream analyses. Then, to further reduce the residual redundancy of the transcriptome datasets in the single-copy gene families, for each gene family the nucleotide sequences of each species were collapsed with CAP3 (Huang & Madan, 1999), using stringent parameters to avoid artefactual creation of chimeras: gap penalty 12 (-g) and overlap percent identity cutoff 98% (-p). This set of 620 quasi single-copy genes was used for the downstream phylogenetic analyses.

Phylogenetic Analysis

Amino acid sequences of the 620 gene families were aligned with MAFFT v. 7.187 (Katoh & Standley, 2013), using accuracy-oriented parameters (--localpair --maxiterate 1,000) and an offset value (--ep) of 0.075. To identify and trim eventual residual in-paralogs, we followed phylogeny-guided approaches. As a first step, a Maximum Likelihood (ML) tree for the resulting alignments was obtained with IQtree v. 1.6.0 (Nguyen *et al.*, 2015), inferring the best model and allowing invariable sites and free rate of heterogeneity across sites (Soubrier *et al.*, 2012; Kalyaanamoorthy *et al.*, 2017), with parameters: number of ultra-fast bootstrap replicates (-bb) 1,000 (Hoang *et al.*, 2018) and SH-aLRT branch test (Guindon *et al.*, 2010) with 1,000 replicates (-alrt). The resulting trees were visualized in MEGA v. 5.1 (Tamura *et al.*, 2011) and the corresponding alignments were inspected and processed in Geneious v. 8.0.5 (Biomatters Ltd., https://www.geneious.com/).

Phylogenetic trees were carefully cured by hand to retain only full length or fragments of co-orthologs single-copy gene families: (1) except for *Mesostigma viride* and *Chlorokybus atmophyticus*, which were allowed to cluster with either the Streptophyta or the Chlorophyta, Streptophyta and Chlorophyta monophyly was enforced for the remaining species, e.g. *Arabidopsis thaliana* sequences clustering within the Trentepohliales (or vice versa) would be excluded, as obvious sign of paralogy or contamination of environmental samples. (2) For a species with two or more overlapping sequence, one in the expected phylogenetic position (according to orthologous sequences of other species in the same taxon), the other one in a conflicting position, the conflicting sequences were removed. In case of overlapping sequences with concordant phylogenetic signal, regardless of their phylogenetic position, but always enforcing Streptophyta and Chlorophyta monophyly, the longest one was retained, if full-length, otherwise both were retained. In case of conflicting but non-overlapping sequences, both sequences were retained and scaffolded (see below), since they could represent *bona fide* single-copy genes with different rates of evolution across sites. (3) Gene family alignments with sequences from less than 30 species (more than 50% of species not represented) were discarded. (4) Alignments composed by two or more near-identical paralogs (e.g.: ribosomal subunit proteins) and where confident segregation of the paralogs was difficult were discarded. After filtering, 539 out of the 620 initial gene families were retained (hereafter referred to as “coreGF”).

Each gene family with more than one sequence per species (398 gene families), was processed independently for orthology-guided scaffolding with a custom script. First, an all-against-all sequence similarity search was performed with BLASTp v. 2.5.0+ (Boratyn *et al.*, 2013), using an e-value cut-off of 10e^{-5}. The sequence with the highest bitscore was selected as reference for that gene family. Then, for each species with more than one sequence, the sequences were concatenated according to their relative position to the reference sequence, based on the BLASTp alignments. This resulted in gene families with only one sequence per species. Moreover, this approach defined a conservative “unscaffold” dataset, where within each gene family species with more than one sequence were discarded, and a more comprehensive but potentially noisier “scaffold” dataset, where multiple sequences of the same species were scaffolded. Unscaffold and scaffold dataset gene families were aligned again with MAFFT, using accuracy-oriented parameters (--localpair --maxiterate 1,000 –ep 0.075). A trimmed version of both datasets was created by trimming the amino acid alignments with TrimAl v. 1.2 (Capella-Gutiérrez *et al.*, 2009), with parameters: gapthreshold 0.75 and simthreshold 0.001. A subset of 355 gene families, that focused on ulvophyceans (hereafter referred to as ulvoGF dataset), was obtained by selecting gene families with at least one representative species for each Order of ulvophyceans. Two pruned datasets, where either *Ignatius tetrasporus* or *Acetabularia acetabulum* and *Scotinosphaera lemnae* were excluded, were generated as well. In this case, the corresponding amino acid sequences were discarded before the sequence alignment step.

For each gene family, ML trees were built with IQtree, inferring the best model and rate of heterogeneity across sites (Figure S3.7). All ML analyses were run using IQtree with 1,000 ultra-fast bootstrap and SH-aLRT branch test replicates. Gene trees were used for the partitioned supermatrix analyses and for the coalescent-based analyses. ML supermatrix analyses were performed using IQtree with two settings: 1) a gene-wise partitioned analysis (Chernomor *et al.*, 2016) was performed, assigning the best substitution model inferred to each partition; 2) an analysis using mixture models was performed using an LG+F+G plus a C20-profile mixture model of substitution rates (Si Quang *et al.*, 2008).

Coalescence-based analyses were run with ASTRAL v. 5.6.1 (Zhang *et al.*, 2018). First, for each ML gene tree, low support branches (ufBS support < 10) were collapsed with Newick Utilities v. 1.6 (Junier & Zdobnov, 2010). Branches contracted in the ML gene trees were removed as well from the pool of the corresponding 1,000 bootstrap trees generated during the ML reconstruction. Then, two independent runs were performed either using the ML tree for each gene (BestML), or using the multilocus bootstrap support (MLBS) approach. For the MLBS analysis, 100 replicates were run (-r) starting from the 1,000 contracted bootstrap trees for each gene, allowing gene and site resampling (--gene-resampling flag).

Statistical tests for rejecting the null hypothesis of polytomies at the branch-level were performed in ASTRAL (-t 10 flag), following Sayyari and colleagues (Sayyari & Mirarab, 2018). Briefly, for each dataset (coreGF and ulvoGF, scaffold-unscaffold, trim and untrim), the coalescence-based MLBS tree was tested (-q flag) by random sampling subsets of ML gene trees representing 1-100% of the total gene families in the dataset, with a minimum of 20 ML gene trees per subset. For each dataset and for each subset, ten independent replicates were generated, and analyzed with ASTRAL on BestML mode with the –t 10 flag –q flag to score the MLBS trees. The support of some key branches was analyzed in each subset of the datasets and the median of the p-values for each subset of trees was calculated.

Clustering of the ML gene trees was performed with the RPANDA v. 1.3 package (Morlon *et al.*, 2015). First, the pairwise Jensen-Shannon distances between the spectral density profiles of the ML gene topologies were computed (Lewitus & Morlon, 2016). Based on the pairwise distances, the topologies were grouped into clusters by hierarchical clustering, and the bootstrap support of each cluster was evaluated (100 BS replicated). The most divergent cluster was considered as an outlier. Then, supermatrix and coalescence-based analyses were performed with IQtree and ASTRAL, respectively, for the gene families of each cluster as described above, as well as for the genes of all the clusters excluding the genes in the outlier cluster.

The species trees with the overall highest support inferred from the previous analyses were collected. Significance of topological incongruences between the well supported topologies (relative position of Chlorodendrophyceae and Pedinophyceae, basal relationships of Ulvophyceae, and position of Ignatiales and the Dasycladales-Scotinosphaerales clade) were tested using Approximately Unbiased (AU) tests (Shimodaira, 2002) implemented in IQtree, with 100,000 RELL resamplings (Kishino *et al.*, 1990).

In addition to the AU test, site-wise and gene-wise log-likelihood scores (Chiari *et al.*, 2012; Shen *et al.*, 2017) were calculated to assess support for alternative topologies. Briefly, for each trimmed amino acid dataset, the ML support of the 22 topologies was inferred with a IQtree run constrained to that topology (-g), using the LG+F+R5 substitution rates model. For each pair of constrained trees, the differences in site-wise log-likelihood scores were calculated in RAxML v. 8.2.9 (Stamatakis, 2014) with PROTGAMMALGF model and without the Broyden–Fletcher–Goldfarb–Shanno parameter optimization (--no-bfgs). Then, the gene-wise log-likelihood scores and the percentage of genes supporting the two compared topologies were calculated as outlined by Shen and colleagues (Chiari *et al.*, 2012; Shen *et al.*, 2017).

Calibrated phylogenetic tree

Due to the high computational cost, the molecular clock analysis was restricted to the 539 coreGF scaffold trim gene set. Clock-likeliness of each gene was assessed with the package SortDate (Smith *et al.*, 2018) against the ML supermatrix and the coalescence-based topologies, scoring the trees on minimal conflict, low root-to-tip variance, and discernible amounts of molecular evolution. The 10 most clock-like genes for each topology were concatenated and subjected to relaxed clock analyses (2,806 and 2,857 amino acid residues for the ML supermatrix and the coalescence-based topology, respectively). Node calibrations were transferred from fossil information and from node age estimates from previous studies (Table S3.4). All analyses were run with the same set of calibration nodes, except for the UB (*Proterocladus*) and the RT (root age, i.e.: Streptophyta-Chlorophyta split) nodes. *Proterocladus*, a Neoproterozoic fossil, is tentatively assigned to the Order of Cladophorales (Porter, 2004), its corresponding calibration node was either included

(UB_1) or excluded (UB_0). Regarding the root age, three different priors were used (RT_1-RT_3). An additional analysis was run without assigning a prior value for the root age (RT_0) (Table S3.4, S3.4).

Relaxed molecular clock analyses were run with PhyloBayes 4.1b (Lartillot *et al.*, 2009). Two sets of analyses were run on two fixed topologies (the ML supermatrix and the coalescence-based topologies), using the set of clock-like genes for each topology. Both lognormal autocorrelated clock (-ln flag) and uncorrelated gamma multiplier clock (-ugam flag) models were tested for each dataset, and the models were run with either LG+Γ4 or CATGTR+Γ4 models of amino acid substitutions. In total, 64 different analyses were run (Table S7) to test the influence of different models, and of root and key prior ages on the age estimations. For each analysis, two distinct MCMC chains were run for at least 10,000 generations. The convergence of the log likelihoods and parameters estimates were tested on PhyloBayes. Chains were summarized after discarding the first 750 generations as burn-in.

The ultrametric trees were used to guide the ancestral state reconstruction of the ecological and cyto-morphological traits in Phytools (Revell, 2011). The posterior probabilities of the ancestral state of each node were calculated from summaries of 1,000 replicates of simulated stochastic character map (make.simmap), using empirical Bayes method under the ADR model, which permits backward and forward rates between states to have different values.

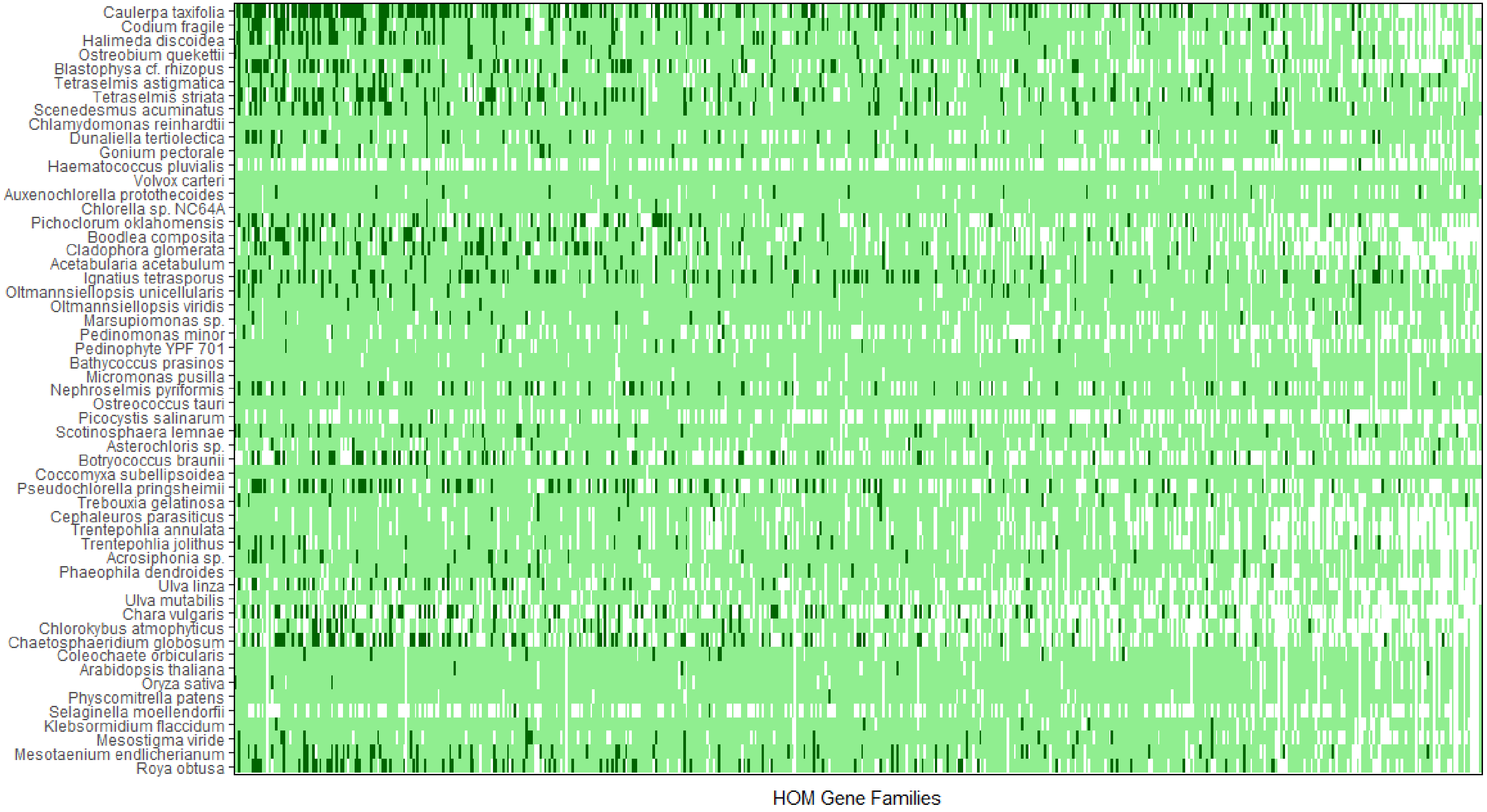


Figure S3.1: Supermatrix occupancy of coreGF.

Supermatrix occupancy of the 539 coreGF single copy-gene families. From the left to the right, the genes with higher taxon occupancy. Dark green: scaffolded sequence (absent from the unscaffold datasets); light green: available sequence; white: missing sequence.

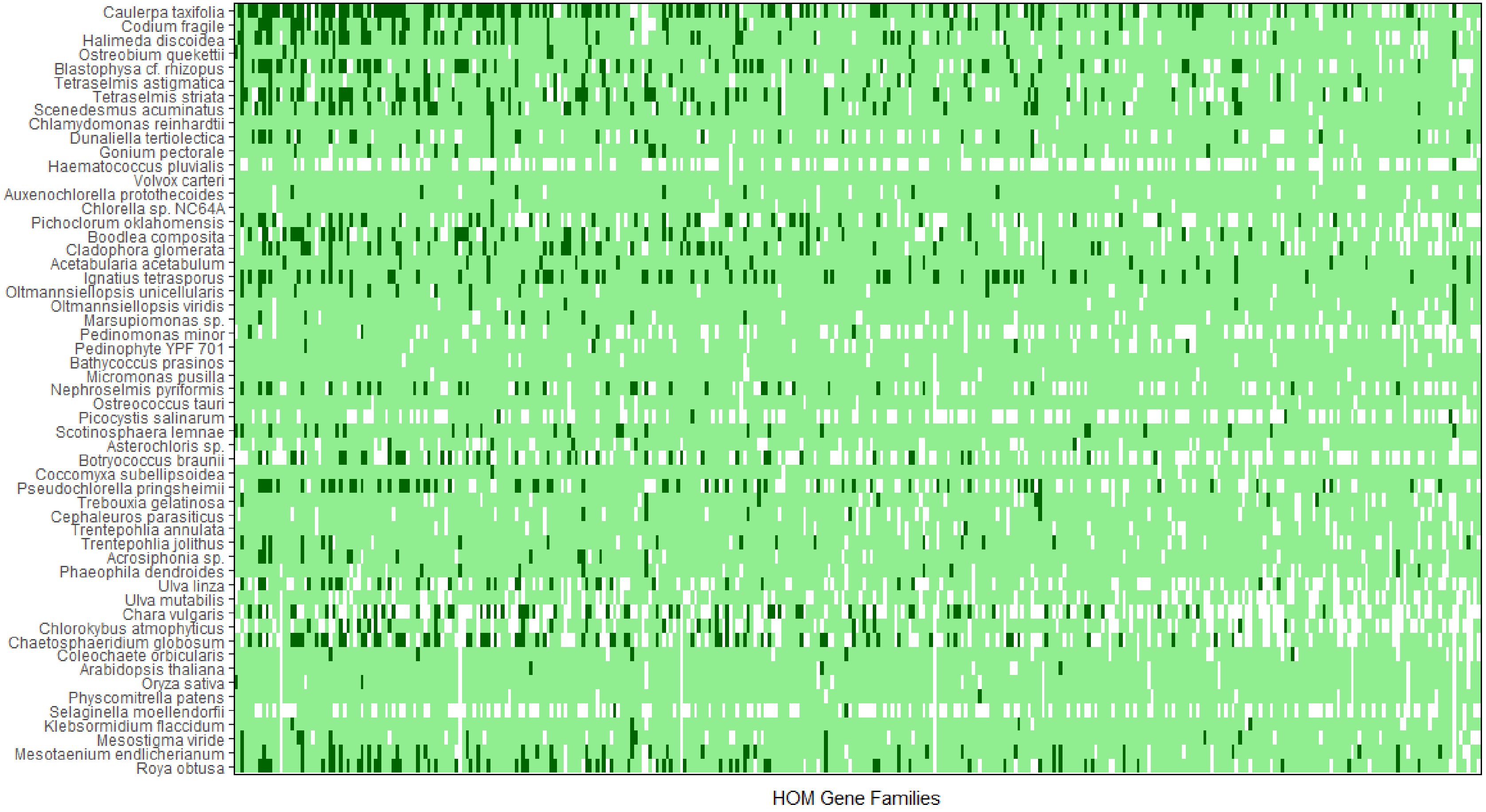


Figure S3.2: Supermatrix occupancy of ulvoGF.

Supermatrix occupancy of the 355 ulvoGF single copy-gene families. From the left to the right, the genes with higher taxon occupancy. Dark green: scaffolded sequence (absent from the unscaffold datasets); light green: available sequence; white: missing sequence.

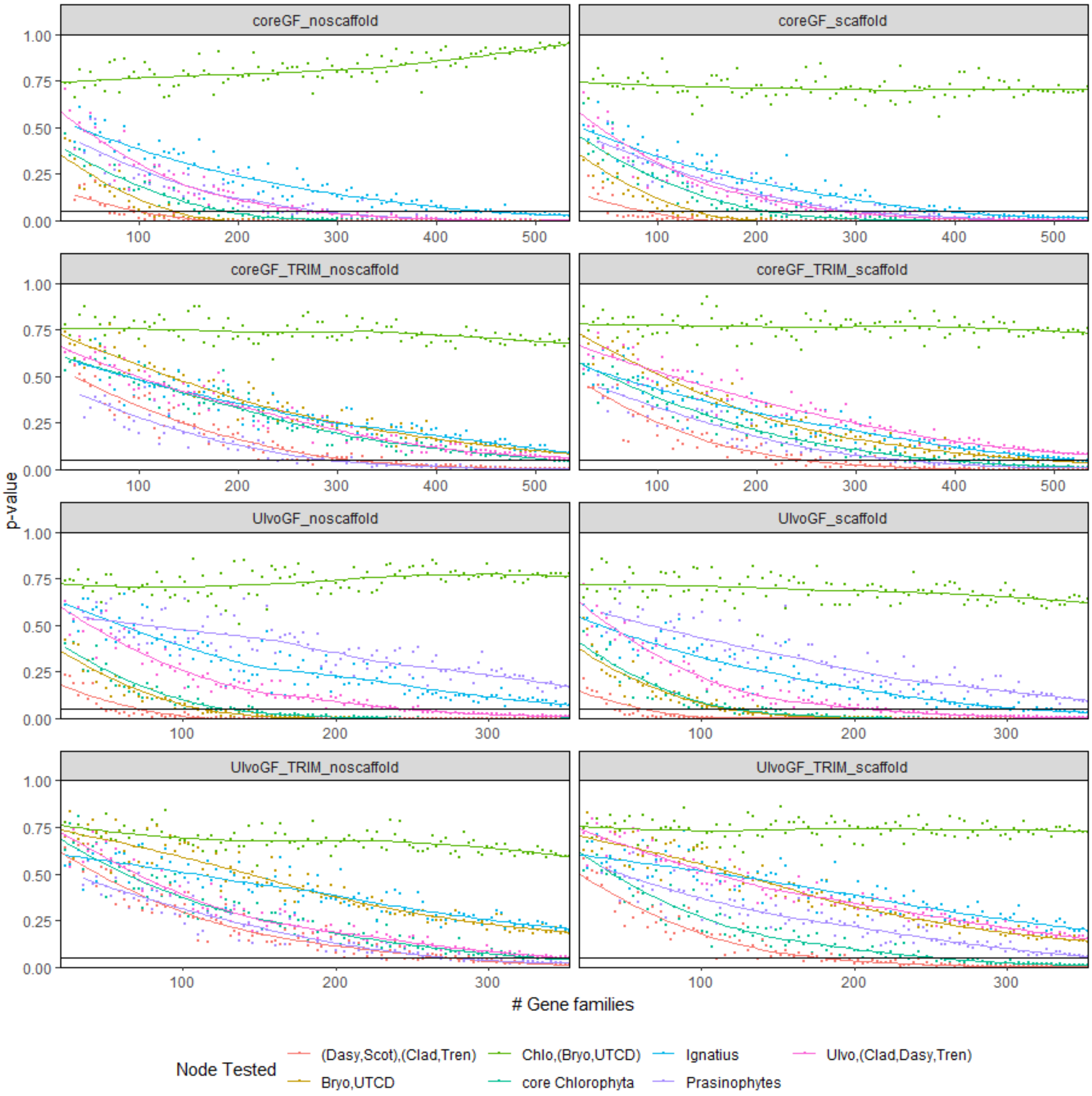


Figure S3.3: Test for polytomy null-hypothesis.

Polytomy test results for selected branches for each dataset. Gene trees were built with random growing subsets of genes (1%, 2%, ... 100%, but not less than 20 genes), 10 replicates were run for each dataset and each subset. The median of the p-values of selected branches for each subset is plotted on the y-axis, against the number of genes in the subset (x-axis). The horizontal black line indicates p-value = 0.05. Increasing gene numbers never reduced the p-value of the rejection.

<u>NO IGNATIUS</u>

		aa supermatrix								aa coalescence															
		partitioned								BestML								MLBS							
		coreGF unscaffold	coreGF scaffold	coreGF unscaffold TRIM	coreGF scaffold TRIM	Ulva unscaffold	ulva scaffold	ulva unscaffold TRIM	ulva scaffold TRIM	coreGF unscaffold	coreGF scaffold	coreGF unscaffold TRIM	coreGF scaffold TRIM	Ulva unscaffold	ulva scaffold	ulva unscaffold TRIM	ulva scaffold TRIM	coreGF unscaffold	coreGF scaffold	coreGF unscaffold TRIM	coreGF scaffold TRIM	Ulva unscaffold	ulva scaffold	ulva unscaffold TRIM	ulva scaffold TRIM
core Chlorophyta relationships	Pedi first/Chld second	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	Chld first/Pedi second	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Treb+Chlr third	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Chlo+Bryo	🟩	🟩	🟨	🟩	🟩	🟩	🟨	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	Chlo-Bryo-Ulvo* radiation	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Bryo+Clado-Tren	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Ulvophytes relationships	Oltm-Ulvo+Clad-Tren	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩	🟩	🟨	🟩	🟨	🟨	🟨	🟩	🟩	🟨	🟩	🟩	🟩	🟨	🟨
	Dasy+Scot	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Dasy-Scot+Oltm	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	Dasy-Scot+Clad-Tren	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟩	🟨	🟨	🟥	🟨	🟥	🟨	🟥	🟩	🟩	🟨	🟥	🟩	🟨	🟨	🟥
	Dasy-Scot+Oltm-Ulvo	🟥	🟥	🟨	🟨	🟥	🟥	🟨	🟨	🟥	🟥	🟥	🟨	🟥	🟨	🟥	🟨	🟥	🟥	🟥	🟨	🟥	🟥	🟥	🟨
	Blastophysa early Clad	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Clad+Tren	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

<u>NO DASY-SCOT</u>

		aa supermatrix								aa coalescence															
		partitioned								BestML								MLBS							
		coreGF unscaffold	coreGF scaffold	coreGF unscaffold TRIM	coreGF scaffold TRIM	Ulva unscaffold	ulva scaffold	ulva unscaffold TRIM	ulva scaffold TRIM	coreGF unscaffold	coreGF scaffold	coreGF unscaffold TRIM	coreGF scaffold TRIM	Ulva unscaffold	ulva scaffold	ulva unscaffold TRIM	ulva scaffold TRIM	coreGF unscaffold	coreGF scaffold	coreGF unscaffold TRIM	coreGF scaffold TRIM	Ulva unscaffold	ulva scaffold	ulva unscaffold TRIM	ulva scaffold TRIM
core Chlorophyta relationships	Pedi first/Chld second	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	Chld first/Pedi second	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Treb+Chlr third	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Chlo+Bryo	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟨	🟨	🟥	🟥	🟨	🟥	🟥	🟥	🟨	🟨	🟥	🟥	🟨	🟥	🟥	🟥	🟨
	Chlo-Bryo-Ulvo* radiation	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟩	🟩	🟥	🟥
	Bryo+Clado-Tren	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟨	🟥
Ulvophytes relationships	Oltm-Ulvo+Clad-Tren	🟩	🟩	🟨	🟨	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟨	🟩	🟩	🟨	🟨	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟨
	Igna early Ulvo	🟩	🟩	🟩	🟨	🟨	🟩	🟩	🟨	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟨	🟩	🟥	🟥
	Igna+Oltm-Ulvo	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟩	🟩
	Blastophysa early Clad	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Clad+Tren	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

Figure S3.4: Ignatiales- and Dasycladales- Scotinosphaerales - pruned phylogenies

Summary of support for hypotheses of Chlorophyta relationships, based on the amino acid and codon-wise alignments of 8 single-copy gene datasets across 24 distinct supermatrix or coalescence-based analyses where *Ignatius* (top panel), or Dasycladales-Scotinosphaerales (bottom panel) were excluded. coreGF unscaffold: 539 single-copy genes, with partial sequences removed. coreGF scaffold: 539 single-copy genes, with partial sequences scaffolded. coreGF unscaffold TRIM: as coreGF unscaffold, but with less conserved regions filtered. coreGF scaffold TRIM: as coreGF scaffold, but with less conserved regions filtered. ulvoGF unscaffold: 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences removed. ulvoGF scaffold: 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences scaffolded. ulvoGF unscaffold TRIM: as ulvoGF unscaffold, but with less conserved regions filtered. ulvoGF scaffold TRIM: as ulvoGF scaffold, but with less conserved regions filtered. Partitioned: gene-wise partition of the supermatrix, with substitution model for each partition inferred by the gene-tree best model, allowing invariable sites and free rate of heterogeneity across sites. BestML: coalescence-based analysis. MLBS: as BestML, but using the Multi-Locus Bootstrap Support approach. Green: strong support. Yellow: low support. Red: no support. Strong support refers to bootstrap values or posterior probabilities > 75% or 0.75, respectively, for the relationships depicted.

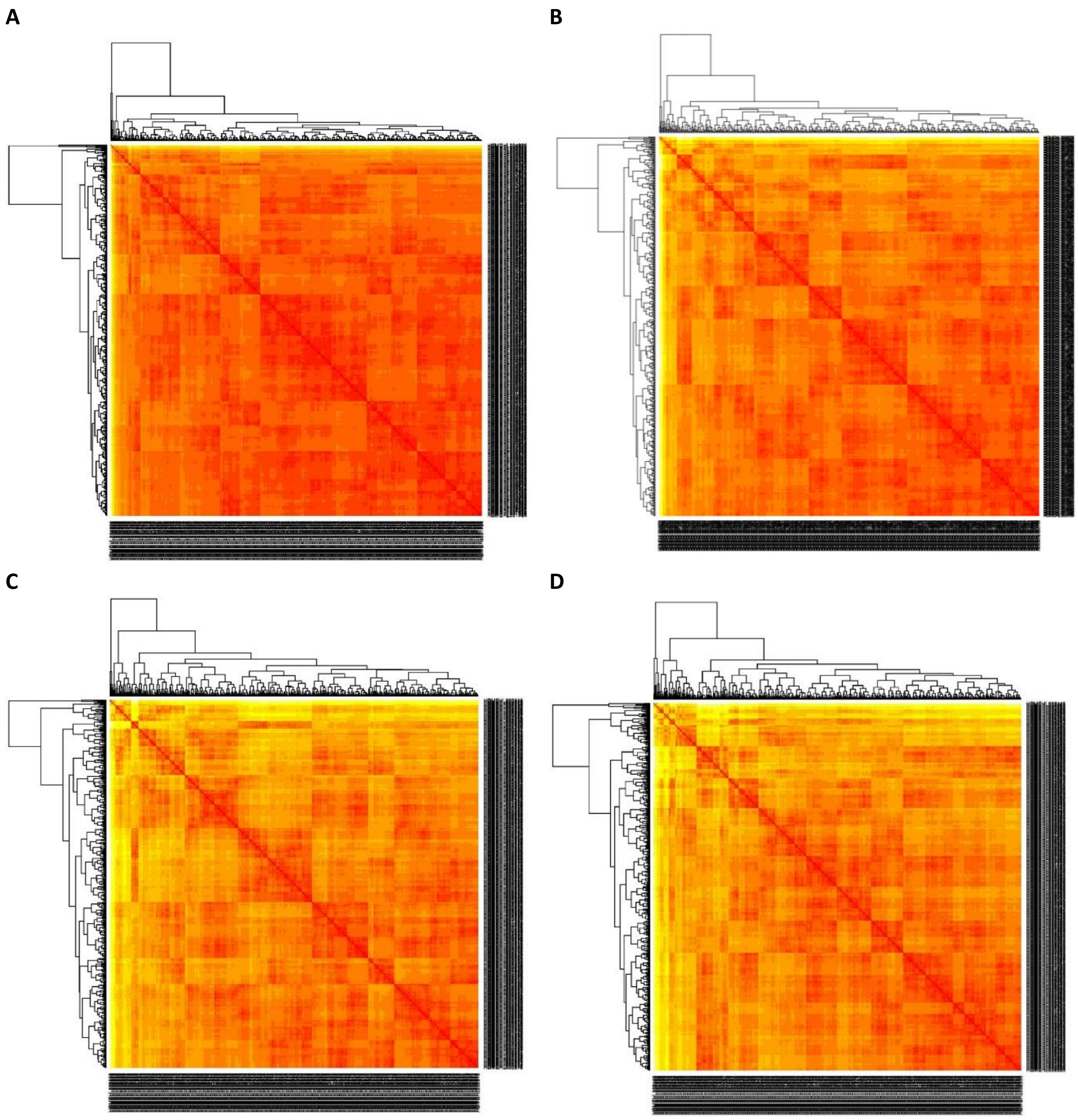


Figure S3.5: Heatmaps showing the pairwise Jensen-Shannon distances between gene trees after hierarchical clustering.

Red indicates low distance, while yellow color indicates high distances. (A) coreGF noscaffold. (B) coreGF scaffold. (C) coreGF TRIM noscaffold. (D) coreGF TRIM scaffold.

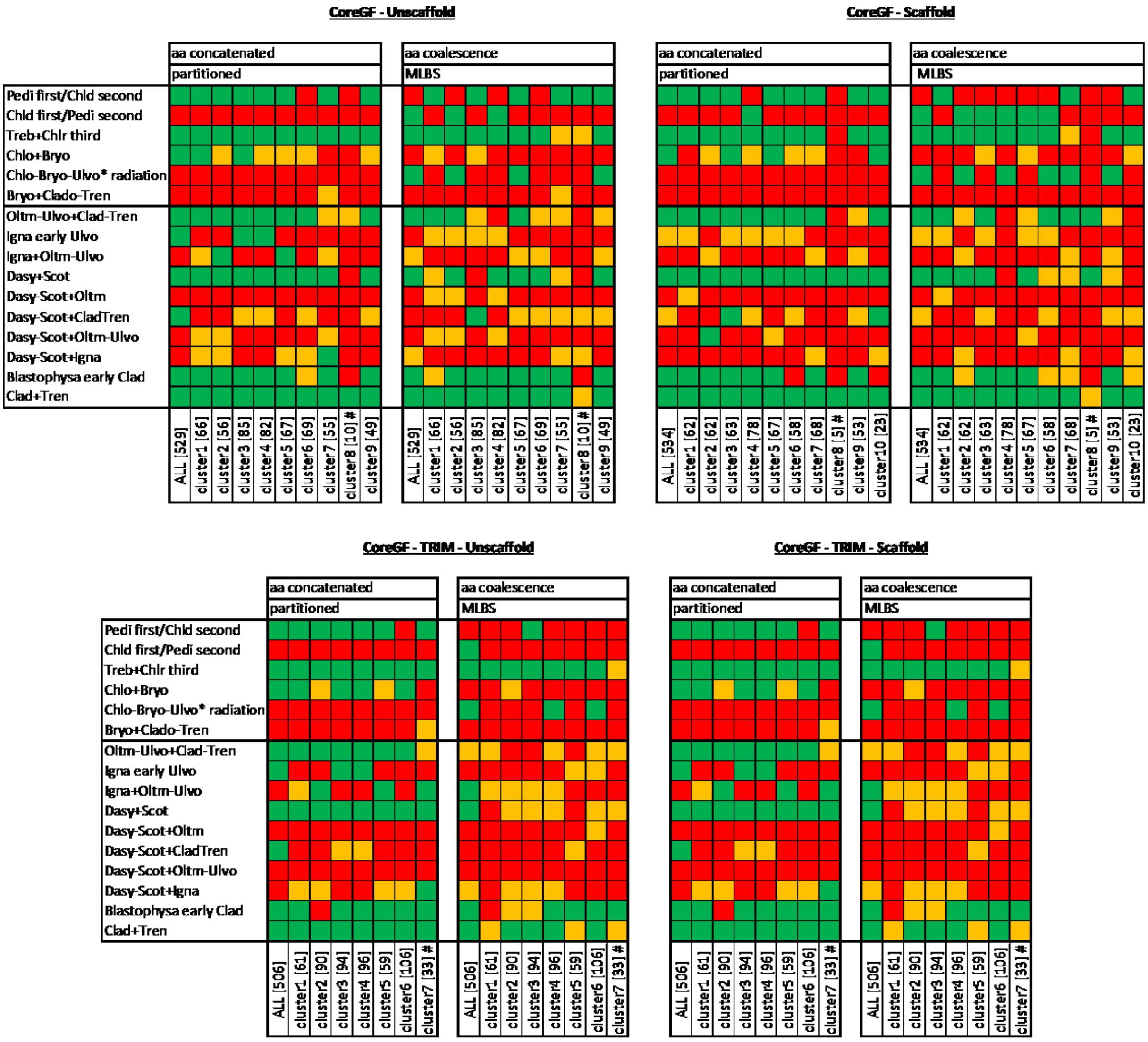


Figure S3.6: Cluster-wise phylogenetic analyses.

Summary of supermatrix or coalescence-based analyses of the clusters identified in Figure S3. Within square brackets are indicated the number of genes in the cluster, # indicates the outlier cluster (i.e.: the one showing the highest Jensen-Shannon distances). ALL indicates analyses for all the genes in all the cluster, except the genes from the outlier cluster. coreGF unscaffold: 539 single-copy genes, with partial sequences removed. coreGF scaffold: 539 single-copy genes, with partial sequences scaffolded. coreGF unscaffold TRIM: as coreGF unscaffold, but with less conserved regions filtered. coreGF scaffold TRIM: as coreGF scaffold, but with less conserved regions filtered. ulvoGF unscaffold: 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences removed. ulvoGF scaffold: 355 single-copy genes subset of coreGF focussing on Ulvophyceae, with partial sequences scaffolded. ulvoGF unscaffold TRIM: as ulvoGF unscaffold, but with less conserved regions filtered. ulvoGF scaffold TRIM: as ulvoGF scaffold, but with less conserved regions filtered. Partitioned: gene-wise partition of the supermatrix, with substitution model for each partition inferred by the gene-tree best model, allowing invariable sites and free rate of heterogeneity across sites. MLBS: coalescence-based analysis using the Multi-

Locus Bootstrap Support approach. Green: strong support. Yellow: low support. Red: no support. Strong support refers to bootstrap values or posterior probabilities > 75% or 0.75, respectively, for the relationships depicted.

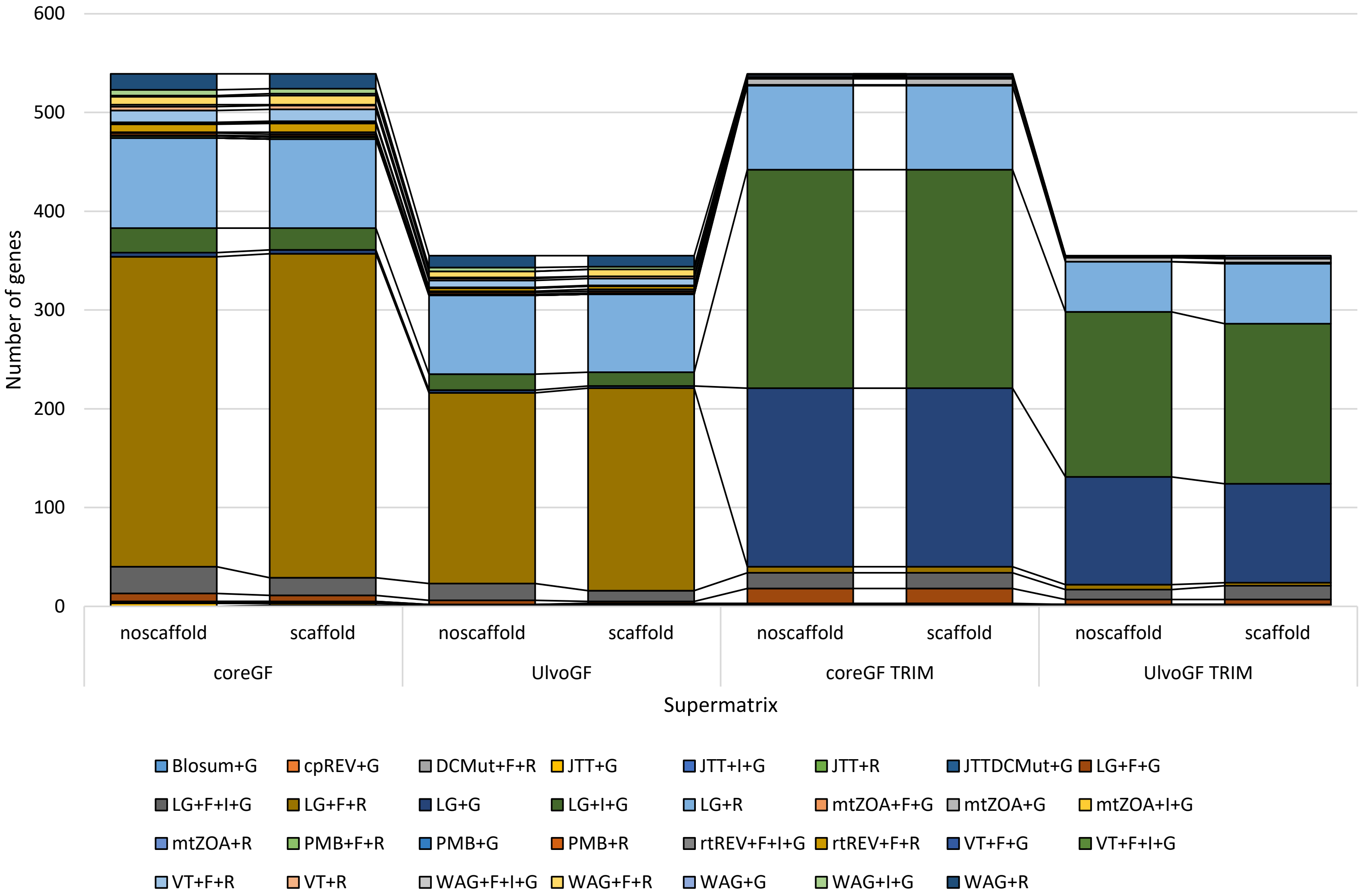


Figure S3.7: Best substitution models inferred for the genes in each dataset.

Histogram showing the best substitution model and rate of heterogeneity across sites inferred for the genes of each dataset.

Table S3.1: Alignments metrics								
	Untrimmed alignments				Trimmed alignments			
	coreGF unscaffold	coreGF scaffold	ulvoGF unscaffold	ulvoGF scaffold	coreGF unscaffold	coreGF scaffold	ulvoGF unscaffold	ulvoGF scaffold
# genes	539	539	355	355	539	539	355	355
# sites (nt)	1,751,925	1,877,544	1,143,789	1,244,520	342, 501	342,501	252,612	252,612
# sites (aa)	583,975	625,848	381,263	414,840	114,167	114,167	84,204	84,204
% missing data	65.6	69.7	43.7	54.8	27.3	19.9	24.0	15.9

Table S3.2: core Chlorophyta and ulvophyceans diversification AU-tests.

AU test: Bryo-Chlo-Ulvo

dataset	Tree	logL	deltaL	bp-RELL	p-KH	p-SH	c-ELW	p-AU
coreGF noscaffold	1	-15,791,662.12	0	0.9877 +	0.9865 +	1.0000 +	0.9877 +	0.9877 +
	2	-15,791,832.34	170.224	0.0123 -	0.0135 -	0.0135 -	0.0123 -	0.0123 -
coreGF scaffold	1	-17,211,164.2	0	0.9869 +	0.9864 +	1.0000 +	0.9869 +	0.9862 +
	2	-17,211,348.3	184.097	0.0131 -	0.0136 -	0.0136 -	0.0131 -	0.0138 -
coreGF noscaffold	1	-3,848,877.26	1.132	0.4969 +	0.4952 +	0.4952 +	0.4970 +	0.4951 +
TRIM	2	-3,848,876.12	0	0.5031 +	0.5048 +	1.0000 +	0.5030 +	0.5049 +
coreGF scaffold	1	-4,182,782.87	0	0.7081 +	0.7080 +	1.0000 +	0.7082 +	0.7136 +
TRIM	2	-4,182,827.07	44.192	0.2918 +	0.2920 +	0.2920 +	0.2918 +	0.2864 +
ulvoGF noscaffold	1	-10,849,295.9	0	0.9916 +	0.9915 +	1.0000 +	0.9916 +	0.9913 +
	2	-10,849,457.1	161.15	0.0084 -	0.0085 -	0.0085 -	0.0084 -	0.0087 -
ulvoGF scaffold	1	-11,946,468	0	0.9893 +	0.9890 +	1.0000 +	0.9893 +	0.9884 +
	2	-11,946,637.9	169.952	0.0107 -	0.0110 -	0.0110 -	0.0107 -	0.0116 -
ulvoGF noscaffold	1	-2,851,993.21	0	0.5974 +	0.5987 +	1.0000 +	0.5975 +	0.5982 +
TRIM	2	-2,852,009.83	16.622	0.4026 +	0.4013 +	0.4013 +	0.4025 +	0.4018 +
ulvoGF scaffold	1	-3,124,270.99	0	0.7499 +	0.7510 +	1.0000 +	0.7499 +	0.7474 +
TRIM	2	-3,124,320.02	49.031	0.2501 +	0.2490 +	0.2490 +	0.2501 +	0.2526 +

AU test: Chld-Pedi

dataset	Tree	logL	deltaL	bp-RELL	p-KH	p-SH	c-ELW	p-AU
coreGF noscaffold	1	-15,791,662	0	1.0000 +	1.0000 +	1.0000 +	1.0000 +	1.0000 +
	2	-15,791,999	337.044	0.0000 -	0.0000 -	0.0000 -	0.0000 -	0.0000 -
coreGF scaffold	1	-17,211,164.2	0	1.0000 +	1.0000 +	1.0000 +	1.0000 +	1.0000 +
	2	-17,211,561.5	397.272	0.0000 -	0.0000 -	0.0000 -	0.0000 -	0.0000 -
coreGF noscaffold	1	-3,848,876.18	0	0.9569 +	0.9563 +	1.0000 +	0.9568 +	0.9559 +
TRIM	2	-3,848,957.15	80.976	0.0431 -	0.0437 -	0.0437 -	0.0432 -	0.0441 -
coreGF scaffold	1	-4,182,782.81	0	0.9016 +	0.9009 +	1.0000 +	0.9013 +	0.9012 +
TRIM	2	-4,182,848.9	66.085	0.0984 +	0.0991 +	0.0991 +	0.0987 +	0.0988 +
ulvoGF noscaffold	1	-10,849,295.9	0	0.9987 +	0.9987 +	1.0000 +	0.9987 +	0.9988 +
	2	-10,849,499.2	203.306	0.0013 -	0.0014 -	0.0014 -	0.0013 -	0.0012 -
ulvoGF scaffold	1	-11,94,6468	0	0.9995 +	0.9994 +	1.0000 +	0.9995 +	0.9996 +
	2	-11,946,705.9	237.929	0.0005 -	0.0006 -	0.0006 -	0.0005 -	0.0004 -
ulvoGF noscaffold	1	-2,851,993.2	0	0.7925 +	0.7911 +	1.0000 +	0.7922 +	0.7870 +
TRIM	2	-2,852,027.89	34.683	0.2075 +	0.2089 +	0.2089 +	0.2078 +	0.2130 +
ulvoGF scaffold	1	-3,124,271.12	0	0.6163 +	0.6159 +	1.0000 +	0.6161 +	0.6113 +
TRIM	2	-3,124,284.86	13.739	0.3836 +	0.3841 +	0.3841 +	0.3839 +	0.3887 +

deltaL: logL difference from the maximal logL in the set.
bp-RELL: bootstrap proportion using RELL method.
p-KH: p-value of one sided Kishino-Hasegawa test.
p-SH: p-value of Shimodaira-Hasegawa test.
c-ELW: Expected Likelihood Weight.
p-AU: p-value of approximately unbiased (AU) test.
Plus signs denote the 95% confidence sets. Minus signs denote significant exclusion. All tests performed 100,000 resamplings using the RELL method.

Table S3.3: *Ignatius* position AU-tests.

AU test: Igna

dataset	Tree	logL	deltaL	bp-RELL	p-KH	p-SH	c-ELW	p-AU
coreGF noscaffold	1	-15,791,661	0	0.4040 +	0.5037 +	1.0000 +	0.4056 +	0.6439 +
	2	-15,791,765	103.93	0.0323 -	0.0674 +	0.2229 +	0.0319 -	0.1019 +
	3	-15,791,661	0	0.4057 +	0.4963 +	0.8447 +	0.4055 +	0.6531 +
	4	-15,791,780	118.92	0.1574 +	0.1665 +	0.2255 +	0.1564 +	0.1879 +
	5	-15,791,909	247.72	0.0005 -	0.0074 -	0.0113 -	0.0005 -	0.0037 -
coreGF scaffold	1	-17,211,164	0	0.4110 +	0.5003 +	0.8368 +	0.4137 +	0.6500 +
	2	-17,211,270	105.62	0.0572 +	0.0917 +	0.2733 +	0.0565 +	0.1357 +
	3	-17,211,164	0	0.4148 +	0.4997 +	1.0000 +	0.4135 +	0.6597 +
	4	-17,211,320	155.99	0.1170 +	0.1279 +	0.1703 +	0.1163 +	0.1463 +
	5	-17,211,550	385.4	0.0000 -	0.0004 -	0.0005 -	0.0000 -	0.0006 -
coreGF noscaffold TRIM	1	-3,848,877	1.134	0.2000 +	0.4971 +	0.7028 +	0.2021 +	0.5484 +
	2	-3,848,899	22.713	0.1342 +	0.4108 +	0.7231 +	0.1329 +	0.3274 +
	3	-3,848,877	1.134	0.2014 +	0.4971 +	0.7028 +	0.2019 +	0.5488 +
	4	-3,848,876	0	0.4643 +	0.5029 +	1.0000 +	0.4630 +	0.5262 +
	5	-3,849,029	152.94	0.0001 -	0.0639 +	0.0820 +	0.0001 -	0.0005 -
coreGF scaffold TRIM	1	-4,182,822	39.404	0.1485 +	0.3463 +	0.4750 +	0.1508 +	0.4015 +
	2	-4,182,833	49.973	0.0987 +	0.2442 +	0.4789 +	0.0977 +	0.2475 +
	3	-4,182,822	39.404	0.1504 +	0.3463 +	0.4750 +	0.1508 +	0.4016 +
	4	-4,182,783	0	0.6024 +	0.6537 +	1.0000 +	0.6007 +	0.6821 +
	5	-4,183,033	250.47	0.0000 -	0.0001 -	0.0017 -	0.0000 -	0.0000 -
ulvoGF noscaffold	1	-10,849,296	0	0.3510 +	0.5000 +	0.8502 +	0.3548 +	0.6308 +
	2	-10,849,366	70.066	0.0662 +	0.1373 +	0.3830 +	0.0653 +	0.1790 +
	3	-10,849,296	0	0.3565 +	0.5000 +	1.0000 +	0.3551 +	0.6373 +
	4	-10,849,374	78.303	0.2259 +	0.2445 +	0.3406 +	0.2246 +	0.2808 +
	5	-10,849,524	228.58	0.0003 -	0.0071 -	0.0110 -	0.0003 -	0.0019 -
ulvoGF scaffold	1	-11,946,468	0	0.3519 +	0.5068 +	1.0000 +	0.3543 +	0.6450 +
	2	-11,946,534	65.504	0.1075 +	0.1851 +	0.4695 +	0.1064 +	0.2546 +
	3	-11,946,468	0	0.3547 +	0.4932 +	0.8458 +	0.3542 +	0.6365 +
	4	-11,946,567	99.317	0.1859 +	0.2162 +	0.2998 +	0.1850 +	0.2378 +
	5	-11,946,819	350.58	0.0000 -	0.0004 -	0.0005 -	0.0000 -	0.0000 -
ulvoGF noscaffold TRIM	1	-2,852,007	14.265	0.1642 +	0.4305 +	0.5863 +	0.1663 +	0.4358 +
	2	-2,852,006	12.597	0.2017 +	0.4186 +	0.7725 +	0.2000 +	0.4275 +
	3	-2,852,007	14.265	0.1651 +	0.4305 +	0.5863 +	0.1663 +	0.4439 +
	4	-2,851,993	0	0.4689 +	0.5814 +	1.0000 +	0.4674 +	0.5709 +
	5	-2,852,131	137.62	0.0000 -	0.0092 -	0.0334 -	0.0000 -	0.0003 -
ulvoGF scaffold TRIM	1	-3,124,319	48.194	0.1721 +	0.3003 +	0.4686 +	0.1721 +	0.3697 +
	2	-3,124,318	47.546	0.1085 +	0.2372 +	0.5131 +	0.1086 +	0.2771 +
	3	-3,124,313	42.563	0.1454 +	0.2881 +	0.5007 +	0.1454 +	0.3350 +
	4	-3,124,271	0	0.5740 +	0.7119 +	1.0000 +	0.5738 +	0.7340 +
	5	-3,124,489	218.71	0.0000 -	0.0005 -	0.0112 -	0.0000 -	0.0001 -

deltaL: logL difference from the maximal logL in the set.

bp-RELL: bootstrap proportion using RELL method.

p-KH: p-value of one sided Kishino-Hasegawa test.

p-SH: p-value of Shimodaira-Hasegawa test.

c-ELW: Expected Likelihood Weight.

p-AU: p-value of approximately unbiased (AU) test.

Plus signs denote the 95% confidence sets. Minus signs denote significant exclusion. All tests performed 100,000 resamplings using the RELL method.

Table S3.4: Node calibrations used for relaxed molecular clock analysis

Node	Node	calibration	Fossil or transfer	period	prior	reference
C	Characeae	C_1	Transfer from other study	Devonian	U[416-∞[	(Magallón *et al.*, 2013)
L	Land plants (embryophytes)	L_1	Cryptospore assemblage from the Middle Ordovician (Dapingian)	Ordovician	U[475-∞[	(Magallón *et al.*, 2013)
V	Vascular plants	V_1	*Baragwanathia longifolia*	Silurian	U[421-∞[	(Magallón *et al.*, 2013)
T_1	*Botryococcus* stem node	T_1	*Botryococcus*	Carboniferous	U[299-∞ [	(Colbath & Grenfell, 1995)
UA	*Ostreobium* stem node	UA_1	Transfer from other study	n/a	U[533-425]	(Verbruggen *et al.*, 2009)
UB	Ulvophyceae/Chlorophyceae stem node	UB_1	*Proterocladus*	Neoproterozoic	U[716-∞ [	(Butterfield *et al.*, 1994)
		UB_0	Absence of calibration	n/a	n/a	n/a
RT	Streptophyta-Chlorophyta split	RT_1	Transfer from other study	n/a	U[1279-1159]	(Herron *et al.*, 2009)
		RT_2	Transfer from other study	n/a	U[1015-863]	(Parfrey *et al.*, 2011)
		RT_3	RT_1/RT_2 hybrid	n/a	U[1279-863]	n/a
		RT_0	n/a	n/a	n/a	n/a

Table S3.5: Relaxed molecular clock analyses results

Tree	clock model	sub model	Root	Proterocladus	root_age	Core Chlorophyta	(Chlo,Bryo),Ulvo	Chlo,(Bryo,Ulvo)
ML	ln	CAT	RT0	UB1	1,021 (1,072-969)	885 (960-810)	801 (879-753)	n/a
ML	ln	CAT	RT1	UB1	1,185 (1,209-1,162)	1,005 (1,079-931)	895 (966-823)	n/a
ML	ln	CAT	RT2	UB1	980 (1,004-956)	858 (904-815)	781 (823-749)	n/a
ML	ln	CAT	RT3	UB1	1,022 (1,073-971)	886 (978-822)	802 (880-753)	n/a
ML	ugam	CAT	RT0	UB1	1,013 (1,088-937)	911 (1,065-814)	834 (974-753)	n/a
ML	ugam	CAT	RT1	UB1	1,198 (1,229-1,168)	1,063 (1,149-978)	967 (1,056-877)	n/a
ML	ugam	CAT	RT2	UB1	961 (995-927)	869 (931-809)	800 (860-749)	n/a
ML	ugam	CAT	RT3	UB1	1,013 (1,086-940)	910 (1,060-815)	835 (971-754)	n/a
ML	ln	LG	RT0	UB1	1,030 (1,081-979)	895 (989-831)	802 (881-755)	n/a
ML	ln	LG	RT1	UB1	1,185 (1,209-1,162)	1,008 (1,081-925)	890 (961-813)	n/a
ML	ln	LG	RT2	UB1	984 (1,007-961)	865 (909-823)	782 (723-751)	n/a
ML	ln	LG	RT3	UB1	1,031 (1,081-980)	896 (989-833)	803 (882-756)	n/a
ML	ugam	LG	RT0	UB1	1,021 (1,096-945)	919 (1,072-820)	837 (976-755)	n/a
ML	ugam	LG	RT1	UB1	1,198 (1,229-1,167)	1,065 (1,152-981)	964 (1,053-876)	n/a
ML	ugam	LG	RT2	UB1	964 (997-931)	872 (931-813)	799 (858-750)	n/a
ML	ugam	LG	RT3	UB1	1,021 (1,094-948)	918 (1,065-821)	836 (969-756)	n/a
CB	ln	CAT	RT0	UB1	1,012 (1,069-955)	833 (931-776)	n/a	759 (845-718)
CB	ln	CAT	RT1	UB1	1,189 (1,215-1,163)	951 (1,023-877)	n/a	854 (927-781)
CB$^{\#}$	ln	CAT	RT2	UB1	973 (1,001-945)	809 (855-772)	n/a	740 (785-717)
CB	ln	CAT	RT3	UB1	1,014 (1,072-957)	833 (934-775)	n/a	758 (846-718)
CB	ugam	CAT	RT0	UB1	993 (1,062-925)	851 (987-775)	n/a	776 (890-718)
CB	ugam	CAT	RT1	UB1	1,196 (1,226-1,166)	1,008 (1,094-922)	n/a	912 (1,000-825)
CB	ugam	CAT	RT2	UB1	955 (990-920)	824 (887-772)	n/a	753 (814-717)
CB	ugam	CAT	RT3	UB1	991 (1,058-924)	851 (984-776)	n/a	775 (896-718)
CB	ln	LG	RT0	UB1	1,022 (1,079-965)	836 (935-777)	n/a	760 (847-718)
CB	ln	LG	RT1	UB1	1,190 (1,217-1,163)	948 (1,021-876)	n/a	851 (925-779)
CB	ln	LG	RT2	UB1	977 (1,003-951)	808 (853-772)	n/a	739 (783-717)
CB	ln	LG	RT3	UB1	1,018 (1,075-962)	834 (932-776)	n/a	760 (846-718)
CB	ugam	LG	RT0	UB1	991 (1,059-923)	849 (986-774)	n/a	776 (898-718)
CB	ugam	LG	RT1	UB1	1196 (1,225-1,166)	1,006 (1,092-918)	n/a	910 (999-821)
CB	ugam	LG	RT2	UB1	955 (991-920)	823 (887-771)	n/a	753 (814-718)
CB	ugam	LG	RT3	UB1	992 (1,059-926)	850 (982-775)	n/a	775 (894-718)
ML	ln	CAT	RT0	UB0	972 (1,038-907)	837 (960-724)	756 (863-657)	n/a
ML	ln	CAT	RT1	UB0	1184 (1,207-1,162)	1,004 (1,077-929)	895 (965-823)	n/a
ML	ln	CAT	RT2	UB0	949 (988-910)	817 (890-736)	739 (809-665)	n/a
ML	ln	CAT	RT3	UB0	978 (1,039-916)	841 (961-739)	860 (864-668)	n/a
ML	ugam	CAT	RT0	UB0	943 (1,039-848)	846 (1,031-697)	774 (942-639)	n/a
ML	ugam	CAT	RT1	UB0	1,199 (1,230-1,168)	1,064 (1,151-978)	968 (1,058-876)	n/a
ML	ugam	CAT	RT2	UB0	937 (979-895)	840 (923-760)	768 (852-691)	n/a
ML	ugam	CAT	RT3	UB0	972 (1,049-895)	870 (1,035-764)	975 (948-693)	n/a
ML	ln	LG	RT0	UB0	983 (1,049-917)	845 (967-734)	757 (862-659)	n/a
ML	ln	LG	RT1	UB0	1,185 (1,208-1,162)	1,005 (1,078-931)	888 (957-815)	n/a
ML	ln	LG	RT2	UB0	953 (991-915)	823(896-740)	738 (808-662)	n/a
ML	ln	LG	RT3	UB0	987 (1,049-925)	852 (970-749)	763 (865-670)	n/a

ML	ugam	LG	RT0	UB0	953 (1,048-858)	854 (1,041-708)	777 (945-643)	n/a
ML	ugam	LG	RT1	UB0	1,199 (1,229-1,168)	1,066 (1,150-982)	965 (1,053-876)	n/a
ML	ugam	LG	RT2	UB0	937 (979-896)	840 (921-763)	765 (845-688)	n/a
ML	ugam	LG	RT3	UB0	977 (1,055-899)	876 (1,041-768)	796 (947-693)	n/a
CB	ln	CAT	RT0	UB0	968 (1,038-898)	792 (914-696)	n/a	721 (829-638)
CB	ln	CAT	RT1	UB0	1,190 (1,216-1,163)	952 (1,026-879)	n/a	854 (928-783)
CB	ln	CAT	RT2	UB0	944 (984-905)	775 (844-709)	n/a	707 (772-646)
CB	ln	CAT	RT3	UB0	974 (1,038-910)	797 (913-712)	n/a	725 (828-650)
CB	ugam	CAT	RT0	UB0	931 (1,019-842)	795 (960-663)	n/a	723 (872-605)
CB	ugam	CAT	RT1	UB0	1,196 (1,226-1,166)	1,007 (1,094-919)	n/a	910 (1,001-821)
CB	ugam	CAT	RT2	UB0	933 (975-892)	798 (879-720)	n/a	725 (807-651)
CB	ugam	CAT	RT3	UB0	963 (1,034-892)	822 (970-725)	n/a	747 (811-655)
CB	ln	LG	RT0	UB0	975 (1,045-905)	795 (916-697)	n/a	724 (833-640)
CB	ln	LG	RT1	UB0	1,190 (1,216-1,163)	948 (1,023-877)	n/a	852 (926-780)
CB	ln	LG	RT2	UB0	949 (988-911)	776 (843-708)	n/a	708 (772-647)
CB	ln	LG	RT3	UB0	983 (1,050-917)	800(919-711)	n/a	727 (832-650)
CB	ugam	LG	RT0	UB0	933 (1,020-846)	797 (958-668)	n/a	736 (873-609)
CB	ugam	LG	RT1	UB0	1,196 (1,226-1,166)	1,007 (1,094-922)	n/a	912 (1,001-825)
CB	ugam	LG	RT2	UB0	933 (975-892)	797 (878-721)	n/a	725 (506-652)
CB	ugam	LG	RT3	UB0	962 (1,032-892)	821 (966-725)	n/a	747 (881-657)

ML: Topology inferred by supermatrix analyses (Figure 3.1)
CB: Topology inferred by coalescence-based analyses (Figure 3.1)
#: this ultrametric tree has been used for illustrating the time-calibrated phylogeny results in Figure 3.3 and the ancestral state reconstruction in Figure 3.4.

Chapter 4 - The plastid genome in Cladophorales green algae is encoded by hairpin chromosomes[6]

Andrea Del Cortona*, Frederik Leliaert*, Kenny A. Bogaert, Monique Turmel, Christian Boedeker, Jan Janouškovec, Juan M. Lopez-Bautista, Heroen Verbruggen, Klaas Vandepoele, and Olivier De Clerck[7]

“One Ring to rule them all, One Ring to find them,

One Ring to bring them all, and in the darkness bind them”

John Ronald Reuel Tolkien -The Lord of the Rings

[6] Published as:
Andrea Del Cortona*, Frederik Leliaert*, Kenny A. Bogaert, Monique Turmel, Christian Boedeker, Jan Janouškovec, Juan M. Lopez-Bautista, Heroen Verbruggen, Klaas Vandepoele, and Olivier De Clerck – The plastid genome in Cladophorales green algae is encoded by hairpin chromosomes – Current Biology 27, 3771–3782, December 18, 2017, https://doi.org/10.1016/j.cub.2017.11.004
* Equal contribution.

[7] Authors contribution: conceptualization and methodology, F.L., M.T., C.B., H.V., K.V., and O.D.C.; resources, F.L., C.B., A.D.C., J.M.L.-B., K.V., and O.D.C.; formal analysis and investigation, A.D.C., F.L., K.A.B., J.J., K.V., and O.D.C.; writing – original draft & visualization, A.D.C., F.L., K.V., and O.D.C.; writing – review & editing, A.D.C., F.L., K.A.B., M.T., J.J., H.V., K.V., and O.D.C..

Abstract

Virtually all plastid (chloroplast) genomes are circular double-stranded DNA molecules, typically between 100-200 kb in size and encoding circa 80-250 genes. Exceptions to this universal plastid genome architecture are very few and include the dinoflagellates where genes are located on DNA minicircles. Here we report on the highly deviant chloroplast genome of Cladophorales green algae, which is entirely fragmented into hairpin chromosomes. Short and long read high-throughput sequencing of DNA and RNA demonstrated that the chloroplast genes of *Boodlea composita* are encoded on 1-7 kb DNA contigs with an exceptionally high GC-content, each containing a long inverted repeat with one or two protein-coding genes and conserved non-coding regions putatively involved in replication and/or expression. We propose that these contigs correspond to linear single-stranded DNA molecules that fold onto themselves to form hairpin chromosomes. The *Boodlea* chloroplast genes are highly divergent from their corresponding orthologs, and display an alternative genetic code. The origin of this highly deviant chloroplast genome likely occurred before the emergence of the Cladophorales, and coincided with an elevated transfer of chloroplast genes to the nucleus. A chloroplast genome that is composed only of linear DNA molecules is unprecedented among eukaryotes and highlights unexpected variation in the plastid genome architecture.

Introduction

Photosynthetic eukaryotes possibly originated 1.9 billion years ago following an endosymbiotic event in which a heterotrophic ancestor of the Archaeplastida engulfed a cyanobacterium that became stably integrated and evolved into a membrane-bound organelle, the plastid (Ponce-Toledo *et al.*, 2017; Sánchez-Baracaldo *et al.*, 2017). Following this primary endosymbiosis, an intricate history of plastid acquisition via eukaryote-eukaryote endosymbioses resulted in the spread of plastids to distantly related eukaryotic lineages (Keeling, 2010).

Plastids have retained a reduced version of the genome inherited from their cyanobacterial ancestor. A core set of genes involved in the light reactions of photosynthesis, ATP generation, and functions related to transcription and translation is typically retained (Green, 2011). Many genes have been lost or transferred to the nuclear genome and, as a result, plastids are dependent on nuclear-encoded, plastid-targeted proteins for the maintenance of essential biochemical pathways and other functions such as genome replication, gene expression, and DNA repair (Kleine *et al.*, 2009). Nearly all plastid genomes consist of a single circular-mapping chromosome, typically between 100-200 kb, encoding circa 80-250 genes (Green, 2011; Lang & Nedelcu, 2012). Diversity in size, gene content, density and organization of plastid genomes among different eukaryotic lineages is by and large limited, especially when compared to mitochondria (Simpson & Stern, 2002; Smith & Keeling, 2015; Muñoz-Gómez *et al.*, 2017).

While fragmented mitochondrial genomes evolved several times independently during the evolution of eukaryotes (Barbrook *et al.*, 2010; Smith & Keeling, 2015), fragmented plastid genomes are only known in dinoflagellates (Howe *et al.*, 2008) and a single green algal species (Watanabe *et al.*, 2016). In peridinin-containing dinoflagellates, the chloroplast genome is fragmented into DNA minicircles of 2-3 kb, most of which carry one gene only (Zhang *et al.*, 1999; Howe *et al.*, 2008). Larger minicircles of up to 12 kb have also been described (Nelson & Green, 2005), as well as minicircles containing two genes (Laatsch *et al.*, 2004), and ‘empty’ minicircles without genes (Hiller, 2001). The genes located on these minicircles mostly encode key components of the major photosynthetic complexes, including subunits of photosystems I and II, the cytochrome b6f complex, and ATP synthase, as well as rRNAs and a few tRNAs (Howe *et al.*, 2008). The only other alga with a fragmented chloroplast genome is the green alga *Koshicola spirodelophila*, but here the level of fragmentation is minor: the plastid genome is divided into three large circular chromosomes totalling 385 kb, with a gene content comparable to other green algae (Watanabe *et al.*, 2016). In addition, plastid minicircles that coexist with a conventional plastid genome have been observed in a few algae, including dinoflagellates with haptophyte-derived plastids (Espelund *et al.*, 2012) and the green alga *Acetabularia* (Green, 1976; Ebert *et al.*, 1985).

Although plastid genomes generally assemble as circular-mapping DNAs, they can take multiple complex conformations *in vivo*, including multigenomic, linear-branched structures with discrete termini (Bendich, 2007; Oldenburg & Bendich, 2016). The alveolate *Chromera velia* is the only known alga with a linear-mapping plastid genome with telomeric arrangement (Janouškovec *et al.*, 2013), and is also atypical in that several core photosynthesis genes are fragmented. Linear plastid genomes, however, may be more widespread, as several plastid genomes currently do not map as a circle (Gabrielsen *et al.*, 2011).

Currently, and in stark contrast to other algae (Turmel *et al.*, 2015; Leliaert *et al.*, 2016; Lemieux *et al.*, 2016; Muñoz-Gómez *et al.*, 2017), little is known about the gene content and structure of the chloroplast genome in the Cladophorales (Ulvophyceae), an ecologically important group of marine and freshwater green algae, which includes several hundreds of species. These macroscopic multicellular algae have giant, multinucleate cells containing numerous chloroplasts (Figure 1A-C). Most attempts to amplify common chloroplast genes have failed (Fučíková *et al.*, 2014), with only one highly divergent *rbcL* sequence published thus far, for *Chaetomorpha valida* (Deng *et al.*, 2013). An atypical plastid genome in the Cladophorales is suggested by the presence of abundant plasmid-like DNA that has been observed in the chloroplasts of several species (La Claire *et al.*, 1997; La Claire & Wang, 2000). These plasmids-like DNA molecules represent a Low Molecular Weight (LMW) DNA fraction, visible on agarose gels of total DNA extracts (Figure 1D). Pioneering work revealed that these structures are single-stranded DNA (ssDNA) molecules of 1.5-3.0 kb that fold in a hairpin configuration and lack sequence similarity to the nuclear DNA (La Claire *et al.*, 1998; La Claire & Wang, 2004). Some of the hairpin-like DNAs contain putatively transcribed sequences with similarity to chloroplast genes encoding subunits of Photosystems I and II (*psaB*, *psbB*, *psbC* and *psbF*) (La Claire *et al.*, 1998). Here, we describe intriguing features of the plastid genome of Cladophorales, focusing on

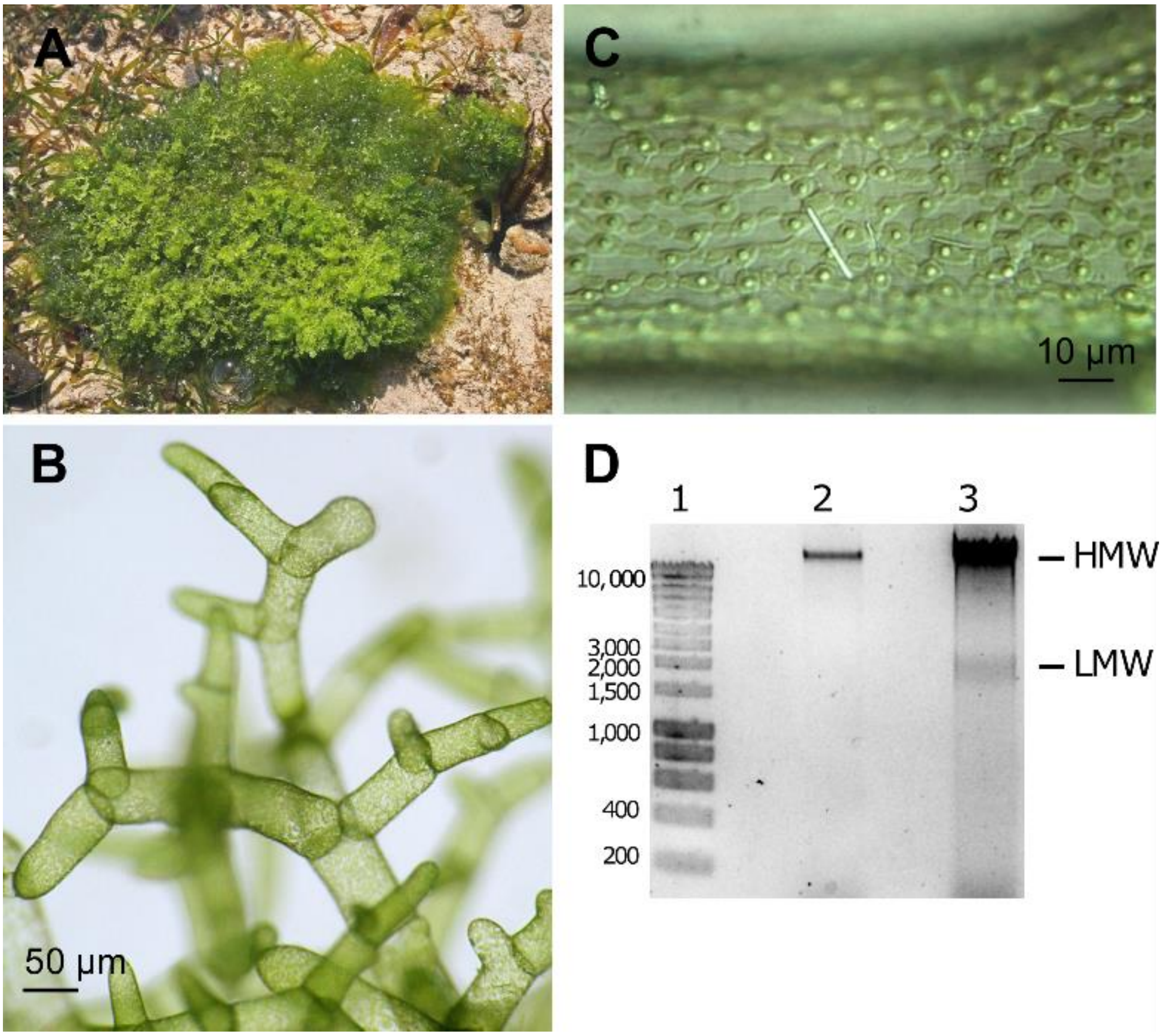


Figure 4.1: *Boodlea composita*.

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

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

Results and Discussion

DNA and RNA-seq data

Our reconstruction of the chloroplast genome of *Boodlea composita* is based on different high-throughput DNA sequencing methods (Figure S4.1, Materials and Methods). The choice of short read DNA sequencing of isolated intact chloroplasts (chloroplast-enriched fraction) using Roche 454 technology was based on comparable sequencing approaches in other plants and algae that successfully resulted in assembly of chloroplast genomes (Lemieux *et al.*, 2016). To overcome possible assembly artefacts in a hypothetical scenario of an inflated chloroplast genome bloated by repetitive elements, long-read sequencing of the High Molecular Weight (HMW) DNA fraction using Pacific Biosciences Single-Molecule Real-Time (SMRT) method was applied, while long read sequencing of the LMW DNA fraction allowed characterization of the previously observed plasmid-like DNA in the chloroplast (La Claire *et al.*, 1997; La Claire & Wang, 2000). To allow comparison of the results of *Boodlea* with other species of Cladophorales, we generated additional DNA sequence data from nine other species using Illumina HiSeq 2000 technology. Finally, two deep-coverage RNA-seq libraries, a total-RNA library and a mRNA library enriched for nuclear transcripts, were generated to confirm the transcription of genes, and to inform whether genes are nuclear versus plastid encoded.

A prodigious chloroplast genome with reduced gene set

Assembly of the chloroplast-enriched DNA reads generated using Roche 454 technology did not result in a typical circular chloroplast genome. Instead, 21 chloroplast protein-coding genes were found on 58 short contigs (1,203-5,426 bp): *atpA*, *atpB*, *atpH*, *atpI*, *petA*, *petB*, *petD*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbJ*, *psbK*, *psbL, psbT* and *rbcL*. All but the *rbcL* gene code for components of the major thylakoid transmembrane protein complexes (ATP synthase, cytochrome b6f, and photosystems I and II). The contigs contained inverted repeats at their termini and, despite high coverage by sequence reads, they could not be extended by iterative contig extension. Sequence similarity searches and a metagenomic binning approach (distribution analysis of 4-mers) demonstrated that the inverted repeats were also found on contigs with no sequence similarity to known proteins, raising the number of contigs of chloroplast origin to 136. These contigs are further referred to as "chloroplast 454 contigs". The length distribution of the chloroplast 454 contigs was consistent with the size of the LMW DNA fraction as estimated by agarose gel electrophoresis of *Boodlea* genomic DNA (Figures 4.1D, S4.6A, S4.6B).

The failure to assemble a circular chloroplast genome might be due to repetitive elements that impair the performance of short-read assemblers (Miller *et al.*, 2010). Inflated chloroplast genomes bloated by repetitive elements have been documented in several green algae (Smith & Lee, 2009; Brouard *et al.*, 2010; de Vries *et al.*, 2013). To overcome assembly artefacts and close putative gaps in the chloroplast 454 contigs, we applied Single-Molecule Real-Time (SMRT) sequencing (Pacific Biosciences) to the HMW and LMW DNA fractions. Only 22 HMW DNA reads (ca. 0.044 %) harboured protein-coding genes commonly present in chloroplast genomes of Archaeplastida (Figure 4.2A). All but three of these genes (*psbA*, *psbB* and *psbC*, which likely correspond to carry-over LMW DNA) contained spliceosomal introns, were absent in the chloroplast 454 contigs, and revealed a high ratio between mapped mRNA and total-RNA reads, altogether suggesting that they are encoded in the nucleus (Figure S4.2A). Conversely, 22 chloroplast genes (that is, the 21 protein-coding genes identified in chloroplast 454 contigs as well as the 16S rRNA gene) were found in the LMW DNA reads (Figure 4.2A). An orthology-guided assembly, where the chloroplast 454 contigs harbouring protein-coding genes guided the assembly of LMW DNA reads with sequence similarity to chloroplast genes, resulted in 34 contigs between 1,179 and 6,925 bp in length, henceforth referred to as "chloroplast genome" (Figure 4.2B, Table S4.1).

Four contigs of the *Boodlea* chloroplast genome (contigs 10, 19, 32 and 33) display long palindromic sequences that include full-length coding sequences (CDSs), and a less conserved tail region (Figure 4.2B). The remaining contigs have similar palindromic structures but appear to be not completely assembled. Such palindromes allow regions of the single-stranded LMW DNA molecules to fold into hairpin-like secondary structures. Additional smaller inverted repeats were identified in many of the contigs (Figure 4.2B), which may result in more complex secondary structures.

Chloroplast 454 contigs could not be scaffolded with long HMW DNA reads, nor did an hybrid assembly between chloroplast 454 contigs and long HMW DNA reads

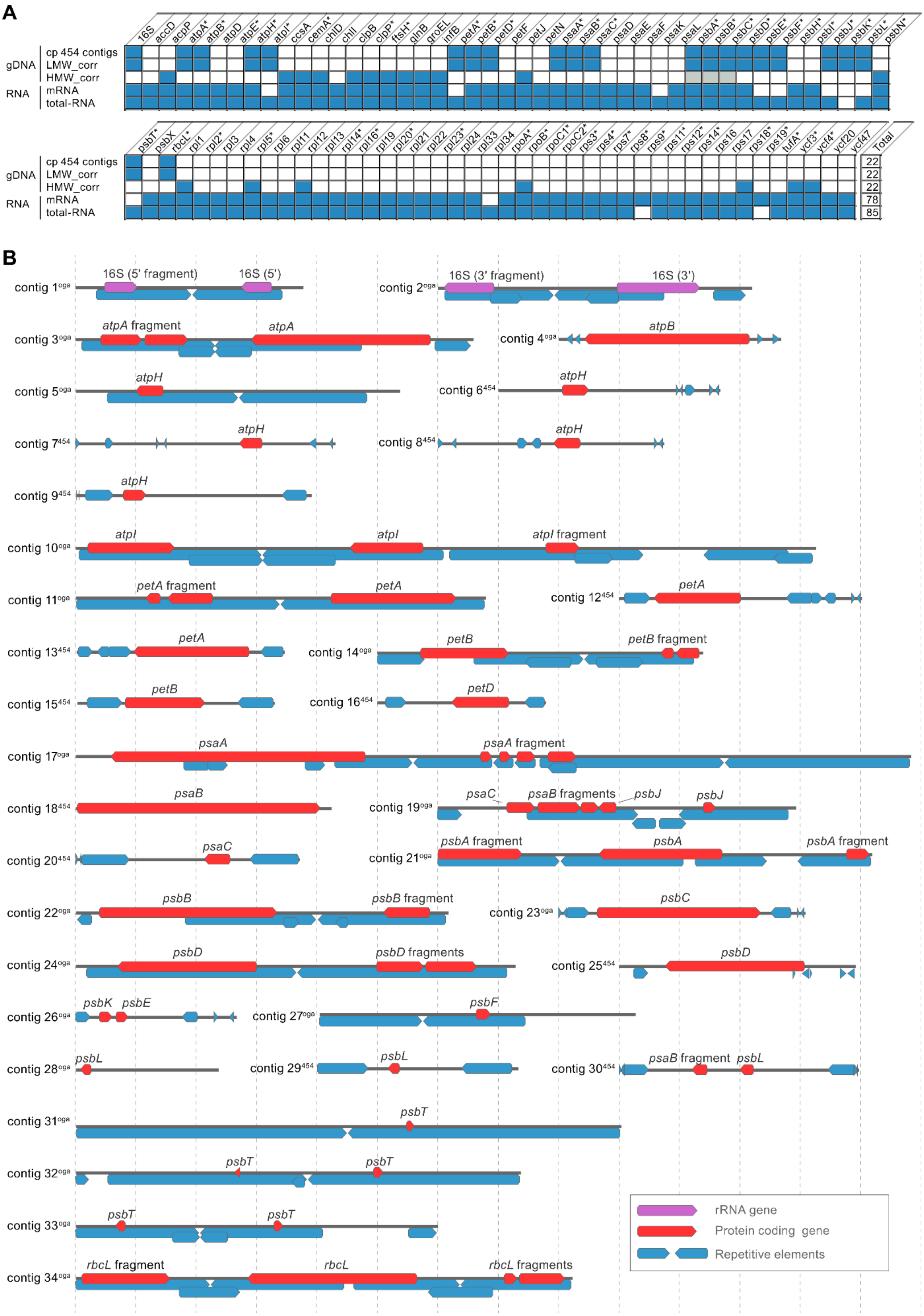

A
gDNA
cp 454 contigs
LMW_corr
HMW_corr
RNA
mRNA
total-RNA
16S accD acpP atpA* atpB* atpD atpE* atpH* atpI* ccsA cemA* chlD chlI clpB clpP* ftsH* glnB groEL infB petA* petB* petD* petF petJ petN psaA* psaB* psaC* psaD psaE psaF psaK psaL psbA* psbB* psbC* psbD* psbE* psbF* psbH* psbI* psbJ* psbK* psbL* psbN*
psbT* psbX rbcL* rpl1 rpl2* rpl3 rpl4 rpl5* rpl6 rpl11 rpl12 rpl13 rpl14* rpl16* rpl19 rpl20* rpl21 rpl22 rpl23* rpl24 rpl33 rpl34 rpoA* rpoB* rpoC1* rpoC2* rps3* rps4* rps7* rps8* rps9* rps11* rps12* rps14* rps16 rps17 rps18* rps19* tufA* ycf3* ycf4* ycf20 ycf47 Total
22
22
22
78
85
B
contig 1oga 16S (5' fragment) 16S (5')
contig 2oga 16S (3' fragment) 16S (3')
contig 3oga atpA fragment atpA
contig 4oga atpB
contig 5oga atpH
contig 6454 atpH
contig 7454 atpH
contig 8454 atpH
contig 9454 atpH
contig 10oga atpI atpI atpI fragment
contig 11oga petA fragment petA
contig 12454 petA
contig 13454 petA
contig 14oga petB petB fragment
contig 15454 petB
contig 16454 petD
contig 17oga psaA psaA fragment
contig 18454 psaB
contig 19oga psaC psaB fragments psbJ psbJ
contig 20454 psaC
contig 21oga psbA fragment psbA psbA fragment
contig 22oga psbB psbB fragment
contig 23oga psbC
contig 24oga psbD psbD fragments
contig 25454 psbD
contig 26oga psbK psbE
contig 27oga psbF
contig 28oga psbL
contig 29454 psbL
contig 30454 psaB fragment psbL
contig 31oga psbT
contig 32oga psbT psbT
contig 33oga psbT psbT
contig 34oga rbcL fragment rbcL rbcL fragments
rRNA gene
Protein coding gene
Repetitive elements

Figure 4.2: Schematic representation of *Boodlea* chloroplast genome.

(A) Distribution of *Boodlea* genes having orthologs in the chloroplast of other Archaeplastida. gDNA (genomic DNA): chloroplast (cp) 454 contigs, HMW and LMW corrected reads; RNA: mRNA and total-RNA assemblies. Asterisks (*) indicate "core" chloroplast genes, i.e. protein-coding genes conserved between chloroplast genomes of Chlorophyta (see Experimental Procedures). The following nine "core" chloroplast genes were not found in any of the *Boodlea* libraries sequenced: *atpF*, *petG*, *petL*, *psaJ*, *psbM*, *psbZ*, *rpl36*, *rps2* and *ycf1*. Grey cells denote putative LMW DNA read contaminants as suggested by the ratios of HMW to LMW DNA reads and mRNA to total-RNA reads (Figures S2B and S2C). (B) Overview of the 34 contigs representing the *Boodlea* chloroplast genome. Purple arrows indicate rRNA genes, red arrows indicate CDSs of protein-coding genes, and blue arrows indicate repetitive elements. For each contig, repetitive elements with similar length indicate similar sequences. Distance between vertical grey lines in the background represents 500 bp. Oga: contig obtained by orthology-guided assembly. 454: chloroplast 454 contig.

generate a circular chloroplast genome (Figure S4.1, Material and Methods). The LMW DNA reads are concordant and consistent with the palindromic sequences of the assembled chloroplast genome, indicating that the palindromes are not a result of assembly artefacts (Figure 4.3, Table S4.1). As a consequence, we conclude that the chloroplast genome is not a single large molecule but that it is instead fragmented in several molecules in the LMW DNA.

A chloroplast genome that is entirely fragmented into hairpin chromosomes is in line with earlier observations of abundant LMW DNA in chloroplasts of several species of Cladophorales (La Claire *et al.*, 1997). The hairpin configuration of the chromosomes derived from our sequence data corresponds with earlier data based on electron microscopy, endo- and exonuclease digestion experiments, acridine orange staining, and denaturing gel electrophoresis (La Claire *et al.*, 1997; La Claire & Wang, 2004). Fluorescence *in situ* hybridization, and Southern blot hybridisation indicated that these plasmid-like DNA molecules are present within the chloroplast only (La Claire & Wang, 2000), supporting the congruence between chloroplast 454 contigs and sequences from the LMW fraction (Figure 4.3, Figure S4.7A).

The chloroplast genome contigs of *Boodlea* feature an exceptionally high GC-content, ranging from 54 to 60 % in the gene-containing contigs (average 57%) (Table S4.1). These values are concordant with the high density of the LMW fraction observed in CsCl/bisbenzimide gradients (La Claire *et al.*, 1997), and also with sequence data from cloned plasmids of *Ernodesmis* (51-59% GC) (La Claire *et al.*, 1998). Plastid genomes are generally AT-rich, and in green algal species, GC-content typically ranges between 26 and 43% (Leliaert *et al.*, 2012; Muñoz-Gómez *et al.*, 2017). GC-rich plastid genomes are very rare, but higher values have been reported for the trebouxiophycean green algae *Coccomyxa subellipsoidea*, *Paradoxia multiseta* (both 51% GC), and Trebouxiophyceae sp. MX-AZ01 (58%) (Turmel *et al.*, 2015). These species, however, feature standard plastid genome architectures.

The size of the *Boodlea* chloroplast genome could not be estimated by inspection of k-mer frequency distributions of the reads in the 454 library, nor from those of the uncorrected and corrected LMW DNA reads (Hozza *et al.*, 2015). Histograms of k-mer frequency distributions revealed several small peaks, indicating a heterogeneous population of molecules present in different stoichiometries, and the signal to noise ratio was too small to make a comfortable estimation of the sizes (Figure S4.3). The cumulative length of the 34 *Boodlea* chloroplast genome contigs is 91 kb (Table S4.3). However, if we would consider the large and heterogeneous population of LMW DNA reads bearing no similarity to protein-coding genes ("empty" hairpin chromosomes, see below) as part of the chloroplast genome, its size could be regarded as much larger.

The largest known circular-mapping chloroplast genomes have been documented in the red algae *Bulboplastis apyrenoidosa* (610 kb) and *Corynoplastis japonica* (1.127 Mb), where the genomes are bloated by group II introns and include transposable elements of possible bacterial origin (Smith & Keeling, 2015). Within green algae, expanded chloroplast genomes have been reported in two distinct clades: the Chlorophyceae and the Ulvophyceae. Inflation of the 521-kb chloroplast genome of *Floydiella terrestris* (Chlorophyceae) resulted mainly from the proliferation of dispersed, heterogeneous repeats (>30 bp) in intergenic regions, representing more than half of the genome length (Brouard *et al.*, 2010). Intergenic regions of the *Volvox carteri* (Chlorophyceae) chloroplast genome, instead, are populated with short palindromic repeats (average size of 50 bp) that constitute ca. 64% of the predicted 525-kb genome (Smith & Lee, 2009). The mechanisms by which such palindromic selfish DNA spread throughout the *Volvox* chloroplast genome are not clear, but the presence of a reverse transcriptase and endonuclease may point toward retrotranscription (Burt & Trivers, 2006; Smith & Lee, 2009). For *Acetabularia acetabulum* (Ulvophyceae), the chloroplast genome was sequenced only partially and

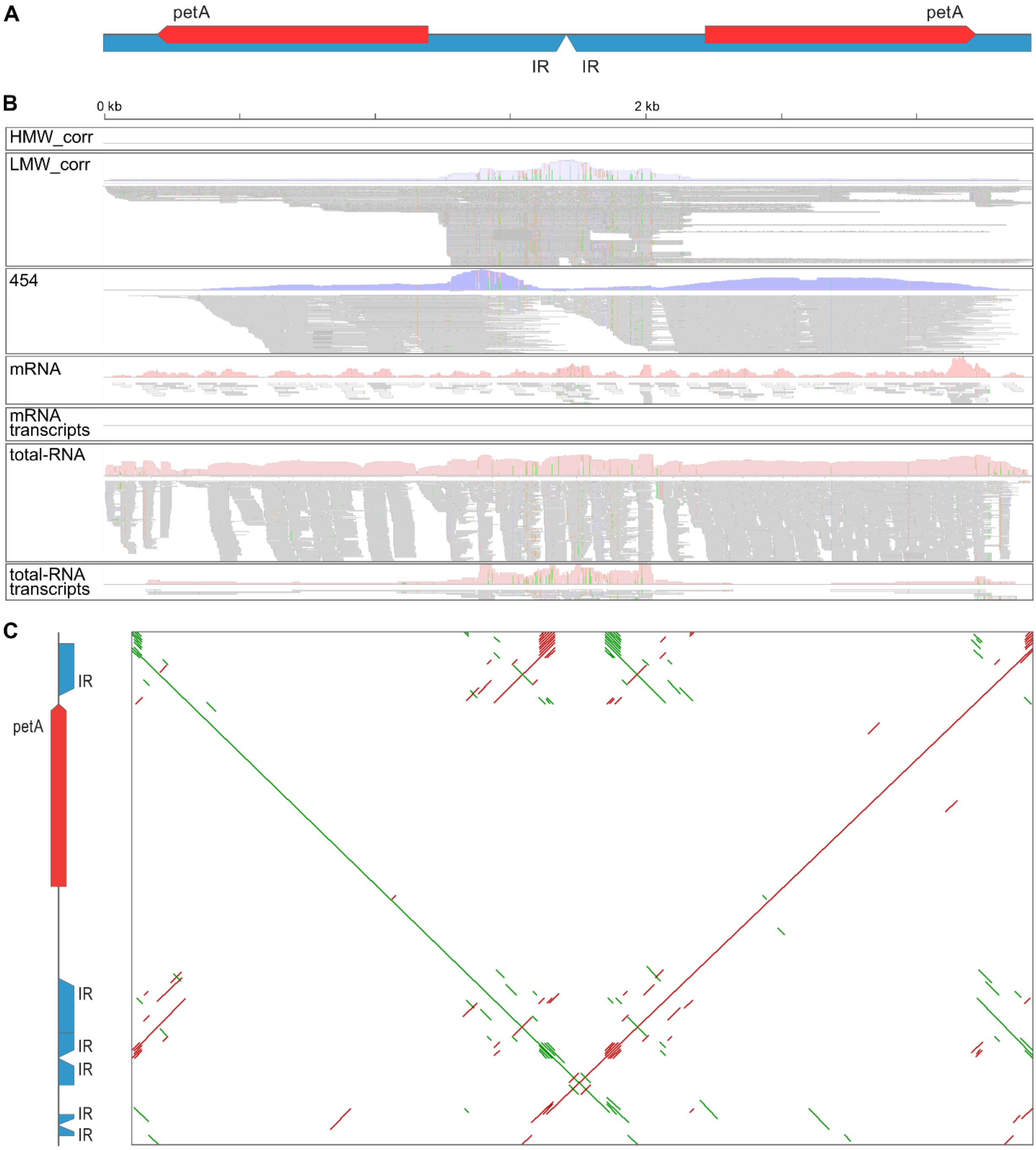


Figure 4.3: LMW DNA reads containing chloroplast genes are expressed, enriched in the total-RNA fraction and congruent to the respective chloroplast 454 contigs.

(A) Representation of *petA* LMW DNA read (3,398 bp). The red arrows indicate CDSs, the blue arrows indicate inverted repeats. (B) Corresponding Genome Browser track, from top to bottom: corrected HMW DNA coverage [0], corrected LMW DNA read coverage [range 0-541], 454 read coverage [range 0-567], mRNA library read coverage [range 0-17], assembled mRNA transcripts mapped [0], total-RNA library read coverage [range 0-7,936], and assembled total-RNA transcripts mapped [range 0-17]. (C) Dotplot showing congruence between *petA* LMW DNA read (x axis) and the corresponding *petA*-containing chloroplast 454 contig (y axis, 2,012 bp). Green lines indicate similar sequences; red lines indicate sequences similar to the respective reverse complements.

its size was estimated to exceed 1 Mb; it has exceptionally long intergenic regions and features long repetitive elements (>10 kb) arranged in tandem (Tymms & Schweiger, 1985; de Vries *et al.*, 2013). The *Boodlea* chloroplast genome is rich as well in non-coding DNA, constituting 92.2% of the 136 chloroplast 454 contigs and 72.8% of the assembled chloroplast genome, comparable to that in inflated chloroplast genomes of other green algae (*Floydiella terrestris,* 82.1%; *Volvox carteri,* ca. 80%; *Acetabularia acetabulum,* ca. 87% of the sequenced chloroplast genome) (Smith & Lee, 2009; de Vries *et al.*, 2013).

The non-coding DNA regions (ncDNA) of the hairpins showed high sequence similarity among one another (52.5-100% sequence similarity). Within the ncDNA, we identified six conserved motifs, 20 to 35 bp in length and with a GC-content ranging from 36 to 84%, which lack similarity to known regulatory elements (Figure 4.4). Motifs 1, 2 and 5 were always present upstream of the start codon of the chloroplast genes, occasionally in more than one copy. Although their distances from the start codon were variable, their orientations relative to the gene were conserved, indicating a potential function as a regulatory element of gene expression and/or replication of the hairpin chromosomes.

These motifs were also present in 1,966 (ca. 1.8 %) LMW DNA reads lacking genes. This observation supports earlier findings of abundant non-coding LMW DNA molecules in the Cladophorales (La Claire *et al.*, 1997; La Claire *et al.*, 1998). In contrast, a very small fraction of the HMW DNA reads (15 corrected reads) displayed the same ncDNA motifs and these were present exclusively on long terminal repeat retrotransposons (RT-LTRs) (Figures S4.2D, S4.2E). RT-LTRs were also abundant in the 454 contigs (Figure S6D). The abundance of RT-LTRs in the 454 contigs and the presence of ncDNA motifs in both the *Boodlea* chloroplast genome and nuclear RT-LTRs is suggestive of DNA transfer between the nucleus and chloroplast and may allude to the origin of the hairpin chromosomes. Hypothetically, an invasion of nuclear RT-LTRs in the chloroplast genome may have resulted in an expansion of the chloroplast genome and its subsequent fragmentation into hairpin chromosomes during replication. Chloroplast genome fragmentation could be caused by recombination between repetitive elements and displacement of the palindromic sequences from the lagging strand during the chloroplast genome replication (Ellis & Day, 1986; Bikard *et al.*, 2010), and it is consistent with the expectation that recombination and cleavage of repetitive DNA will produce a heterogeneous population of molecules, as observed in dinoflagellates plastid genomes (Howe *et al.*, 2008), and in the *Boodlea* LMW DNA.

A fragmented chloroplast genome is a common feature of Cladophorales

DNA sequence data were obtained from 9 additional Cladophorales species, representing the main lineages of the order: *Chaetomorpha aerea*, *Cladophora albida*, *C. socialis*, *C. vadorum*, *Dictyosphaeria cavernosa*, *Pithophora* sp., *Siphonocladus tropicus*, *Struvea elegans*, and *Valonia utricularis* (Tables S4.2 and S4.4). Although comparable sequencing approaches resulted in the assembly of circular chloroplast genomes for other algae, including green seaweeds (Leliaert & Lopez-Bautista, 2015; Marcelino *et al.*, 2016), only short chloroplast contigs (ca. 200-8,000 bp) were assembled from these libraries, similar to *Boodlea composita*. Interestingly, a similar set of chloroplast genes was identified in all sequenced Cladophorales species (Table S4.5). In contrast to the genes found in the *Boodlea* hairpin chromosomes, however, most of the chloroplast genes identified in the additional Cladophorales libraries were fragmented, possibly due to assembly of the shorter Illumina reads (Table S4.3). These findings support the idea that fragmentation of the chloroplast genome occurred before or early in the evolution of the Cladophorales.

Highly divergent chloroplast genes

The 21 chloroplast protein-coding genes of *Boodlea* and the other species of Cladophorales display extremely high sequence divergence compared to orthologous genes in other photosynthetic organisms (Figure 4.5). A maximum likelihood phylogenetic tree based on a concatenated amino acid alignment of 19 chloroplast genes from Archaeplastida and Cyanobacteria species (Figure 4.5) shows that despite their high divergence, the Cladophorales sequences form a monophyletic group within the core Chlorophyta (Figure S4.4), a position that is supported by phylogenetic analyses of nuclear genes (Cocquyt *et al.*, 2010b). The high sequence divergence of chloroplast genes in the Cladophorales supports the notion that organellar genomes with extremely derived architectures, including those of peridinin-containing

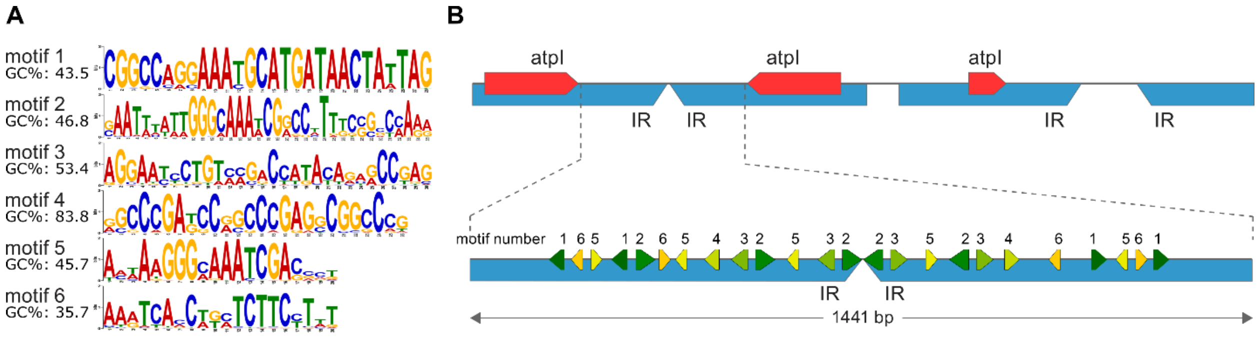
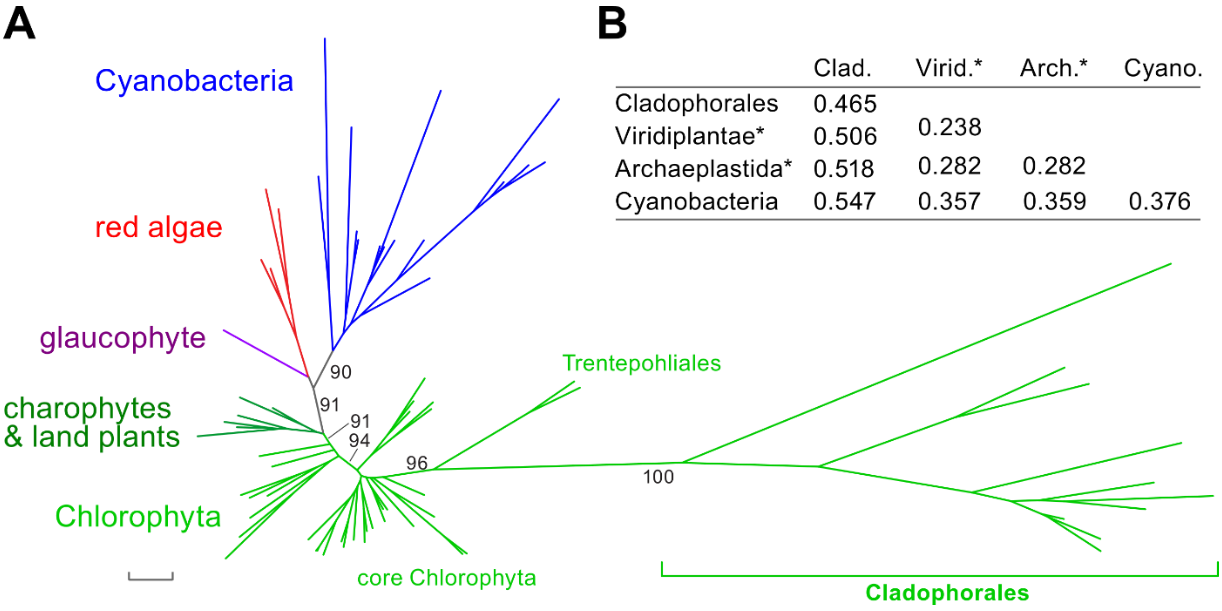


Figure 4.4: Conserved non-coding motifs in *Boodlea* LMW DNA.

(A) Sequence logos and GC contents of the conserved motifs predicted in the *Boodlea* chloroplast genome. The relative sizes of the nucleotides indicate their frequency in the sequences. (B) Schematic representation of the distribution of the motifs in the 1,441 bp ncDNA region from the *atpI* group A read used for the identification of additional chloroplast reads in the LMW DNA library. Motifs with conserved orientation relative to the downstream genes are represented by green arrows, while motifs without conserved orientation to the downstream genes are represented by yellow arrows. CDSs are represented by red arrows, inverted repeats are represented by blue arrows.

dinoflagellates, also tend to fall at the extreme ends of the range observed at the mutation rate (or gene sequence divergence) level (Zhang *et al.*, 2000; Simpson & Stern, 2002). For some *Boodlea* chloroplast genes, the identification of start and stop codons was uncertain and a non-canonical genetic code was identified (Figure 4.6). The canonical stop codon UGA was found 11 times internally in six genes (*petA*, *psaA*, *psaB*, *psaC*, *psbC* and *rbcL*), but was also present as a genuine termination codon in several genes, *petA* and *psaA* included. At seven of these 11 positions, the corresponding amino acid residue in orthologous genes was conserved (i.e. present in more than 75% of the taxa in the alignment), but different amino acids were observed at these positions: V, S, I, L and C (Figures 4.6A and 4.6B). The reassignment of the stop codon UGA to C has been documented in the nuclear genetic code of several species of ciliates (Heaphy *et al.*, 2016). For the remaining positions, the amino acid in the alignment was not conserved, and therefore the amino acid coded by the UGA codon could not be determined with certainty.

Deviations from the universal genetic code are widespread among mitochondrial genomes, and include loss of start and stop codons in some groups, including



	Clad.	Virid.*	Arch.*	Cyano.
Cladophorales	0.465			
Viridiplantae*	0.506	0.238		
Archaeplastida*	0.518	0.282	0.282	
Cyanobacteria	0.547	0.357	0.359	0.376

Figure 4.5: *Boodlea* chloroplast genes have large sequence divergence.

(A) Maximum likelihood phylogenetic tree, with indication of relevant bootstrap values (see also Figure S4). The scale represents 0.5 substitution per amino acid position. (B) Maximum pairwise amino acid sequence distances of the concatenated amino acid alignment within and between clades (* excluding Cladophorales).

dinoflagellates (Slamovits *et al.*, 2007; Howe *et al.*, 2008; Waller & Jackson, 2009). In contrast, non-canonical genetic codes are much rarer in plastid genomes, and up to now have only been detected in the apicomplexans *Neospora caninum* (Lang-Unnasch & Aiello, 1999), *Chromera velia* (Janouškovec *et al.*, 2013), and the dinoflagellate *Lepidodinium chlorophorum* (Matsumoto *et al.*, 2011). In genomes of primary plastids, a non-canonical genetic code is unprecedented.

Dual meaning of UGA as both stop and sense codons has recently been reported from a number of unrelated protists (Heaphy *et al.*, 2016; Swart *et al.*, 2016; Zahonova *et al.*, 2016). While in *Saccharomyces cerevisiae*, the tetranucleotide UGA-C allows increased incorporation of the near-cognate Cys-tRNA for the UGA premature termination codon (Beznoskova *et al.*, 2016), such preference was not observed in the *Boodlea* chloroplasts. Importantly, a non-canonical genetic code has also been described for Cladophorales nuclear genes, where UAG and UAA codons are reassigned to glutamine (Cocquyt *et al.*, 2010a), which implies two independent departures from the standard genetic code in a single organism.

Unexpectedly, we found that the 16S rRNA gene in the *Boodlea* chloroplast genome is split across two distinct hairpin chromosomes and that its size is much smaller compared to its algal and bacterial homologs and (Figures 4.2B and 4.7). Fragmentation of rRNA genes has been observed in organellar genomes, including Apicomplexa, dinoflagellates, and many green algae (Barbrook *et al.*, 2010). In general, fragmentation of protein-coding and rRNA genes is more common in mitochondrial genomes than in plastid genomes (Smith *et al.*, 2010; Espelund *et al.*, 2012; Janouškovec *et al.*, 2013). Despite considerable effort, we could not detect the 23S rRNA gene nor the 5S rRNA gene.

The transcription of the aberrant chloroplast genes was confirmed using RNA-seq, and is concordant with previous results of Northern blots (La Claire *et al.*, 1998). Transcripts of 21 chloroplast genes (that is, 20 protein-coding genes as well as the 16S rRNA gene) were identical to the genes encoded by the chloroplast 454 contigs (Figure 4.2A; 4.3 and S4.5), providing evidence for the absence of RNA editing and corroborating the use of a non-canonical genetic code (Figure S4.5). Lack of RNA editing was also evidenced for the 11 internal occurrences of UGA (Figure S4.5). This observation, in combination with conservation of the sequence after the UGA codon, serves as evidence that it is not a termination codon but an alternative code. The high total-RNA

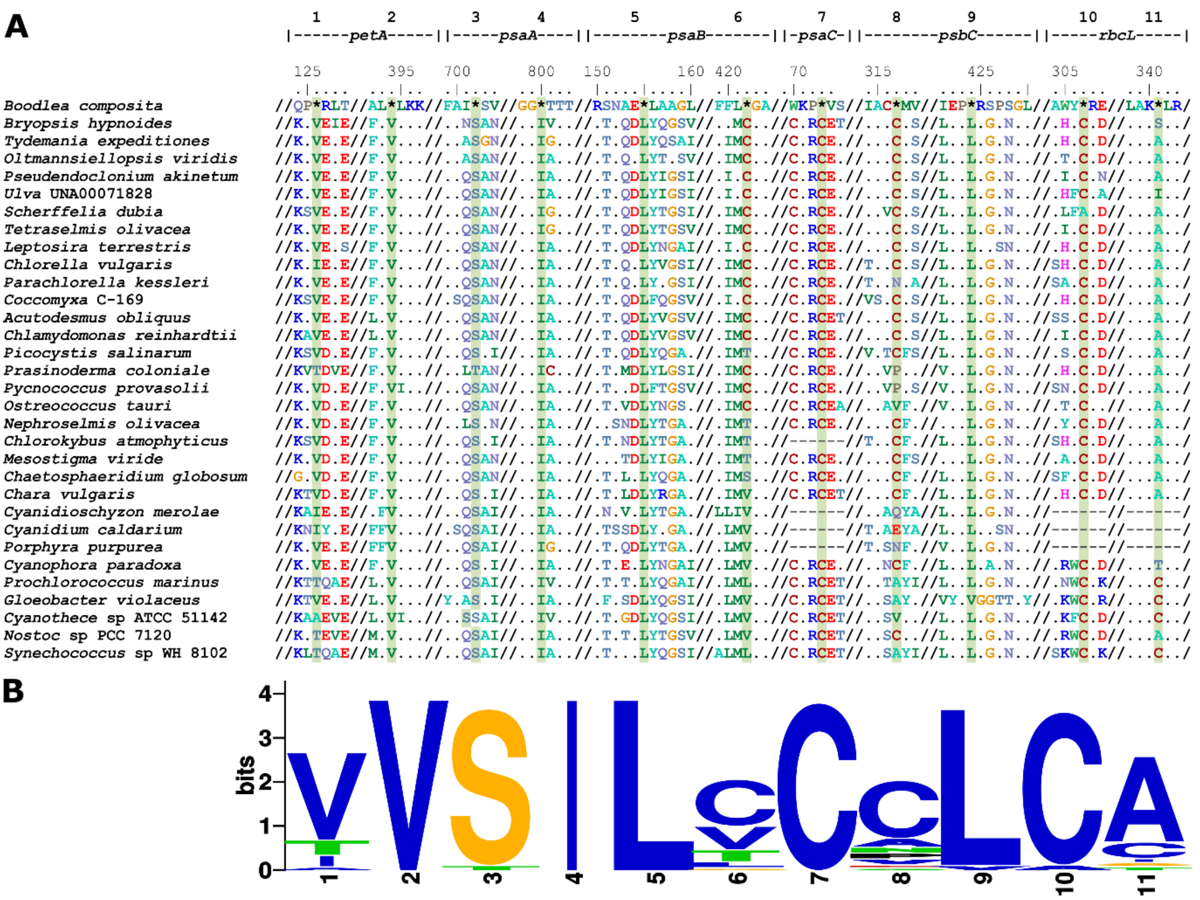

A
1 2 3 4 5 6 7 8 9 10 11
|------petA-----||-----psaA----||--------psaB-------||-psaC-||--------psbC------||-----rbcL-----|
125 395 700 800 150 160 420 70 315 425 305 340
Boodlea composita //QP*RLT//AL*LKK//FAI*SV//GG*TTT//RSNAE*LAAGL//FFL*GA//WKP*VS//IAC*MV//IEP*RSPSGL//AWY*RE//LAK*LR/
Bryopsis hypnoides //K.VEIE//F.V...//..NSAN//..IV..//.T.QDLYQGSV//..MC..//C.RCET//...C.S//L..L.G.N..//.H.C.D//...S../
Tydemania expeditiones //K.VE.E//F.V...//..ASGN//..IG..//.T.QDLYQSAI//.IMC..//C.RCE.//...C.S//L..L.G.N..//.H.C.D//...A../
Oltmannsiellopsis viridis //K.VE.E//F.V...//..ASAN//..IA..//.T.Q.LYT.SV//.IMC..//C.RCE.//...C.S//L..L.G.N..//.T.C.D//...A../
Pseudendoclonium akinetum //K.VE.E//F.V...//..QSAN//..IA..//.T.QDLYIGSI//.I.C..//C.RCE.//...C.S//L..L.G.N..//.I.C.N//...A../
Ulva UNA00071828 //K.VE.E//F.V...//..QSAN//..IA..//.T.QDLYIGSI//.I.C..//C.RCE.//...C.S//L..L.G.N..//.HFC.A//...I../
Scherffelia dubia //KSVE.E//F.V...//..QSAN//..IG..//.T.QDLYTGSI//.IMC..//C.RCE.//..VC.S//L..L.G.N..//.LFA.D//...A../
Tetraselmis olivacea //K.VE.E//F.V...//..QSAN//..IG..//.T.QDLYTGSV//.IMC..//C.RCE.//...C.S//L..L.G.N..//.I.C.D//...A../
Leptosira terrestris //K.VE.S//F.V...//..QSAN//..IA..//.T.QDLYNGAI//.I.C..//C.RCE.//...C.S//L..L..SN..//.H.C.D//...A../
Chlorella vulgaris //K.IE.E//F.V...//..QSAN//..IA..//.T.Q.LYVGSI//.IMC..//C.RCE.//T..C.S//L..L.G.N..//SH.C.D//...A../
Parachlorella kessleri //K.VE.E//F.V...//..QSAN//..IA..//.T.Q.LY.GSI//.IMC..//C.RCE.//T..N.A//L..L.G.N..//SA.C.D//...A../
Coccomyxa C-169 //KSVE.E//F.V...//.SQSAN//..IA..//.T.QDLFQGSV//.I.C..//C.RCE.//VS.C.S//L..L.G.N..//.H.C.D//...A../
Acutodesmus obliquus //K.VE.E//L.V...//..QSAN//..IA..//.T.QDLYVGSV//.IMC..//C.RCET//...C.S//L..L.G.N..//SS.C.D//...A../
Chlamydomonas reinhardtii //KAVE.E//L.V...//..QSAN//..IA..//.T.QDLYVGSV//.IMC..//C.RCE.//...C.S//L..L.G.N..//.I.C.D//...A../
Picocystis salinarum //KSVD.E//F.V...//..QS.I//..IA..//.T.QDLYQGA.//.IMT..//C.RCE.//V.TCFS//L..L.G.N..//.S.C.D//...A../
Prasinoderma coloniale //KVTDVE//F.V...//..LTAN//..IC..//.T.MDLYLGSI//.IMC..//C.RCE.//..VP..//V..L.G.N..//.H.C.D//...A../
Pycnococcus provasolii //K.VD.E//F.VI..//..QSAN//..IA..//.T..DLFTGSV//.IMC..//C.RCE.//..VP.S//V..L.G.N..//SN.C.D//...A../
Ostreococcus tauri //K.VD.E//F.V...//..QSAN//..IA..//.T.VDLYNGS.//.IMC..//C.RCEA//..AVF.//V..L.G.N..//.T.C..//...A../
Nephroselmis olivacea //K.VD.E//F.V...//..LS.N//..IA..//..SNDLYTGA.//.IMT..//C.RCE.//...CF.//...L.G.N..//.Y.C.D//...A../
Chlorokybus atmophyticus //KSVD.E//F.V...//..QS.I//..IA..//.T.NDLYTGA.//.IMT..//------//T..CF.//L..L.G.N..//SH.C.D//...A../
Mesostigma viride //K.VD.E//F.V...//..QSAN//..IA..//...TDLYIGA.//.IMT..//C.RCE.//...CFS//...L.G.N..//.A.C.D//...A../
Chaetosphaeridium globosum //G.VD.E//F.V...//..QSAN//..IA..//.T.L.LYQGA.//.IMS..//C.RCE.//...CF.//L..L.G.N..//SF.C.D//...A../
Chara vulgaris //KTVD.E//F.V...//..QS.I//..IA..//.T.LDLYRGA.//.IMV..//C.RCET//...CF.//L..L.G.N..//.H.C.D//...A../
Cyanidioschyzon merolae //KAIE.E//.FV...//..QSAI//..IA..//.N.V.LYTGA.//LLIV..//------//..AQYA//L..L.G.N..//------//------/
Cyanidium caldarium //KNIY.E//FFV...//.SQSAI//..IA..//.TSSDLY.GA.//.LMV..//------//T.AEYA//L..L..SN..//------//------/
Porphyra purpurea //K.VE.E//FFV...//..QSAI//..IG..//.T.QDLYTGA.//.LMV..//------//T.SNF.//V..L.G.N..//------//------/
Cyanophora paradoxa //K.VE.E//F.V...//..QS.I//..IA..//.T.E.LYNGAI//.LMV..//C.RCE.//..NCF.//L..L.A.N..//.RWC.D//...T../
Prochlorococcus marinus //KTTQAE//L.V...//..QSAI//..IV..//.T.T.LYQGAI//.LML..//C.RCET//..TAYI//L..L.G.N..//.NWC.K//...C../
Gloeobacter violaceus //KTVE.E//L.V...//Y.AS.I//..IA..//.F.SDLYQGSI//.LMV..//C.RCET//..SAY.//VY.VGGTT.Y//.KWC.R//...C../
Cyanothece sp ATCC 51142 //KAAEVE//L.VI..//..SSAI//..IV..//.T.GDLYQGSI//.LMV..//C.RCET//..SV..//L..L.G.N..//.KFC.D//...C../
Nostoc sp PCC 7120 //K.TEVE//M.V...//..QSAI//..IA..//.T.T.LYTGSV//.LMV..//C.RCET//..SC..//L..L.G.N..//.RWC.D//...A../
Synechococcus sp WH 8102 //KLTQAE//M.V...//..QSAI//..IA..//.T...LYQGSI//ALML..//C.RCET//..SAYI//L..L.G.N..//SKWC.K//...C../
B
bits
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11

Figure 4.6: Non-canonical genetic code in *Boodlea* chloroplast genes.

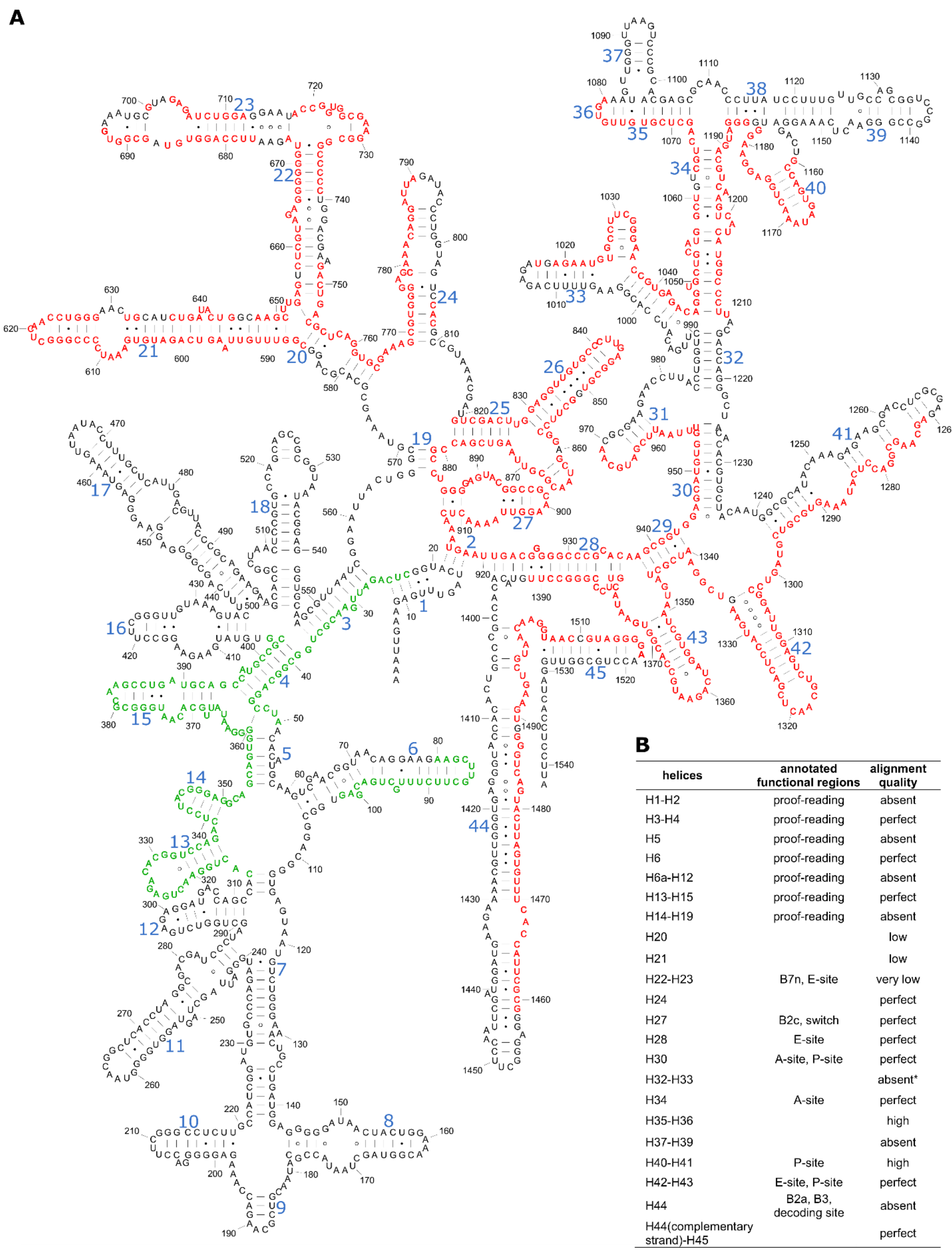
Boodlea chloroplast protein-coding genes were aligned with the respective orthologs of 43 Archaeplastida and 14 Cyanobacteria. (A) Relevant parts of amino acid sequence alignment for six chloroplast genes of *Boodlea* and representatives of Archaeplastida and Cyanobacteria. Positions corresponding to UGA codons in *Boodlea* are indicated by an asterisk. Slashes represent regions of the sequence alignment that were omitted for simplicity. Dots indicate amino acid identity with the top-most sequence. For each gene, position in the alignment is indicated by the numbers shown above the sequence alignment. The numbers below the gene names indicate the eleven positions where UGA was identified as premature termination codon in the six *Boodlea* genes. (B) Sequence logo of the Position Weight Matrix reporting the relative amino acid frequencies in the alignment for each premature termination UGA position in *Boodlea*.

to mRNA ratio observed for reads that mapped to the chloroplast 454 contigs confirmed that these genes were not transcribed in the nucleus (Figure S4.6C). All coding sequences of the same protein-coding genes found on different contigs of the *Boodlea* chloroplast genome were expressed, despite minor differences in their nucleotide sequences (Table S4.1).

Additional transcripts of 66 genes that have been located in the chloroplast in other Archaeplastida were identified (Figure 4.2A). Although their subcellular origin was not determined experimentally, they are probably all nuclear-encoded, based on high mRNA to total-RNA reads ratio and their presence on High Molecular Weight (HMW) DNA reads.

Conclusions

We collected several lines of evidence indicating that *Boodlea composita* lacks a typical large circular chloroplast genome. The chloroplast genome is instead fragmented into multiple linear hairpin chromosomes, and has a highly reduced gene repertoire compared to other chloroplast genomes. Thirty-four hairpin chromosomes were identified, harbouring 21 protein-coding genes and the 16S rRNA gene, which are highly divergent in sequence compared to orthologs in other algae, and display an alternative genetic code. The exact set of *Boodlea* chloroplast genes remains elusive, but at least 19 genes coding for chloroplast products appear to be nuclear-encoded, of which nine are always chloroplast-encoded in related green algae (Figure 4.2A). This suggests that fragmentation of a conventional chloroplast genome in the Cladophorales has been accompanied with an elevated transfer of genes to the nucleus, similarly to the situation in peridinin-containing dinoflagellates (Howe *et al.*, 2008), with plastid genomes encoding about 12 genes or less (Howe *et al.*, 2008; Barbrook *et al.*, 2014). Notably, the two distantly related algal groups have converged on a very similar gene distribution: chloroplast genes code only for the subunits of photosynthetic complexes (and also for Rubisco in *Boodlea*), whereas the expression machinery appears to be fully nucleus-encoded (Figure 4.2A). Other nonstandard chloroplast genome architectures have recently been observed, such as a monomeric linear chromosome in the alveolate microalga *Chromera velia* (Janouškovec *et al.*, 2013) and three circular chromosomes in the green alga *Koshicola spirodelophila* (Watanabe *et al.*, 2016), but these represent relatively small deviations from the paradigm, when compared to the chloroplast genome of the Cladophorales. The highly fragmented chloroplast genome in the Cladophorales is wholly unprecedented and will be of significance to understanding processes driving organellar genome fragmentation and gene reduction, endosymbiotic gene transfer, and the minimal functional chloroplast gene set.



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

Figure 4.7: The *Boodlea* chloroplast 16S rRNA is fragmented and reduced compared to its algal and bacterial homologs.

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

Acknowledgments

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

Material and Methods

Experimental model and subject details

Clonal cultures of *Boodlea composita* FL1110, *Chaetomorpha aerea* UTEX799, *Cladophora albida* Calb2, *Cladophora socialis* Csoc2, *Cladophora vadorum* Cvad2, *Dictyosphaeria cavernosa* FL1134, *Pithophora* sp. UTEX787, *Siphonocladus tropicus* Siph3, *Struvea elegans* Sele1, *Valonia utricularis* Vutric3 and *Valonia ventricosa* UTEX 2260 are maintained in the algal culture collection of the Phycology Research Group, Ghent University. The specimens were grown in enriched sterilized natural seawater at 22°C under 12:12 (light:dark) cool white fluorescent light at 60 μmol photons m^{-2} s^{-1}. To prepare the enriched natural seawater, 20 mL of enriched solution is added to 980 mL of filtered and sterilized natural seawater. The enriched solution consists of: Tris base 5.0 g/L; $NaNO_3$ 3.5 g/L; Na_2 β-glycerophosphate · H_2O; $Na_2EDTA \cdot 2\ H_2O$ 0.529 g/L; $Fe(NH_4)_2(SO_4)_2 \cdot 6\ H_2O$ 0.176 g/L; $FeCl_3 \cdot 6\ H_2O$ 12.1 mg/L; H_3BO_3 0.286 g/L; $MnSO_4 \cdot 4\ H_2O$ 40.6 mg/L; $ZnSO_4 \cdot 7\ H_2O$ 5.5 mg/L; $CoSO_4 \cdot 7\ H_2O$ 1.2 mg/L; Thiamine–HCl 0.5 mg/L; Biotin 5.0 mg/L; Cyanocobalamin 10.0 mg/L (Andersen, 2005).

Genomic DNA sequencing

Total genomic DNA from fresh *Boodlea* cultures was isolated by using a modified CTAB extraction protocol (Doyle & Doyle, 1987). Briefly, 100 mg of fresh algal material was blotted dry on paper, placed inside a 1.5 ml test tube and immediately frozen in liquid nitrogen. Samples were ground with a pestle that fits the 1.5 mL tubes and resuspended in 500 μL of CTAB isolation buffer (2% w/v cetyltrimethylammonium bromide, 1.4 M NaCl, 100 mM Tris-HCl pH 8.0, 20 mM EDTA pH 8.0, 1% w/v polyvinylpyrrolidone) with 5 μL of Proteinase K (QUIAGEN, Germany). The samples were then incubated at 60°C for 40 min. After 30 min, 5 μL of RNAse A (QUIAGEN) was added to each sample. Cellular debris were spun down and the aqueous layer was extracted first with phenol:chloroform:isoamylic alcohol (25:24:1 v/v) and then with chloroform:isoamylic alcohol (24:1 v/v). Genomic DNA was precipitated with the addition of two volumes of ice-cold absolute ethanol and 0.3 M of sodium acetate pH 5.5 to each sample and overnight incubation at -20°C. The genomic DNA was washed with ice-cold 70% ethanol, air-dried and dissolved in 50 μL TE buffer (10 mM Tris-HCl pH 8.0, 1 mM Na_2-EDTA). HMW and LMW DNA bands were size-selected using a BluePippin™ system (Sage Science, USA). The HMW DNA band was isolated with a cut-off range of 10 kb to 50 kb, while the LMW DNA band was isolated with a cut-off range of 1.5 kb to 2.5 kb. The quantity, quality and integrity of the extracted DNA were assessed with Qubit (ThermoFisher Scientific, USA), Nanodrop spectrophotometer (ThermoFisher Scientific), and Bioanalyzer 2100 (Agilent Technologies, USA).

Intact chloroplasts were isolated from living *Boodlea* cells following the protocol of Palmer et al. (Palmer, 1982). In short, 200 g of *Boodlea* filaments were placed in 400 ml of ice-cold isolation buffer (0.35 M sorbitol, 50 mM Tris-HCl pH 8.0, 5 mM EDTA, 0.1 % BSA, 1.5 mM β-mercaptoethanol), homogenized in a blender at 4 ˚C, and filtered through miracloth (Calbiochem). The filtrate was centrifuged at 1000 g for 15 min at 4 ˚C, the supernatant was poured off, and the pellet resuspended in 8 mL of ice-cold wash buffer (0.35 M sorbitol, 50 mM Tris-HCl pH 8.0, 25 mM EDTA). The resuspended pellet was loaded on a step gradient consisting of 18 mL of 52% w/v sucrose, over-layered with 7 mL of 30% w/v sucrose, and centrifuged at 25,000 rpm for 40 min at 4 ˚C. The chloroplast band was removed from the 30%-52% interface using a Pasteur pipette, diluted with 6 volumes of wash buffer, centrifuged at 1,500 g for 15 min at 4 ˚C, and resuspended in wash buffer to a final volume of 10 mL. This fraction of isolated chloroplasts is further referred to as “chloroplast-enriched fraction”. DNA from the chloroplast-enriched fraction was sequenced with Roche 454 GS FLX at GATC Biotech, Germany. The HMW and LMW DNA fractions were sequenced on two SMRT cells on a PacBio RS II (VIB Nucleomics Core facilities, Leuven, Belgium) using PacBio P5 polymerase and C3 chemistry combination (P5-C3). For the HMW DNA fraction, a 20-kb SMRT-bell library was constructed, while for the LMW DNA fraction, a 2-kb SMRT-bell library was constructed.

Chloroplast DNA assembly and annotation

Quality of the reads from the 454 library was assessed with FastQC v.0.10.1 (http://www.bioinformatics.babraham.ac.uk, last accessed March 01, 2017) (Table S4.2). Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' ends of the reads were trimmed with Fastx v.0.0.13 (https://github.com/agordon/fastx_toolkit, last accessed March 01, 2017). After trimming, reads shorter than 50 bp were discarded. *De novo* assembly of the trimmed reads was performed with MIRA v. 4.0rc5 (Chevreux *et al.*, 2004). The assembly resulted in 3,735 contigs, which will be further referred to as “454 contigs” (Table S4.3). Length distribution of the 454 contigs is reported in Figure S4.6B.

After the assembly, putative chloroplast contigs were identified by comparing their translated sequences against the NCBI non-redundant protein database using BLAST 2.2.29+ (Boratyn *et al.*, 2013), resulting in the identification of 58 contigs harbouring fragments or full-length chloroplast genes by sequence similarity search. These contigs had long stretches of conserved repetitive sequences at their 5' and 3' extremities. The conserved inverted repeats were used in a sequence similarity search with high stringency (high mismatch cost, high cost for gap opening and gap extension, long minimal word-size) against the 454 contigs to identify 18 additional contigs of putative chloroplast origin. This initial set of 76 contigs had a mean coverage of 84×, ranging between 11× and 191×. 17 of the 76 contigs had internal inverted repeats, with a sudden drop in read coverage. These contigs were regarded as chimeric contigs and were cleaved at the sites of coverage drop, raising the number of contigs of chloroplast origin to 89.

Additional chloroplast contigs without similarity to protein-coding genes were identified by metagenomic binning (distribution analysis of 4-mers) with MyCC (Lin & Liao, 2016), resulting in 21 clusters of 454 contigs (Figure S4.6D). The initial set of 89 chloroplast 454 contigs was present in three neighbouring clusters: Cluster 14, Cluster 17 and Cluster 21. These clusters contained 122, eight and six contigs, respectively, raising the number of identified chloroplast contigs assembled from the chloroplast-enriched fraction from 89 to 136 (“chloroplast 454 contigs” in Table S4.3). Of these, 71 contigs had no sequence similarity to known protein-coding genes, 29 contigs harboured only full-length chloroplast genes, 29 contigs harboured only fragments of chloroplast genes, and 7 contigs harboured both fragments and full-length CDSs of different chloroplast genes.

Contigs potentially coding for chloroplast tRNAs and rRNAs were identified using Infernal 1.1 (Nawrocki & Eddy, 2013). The chloroplast 454 contigs served as seeds for iterative contig extension with PRICE 1.0.1 (Ruby *et al.*, 2013). Single-end 454 reads were used as false paired-end reads with expected insert size equal to the median length of the 454 reads. 141 different combinations of parameters were tested in order to optimize the contig extension. None of the selected assemblies showed a length improvement for the initial set of chloroplast 454 contigs. The length distribution of the chloroplast 454 contigs was consistent with the size of the LMW DNA fraction as estimated by agarose gel electrophoresis of *Boodlea* genomic DNA (Figure 4.1D, Figure S4.6B).

Repetitive regions in the contigs were identified with ‘einverted’, ‘etandem’ and ‘palindrome’ from the EMBOSS 6.5.7 (Rice *et al.*, 2000) package. Dotplots for all contigs were generated with YASS v. 1.14, using standard parameters (Noé & Kucherov, 2005). Coverage of the chloroplast 454 contigs was evaluated by mapping the 454 reads, the mRNA and the total-RNA libraries to these contigs with CLC Genomics Workbench 7.0 (Qiagen) (Figure S4.6C).

The chloroplast 454 contigs were used together with HMW DNA reads for two independent hybrid assemblies. First, we tried to close hypothetical gaps between the chloroplast 454 contigs with the pbahaScaffolder.py script integrated in the smrtanalysis 2.3.0 pipeline (Bashir *et al.*, 2012). Secondly, the pre-assembled chloroplast 454 contigs were used as anchors for HMW DNA reads in a round of hybrid assembly with dbg2olc (Ye *et al.*, 2016). These analyses failed to close the hypothetical gaps between the short chloroplast 454 contigs and did not yield longer contigs. These results stand in stark contrast to the mitochondrial 454 contigs, where the same approaches yielded markedly longer contigs (Figures S4.2F and S4.2G).

Since the hybrid assemblies with uncorrected reads could not reconstruct a circular chloroplast genome, HMW DNA reads were further characterized after error correction. The high-noise HMW DNA reads were corrected by applying a hybrid correction with proovread 2.12 (Hackl *et al.*, 2014) using 454 reads and reads from Illumina RNA-seq libraries (see below). Corrected reads encoding chloroplast genes were identified by aligning them against a custom protein database, named Chloroprotein_db, including genes from the pico-PLAZA protein database (Vandepoele *et al.*, 2013) and protein-coding genes from published green algal chloroplast genomes (Chlorophyta sensu Bremer 1985, NCBI Taxonomy id: 3041).

LMW DNA reads were self-corrected with the PBcR pipeline (Koren *et al.*, 2013). Since the LMW DNA size is unknown and PBcR requires an estimate of the genome size for proper read correction, six different putative genome sizes were tested (100 kb, 1 Mb, 2.24 Mb, 10 Mb, 100 Mb). The best performance in terms of number of corrected reads was obtained by the combination of "10 Mb" for the estimated genome size and the *–sensitive* flag turned on; these corrected reads were used for the downstream analysis. After error correction, the number of reads was reduced from 154,852 to 106,428 (Table S4.2), with a similar length distribution as the uncorrected reads library (Figure S4.6A).

In order to estimate the *Boodlea* chloroplast genome size, k-mer frequency distributions were calculated with jellyfish 2.0 (Marçais & Kingsford, 2011). K-mers ranging from 11 toto 47 were analyzed for uncorrected and corrected LMW DNA reads, for the filtered 454 reads and for the 454 reads that could be mapped on the chloroplast 454 contigs (Figure S4.3).

De novo genome assembly of corrected LMW DNA reads was performed with the Celera WGS assembler version 8.3rc2 (Berlin *et al.*, 2015). The resulting assembly, hereafter called the Celera Assembly, consisted of 558 contigs (Table S4.3). Corrected and uncorrected reads as well as assembled contigs potentially encoding chloroplast genes were identified by aligning them with BLAST 2.2.29+ against Chloroprotein_db. In order to identify additional short protein-coding genes, HMM profiles were generated from alignments of chloroplast genes present in Chloroprotein_db and used to search the 6-frame translations of 454 contigs and corrected and uncorrected HMW and LMW DNA reads with HMMer3 (Eddy, 2011). To prevent assembly artefacts caused by repetitive elements and palindromic sequences, we also performed an orthology-guided assembly, in which the LMW DNA reads harbouring chloroplast CDSs were re-assembled together with the respective chloroplast 454 contigs. First, LMW DNA corrected reads and chloroplast 454 contigs were grouped according to their best BLAST hit. The corrected reads and contigs belonging to the same group were assembled using Geneious v. 8.1.7 (Biomatters, http://www.geneious.com/, last accessed March 01, 2017) with parameters "High Sensitivity/Medium", and each assembly (or lack of assembly) was visually screened to exclude potential chimeric contigs (e.g. palindromic corrected subreads should be collapsed in the same locus rather than being concatenated). Where possible, LMW DNA reads and chloroplast 454 contigs were assembled as larger molecules (Figure S4.7). The orthology-guided assembly yielded 21 contigs, 2 belonging to group A, 15 to group B and 4 to group E (Table S4.1). Two groups of reads could not be assembled into longer molecules, and for them, the corresponding chloroplast 454 contigs were retained. Eleven additional chloroplast 454 contigs were retained (Group E), since they were not congruent with the LMW DNA reads and could not be included in the assembly. This resulted in a total of 32 contigs containing chloroplast protein-coding genes, which together with the two later identified Group B contigs encoding the 16S rRNA gene, are regarded as the *Boodlea* chloroplast genome contigs (Figure 4.2B, Table S4.3).

Protein-coding genes in the *Boodlea* chloroplast genome contigs were identified with a sequence similarity search against the NCBI non-redundant protein database with BLAST 2.2.29+. Their annotation was manually refined in Geneious and Artemis 16.0.0 (Rutherford *et al.*, 2000) based on the BLAST search results. rRNAs were identified using Infernal 1.1 (Nawrocki & Eddy, 2013). Repetitive elements were mapped on the *Boodlea* chloroplast genome by aligning the contigs with themselves using BLAST 2.2.29+. Non-coding RNAs were identified with infernal 1.1 (cut-off value 10^{-5}). Conserved motifs were predicted with

MEME suite (Bailey *et al.*, 2009), and the discovered motifs were clustered with RSAT (Medina-Rivera *et al.*, 2015). The motifs were compared with the JASPAR-2016 (Mathelier *et al.*, 2013) database using TOMTOM (Gupta *et al.*, 2007) (p-value cut-off 10^{-3}).

Boodlea chloroplast genome coverage was evaluated by mapping the 454 reads with gsnap v.2016-04-04 (Wu *et al.*, 2016). Corrected and uncorrected LMW DNA subreads and chloroplast 454 contigs resulting from the MIRA assembly were mapped against the *Boodlea* chloroplast genome with gmap v. 2014-12-06 (Wu *et al.*, 2016) using the –*nosplicing* flag. Due to the high number of repetitive sequences in LMW DNA reads and 454 contigs, the resulting annotated *Boodlea* chloroplast genome was carefully inspected in order to exclude sequencing and assembly artefacts.

Completeness of the chloroplast genome was evaluated by comparing the annotated chloroplast genes to a set of 60 "core" chloroplast protein-coding genes, defined as protein-coding genes conserved among the chloroplast genomes of the following representative species of Chlorophyta: *Bryopsis plumosa*, *Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Coccomyxa subellipsoidea*, *Gonium pectorale*, *Leptosira terrestris*, *Nephroselmis olivacea*, *Oltmannsiellopsis viridis*, *Parachlorella kessleri*, and *Tupiella akineta,* and the streptophyte *Mesostigma viride* (Figure 4.3A).

RNA sequencing

Total RNA was isolated using a modified CTAB extraction protocol (Le Bail *et al.*, 2008). RNA quality and quantity were assessed with Qubit and Nanodrop spectrophotomete, and RNA integrity was assessed with a Bioanalyzer 2100. Two cDNA libraries for NextSeq sequencing were generated using TruSeq™ Stranded RNA sample preparation kit (Illumina, USA): one library enriched in poly(A) mRNA due to oligo-(dT) retrotranscription and one total RNA library depleted in rRNAs with Ribo-Zero Plant kit (Epicentre, USA). The two libraries were sequenced on one lane of Illumina NextSeq 500 Medium platform at 2x76 bp by VIB Nucleomics Core facilities (Leuven, Belgium) (Table S4.2).

Transcriptome assembly and annotation

Quality of the reads from the two RNA-seq libraries was assessed with FastQC. Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' ends of the reads were trimmed with Fastx. After trimming, reads shorter than 30 bp were discarded. Read normalization and *de novo* assembly of the libraries were performed with Trinity 2.0.4 (Grabherr *et al.*, 2011). The resulting contigs (hereafter, transcripts) were compared using sequence similarity searches against the NCBI non-redundant protein database using Tera-BLAST™ DeCypher (Active Motif, USA). Taxonomic profiling of the transcripts was performed using the following protocol: for each transcript, sequence similarity searches were combined with the NCBI Taxonomy information of the top ten BLAST hits in order to discriminate between eukaryotic and bacterial transcripts, or transcripts lacking similarity to known protein-coding genes (Table S4.3). Transcripts classified as "eukaryotic" were further examined to assess transcriptome completeness and to identify chloroplast transcripts. These transcripts were analysed using Tera-BLAST™ DeCypher against Chloroprotein_db. Transcriptome completeness was evaluated with a custom Perl script that compared gene families identified in the *Boodlea* transcriptome to a set of 1,816 "core" gene families shared between Chlorophyta genomes present in pico-PLAZA 2.0 (Vandepoele *et al.*, 2013), following Veeckman et al. guidelines to estimate the completeness of the annotated gene space (Veeckman *et al.*, 2016) (mRNA 1,741; total-RNA 1,724 out of 1,816 core gene families identified respectively).

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

Cladophorales genomic DNA sequencing

Sequence data were obtained from 9 additional Cladophorales species, representing the main lineages of the order (Tables S4.2 and S4.4). Total genomic DNA was extracted using a modified CTAB extraction protocol as described above, and sequenced using Illumina HiSeq 2000 technology (2×100 bp paired-end reads) on 1/5th of a lane by Cold Spring Harbor Laboratory (Cold Spring Harbor, NY, USA). Quality of the reads from the sequenced libraries was assessed with FastQC 0.10.1. Low-quality reads (average Phred quality score below 20) were discarded and low-quality 3' ends of the reads were trimmed with Fastx 0.0.13 toolkit. After trimming, reads shorter than 50 bp were discarded. Trimmed reads were assembled with CLC Genomics Workbench, MIRA and SPAdes 3.6.2 (Bankevich *et al.*, 2012).

The taxonomic profiling of the contigs was performed with the following protocol: for each contig, sequence similarity searches were combined with the NCBI Taxonomy ID's of the top ten BLAST hits in order to discriminate between eukaryotic and bacterial contigs and contigs with no similarity to known proteins ("NoHit"). Contigs classified as eukaryotic were further analysed to identify chloroplast contigs with a sequence similarity search using Tera-BLAST™ (DeCypher, www.timelogic.com) against Chloroprotein_db. After chloroplast contig identification, the assembly that allowed the reconstruction of the highest number of full-length chloroplast genes was retained. An overview of the assembly metrics is reported in Table S4.3.

Phylogenetic analysis.

Phylogenetic analysis was based on a concatenated alignment of 19 chloroplast protein-coding genes (*atpA*, *atpB*, *atpH*, *atpI*, *petA*, *petB*, *petD*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE, psbF*, *psbJ*, *psbL*, and *rbcL*) from *Boodlea*, nine other Cladophorales species, 41 additional species of Archaeplastida, and 14 Cyanobacteria species (Table S4.6). For each gene, DNA sequences were translated to amino acid sequences and aligned using ClustalW in Geneious using the BLOSUM weight matrix, with gap open penalty 10 and gap extension penalty 0.1. The 19 alignments were concatenated and poorly aligned positions were removed using Gblocks server (Talavera & Castresana, 2007), using the least stringent settings, resulting in an amino acid alignment of 5,704 positions. A maximum likelihood (ML) phylogenetic tree was inferred from the amino acid alignment using RAxML with the cpREV + Γ model (Stamatakis, 2014). Branch support was assessed by bootstrapping with 500 replicates. Phylogenetic analysis was run on the CIPRES Science Gateway v3.3 (Miller *et al.*, 2011).

Data and software availability

DNA and RNA sequence data have been deposited to the NCBI Sequence Read Archive as BioProject PRJNA384503. The annotated chloroplast and mitochondrial contigs of *Boodlea composita* were deposited to GenBank under accession numbers MG257795 – MG257880. Chloroplast genes from additional Cladophorales species were made available on Mendeley Data (http://dx.doi.org/10.17632/7dyphg7pbk.1). Phylogenetic data (sequence alignments, analyses and phylogenetic tree) were deposited in TreeBase under accession number 21737 (http://purl.org/phylo/treebase/phylows/study/TB2:S21737).

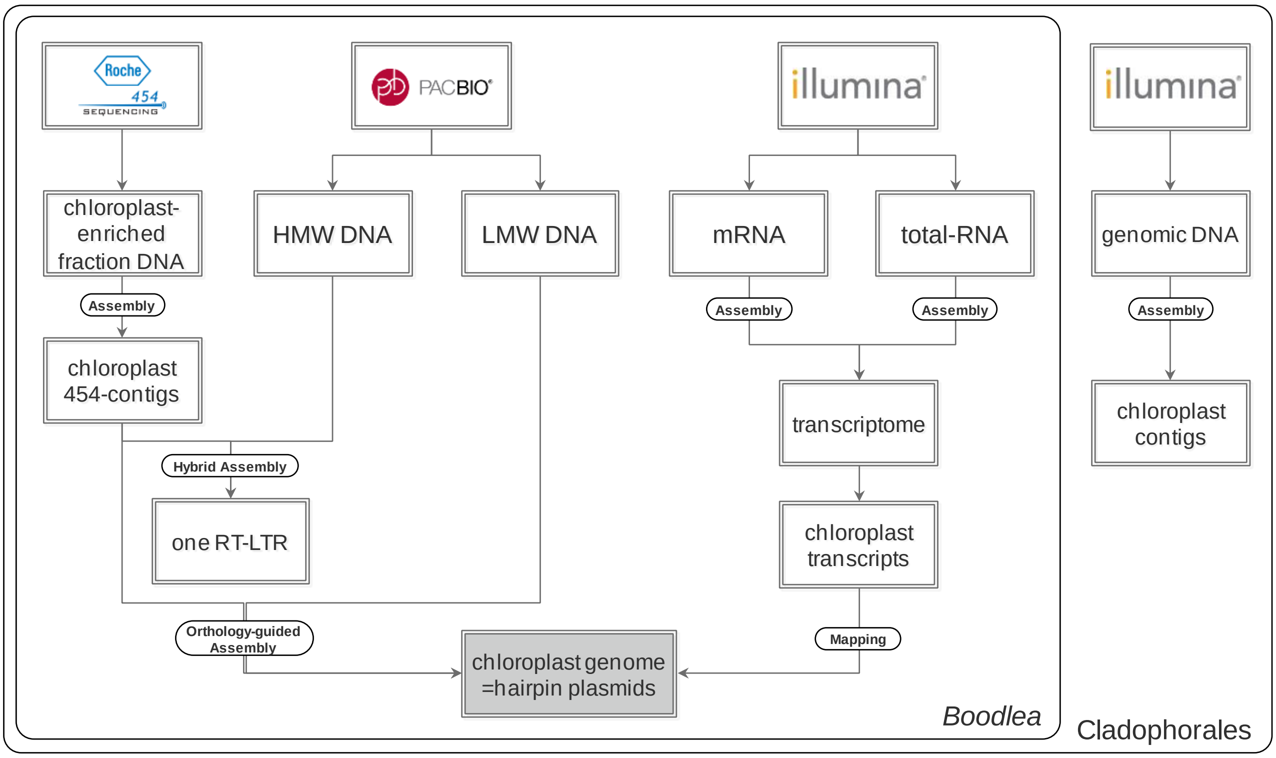


Figure S4.1: Sequence datasets generated in this study and their main analyses.

The diagram in the inner box “*Boodlea*” describes the workflow used to characterize the chloroplast genome structure and organization. The datasets represented in the outer box “Cladophorales” were used for phylogenetic inference and confirmation that the chloroplast genome of the entire order Cladophorales is distributed over hairpin plasmids. HMW DNA: High Molecular Weight DNA; LMW DNA: Low Molecular Weight DNA, RT-LTR: Long Terminal Repeat Retrotransposon.

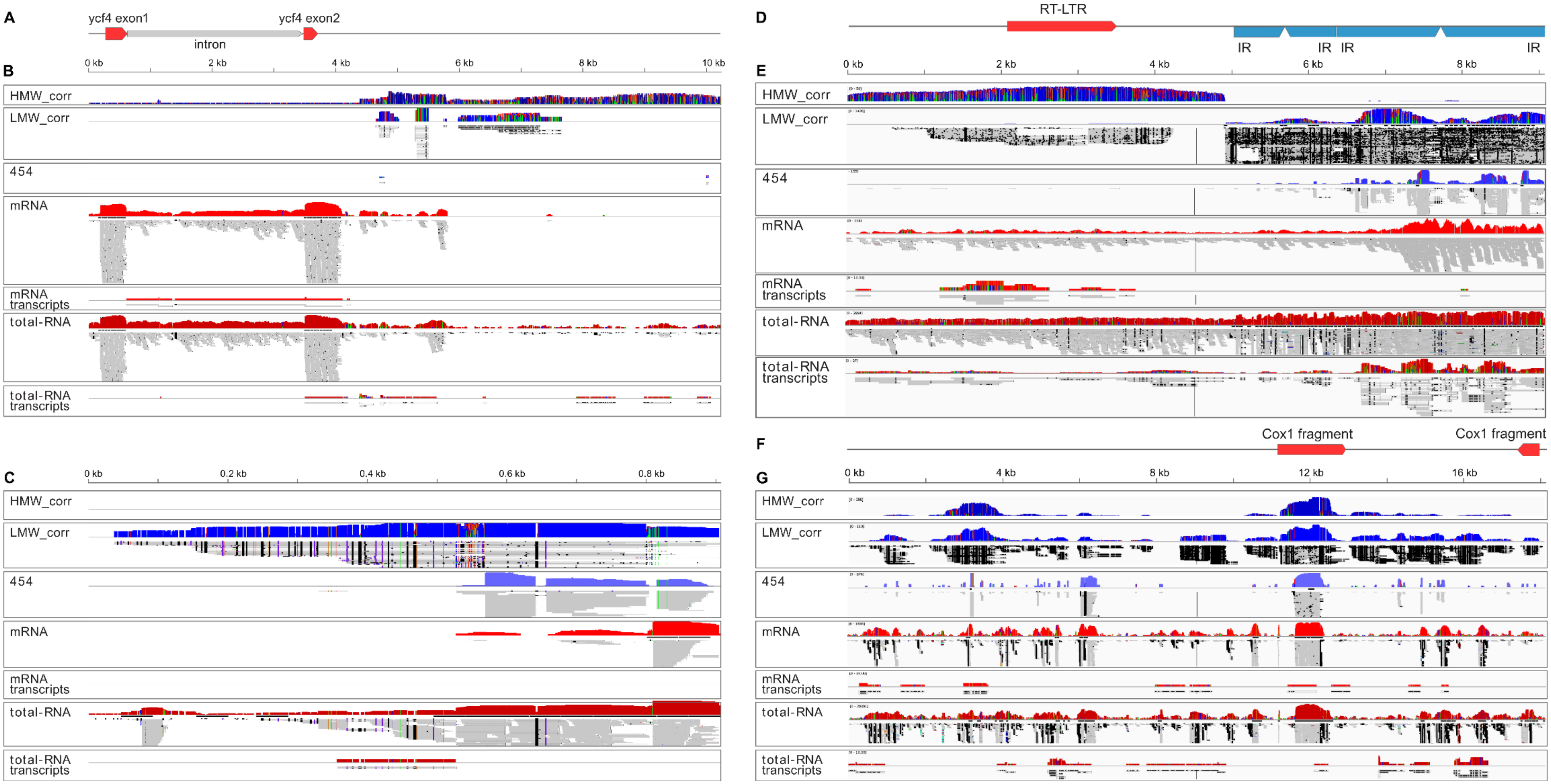


Figure S4.2: HMW DNA reads contain RT-LTR and nuclear-encoded protein-coding genes that are normally present in the chloroplast genome.

(A) Representation of the HMW DNA read encoding ycf4 (10,517 bp), one of the 19 nuclear-encoded genes identified in the HMW DNA fraction. The red arrow indicates the spliced CDS, the grey dash indicates the intron. (B) Corresponding Genome Browser track showing from top to bottom: HMW DNA read coverage [range 0-13], LMW DNA read coverage [range 0-22], 454 read coverage [range 0-2], mRNA library read coverage [range 0-5,131], assembled mRNA transcripts mapped [range 0-2], total-RNA library read coverage [range 0-896], and assembled total-RNA transcripts mapped [range 0-3]. (C) psbA, psbB, psbC HMW DNA reads are not transcribed by the nuclear machinery and they are possibly LMW DNA carry-over contaminants, rather than genuine genes transferred to the nucleus. HMW to LMW DNA reads ratio and mRNA to total-RNA reads ratio suggest and support the first hypothesis. Genome Browser track of the psbA HMW DNA read (909 bp) showing from top to bottom: corrected HMW DNA read coverage [range 0-1], corrected LMW DNA read coverage [range 0-58], 454 read coverage [range 0-27], mRNA library read coverage [range 0-795], assembled mRNA transcripts mapped [0], total-RNA library read coverage [range 0-408,934], and assembled total-RNA transcripts mapped [range 0-2]. (D) Representation of the HMW DNA read encoding RT-LTR (9,098 bp). This read (p0/144332), which was assembled together with the empty chloroplast 454 contig c474, presented similarity to ncDNA in the hairpin plasmids and potentially encoded a retrotranscriptase gene. The red arrow indicates the RT-LTR CDS, the blue arrows indicate inverted repeats with sequence similarity to the inverted repeats and the ncDNA conserved motifs of Boodlea chloroplast genome. (E) Corresponding Genome Browser track showing from top to bottom: coverage of the HMW DNA [range 0-50] and LMW DNA [range 0-1,435] corrected reads respectively; coverage of 454 reads [0-155], coverage of the mRNA library [range 0-104] and the corresponding assembled transcripts [range 0-6], respectively; coverage of the total-RNA library [range 0-3,884] and the corresponding assembled transcripts, respectively [range 0-27]. (F) Representation of one of the 52 contigs obtained by hybrid assembly between HMW DNA reads and 454 mitochondrial contigs (17,353 bp). The red arrows indicate two distinct cox1 fragments. Metagenomic binning of the 454 contigs revealed a cluster of 102 contigs with sequence similarity to mitochondrial protein-coding genes (Cluster 4 in Figure S4.6D; see also Material and Methods). Potential genes coding for mitochondrial proteins were identified as well on 346 HMW DNA reads. In stark contrast with the chloroplast 454 contigs, a hybrid assembly between mitochondrial 454 contigs and HMW DNA reads resulted into 52 contigs, with an N50 length of 15,105 bp and a cumulative length of 732 kb and a GC content of 57.6%. A similarity search for mitochondrial genes commonly present in the mitochondrial genomes of Chlorophyta revealed the presence of abundant fragments of 8 protein coding genes (*atp8*, *atp9*, *cob*, *cox1*, *nad3*, *nad4*, *nad5*, *nad6*), scrambled and/or repeated in tandem. The assembled mitochondrial contigs are rich in direct and inverted repeats, however, they bear no sequence similarity to conserved ncDNA motifs identified in Boodlea chloroplast genome. (G) Corresponding Genome Browser track showing from top to bottom: coverage of the HMW DNA [range 0-396] and LMW DNA [range 0-110] corrected reads respectively; coverage of the 454 reads [0-876], coverage of the mRNA library [range 0-1,695] and the corresponding assembled transcripts [range 0-2], respectively; coverage of the total-RNA library [range 0-29,391] and the corresponding assembled transcripts [range 0-5], respectively.

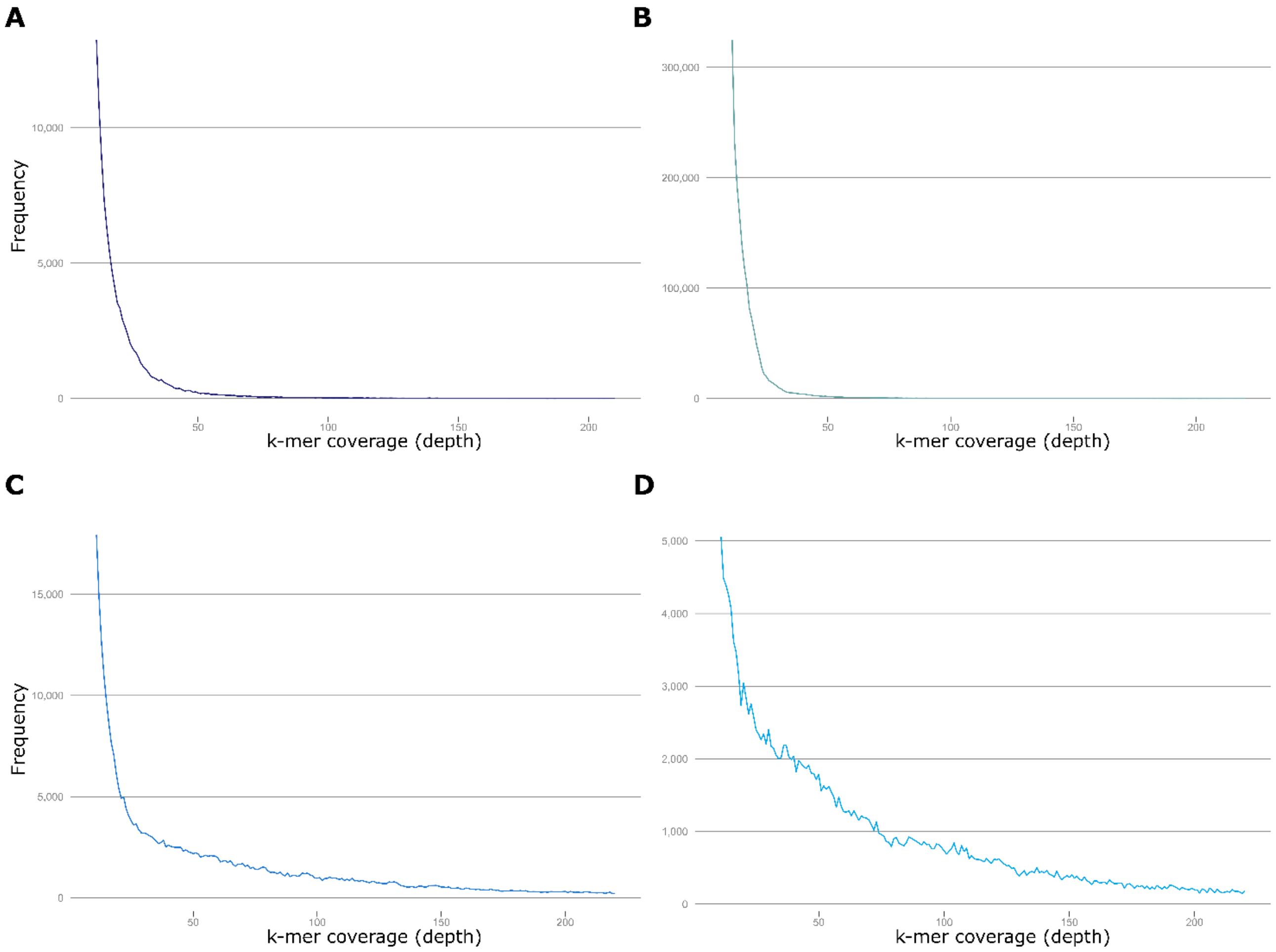


Figure S4.3: k-mer frequency distributions of *Boodlea* 454 and LMW DNA libraries.

On the x-axis is reported the k-mer coverage depth, while on the y axis is reported the frequency (A) k-mer distribution frequencies of LMW DNA reads. (B) k-mer frequency distribution of corrected LMW DNA reads. (C) k-mer frequency distribution of 454 reads. (D) k-mer frequency distribution of 454 reads that map on the chloroplast 454 contigs.

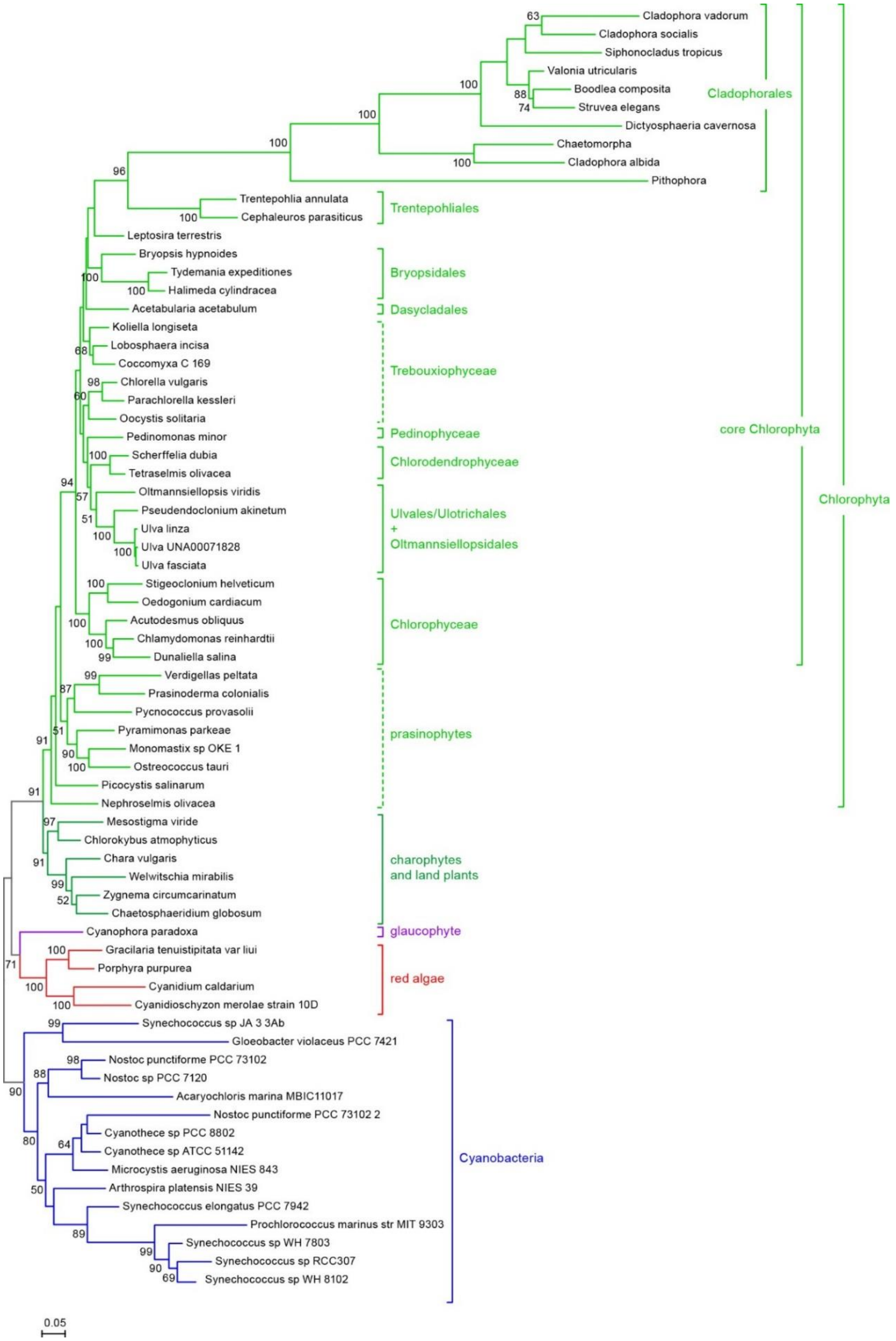


Figure S4.4: Maximum likelihood phylogenetic tree.

Maximum likelihood phylogenetic tree inferred from a concatenated amino acid alignment of 19 chloroplast genes, including *atpA*, *atpB*, *atpH*, *atpI*, *petA*, *petB*, *petD*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE, psbF*, *psbJ*, *psbL* and *rbcL*. The scale represents 0.1 substitution per amino acid position.

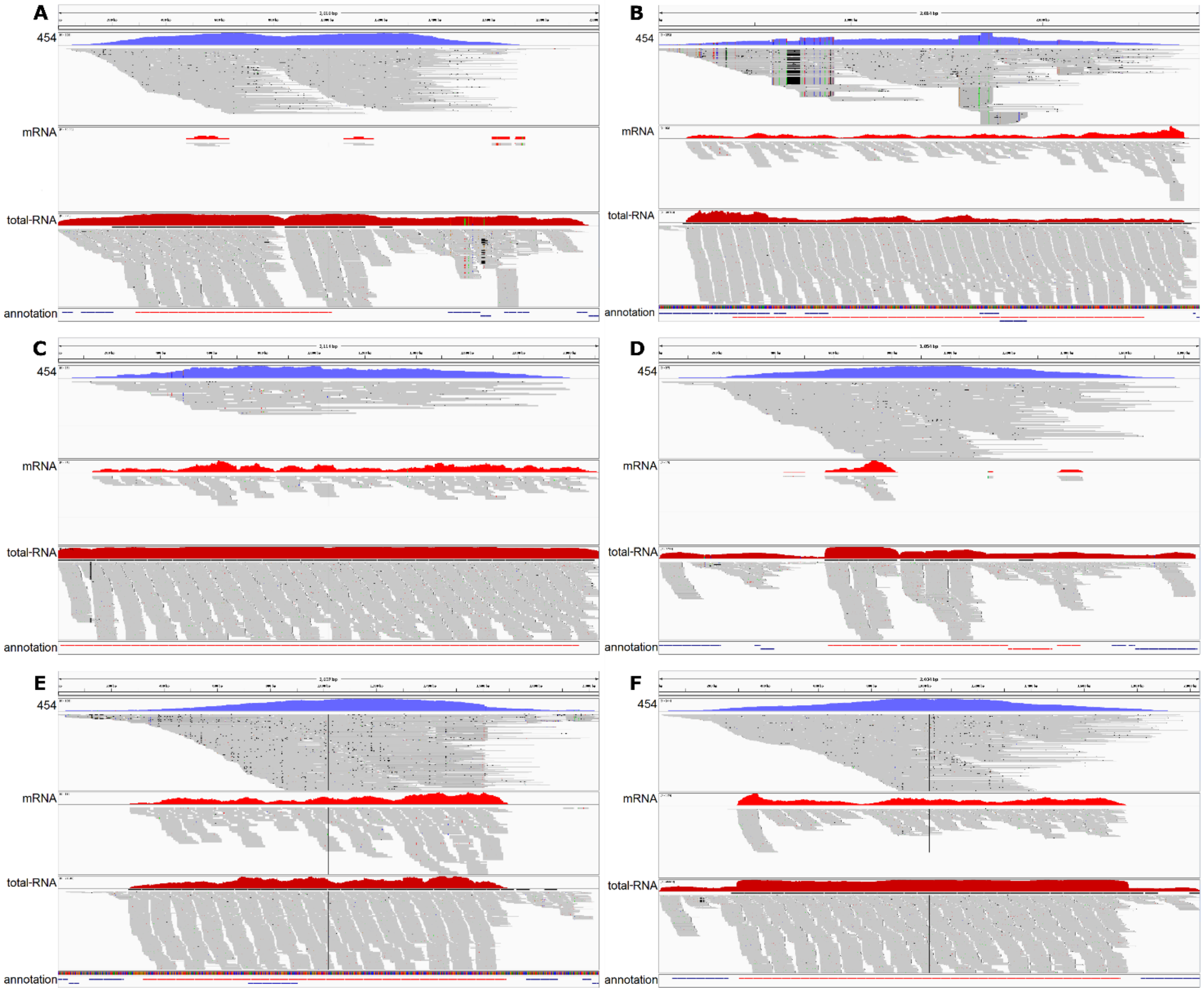


Figure S4.5: Non-canonical genetic code in *Boodlea* chloroplast genes.

(A) Concordance between DNA contigs and mapped total-RNA reads for the chloroplast 454 contig containing the *petA* gene. From top to bottom: coverage of 454 reads [range 0-208]; coverage of the mRNA library [range 0-3]; coverage of the total-RNA library [range 0-720]. The bottom track shows the annotated regions (CDS in red, inverted repeats in blue). (B) Concordance between DNA contigs and mapped total-RNA reads of the chloroplast 454 contig containing the *psaA* gene. From top to bottom: coverage of 454 reads [range 0-252]; coverage of the mRNA library [range 0-93]; coverage of the total-RNA library [range 0-30,713]. The bottom track shows the annotated regions (CDS in red, inverted repeats in blue). (C) Concordance between DNA contigs and mapped total-RNA reads of the chloroplast 454 contig containing the *psaB* gene. From top to bottom: coverage of 454 reads [range 0-27]; coverage of the mRNA library [range 0-24]; coverage of the total-RNA library [range 0-19,077]. The bottom track shows the annotated regions (CDS in red). (D) Concordance between DNA contigs and mapped total-RNA reads of the chloroplast 454 contig containing the *psaC* gene. From top to bottom: coverage of 454 reads [range 0-97]; coverage of the mRNA library [range 0-19]; coverage of the total-RNA library [range 0-7,255]. The bottom track shows the annotated regions (CDS in red, inverted repeats in blue). (E) Concordance between DNA contigs and mapped total-RNA reads of the chloroplast 454 contig containing the *psbC* gene. From top to bottom: coverage of 454 reads [range 0-306];

coverage of the mRNA library [range 0-54]; coverage of the total-RNA library [range 0-21,340]. The bottom track shows the annotated regions (CDS in red, inverted repeats in blue). (F) Concordance between DNA contigs and mapped total-RNA reads of the chloroplast 454 contig containing the *rbcL* gene. From top to bottom: coverage of 454 reads [range 0-244]; coverage of the mRNA library [range 0-135]; coverage of the total-RNA library [range 0-46,618]. The bottom track shows the annotated regions (CDS in red, inverted repeats in blue).

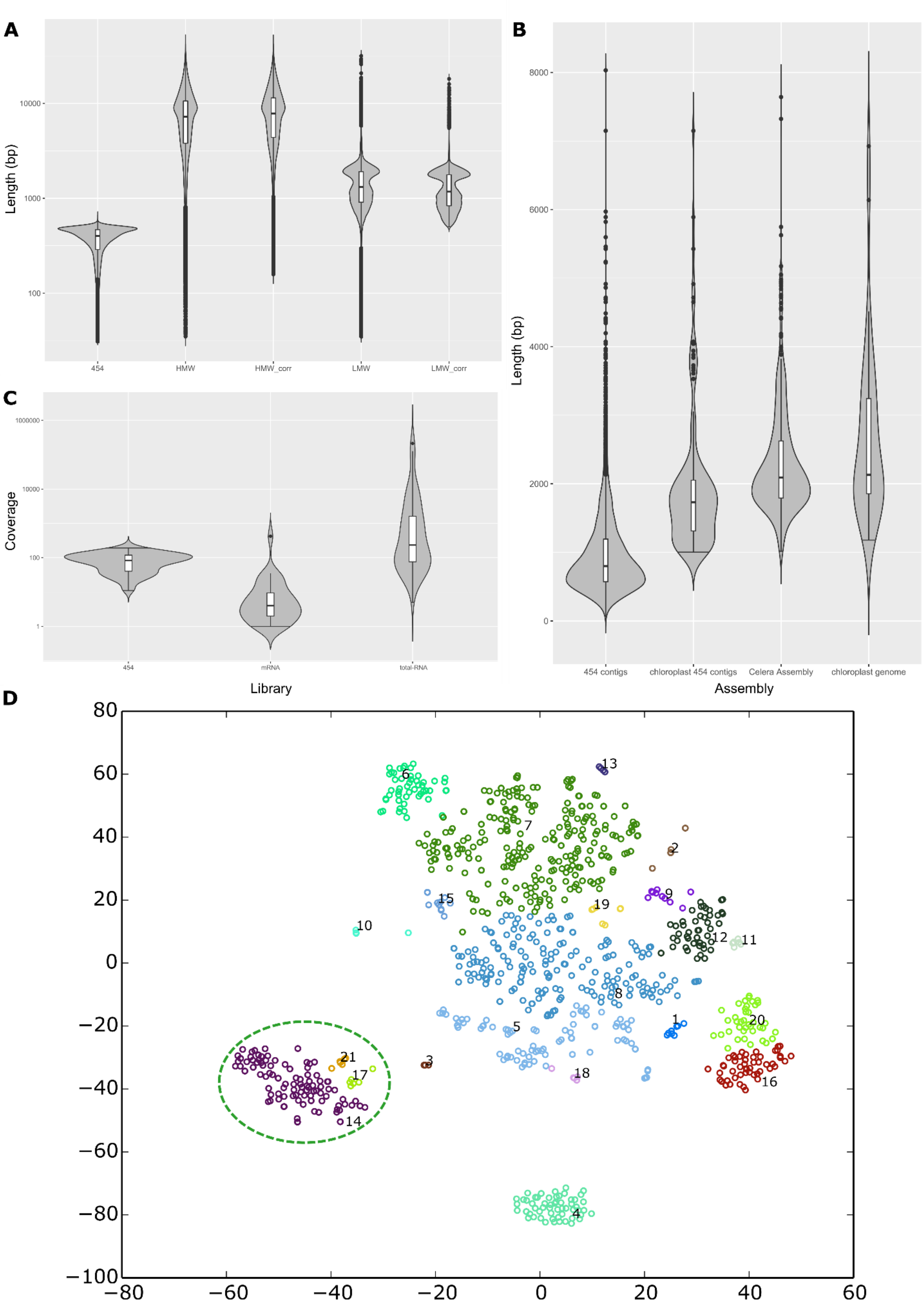

A
Length (bp)
10000
1000
100
454
HMW
HMW_corr
LMW
LMW_corr
B
Length (bp)
8000
6000
4000
2000
0
454 contigs
chloroplast 454 contigs
Celera Assembly
chloroplast genome
Assembly
C
Coverage
1000000
10000
100
1
454
mRNA
total-RNA
Library
D
80
60
40
20
0
−20
−40
−60
−80
−100
−80
−60
−40
−20
0
20
40
60
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Figure S4.6: Length distributions of *Boodlea* genomic libraries and genomic binning of the 454 contigs.

(A) Violin boxplot indicates the length distributions of the reads for each library. Inner boxplot indicates the median and the first and third quartiles of the distributions. Black circles indicate outliers. 454: chloroplast-enriched fraction library; HMW: uncorrected HMW DNA library; HMW corr: corrected HMW DNA library; LMW: uncorrected LMW DNA library; LMW corr: corrected LMW DNA library. Details on the genomic libraries are provided in Table S4.2. (B) Violin boxplots indicate the length distributions of the contigs for each *Boodlea* genomic assembly. Inner boxplots indicate the median and the first and third quartiles of the distributions. Black circles indicate outliers. Genomic assembly metrics are provided in Table S4.3. (C) Violin boxplot indicating the mean coverage of the 454 reads, and RNA (mRNA and total RNA) reads mapping to the 136 chloroplast 454 contigs. The higher coverage of the total-RNA library compared to the mRNA library confirms that the chloroplast 454 contigs are not transcribed in the nucleus. The Y axis reports the mean coverage of the chloroplast contigs. (D) Metagenomic binning of the 454 contigs. Only contigs longer than 1,000 bp were included in the analysis. Each circle represents a contig. Contigs were grouped into 21 clusters (indicated with different colours) based on their genomic signature and coverage profile. Sequence similarity searches against NCBI non-redundant protein database showed that clusters 14, 17 and 21 are composed by chloroplast contigs (indicated by a dashed green ellipse). Cluster 1, Cluster 11, Cluster 13 and Cluster 18 (15, 9, 7, 8 contigs, respectively) had no similarity to known proteins. Cluster 2 contained 7 contigs of possible bacterial origin. Cluster 4 contained 102 contigs with sequence similarity to mitochondrial proteins. Contigs in the remaining clusters had sequence similarity to Long-Terminal Repeats retrotransposon proteins (RT-LTRs) and to hypothetical and predicted proteins with domains typical of RT-LTRs.

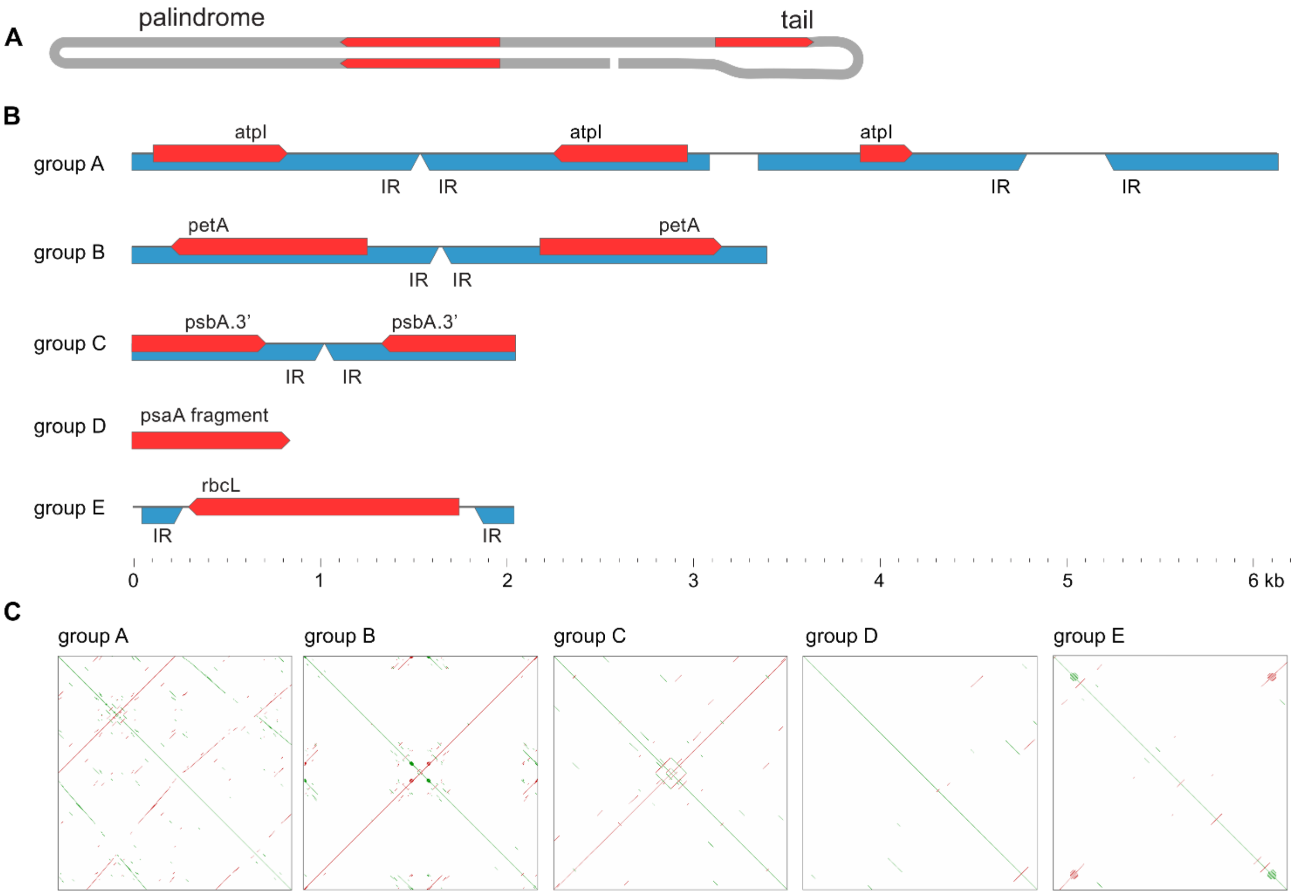


Figure S4.7: Schematic representation of assembled *Boodlea* chloroplast DNA contigs.

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

Table S4.1: Features of the 34 contigs constituting the *Boodlea* chloroplast genome and unassembled LMW DNA reads with sequence similarity to chloroplast genes.

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

-	**psaB**	-	-	-	**57.7**; -; -	337	-; 523; 2,493; 776	○	○	○	0	0	8	329	0	Unassembled. shorter (5' 30 aa)
ctg18 *	**psaB**	95.9	?	TAA	**58.2**; 58.5; 51.1	-	**2,116**	-	-	-	0	0	0	0	1	lacks 5'. TGA → C/V
ctg19	**psaC** **psbJ**	16.6	ATG TTG	TAG TGA	**57.7**; 55.2; 58.2	92	**2,971**; 503; 2,346; 726	●	○	●	0	6	0	86	0	Group B. TGA → C
ctg20 *	**psaC**	12.5	?	?	**56.1**; 56.3; 56.1	-	**1,850**	-	-	-	0	0	0	0	1	longer 5'?
ctg21	**psbA**	54.5	ATG	TAG	**53.7**; 52.5; 55.5	136	**3,599**; 551; 2,636; 893	○	●	○	0	0	37	109	0	Group A
ctg22	**psbB**	61.1	ATG	TAG	**60.4**; 60.4; 60.4	152	**3,090**; 552; 4,273; 1,757	○	●	○	0	0	64	88	0	Group B
ctg23	**psbC**	67.0	ATG	TAA	**59.3**; 59.3; 59.3	187	**2,041**; 500; 3,404; 646	-	-	-	0	0	5	171	11	Group E ; TGA → L
ctg24	**psbD**	52.0	ATG	TAA	**56.3**; 56.7;56.0	90	**3,643**; 558; 2753; 751	●	○	●	0	0	3	78	4	Group B, lacks 5' ?
ctg25 *	**psbD**	59.5	ATG	TAA	**56.5**; 56.4; 56.7	-	**1,960**	-	-	-	0	0	0	0	1	longer 5'
ctg26	**psbE** **psbK**	16.9	? ?	? TAG	**55.8**; 53.6; 56.2	10	**1,327**; 689; 2,077; 1,405	●	○	●	0	0	0	7	2	Group E; no 5', no stop codon
ctg27	**psbF**	5.2	CTC	TAG	**56.8**; 54.8; 57.0	5	**2,617**; 1,715; 1,726; 1,718	●	○	●	0	4	0	0	0	Group B, uncorrected
ctg28	**psbL**	8.4	ATG	TAG	**55.9**; 48.5; 56.6	1	**1,179**	●	○	●	0	0	0	0	1	Group B; uncorrected
ctg29 *	**psbL**	6.3	ATG	TAG	**57.8**; 49.5; 58.5	-	**1,666**	-	-	-	0	0	0	0	1	
ctg30 *	**psbL**	6.2	TTG	?	**56.9**; 50.4; 57.3	-	**1,986**	-	-	-	0	0	0	0	1	arbitrary
ctg31	**psbT**	1.6	GTG	?	**57.9**; 47.2; 58.0	70	**4,515**	●	○	●	0	9	0	42	2	Group B
ctg32	**psbT**	4.3	GTG	TAG	**57.2**; 53.5; 57.4	43	**3,687**	●	○	●	0	10	0	33	0	Group B
ctg33	**psbT**	6.0	CAT	TAA	**56.9**; 50.8; 57.3	33	**3,000**	●	○	●	0	12	0	21	0	Group B
ctg34	**rbcL**	65.1	ATG	TAA	**57.3**; 57.8; 56.1	411	**4,116**; 503; 3,120; 909	○	●	○	0	0	26	385	0	Group B, TGA → C

*****chloroplast 454 contigs that could not be assembled together with LMW DNA reads

● yes

○ no

- not applicable

length: in bold is indicated the length of the contig resulting from the orthology-guided assembly (chloroplast-enriched fraction + LMW DNA); minimum, maximum and N50 lengths in bp of the LMW DNA reads. With regards to the chloroplast 454 contigs that could not be assembled together with the LMW DNA reads, their lengths are also reported in bold.

% coding: percentage of coding sequence of the orthology-guided contig or of the chloroplast 454 contigs.

GC%: GC contents of orthology-guided/chloroplast 454 contigs, of the respective CDS and of the non-coding region.

LMW full-CDS: whether a full-length CDS could be detected in the LMW DNA reads.
LMW assembled with 454: whether a full-length CDS could be reconstructed by orthology-guided assembly of LMW DNA reads and chloroplast 454 contigs.

LMW congruent with 454: whether LMW DNA reads harboring full-length CDSs have corresponding chloroplast 454 contigs.

groupA-E: distribution of LMW DNA reads in different Groups (Figure S4.7). All unassembled chloroplast 454 contigs belong to Group E molecules.

start: the identified start codon of the CDS. A question mark indicates that the start codon could not be univocally identified.

stop: the identified stop codon of the CDS. A question mark indicates that the stop codon could not be univocally identified.

features: in this column are reported the Groups to which the orthology-guided contigs (ctg) with full-length CDSs belong, the alternative codons identified if it was possible to assemble a full-length CDS, and the eventual differences with the orthologous sequences of other green algae.

Table S4.2: Features Genomic and transcriptomic libraries analysed in this study.

Number of reads, total length, N50 length and GC content are reported. For the long noisy reads from HMW and LMW DNA libraries, information before and after read corrections are included.

	# of reads	**total length (bp)**	**N50 (bp)**	**GC%**
454	261,577	96,038,799	441	53.38
HMW	67,706	535,096,685	10,357	50.03
HMW corrected reads	50,119	432,837,312	11,021	50.21
LMW	154,852	224,767,151	1,798	50.90
LMW corrected reads	106,428	139,581,606	1,658	51.07
mRNA	32,336,598	4,871,957,016	-	48.03
total-RNA	73,362,835	11,082,500,196	-	53.19
C. aerea	3,977,613	803,477,826	-	54.70
C. albida	13,907,529	2,809,320,858	-	44.14
C. socialis	60,823,223	12,286,291,046	-	50.21
C. vadorum	61,955,110	12,514,932,220	-	50.53
D. cavernosa	25,664,333	5,184,195,266	-	42.41
***Pithophora* sp.**	1,682,879	339,941,558	-	51.82
S. tropicus	7,742,251	1,563,934,702	-	57.02
S. elegans	6,291,774	1,270,938,348	-	51.72
V. utricularis	11,073,789	2,236,905,378	-	59.25
V. ventricosa	3,303,983	667,404,566	-	55.53

Table S4.3: Genomic and transcriptomic assembly metrics.

Number of contigs, total length, N50 length, GC content and approximate length of the longest contig are reported.

	# contigs	**total length (bp)**	**N50 (bp)**	**GC%**	**longest contig (bp)**
454 contigs	3,735	3,696,003	1,138	51.08	8,000
chloroplast 454 contigs	136	268,038	1,964	57.55	7,100
Celera Assembly	558	1,299,546	2,287	54.51	7,800
chloroplast genome	34	91,391	2,971	56.52	6,900
C. aerea	86,127	56,059,813	877	55.55	639,600
C. albida	58,487	20,927,231	345	44.14	46,500
C. socialis	1,047,600	575,965,604	630	50.21	106,200
C. vadorum	638,917	404,812,870	781	50.53	183,700
D. cavernosa	39,589	14,858,121	333	42.41	104,000
***Pithophora* sp.**	42,095	14,710,190	344	60.83	17,000
S. tropicus	13,325	4,660,819	337	57.02	8,200
S. elegans	7,725	3,159,309	414	51.72	5,400
V. utricularis	26,338	8,907,076	325	59.25	9,300
V. ventricosa	30,821	24,976,680	1,359	55.53	784,900
mRNA	91,362	63,341,422	1,198	50.70	9,600
mRNA euk	24,790	26,403,806	1,729	51.65	9,600
mRNA nohit	64,078	34,837,854	741	49.90	9,100
total-RNA	174,989	107,076,485	797	50.16	13,700
total-RNA euk	32,021	34,677,513	1,721	51.89	13,700
total-RNA nohit	136,680	69,059,114	552	49.19	11,600

Table S4.4: Collection details of *Boodlea composita*, and the nine additional Cladophorales sequenced and used in the phylogenetic analysis.

Species	Collection locality, collector, and date	Culture / voucher number
Boodlea composita	Dumaguete, Negros Oriental, Philippines (Leliaert, 14 Nov 2007)	FL1110 *
Chaetomorpha aerea	Woods Hole, Massachusetts, USA (H.C. Bold, Summer 1956)	UTEX799 **
Cladophora albida	Swan River, W Australia, (1985)	Calb2 (= A85.23) *
Cladophora socialis	Rottnest Isl., Australia (1988)	Csoc2 (= CPS7A) *
Cladophora vadorum	Punta del Hidalgo, Tenerife (1988)	Cvad2 (= CvadoPH) *
Dictyosphaeria cavernosa	Dapdap, Siquijor, Philippines (F. Leliaert, 16 Nov 2007)	FL1134 *
Pithophora sp.	Brooklyn, Indiana, USA (C. Kelly)	UTEX787 **
Siphonocladus tropicus	Arinaga, Gran Canaria	Siph3 (= StGC) *
Struvea elegans	Bahamas (S. Brawley, 1975)	Sele1 (=SE 1572 = West 1572 = UTEX LB 2372) *
Valonia utricularis	Punta Carnero, Spain (1996)	Vutric3 (= VUSC) *
Valonia ventricosa	Coconut Island, Kaneohe Bay, Oahu, Hawaii, USA (J.A. West, 1970)	UTEX2260 **

* Algal culture collection of the Phycology Research Group, Ghent University, Belgium

** UTEX Culture Collection of Algae at the University of Texas at Austin, United States of America

Table S4.5: Chloroplast protein-coding genes identified in the Cladophorales genomic libraries.

	atpA	*atpB*	*atpE*	*atpH*	*atpI*	*petA*	*petB*	*petD*	*psaA*	*psaB*	*psaC*	*psbA*	*psbB*	*psbC*	*psbD*	*psbE*	*psbF*	*psbH*	*psbJ*	*psbK*	*psbL*	*psbT*	*rbcL*
B. composita	●	●	-	●	●	●	●	●	●	●	●	●	●	●	●	●	●	-	●	●	●	●	●
C. aerea	●	●	-	●	●	-	●	-	●	●	-	●	-	●	●	-	-	-	-	-	-	-	●
C. albida	○	○	-	●	●	○	-	-	○	●	-	●	○	●	○	○	-	-	-	-	-	-	○
C. socialis	-	●	●	●	○	-	○	-	○	○	●	●	○	●	●	●	●	●	-	-	-	-	○
C. vadorum	○	○	●	●	○	-	○	-	○	○	-	○	○	○	○	-	○	-	-	-	-	-	○
D. cavernosa	○	○	-	●	●	-	●	-	●	●	○	●	●	●	○	-	-	-	-	-	-	-	●
Pitophora sp.	○	●	-	●	-	○	●	-	●	○	-	●	○	○	●	-	-	-	-	-	-	-	●
S. tropicus	-	●	-	-	○	-	○	○	○	●	●	●	○	○	●	●	-	-	-	-	-	-	●
S. elegans	○	○	-	●	○	○	○	○	○	○	●	-	○	○	○	-	-	-	-	-	-	-	○
V. utricularis	○	●	-	○	○	-	○	-	○	○	●	-	○	○	○	●	●	-	-	-	-	-	●

● full-length gene reconstructed

○ gene detected but partial

- gene not detected

Table S4.6: Species used in the phylogenetic analysis, along with GenBank accession numbers.

Species	GenBank accession number(s)	Species	GenBank accession number(s)
Acaryochloris marina MBIC11017	NC_009925.1	*Nostoc* sp. PCC 7120	NC_003272
Acetabularia acetabulum	HG518455-HG518469	*Oedogonium cardiacum*	NC_011031
Acutodesmus obliquus	NC_008101	*Oltmannsiellopsis viridis*	NC_008099
Arthrospra platensis NIES 39	NC_016640	*Oocystis solitaria*	FJ968739
Boodlea composita FL1110	this study	*Ostreococcus tauri*	NC_008289
Bryopsis plumosa	NC_026795	*Parachlorella kessleri*	NC_012978
Cephaleuros parasiticus	KM464687-KM504519	*Pedinomonas minor*	NC_016733
Chaetomorpha sp	this study	*Picocystis salinarum*	NC_024828
Chaetospaeridium globosum	NC_004115	*Pithophora* sp.	this study
Chara vulgaris	NC_008097	*Porphyra purpurea*	NC_000925
Chlamydomonas reinhardtii	NC_005353	*Prasinoderma coloniale*	NC_024817
Chlorella vulgaris	NC_001865	*Prochlorococcus marinus* str MIT 9303	CP000554
Chlorokybus atmophyticus	NC_00882	*Tupiella akineta*	NC_008114
Cladophora albida	this study	*Pycnococcus provasolii*	NC_012097
Cladophora socialis	this study	*Pyramimonas parkeae*	NC_012099
Cladophora vadorum	this study	*Scherffelia dubia*	NC_029807
Coccomyxa C-169	NC_015084	*Struvea elegans*	this study
Cyanidioschyzon merolae strain 10D	NC_004799	*Siphonocladus tropicus*	this study
Cyanidium caldarium	NC_001840	*Stigeoclonium helveticum*	NC_008372
Cyanophora paradoxa	NC_001675	*Synechococcus elongatus* PCC 7942	NC_007604
Cyanothece sp ATCC 51142	NC_015047	*Synechococcus* sp JA 3 3Ab	NC_007775
Cyanothece sp PCC 8802	NC_013161	*Synechococcus* sp RCC307	NC_009482
Dictyospaeria cavernosa	this study	*Synechococcus* sp WH 7803	NC_009481
Dunaliella salina	NC_016732	*Synechococcus* sp WH 8102	NC_005070
Gloeobacter violaceus PCC 7421	NC_005125	*Tetraselmis olivacea*	KU167097
Gracilaria tenuistipitata var *liui*	NC_006137	*Trentepohlia annulata*	KM464689-KM491845
Halimeda cylindracea	KM820107-KM820166	*Tydemania expeditionis*	NC_026796
Koliella longiseta	NC_025531	*Ulva fasciata*	NC_029040
Leptosira terrestris	NC_009681	*Ulva linza*	NC_030312
Lobospaera incisa	NC_025533	*Ulva* UNA00071828	KP_720616
Mesostigma viride	NC_002186	*Valonia utricularis*	this study
Microcystis aeruginosa NIES 843	NC_010296	*Verdigellas peltata*	NC_030220
Monomastix sp. OKE 1	NC_012101	*Welwitschia mirabilis*	NC_010654
Nephroselmis olivacea	NC_000927	*Zygnema circumcarinatum*	NC_008117
Nostoc punctiforme PCC 73102	NC_010628		

Chapter 5 - General Discussion

Andrea Del Cortona[8]

“I disapprove of what you say, but I will defend to the death your right to say it.”

Evelyn Beatrice Hall - The Friends of Voltaire

[8] Authors contribution: A.D.C.: manuscript conceptualization, drafting and writing.

Ulvophyceans evolution: chasing shadows

Are green algal phylogenetic relationships resolved?

Wherever there is sunlight, photosynthetic eukaryotes with green plastids can be found. Both in species numbers as well as biomass, the overwhelming majority of green plastids is found in *Viridiplantae*, while euglenids, chlorarachniophytes, and green dinoflagellates only contain a tiny fraction of the green plastid diversity. The*Viridiplantae* includes the land plants, which dominate terrestrial habitats, and green algae, which are widespread in freshwater environments. One group of green algae, the green seaweeds, form macroscopic thalli and are abundant in shallow benthic habitats along coastlines, and to a lesser extent the open oceans (Falkowski *et al.*, 2004b). Yet, little is known about their evolutionary history. *Viridiplantae* are a natural group (Cavalier-Smith, 1981; Bremer, 1985; Adl *et al.*, 2005) composed of two major lineages which followed two separate paths of evolution: the Streptophyta and the Chlorophyta (Lewis & McCourt, 2004; Becker & Marin, 2009; Brocks *et al.*, 2017; Jackson *et al.*, 2018). The divergence of the Streptophyta and Chlorophyta is ancient, and likely took place before the the Paleozoic (570 mya), possibly in the the Mesoproterozoic (1600-1000 mya), although convincing fossils are largely absent (Becker & Marin, 2009; Wodniok *et al.*, 2011; Sánchez-Baracaldo *et al.*, 2017).

The relationships between the major clades of Streptophyta are generally well-resolved, however, some uncertainties remain, such as the relationships of hornworts, liverworts and mosses at the base of the Embryophyta despite analyses of phylogenomic datasets with hundreds of genes. Conflicting topologies based on different markers and evolutionary models have been ascribed to incomplete lineage sorting (ILS) and lineage-specific heterogeneity (Puttick *et al.*, 2018; Rensing, 2018). Unfortunately, the long and intricate evolutionary history of Streptophyta probably goes hand in hand with convergent loss and gain of key features (e.g.: independent evolution of stomata), preventing the parsimonious use of synapomorphies to solve the uncertainties (Duckett & Pressel, 2018).

Historically, the evolution of green algae was interpreted based on thallus organization, whereby clades with increasing size and complexity were thought to have originated from a simple unicellular ancestor (Fott, 1971). Later, analyses of ultrastructural data related to mitosis (e.g.: mitotic spindle), the arrangement of flagellar basal bodies, and cytokinesis resulted in a thorough reevaluation of the classification of green algae. These features were believed to better reflect phylogenetic relationships because of their involvement in fundamental processes of cell replication and cell motility, and thus to be less liable to convergent evolution than thallus form. More recently, molecular phylogenetic data provided a new framework for reconstructing the evolutionary history of the green algae. Within the Chlorophyta, four major groups are recognized: a paraphyletic assemblage of prasinophytes, which are the earliest diverging Chlorophyta, and the morphologically diverse Chlorophyceae, Trebouxiophyceae and Ulvophyceae, which form a clade defined as 'core Chlorophyta' (Lewis & McCourt, 2004; Leliaert *et al.*, 2012; Turmel & Lemieux, 2018). Recently, Chlorodendrophyceae and Pedinophyceae have been elevated to the class-level and included in the core Chlorophyta based on molecular data (Marin, 2012; Fučíková *et al.*, 2014).

The relationships between the core Chlorophyta, however, have been the subject of a long-standing debate, and the monophyly of both Trebouxiophyceae and Ulvophyceae have been questioned (Fučíková *et al.*, 2014; Lemieux *et al.*, 2014a; Leliaert & Lopez-Bautista, 2015; Melton *et al.*, 2015; Sun *et al.*, 2016; Turmel *et al.*, 2016a; Turmel *et al.*, 2017; Fang *et al.*, 2018). Although a progression from single-gene analyses to analyses using entire chloroplast or mitochondrial genomes has generally resulted in better resolved phylogenies, several topological uncertainties among the main lineages in the core Chlorophyta remain.

Based on a phylotranscriptomic dataset we present for the first time a highly supported topology for the core Chlorophyta (Chapter 3). Confirming previous analyses, Pedinophyceae and Chlorodendrophyceae represent the earliest diverging lineages of the core Chlorophyta, followed by the monophyletic Trebouxiophyceae, which includes the core Trebouxiophyceae and the Chlorellales. Our results confirmed several relationships that are robust and supported by previous studies: sister relationships between Chlorellales and core trebouxiophytes, a monophyletic Chlorophyceae and two major clades within the ulvophyceans. The first clade contains Oltmannsiellopsidales, Ignatiales, Ulvales and Ulotrichales. The second clade comprises Cladophorales, Dasycladales, Scotinosphaerales, Trentepohliales, and *Blastophysa*. The latter clade is characterized by an alternative nuclear genetic code, which is unique among green algae. Our results support the idea that multicellularity evolved independently several times during the evolution of Chlorophyta, and likely also within the ulvophyceans (Watanabe & Nakayama, 2007; Cocquyt *et al.*, 2010b). However, despite the considerable progress in clarifying the backbone of the core Chlorophyta, some relationships remain notoriously difficult to resolve, such as the affinities between the Chlorophyceae, Bryopsidales and the rest of the Ulvophyceae.

Monophyletic, paraphyletic or polyphyletic Ulvophyceans: tale of an early radiation

Chloroplast phylogenomic analyses yielded several conflicting topologies for the ulvophyceans, with the leitmotiv of ulvophyceans being polyphyletic (Fučíková *et al.*, 2014; Lemieux *et al.*, 2014a; Leliaert & Lopez-Bautista, 2015; Melton *et al.*, 2015; Sun *et al.*, 2016; Turmel *et al.*, 2016a; Turmel *et al.*, 2017; Fang *et al.*, 2018). Similar analyses using nuclear markers, be it with shorter alignments, instead pointed towards a monophyletic Ulvophyceae (Watanabe & Nakayama, 2007; Cocquyt *et al.*, 2010b). Our analyses which are based on more than 500 single-copy nuclear genes, converged on two very similar topologies with respect to the Ulvophyceae. Supermatrix analyses supported a scenario whereby Chlorophyceae are sister to the Bryopsidales, leaving the Ulvophyceae paraphyletic. Coalescence-based analyses, on the other hand, indicated a hard polytomy of the Chlorophyceae, Bryopsidales, and the rest of the Ulvophyceae.

Shared cytological and ultrastructural features suggest a close affiliation between Bryopsidales and Dasycladales: a siphonous morphology, cytoplasmic streaming, closed mitosis with a prominent persistent telophase spindle and a 11 o'clock – 5 o'clock configuration of the flagellar apparatus in the gametes (Sluiman, 1989b). This relationship is supported by some studies based on chloroplast and nuclear gene sequences (Cocquyt *et al.*, 2010b; Fučíková *et al.*, 2014; Sun *et al.*, 2016), but it is not supported by other studies (Melton *et al.*, 2015; Fang *et al.*, 2018), including our study, where the Bryopsidales forms a separate clade from the rest of the ulvophyceans. Perhaps the strongest clue to help resolving Bryopsidales and Dasycladales relationships comes from the distribution of the nuclear alternative genetic code in the Ulvophyceae. The Bryopsidales and Dasycladales sister relationship would imply a stepwise acquisition model for the alternative nuclear genetic code, where the

Bryopsidales represent a sort of intermediate situation where eventually the alternative code has not been established (Cocquyt *et al.*, 2010a). Instead, our results support a more parsimonious hypothesis in which an alternative genetic code evolved once in the clade containing the Cladophorales, Dasycladales, Scotinosphaerales and Trentepohliales.

Other molecular features, such as the dispersed distribution of the eukaryotic elongation factors, are more difficult to explain by the new phylogeny (Figure 1.2). Trebouxiophyceae and Chlorophyceae present the EFL relongation factor, while Bryopsidales and the members of the TCD clade (Cladophorales, Dasycladales, Scotinosphaerales) possess the EF-1α elongation factor. In contrast, Oltmannsiellopsidales, Ulvales, Ulotrichales have the ELF elongation factor (similarly to Chlorophyceae and Trebouxiophyceae), while *Ignatius*, which has been recovered as a member of this clade by our phylogenetic analyses, presents the elongation factor EF-1α. This punctuate distribution of the elongation factors could be explained by horizontal gene-transfer, or both elongation factors may have been present in the ancestor of all core Chlorophyta and differentially lost in the different lineages. The latter scenario is less likely, however, given that a green alga where both elongation factors co-occur still have to be discovered (Kamikawa *et al.*, 2013). In the end, the most plausible scenario seems to be a complex evolution history with multiple events of elongation factor loss and gains.

Therfore, more questions still remain to be addressed: which of the two topologies inferred by our analysis is the one reflecting the evolutionary history of Chlorophyceae and Ulvophyceae? Are Ulvophyceans paraphyletic or monophyletic? Two conserved ultrastructural features may support the monophyletic scenario. The phycoplast, a microtubule structure mediating cell division that evolved early during the diversification of core Chlorophyta, is absent in Bryopsidales and in the remaining ulvophyceans. Second, although many Chlorophyceae and Trebouxiophyceae present a 1 o'clock 7 o'clock clockwise orientation or a direct-opposite orientation of the basal bodies in the zoids, all Ulvophyceae zoids, Bryopsidales included, have an 11 o'clock 5 o'clock anticlockwise orientation of the basal bodies (Van den Hoek *et al.*, 1995). While, the phycoplast loss may have occurred independently in Bryopsidales and remaining ulvophyceans, the alternation between basal bodies orientation under the paraphyletic ulvophyceans hypothesis is unlikely.

Soft or hard-boiled: how to cook a multifurcation

An alternative scenario for Chlorophyceae, Bryopsidales and ulvophyceans diversification is a multifurcation, hinted by the very short branches at the base of these three clades inferred by the coalescence-based analyses. We might be dealing with a hard polytomy which implies that the processes generating the main lineages of Chlorophyceae and Ulvophyceae were essentially non-bifurcating. While the phylogenetic theory does not imply it, phylogenetic reconstruction is mostly presented as a series of bifurcations. Multifurcations, or polytomies, are often used to indicate unresolved nodes in a phylogenetic tree due to lack of signal in the available data, with the underlying concept that a polytomy is a temporary artifact and not a realistic topology (Maddison, 1989). A “soft” polytomy represents an unresolved node due to ambiguous or scarce data, while a “hard” polytomy indicates a real multifurcation in the phylogeny. Although the probability of a hard polytomy in a gene tree is negligible (Hudson, 1990), several authors argued in favor for the concreteness of a hard polytomy for species trees representing multiple, simultaneous divergence events (Hoelzer & Meinick, 1994; Slowinski, 2001; Suh, 2016).

Several scenarios have been postulated for simultaneous divergence of populations and resulting speciation events. Large-scale environmental changes, such as the fast rise or fall of sea levels or freshwater levels, can causes terrestrial and aquatic environments to be divided into multiple isolated parts. Moreover, a species range can be fragmented into multiple distinct diverging populations when it is declining (Hoelzer & Meinick, 1994). Presence of hard polytomies have been confirmed in several modern species, e.g.: Chinese macaque monkeys (Fooden, 1980; Melnick *et al.*, 1993), fruit flies and cichlid fishes in the great African rift (Sturmbauer & Meyer, 1993; Kliman *et al.*, 2000; Takahashi *et al.*, 2001), and Australian sittellas birds (Schodde & Mason, 1999). Our molecular clock analysis suggests that the split between Chlorophyceae, Bryopsidales and remaining ulvophyceans occurred during a timespan encompassing the middle and the end of Neoproterozoic, in the Cryogenian era before the transition to the Ediacaran period (750-650 mya, Figure 5.1). The Cryogenian was characterized by global-scale glaciations that lasted for millions of years (Hoffman *et al.*, 1998; Kennedy *et al.*, 2008; Macdonald *et al.*, 2010). Fossil record indicating persistence of marine life during the global scale glaciations was limited (Knoll *et al.*, 2006; Brocks *et al.*, 2017). Most photosynthetic eukaryotes allegedly survived in isolated refugia in brine channels within the ice blocks, followed by recolonization of marine environments after global meltdown (Shields, 2008; Bechstädt *et al.*, 2018). These events are compatible with a rapid radiation and hard phylogenetic polytomy. Theoretically it is impossible to distinguish between soft and hard polytomies in molecular datasets. Increasing the amount of data should manage to identify and solve a soft polytomy (Walsh *et al.*, 1999; Sayyari & Mirarab, 2018). Our results, however, indicate that the null hypothesis for the Chlorophyceae, Bryopsidales and ulvophyceans forming a polytomy cannot be rejected by increasing the amount of positions aligned. These considerations indicate that a rapid series of dichotomous branching leading to Ulvophyceae monophyly and a hard polytomy scenario represent both acceptable and likely hypotheses for the Ulvophyceae radiation.

Uncertainties concerning the radiation of Ulvophyceans thus remain, despite the tremendous increase in the quantity of data and the use of evolutionary models that best fit the data, reflecting the long and complex evolutionary history of green algae. At last, independent evolution of Bryopsidales and the rest of the ulvophyceans was suggested as well as a result of a thorough revision of the evolution of green algal morphological, cytological (fine details of mitosis and cell division, zoid architecture, cell wall composition) and life history characters. Based on these unique characteristics and on the long evolutionary history that separated ulvophyceans orders, van den Hoek proposed already more than 20 years ago to elevate the ulvophycean orders to the class level, thus dividing the Ulvophyceae *sensu* Bremer in Bryopsidophyceae, Cladophorophyceae, Dasycladophyceae, Trentepohliophyceae and Ulvophyceae (Van den Hoek *et al.*, 1995).

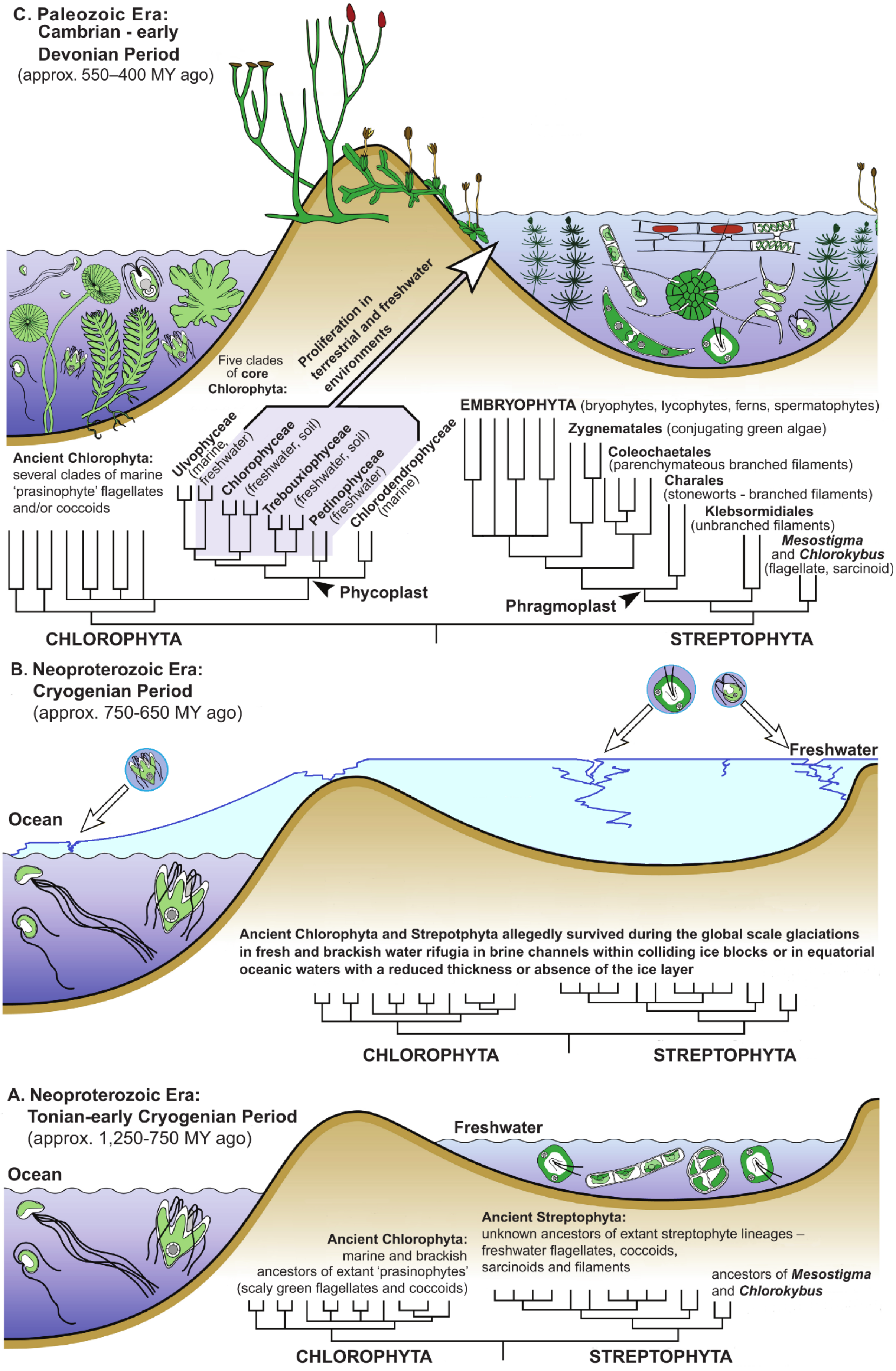

C. Paleozoic Era:
Cambrian - early
Devonian Period
(approx. 550–400 MY ago)
Proliferation in terrestrial and freshwater environments
Five clades of core Chlorophyta:
Ulvophyceae (marine, freshwater)
Chlorophyceae (freshwater, soil)
Trebouxiophyceae (freshwater, soil)
Pedinophyceae (freshwater)
Chlorodendrophyceae (marine)
Ancient Chlorophyta:
several clades of marine 'prasinophyte' flagellates and/or coccoids
Phycoplast
EMBRYOPHYTA (bryophytes, lycophytes, ferns, spermatophytes)
Zygnematales (conjugating green algae)
Coleochaetales (parenchymateous branched filaments)
Charales (stoneworts - branched filaments)
Klebsormidiales (unbranched filaments)
Mesostigma and Chlorokybus (flagellate, sarcinoid)
Phragmoplast
CHLOROPHYTA
STREPTOPHYTA
B. Neoproterozoic Era:
Cryogenian Period
(approx. 750-650 MY ago)
Ocean
Freshwater
Ancient Chlorophyta and Strepotphyta allegedly survived during the global scale glaciations in fresh and brackish water rifugia in brine channels within colliding ice blocks or in equatorial oceanic waters with a reduced thickness or absence of the ice layer
CHLOROPHYTA
STREPTOPHYTA
A. Neoproterozoic Era:
Tonian-early Cryogenian Period
(approx. 1,250-750 MY ago)
Ocean
Freshwater
Ancient Chlorophyta:
marine and brackish ancestors of extant 'prasinophytes' (scaly green flagellates and coccoids)
Ancient Streptophyta:
unknown ancestors of extant streptophyte lineages – freshwater flagellates, coccoids, sarcinoids and filaments
ancestors of Mesostigma and Chlorokybus
CHLOROPHYTA
STREPTOPHYTA

Figure 5.1: Diversification of green algae.

(A) Early evolution of chlorophytan (light green) and streptophytan (dark green) algae in sea and fresh waters respectively during the Neoproterozoic era. (B) During the global-scale glaciations that characterized the Cryogenian period (750-650 mya), the ancestors of green algal lineages survived in fresh and brackish water refugia in brine channels, or in equatorial ocean waters, where the ice layer was thinner or absent. (C) After the subsequent global meltdown, during the Paleozoic era, Chlorophyceae, Pedinophyceae, Streptophyta, Trebouxiophyceae and some Ulvophyceae lineages proliferated in terrestrial and freshwater environments, while additional ulvophyceans and Prasinophytes re-colonized and differentiated in seawater habitats. Adapted from Becker & Main (2009).

Organellar genome evolution and repetitive elements proliferation

Organellar genomes architecture in green algae

The long and convoluted evolutionary history of green algae is mirrored as well by their organellar genomes. Mitochondria and chloroplasts derive from two distinct events involving endosymbiosis of a *Ricketsia*-like bacterium and Cyanobacterium, respectively. It is widely accepted that all modern mitochondria and all but one chloroplast (*Paulinella* being the exception) have common origins (Boxma *et al.*, 2005; Cox *et al.*, 2008; Zimorski *et al.*, 2014; Ponce-Toledo *et al.*, 2017). The organellar genomes resemble the genome of their bacterial ancestors. They have been classically depicted as circular mapping DNA molecules retaining only a fraction of the original gene content, with most of the genes being lost or transferred to the host nucleus. Following closer inspection, however, chloroplast and mitochondrial genomes display a wide array of sizes and architectures (Burger *et al.*, 2003; Smith & Keeling, 2015).

Chloroplast genomes (cpDNA) have been considered for a long time to be more conserved that their mitochondrial counterparts (Smith & Keeling, 2015; Smith, 2017; de Vries & Archibald, 2018). While circular mapping mitochondrial genomes (mtDNA) are most commonly reported in green algae (Pombert *et al.*, 2006a; Smith *et al.*, 2011; Melton *et al.*, 2015), alternative architectures are well known, especially within the

Chlorophyceae, e.g.: *Chlamydomonas reinhardtii*, *Scenedesmus obliquus* and *Pandorina morum* mitochondrial genomes are reported to possess linear chromosomes (Gray & Boer, 1988; Nedelcu *et al.*, 2000; Hamaji *et al.*, 2017). In the non-photosynthetic genus *Polytomella*, not only the mitochondrial genome is linear, but in some species it is fragmented into two distinct molecules (Smith *et al.*, 2013). In comparison, virtually all green algal chloroplast genomes are circular, gene-rich molecules between 100-200 kb, encoding 80-100 genes (Figure 5.2). This paradigm held true despite the evidence of linear and branched DNA molecules in chloroplasts of land plants (Bendich, 2004; Oldenburg & Bendich, 2016). Two recent studies, however, suggested that green algal chloroplast genomes may have a more intricate genomic architecture that has been overlooked so far. The chloroplast genome of the epiphytic green alga *Koshicola spirodelophila* is in fact fragmented over three large circular DNA molecules, for a total length of ca. 385 kb (Watanabe *et al.*, 2016). The most deviant chloroplast genome so far has been described for the order of the Cladophorales (Chapter 4). The Cladophorales chloroplast genome is in fact fragmented over multiple palindromic single-stranded DNA chromosomes, which fold intramolecularly into hairpin chromosomes. The gene content seems highly reduced and many genes commonly found in green algal chloroplast genomes have been transferred to the nucleus. Abundant non-coding hairpin chromosomes suggest frequent events of recombination during the evolution of this deviant chloroplast genome, Chapter 4 (Del Cortona *et al.*, 2017).

Proliferation of repetitive elements: expansion and fragmentation of organellar genomes

In addition to the unexpected variability in green algal organellar genome architecture, one can also note a high variability in genome size. The latter is often due to proliferation of repetitive elements. Inflated chloroplast genomes emerged several times independently within the Chlorophyceae. The *Tetrabaena socialis* chloroplast genome is ca. 405 kb, the *Volvox carteri* chloroplast genome is larger than 420 kb and is composed for more than 80% by non-coding DNA packed with 14-79 bp palindromic repeats (Smith & Lee, 2009; Featherston *et al.*, 2016). Instead, the expansion of the *Dunaliella salina* chloroplast genome is to be ascribed mainly to intron proliferation (ca. 30% of the 269 kb cpDNA), with a mean of 1 intron every two genes (Smith *et al.*, 2010). The largest Chlorophycean chloroplast genome sequenced so far belongs to

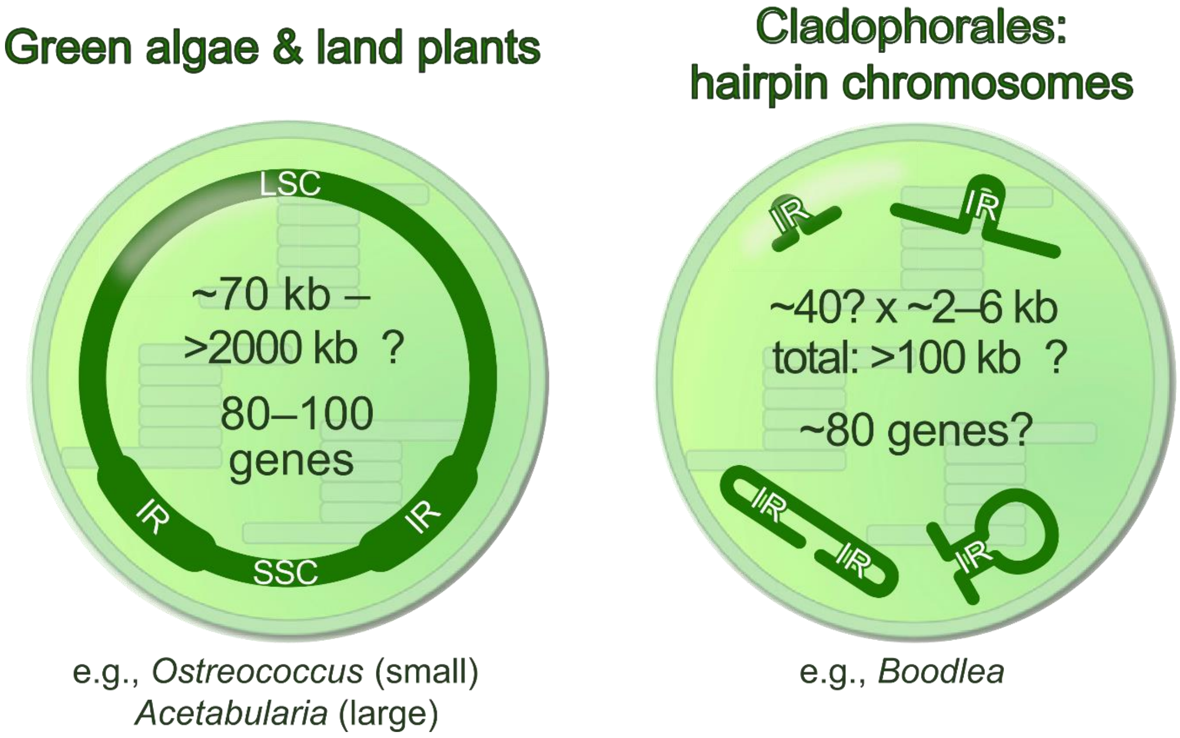


Figure 5.2: Green algal chloroplast genomes.

The majority of green algal chloroplast genomes sequenced so far are circular-mapping molecules between 100-200 kb in size and coding for 80-100 genes (left). Often they display a quadripartite structure where two inverted repeats (IR) divide a large single copy region (LSC) from a small single copy region (SSC). Smaller (e.g.: *Ostreococcus tauri*) and larger (*Acetabularia acetabulum*) chloroplast genomes have been described. The chloroplast genome of Cladophorales green algae, instead, is fragmented over multiple hairpin chromosomes (right). Adapted from de Vries & Archibald (2018).

Floydiella terrestris (521 kb), with almost 50% of its cpDNA composed of repetitive elements. The *Floydiella* chloroplast genome harbors more than 1,000 copies of highly conserved (>95% sequence similarity) 30 bp repeats (Brouard *et al.*, 2010).

Despite that the proliferation of introns and repetitive elements have bloated the chloroplast genomes of several clades of Chlorophyceae, the most extreme scenarios evolved within the Ulvophyceae. In the Dasycladales, *Acetabularia acetabulum* cpDNA has been estimated to be larger than 2 Mb in length (Burton & Hugh, 1970; Padmanabhan & Green, 1978; Tymms & Schweiger, 1985). The *Acetabularia* chloroplast genome is inflated by long (ca. 10 kb) tandem repeats, which makes the assembly of the cpDNA by short reads a challenge (Tymms & Schweiger, 1985; de Vries *et al.*, 2013). Only 300 kb of the *Acetabularia* cpDNA has been sequenced. The assembled contigs show extremely long intergenic regions and open reading frames with no similarity to known protein-coding genes (de Vries *et al.*, 2013). The *Boodlea composita* (Cladophorales) chloroplast genome size is unknown. Only 34 hairpins chromosomes harboring protein-coding genes have been characterized so far, resulting in a total length of 91 kb, 77% of it being inverted repeats. However, the abundancy of empty hairpin chromosomes and long terminal repeat retrotransposons (RT-LTRs) in the chloroplast DNA suggests that the actual cpDNA is much larger (Del Cortona *et al.*, 2017).

An interesting observation is that bloated chloroplast and mitochondrial genomes often co-occur within the same species (Smith & Keeling, 2015). In addition to the inflated chloroplast genomes, mitochondrial genomes both in *Volvox carteri* and *Dunaliella salina* mtDNAs are inflated by the proliferation of short palindromic sequences and introns, respectively (Smith & Lee, 2009; Smith *et al.*, 2010). Despite considerable effort, we did not manage to assemble the mitochondrial genome of *Boodlea composita.* Partial sequences, however, suggest a genome inflated by arrays of tandem repeats. The entire genome organization is unclear, but 52 contigs were assembled for a total length of more than 730 Kb. Interestingly, inverted repeats and RT-LTRs abundant in the chloroplast genome were absent in mitochondrial DNA, suggesting that the expansion of these two organellar genomes was caused by the proliferation of two distinct repetitive elements, Chapter 4 (Del Cortona *et al.*, 2017).

The dependency of chloroplasts and mitochondria on nuclear-encoded proteins for crucial repair-, replication-, and expression-related functions is probably accountable for the intraspecific common traits shared between the organellar genomes. Molecular cross talk and DNA transfer between chloroplasts, mitochondria and the nucleus, together with dual targeting of DNA maintenance proteins to both organelles, plays a huge role in the proliferation of repetitive elements in the cpDNA and mtDNA within the same organism (Leister, 2005; Carrie *et al.*, 2009; Kleine *et al.*, 2009). Evidence of transfer of DNA from one organellar genome to the other have been found in Ulotrichales (Turmel *et al.*, 2016b). In *Volvox carteri*, cpDNA and mtDNA repetitive

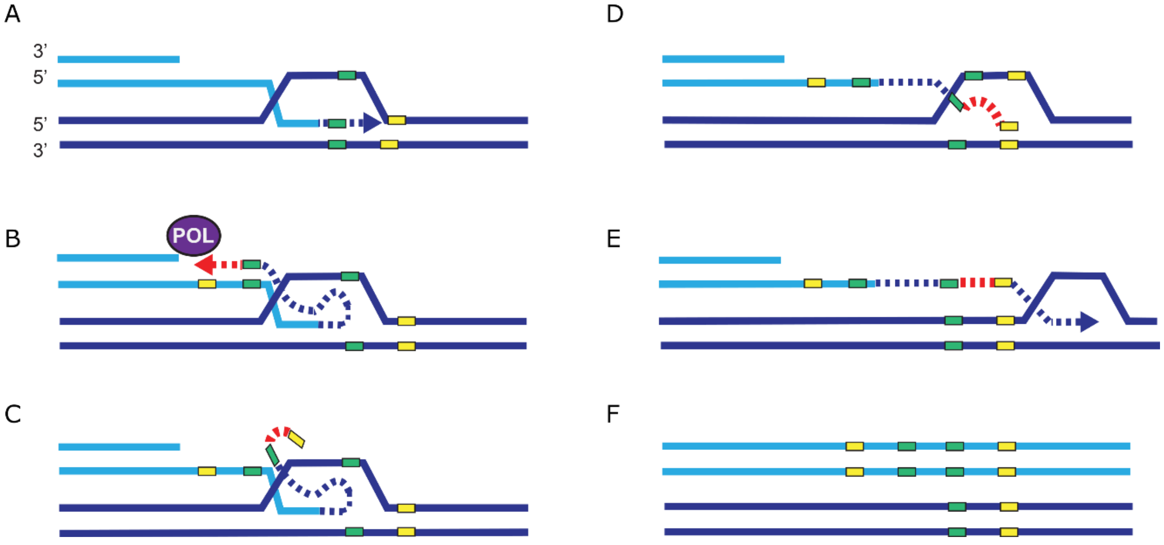


Figure 5.3: Amplification of repeats mediated by break-induced replication.

(A) Break-Induced Replication (BIR) repair causes a ssDNA strand from a broken organellar genome (light blue) to align to a homologous region in an integer genome (drak blue), which acts as a template. (B) Dissociation of 3' end from the template and annealing at an ectopic homologous region (green) in the ssDNA accumulated behind the replication repair bubble. The annealed region can be used by a DNA polymerase to initiate DNA synthesis that eventually disrupts. In case of repetitive elements in the broken strand, the chances of ectopic homologous pairing increases dramatically. (C) A second template switch reanneals the *de novo* synthetized ssDNA to the original template. (D), (E) BIR replication completes the synthesis of the broken strand following the integer template. (F) As a consequence of the faulty BIR repair, the repeat elements in the repaired genome are amplified. Green and yellow boxes: repetitive elements that can consist of 1–10 bp. POL: DNA polymerase. Adapted from Sakofsky & Malkova (2017).

elements were both identified in the nuclear genome as well (Smith & Lee, 2009). The presence of the same inverted repeats and RT-LTRs in *Boodlea* nuclear and chloroplast genome suggests a RT-LTR driven invasion from nuclear repeats. A working hypothesis is that first RT-LTRs inflated the chloroplast genomes of the Cladophorales ancestor; the cpDNA subsequently degenerated into hairpin chromosomes through recombination and displacement of the repeats from the lagging strand during replication. Repeat proliferation probably is a common theme within the Cladophorales. Based on cytofluorometry, their predicted nuclear genomes suggest that Cladophorales genome sizes are huge when compared to the other green algae, ca. 880 Mb-2,000 Mb (Kapraun, 2007).

The mechanisms responsible for the duplication of the non-coding sequences and the consequent proliferation of repetitive elements and inflation of organellar genomes are possibly DNA slippage events during replication and break-induced replication repair (BIR). Polymerase slippage during DNA replication occurs due to DNA misalignment or formation of hairpin structures in correspondence to tandem and inverted repeats (Massouh *et al.*, 2016). BIR repair system occurs at breaks in the DNA and it involves the correction of a broken DNA molecule by the joining of one of the two broken strands with a homologous sequence in another DNA molecule (Figure 5.3). If the break is located in a repetitive region, chances are high that the homologous pairing is faulty. Since BIR is an error-prone repair mechanism, BIR-induced non-homologous joining within coding sequences would result in deleterious mutations (Christensen, 2013). Organellar genomes therefore require an additional, more accurate mechanism of repair to retain functional genes. The low substitution rates observed in the coding regions of most green algal organellar genomes is ascribed to gene conversion. This mechanism is responsible for accurate homology-guided repair in coding regions, where breaks in a copy of the organellar genome are corrected by using another copy of the genome as template. In gene conversion, both strands of the broken molecules are corrected simultaneously, preventing the misplacing that could happen in BIR-mediated repairs. Therefore, while fidelity of the coding sequences is maintained by gene conversion, the non-coding regions can accumulate repeats by BIR-mediated repairs, resulting in expansion of the organellar genomes (Christensen, 2013; Smith & Keeling, 2015; Smith, 2016).

Future perspectives

How sharp is phylotranscriptomics for green algae?

Chloroplast phylogenomic studies contributed considerably toward resolving phylogenetic relationships in several groups of photosynthetic eukaryotes, including land plants (Zhong *et al.*, 2013; Ruhfel *et al.*, 2014; Zhong *et al.*, 2014; Lemieux *et al.*, 2016), green algae (Lemieux *et al.*, 2014b; Lemieux *et al.*, 2014a; Lemieux *et al.*, 2015; Turmel *et al.*, 2015; Leliaert *et al.*, 2016; Turmel *et al.*, 2016a), and red algae (Costa *et al.*, 2016; Díaz-Tapia *et al.*, 2017; Muñoz-Gómez *et al.*, 2017). Despite considerable effort, however, several chloroplast phylogenomic studies did not converge toward a consensus topology in the core Chlorophyta, nor were they congruent with nuclear-based marker studies (Cocquyt *et al.*, 2010b; Sun *et al.*, 2016; Fang *et al.*, 2018). The debate on the branching order of the UTC clade within the core Chlorophyta and even on the monophyly of trebouxiophytes and ulvophyceans is still ongoing.

Transcriptome sequencing by high-throughput sequencing technologies opened the treasure trove of nuclear markers for phylogenetic studies. Shotgun sequencing of cDNAs released the need of primer-specific amplification, and allowed the simultaneous sequencing of thousands of transcripts. Moreover, phylotranscriptomics allowed the analysis of more and unlinked markers coming from distinct chromosomes (important to study lineage sorting) than chloroplast or mitochondrial phylogenomic studies. The relatively low cost of transcriptome sequencing allows the profiling of lineages neglected by genome sequencing projects and lineages where genome sequencing seems an unaccountable challenge, e.g.: the huge dinoflagellates genomes composed by liquid crystalline chromosomes (Shoguchi *et al.*, 2013; Maeshima *et al.*, 2016). As a consequence, the number of transcriptome data publicly available is growing exponentially (van Dijk *et al.*, 2014; Muir *et al.*, 2016). Furthermore, state-of-the-art applications, such as single cell sequencing and metatranscriptome sequencing from environmental samples, allows sequencing of undescribed species and species that cannot be cultured (Pesant *et al.*, 2015; Carradec *et al.*, 2018). Phylotranscriptomics approaches already resulted in resolving long standing phylogenetic problems, such as disentangling the relationships between charophyte algae and land plants (Wickett *et al.*, 2014), reconstructing the ancient divergence of several animal groups (Bazinet Adam *et al.*, 2016) and the evolution of thecate dinoflagellates (Janouškovec *et al.*, 2017).

Despite its phylogenetic power and its broad-spectrum adaptability, phylotranscriptomics suffer from a number of shortcomings inherent to sequencing technology and sample preparation. In absence of a reference genome, transcriptomes only offer a partial coverage of the full gene space of an organism, and assembled transcripts are often fragmented or incomplete. Despite we focused on single-copy gene families, orthology inference can be complicated by redundancy in the transcriptome assembly, allelic variants and undetected contaminant sequences (Chapter 2).

Our phylotranscriptomic analyses based on green algal datasets represent a big leap forward to resolve the relationships between the core Chlorophyta lineages and to unravel the evolution and diversification of green seaweeds. A vast nuclear dataset was provided for 55 *Viridiplantae* representative species, respectively 50 and 5-10 times larger than previous nuclear and chloroplast marker based studies. To obtain an unbiased and highly supported phylogeny of green algae, multiple filtering methods, partition strategies, evolutionary models and complementary phylogenetic analyses were used. More importantly, the topology inferred was mostly congruent with the distribution of several ultrastructural characters. Phylotranscriptomics revealed to be a more powerful and versatile tool than organellar-based phylogenomics, however, there is still space for improvement and phylotranscriptomics does not represent the remedy for resolving all the phylogenetic debates that are still standing.

On the other hand, our study suffered as well from inherent disadvantages of *de novo* transcriptomic sequencing. The removal of contaminant sequenced from other green algae is a challenging task which might not have been properly addressed in our pipeline. For certain ulvophyceans orders, discrimination between a *bona fide* sequence and a green algal contaminant is hampered by the lack of reference sequences from closely related species. This problem is exacerbated for samples collected in the field, which are often composed by a mixture of micro- and macroscopic organisms where the species of interest represents only the most abundant eukaryote. In addition to increase the noise in the phylogenetic signal, contaminant sequences can lead to wrong phylogenetic inference (Laurin-Lemay *et al.*, 2012). In an ideal genomic study, such kind of problem would be solved by filtering the sequences by their k-mer frequencies distributions, since in each sequenced library each genome is supposed to have a peculiar k-mer frequency distribution (Lin & Liao, 2016). Unfortunately, this approach does not extend to transcriptomes. Recently, isoform abundance has been proposed to dereplicate redundant sequences in *de novo* assembled transcriptomes. These techniques are proposed to identify also potential contaminant, due to the assumption that contaminating sequences should be expressed at a lower level than genuine sequences, and may help to resolve the Bryopsidales-Chlorophyceae-Ulvophyceae radiation (Simion *et al.*, 2017; Schvartzman *et al.*, 2018; Simion *et al.*, 2018).

Toward understanding green seaweeds and green algae evolution

It may sound like a paradox, but in the era of Big Data, what is needed to obtain an exhaustive answer is even more data. To fully resolve the relationships between the core Chlorophyta lineages we will need a much denser, broader and balanced taxon sampling, especially for those clades where only data from one species are currently available (e.g.: Dasycladales and Ignatiales). The availability of one or more reference genomes for each Order of green algae will be without any doubt beneficial. Reference genomes from closely related organisms are fundamental both for orthology inference and for the discovery of molecular innovations driving the diversification of Chlorophyta and green seaweeds in particular. In addition, genomes carry features in their non-coding regions, such as position of transposable elements (Takahashi *et al.*, 2001), introns and ultraconserved elements flanking coding regions (Jarvis *et al.*, 2014), that are potentially phylogenetic informative and could be the key to solve topological uncertainties.

Employing comparative transcriptomics to unravel the molecular innovations behind green seaweeds diversity may be a short-term solution, since a complete overview will probably be based only on the full understanding of the relationships between green seaweeds lineages and on the dynamics of their genomes. Unless directed toward the detection of specific gene families, as shown for the evolution of thecate dinoflagellates (Janouškovec *et al.*, 2017), incompleteness of a transcriptome does not allow a safe statement on the loss of a gene or gene family, or the expansion of gene families. Moreover, the intrinsic redundancy of the transcriptomes and the high error margin in gene family size estimation affect comparative analyses on gene family expansions.

Nevertheless, transcriptome sequencing can be used to address several fundamental questions. Almost one million protein-coding sequences and 70 thousand gene families have been described in this study. Despite the intrinsic limitations of the transcriptomic data discussed previously, this dataset should consent the analysis of lineage-specific gene families and gene family expansion, at least where multiple species are present. RNA libraries generated during different stages in the life cycle of green seaweeds may help to unveil the genes behind seaweeds development and to discover regulators in the cell growth and division. Recently, an atlas of *Caulerpa racemosa* (Bryopsidales) transcripts indicated differential patterns of transcripts distribution and accumulation between the different subcellular structures of the giant cell (Coneva & Chitwood, 2015; Ranjan *et al.*, 2015). With a similar approach, we generated RNA-seq libraries from *Acetabularia acetabulum* rhizoid, stalks, hairs and caps, with the aim to perform a differential expression analysis within these subcellular compartments. In addition to that, a comparison of the transcripts expressed or targeted to similar subcellular compartments in *Caulerpa* and in *Acetabularia* should reveal if similar molecular mechanisms have been recruited (perhaps through convergent evolution), or if completely different strategies are responsible for subcellular specialization of the siphonous cell in Bryopsidales and in Dasycladales.

A last, additional observation is that only 620 single-copy genes are conserved among the genomes that populated picoPLAZA database (of which, only 539 were usable for our study, Chapter 3). This is in stark contrasts with the number of single-copy gene families identified in similar studies: 2,022 single-copy gene families were used to solve the bird phylogeny (Jarvis *et al.*, 2014); 1,719 gene families for Metazoa (Simion *et al.*, 2017); 1,478 for an insect phylogeny (Misof *et al.*, 2014) and 844 quasi single-copy gene families to disentangle the origin of land plants (Wickett *et al.*, 2014). While avian diversification is relatively recent (~65 mya) – which may explain the high number of shared single-copy genes – the insect diversification is estimated to be more ancient (~479 mya) (Jarvis *et al.*, 2014; Misof *et al.*, 2014). The lower number of single-copy genes identified in land plants instead may reflect its evolutionary history populated by subsequent events of whole genome duplications (Vanneste *et al.*, 2014; Wickett *et al.*, 2014; Lohaus & Van de Peer, 2016; Clark & Donoghue, 2018). Why such scarcity of single-copy gene families in green algae when compared to other species? One may look at microscopic green algae as very simple organisms and postulate that they may have a simple genome and with a reduced gene space and conserved single-copy gene set. However, the microscopic *Chlamydomonas reinhardtii*, model and representative of green algae, has a genome and gene space size on the same order of magnitude of *Arabidopsis thaliana* (genome size of 111 Mb and 135 Mb, coding for ~17 thousand and ~27 thousand genes, respectively).

The extremely long evolutionary distances may account for the low number of single-copy gene families shared among green algae. Green algae experienced the extraordinary conditions of the Cryogenian era and survived all the mass extinction events (Raup & Sepkoski, 1982; Raup & Sepkoski, 1984; Butterfield, 2007). Several key lineages that could break some long branches artefacts and increase the resolution of the phylogenetic signal gone extinct, hiding addition critical data. Furthermore, green algal genomes have adapted and diverged during the last billion years. Green algae followed independent parallel paths toward macroscopic growth coupled to genome expansion (like in Cladophorales) and toward size and genome reduction, like in Mamiellales (Derelle *et al.*, 2006; Kapraun, 2007; Palenik *et al.*, 2007). These dramatic genome evolution events could make the circumscription of single-copy genes non-trivial or unbalanced in the lineages sequenced so far.

This study represents an additional step toward the understanding of green seaweeds evolution and differentiation. Yet, fundamental questions are still standing. There is so little known and still so much to be understood that a transcriptome is not suitable to capture the whole complexity of such intriguing organisms, but undoubtedly it can provide precious insights. I am confident that many groundbreaking discoveries on green algae and green seaweeds evolution are out there waiting to surface. Especially because, in addition to their sometimes overlooked complex evolution, green algae have the habit to hide in plain sight (Verbruggen & Tribollet, 2011; Leliaert *et al.*, 2016; Watanabe *et al.*, 2016).

“Ed ogni via conserva in sé la stessa eco di un tempo

Quando fiori blu crescevan sul cemento

Sbarbi nei rioni, guerra dei bottoni, tutti campioni

Giocavamo ad inventarci i nomi

Sapendo già che ce ne saremo andati prima o poi

Portando terra, vento e fuoco via con noi”

Chico MD con Fritz – Dopo di noi la quiete

Curriculum Vitae

First name/ Surname	Andrea Del Cortona
Telephone	+32 470 255667
E-mail	andrea.delcortona@gmail.com
Website	http://www.linkedin.com/pub/andrea-del-cortona/74/754/86a

At a Glance: Comparative and Functional Genomics and Transcriptomics, Biotechnology, Molecular Biology, Biochemistry

KEY COMPETENCES:

- BIOINFORMATICS: data mining, data integration, single-cell sequencing, NGS and SMRT sequencing, genome and transcriptome *de novo* assembly, genome and transcriptome functional annotation, comparative transcriptomics, functional genomics, phylogenomics, programming (sh, python, R, MySQL)
- MOLECULAR BIOLOGY: PAGE/agarose electrophoresis, GATEWAY® technology, recombinant DNA techniques, molecular genetics, fusion proteins, isolation and purification of high quality nucleic acids and proteins, (semi-) quantitative PCR, cell culturing
- BIOCHEMISTRY: transient transformation assays, protein (co)-immunoprecipitation, mass spectrometry, protein targeting, protein engineering, (fluorescence) microscopy, Western analysis

PROJECTS:

- Intracellular comparative transcriptomics of green algae
- Function and characterization of PI-PLCs in immune signaling
- *Ulva mutabilis* genome sequencing project
- A two-hybrid system to detect transient protein-protein interactions at the plasma membrane
- Comparative phylotranscriptomics and molecular evolution of green algae
- Molecular characterization of the elusive Cladophorales chloroplast genome
- Batrachospermales organellar genomes sequencing project
- Genetic improvement of broad beans and witloof chicory

EDUCATION AND TRAINING

Title of qualification awarded	**Doctor of Philosophy (PhD): Biology** Research Group of Comparative Network Biology lab (Prof. Klaas Vandepoele, VIB-UGent) Research Group of Phycology (Prof. Olivier De Clerck)
Dates	From Dec 2013 - ongoing
Organization	**Plant System Biology-VIB, Ghent University**
Thesis	"Molecular evolution and development of green algae"
Principal subjects	**Comparative genomics, comparative transcriptomics, functional genomics, phylogenomics, bioinformatics, molecular biology, programming, marine biology, cell cultures**

Title of qualification awarded	**Master of Science (MSc): Plant Biotechnology** specialization: Molecular Plant Breeding and Phytopathology
Dates	From Feb 2011 to Jul 2013
Organization	**Wageningen University and Research centre (WUR)**
Theses	“Function of Phospholipases C (PLC) in immune signaling” "Identification of PLC-interacting proteins using a Plant two-Hybrid system”
Principal subjects	**Plant-microbe interaction, plant-parasite interaction, biotechnology, molecular biology, biochemistry**

Title of qualification awarded	**Bachelor degree (BSc): Biotechnology** curriculum: Biomolecular Sciences
Dates	From Sep 2006 to Mar 2010
Organization	**University of Bologna**
Thesis	“An evaluation of gene and protein expression in the skeletal muscle of rats subjected to physical exercise”
Principal subjects	**Molecular biology, molecular genetics, molecular microbiology and virology, biochemistry**

LANGUAGES

Italian	native or bilingual proficiency
English	full professional proficiency
Spanish	limited working proficiency

REFERENCES

Prof. Frederik Leliaert
Meise Botanic Garden
Nieuwelaan 38
1860 Meise
Belgium
T: +32 2 260 0939
frederik.leliaert@gmail.com

Prof. Klaas Vandepoele
Comparative Network Biology
VIB / Ghent University
Bioinformatics & Systems Biology
Technologiepark 927
9052 Gent
Belgium
T: +32 9 331 3822
klpoe@psb.vib-ugent.be

Prof. Olivier De Clerck
Phycology Research Group
Ghent University
Krijgslaan 281, building S8
9000 Gent
Belgium
T: +32 9 264 8500
olivier.declerck@ugent.be

List of scientific publications

1. **Andrea Del Cortona**, François Bucchini, Chris Jackson, Michiel Van Bel, Endymion Cooper, Pavel Škaloud, Sofie D'hondt, Heroen Verbruggen, Charles Delwiche, Frederik Leliaert*, Klaas Vandepoele*, Olivier De Clerck* – Reconstruction of the early diversification of green seaweeds – manuscript in preparation

2. **Andrea Del Cortona**, Klass Vandepoele, Olivier De Clerck – Intracellular comparative transcriptomics of siphonous green algae *Acetabularia acetabulum*, the mermaid's wineglass – manuscript in preparation

3. Olivier De Clerck, Shu-Min Kao, Kenny A. Bogaert, Jonas Blomme, Fatima Foflonker, Michiel Kwantes, Emmelien Vancaester, Emmelien Vanderstraten, Eylem Aydogdu, Jens Boesger, Giammaria Califano, Bénédicte Charrier, Rachel Clewes, **Andrea Del Cortona**, Sofie D'hondt, Noe Fernandez-Pozo, Claire M. Gachon, Marc Hanikenne, Linda Latterman, Frederik Leliaert, Xiaojie Liu, Christine A. Maggs, Zoe A. Popper, John A. Raven, Michiel Van Bel, Per K. I. Wilhelmsson, Juliet C. Coates, Stefan A. Rensing, Dominique Van Der Straeten, Assaf Vardi, Lieven Sterck, Klaas Vandepoele, Yves Van De Peer, Thomas Wichard, John H. Bothwell − Insights into the Evolution of Multicellularity from the Sea Lettuce Genom − Current Biology 28, 2921-2933, September 24, 2018, https://doi.org/10.1016/j.cub.2018.08.015.

4. **Andrea Del Cortona**, Frederik Leliaert – Molecular evolution and morphological diversification of Ulvophytes (Chlorophyta) – Perspectives in Phycology 5(1), 27-43, June 1, 2018, https://doi.org/10.1127/pip/2017/0075

5. Monica Orlandi Paiano*, **Andrea Del Cortona***, Olivier De Clerck, Orlando Necchi Jr – Evolutionary insights from the mitochondrial genomes of Batrachospermales (Rhodophyta) – Mitochondrial DNA Part B: Resources 3(2), 607–610, May 23, 2018, https://doi.org/10.1080/23802359.2018.1473734

6. Monica Orlandi Paiano*, **Andrea Del Cortona***, Joana F. Costa, Shao-Lun Liu, Heroen Verbruggen, Olivier De Clerck, Orlando Necchi Jr – Organization of plastid genomes in the freshwater red algal order Batrachospermales (Rhodophyta) – Journal of Phycology 54, 25–33, February 2, 2018, https://doi.org/10.1111/jpy.12602

7. **Andrea Del Cortona***, Frederik Leliaert*, Kenny A. Bogaert, Monique Turmel, Christian Boedeker, Jan Janouškovec, Juan M. Lopez-Bautista, Heroen Verbruggen, Klaas Vandepoele, and Olivier De Clerck – The plastid genome in Cladophorales green algae is encoded by hairpin chromosomes – Current Biology 27, 3771–3782, December 18, 2017, https://doi.org/10.1016/j.cub.2017.11.004

*** equal contribution / shared authorship.**

References

Abnizova, I., Boekhorst Te, R., Orlov, Y.L. (2017): Computational Errors and Biases in Short Read Next Generation Sequencing. − J Proteomics Bioinform.

Adl, S.M., Simpson, A.G.B., Farmer, M.A., Andersen, R.A., Anderson, R.O., Barta, J.R., Bowser, S.S., Brugerolle, G.U.Y., Fensome, R.A., Fredericq, S., James, T.Y., Karpov, S., Kugrens, P., Krug, J., Lane, C.E., Lewis, L.A., Lodge, J., H., L.D., Mann, D.G., Mccourt, R.M., Mendoza, L., Moestrup, Ø., Mozley-Standridge, S.E., Nerad, T.A., Shearer, C.A., Smirnov, A.V., Spiegel, F.W., Taylor, M.F.J.R. (2005): The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists. − Journal of Eukaryotic Microbiology 52: 399-451.

Aires, T., Moalic, Y., Serrao, E.A., Arnaud-Haond, S. (2015): Hologenome theory supported by cooccurrence networks of species-specific bacterial communities in siphonous algae (Caulerpa). − FEMS Microbiology Ecology 91: fiv067-fiv067.

An, D., Cao, X.H., Li, C., Humbeck, K., Wang, W. (2018): Isoform Sequencing and State-of-Art Applications for Unravelling Complexity of Plant Transcriptomes. − Genes 9.

Andersen, R.A. (2005): Algal culturing techniques. Elsevier, Amsterdam.

Andrews, S. (2010): FastQC: a quality control tool for high throughput sequence data. − available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S. (2009): MEME Suite: tools for motif discovery and searching. − Nucleic Acids Research 37: W202-W208.

Baker, M. (2012): De novo genome assembly: what every biologist should know. − Nat Meth 9: 333-337.

Ballesteros, J.A., Hormiga, G. (2016): A new orthology assessment method for phylogenomic data: unrooted phylogenetic orthology. − Molecular Biology and Evolution 33: 2117-2134.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A. (2012): SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. − Journal of Computational Biology 19: 455-477.

Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., Depamphilis, C., Albert, V.A., Aono, N., Aoyama, T., Ambrose, B.A., Ashton, N.W., Axtell, M.J., Barker, E., Barker, M.S., Bennetzen, J.L., Bonawitz, N.D., Chapple, C., Cheng, C., Correa, L.G.G., Dacre, M., Debarry, J., Dreyer, I., Elias, M., Engstrom, E.M., Estelle, M., Feng, L., Finet, C., Floyd, S.K., Frommer, W.B., Fujita, T., Gramzow, L., Gutensohn, M., Harholt, J., Hattori, M., Heyl, A., Hirai, T., Hiwatashi, Y., Ishikawa, M., Iwata, M., Karol, K.G., Koehler, B., Kolukisaoglu, U., Kubo, M., Kurata, T., Lalonde, S., Li, K., Li, Y., Litt, A., Lyons, E., Manning, G., Maruyama, T., Michael, T.P., Mikami, K., Miyazaki, S., Morinaga, S.I., Murata, T., Mueller-Roeber, B., Nelson, D.R., Obara, M., Oguri, Y., Olmstead, R.G., Onodera, N., Petersen, B.L., Pils, B., Prigge, M., Rensing, S.A., Riaño-Pachón, D.M., Roberts, A.W., Sato, Y., Scheller, H.V., Schulz, B., Schulz, C., Shakirov, E.V., Shibagaki, N., Shinohara, N., Shippen, D.E., Sørensen, I., Sotooka, R., Sugimoto, N., Sugita, M., Sumikawa, N., Tanurdzic, M., Theißen, G., Ulvskov, P., Wakazuki, S., Weng, J.K., Willats, W.W.G.T., Wipf, D., Wolf, P.G., Yang, L., Zimmer, A.D., Zhu, Q., Mitros, T., Hellsten, U., Loqué, D., Otillar, R., Salamov, A., Schmutz, J., Shapiro, H., Lindquist, E., Lucas, S., Rokhsar, D., Grigoriev, I.V. (2011): The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants. − Science 332: 960.

Barbrook, A.C., Howe, C.J., Kurniawan, D.P., Tarr, S.J. (2010): Organization and expression of organellar genomes. − Philosophical Transactions of the Royal Society B-Biological Sciences 365: 785-797.

Barbrook, A.C., Voolstra, C.R., Howe, C.J. (2014): The chloroplast genome of a *Symbiodinium* sp. clade C3 isolate. − Protist 165: 1-13.

Bashir, A., Klammer, A.A., Robins, W.P., Chin, C.S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., Sebra, R., Sorenson, J., Bullard, J., Yen, J., Valdovino, M., Mollova, E., Luong, K., Lin, S., Lamay, B., Joshi, A., Rowe, L., Frace, M., Tarr, C.L., Turnsek, M., Davis, B.M., Kasarskis, A., Mekalanos, J.J., Waldor, M.K., Schadt, E.E. (2012): A hybrid approach for the automated finishing of bacterial genomes. − Nature Biotechnology 30: 701-707.

Bazinet Adam, L., Mitter Kim, T., Davis Donald, R., Nieukerken Erik, J., Cummings Michael, P., Mitter, C. (2016): Phylotranscriptomics resolves ancient divergences in the Lepidoptera. − Systematic Entomology 42: 305-316.

Bechstädt, T., Jäger, H., Rittersbacher, A., Schweisfurth, B., Spence, G., Werner, G., Boni, M. (2018): The Cryogenian Ghaub Formation of Namibia – New insights into Neoproterozoic glaciations. − Earth-Science Reviews 177: 678-714.

Becker, B., Marin, B. (2009): Streptophyte algae and the origin of embryophytes. − Annals of Botany 103: 999-1004.

Bendich, A.J. (2004): Circular Chloroplast Chromosomes: The Grand Illusion. − The Plant Cell 16: 1661.

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

Berger, S., Kaever, M.J. (1992): Dasycladales: an illustrated monograph of a fascinating algal order. Thieme, Stuttgart.

Berlin, K., Koren, S., Chin, C.-S., Drake, J.P., Landolin, J.M., Phillippy, A.M. (2015): Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. − Nature Biotechnology 33: 623-630.

Berney, C., Pawlowski, J. (2006): A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. − Proceedings of the Royal Society B: Biological Sciences 273: 1867.

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

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

Bischoff, H.W.B., H. C. . (1963): Phycological Studies IV. Some Soil Algae From Enchanted Rock and Related Algal Species. University of Texas, Austin.

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D.D., Gurnon, J., Ladunga, I., Lindquist, E., Lucas, S., Pangilinan, J., Pröschold, T., Salamov, A., Schmutz, J., Weeks, D., Yamada, T., Lomsadze, A., Borodovsky, M., Claverie, J.-M., Grigoriev, I.V., Van Etten, J.L. (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., Salamov, A., Terry, A., Yamada, T., Dunigan, D.D., Grigoriev, I.V., Claverie, J.-M., Van Etten, J.L. (2010): The *Chlorella variabilis* NC64A Genome Reveals Adaptation to Photosymbiosis, Coevolution with Viruses, and Cryptic Sex. − The Plant Cell 22: 2943.

Bolger, A.M., Lohse, M., Usadel, B. (2014): Trimmomatic: a flexible trimmer for Illumina sequence data. − Bioinformatics 30: 2114-2120.

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., Mcginnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., Zaretskaya, I. (2013): BLAST: a more efficient report with usability improvements. − Nucleic Acids Research 41: W29-W33.

Boussau, B., Szöllősi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V. (2013): Genome-scale coestimation of species and gene trees. − Genome Research 23: 323-330.

Boxma, B., De Graaf, R.M., Van Der Staay, G.W.M., Van Alen, T.A., Ricard, G., Gabaldón, T., Van Hoek, A.H.a.M., Moon-Van Der Staay, S.Y., Koopman, W.J.H., Van Hellemond, J.J., Tielens, A.G.M., Friedrich, T., Veenhuis, M., Huynen, M.A., Hackstein, J.H.P. (2005): An anaerobic mitochondrion that produces hydrogen. − Nature 434: 74.

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L. (2016): Near-optimal probabilistic RNA-seq quantification. − Nature Biotechnology 34: 525.

Breinholt, J.W., Kawahara, A.Y. (2013): Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on Next-Gen data. − Genome Biology and Evolution 5: 2082-2092.

Bremer, K. (1985): Summary of green plant phylogeny and classification. − Cladistics 1: 369-385.

Brocks, J.J., Jarrett, A.J.M., Sirantoine, E., Hallmann, C., Hoshino, Y., Liyanage, T. (2017): The rise of algae in Cryogenian oceans and the emergence of animals. − Nature 548: 578.

Brodie, J., Maggs, C.A., John, D.M., Blomster, J. (2007): The green seaweeds of Britain and Ireland. British Phycological Society, United Kingdom.

Brooks, F., Rindi, F., Suto, Y., Ohtani, S., Green, M. (2015): The Trentepohliales (Ulvophyceae, Chlorophyta): an unusual algal Order and its novel plant pathogen—*Cephaleuros*. − Plant Disease 99: 740-753.

Brouard, J.-S., Otis, C., Lemieux, C., Turmel, M. (2010): The exceptionally large chloroplast genome of the green alga *Floydiella terrestris* illuminates the evolutionary history of the Chlorophyceae. − Genome Biology and Evolution 2: 240-256.

Brown, J.W., Sorhannus, U. (2015): A Molecular Genetic Timescale for the Diversification of Autotrophic Stramenopiles (Ochrophyta): Substantive Underestimation of Putative Fossil Ages. − PLoS ONE 5.

Buchfink, B., Xie, C., Huson, D.H. (2014): Fast and sensitive protein alignment using DIAMOND. − Nature Methods 12: 59.

Burger, G., Gray, M.W., Franz Lang, B. (2003): Mitochondrial genomes: anything goes. − Trends in Genetics 19: 709-716.

Burr, F.A., West, J.A. (1971): Comparative ultrastructure of the primary nucleus in *Bryopsis* and *Acetabularia*. − Journal of Phycology 7: 108-113.

Burt, A., Trivers, R. (2006): Genes in conflict: the biology of selfish genetic elements. Harvard University Press, Cambirdge.

Burton, B.R.G., Hugh. (1970): *Acetabularia* chloroplast DNA: electron microscopic visualization. − Science 168: 981-982.

Butterfield, N., J., Knoll, A., H., Swett, K. (1994): Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen. − Lethaia 27: 76-76.

Butterfield, N.J. (2007): Macroevolution and Macroecology Through Deep Time. − Palaeontology 50: 41-55.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T. (2009): trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. − Bioinformatics 25: 1972-1973.

Carniel, F.C., Gerdol, M., Montagner, A., Banchi, E., De Moro, G., Manfrin, C., Muggia, L., Pallavicini, A., Tretiach, M. (2016): New features of desiccation tolerance in the lichen photobiont *Trebouxia gelatinosa* are revealed by a transcriptomic approach. − Plant Molecular Biology 91: 319-339.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D.J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M.B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., De Vargas, C., Iudicone, D., Bowler, C., Wincker, P. (2018): A global ocean atlas of eukaryotic genes. − Nature Communications 9: 373.

Carrie, C., Giraud, E., Whelan, J. (2009): Protein transport in organelles: Dual targeting of proteins to mitochondria and chloroplasts. − The FEBS Journal 276: 1187-1195.

Caspersson, T., Schultz, J. (1939): Pentose nucleotides in the cytoplasm of growing tissues. − Nature 143: 602-603.

Cavalier-Smith, T. (1981): Eukaryote kingdoms: Seven or nine? − Biosystems 14: 461-481.

Chapman, R.L., Borkhsenious, O., Brown, R.C., Henk, M.C., Waters, D.A. (2001): Phragmoplast-mediated cytokinesis in *Trentepohlia*: results of TEM and immunofluorescence cytochemistry. − International Journal of Systematic and Evolutionary Microbiology 51: 759-765.

Chapman, R.L., Henk, M.C. (1985): Observations on the habit, morphology and ultrastructure of *Cephaleuros parasiticus* (Chlorophyta) and comparison with *C. virescens*. − Journal of Phycology 21: 513-522.

Chappell, D.F., O'kelly, C.J., Floyd, G.L. (1991): Flagellar apparatus of the biflagellate zoospores of the enigmatic marine green alga *Blastophysa rhizopus*. − Journal of Phycology 27: 423-428.

Chen, F., Mackey, A.J., Stoeckert, C.J., Roos, D.S. (2006): OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. − Nucleic Acids Res 34: D363-368.

Chernomor, O., Von Haeseler, A., Minh, B.Q. (2016): Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. − Systematic Biology 65: 997-1008.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Muller, W.E., Wetter, T., Suhai, S. (2004): Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. − Genome Res 14: 1147-1159.

Chiari, Y., Cahais, V., Galtier, N., Delsuc, F. (2012): Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). − BMC Biology 10: 65.

Chihara, M., Inouye, I., Takahata, N. (1986): *Oltmannsiellopsis*, a new Genus of marine flagellate (Dunaliellaceae, Chlorophyceae). − Archiv für Protistenkunde 132: 313-324.

Chisholm, J.R.M., Dauga, C., Ageron, E., Grimont, P.a.D., Jaubert, J.M. (1996): 'Roots' in mixotrophic algae. − Nature 381: 382-382.

Christensen, A.C. (2013): Plant Mitochondrial Genome Evolution Can Be Explained by DNA Repair Mechanisms. − Genome Biology and Evolution 5: 1079-1086.

Clark, J.W., Donoghue, P.C.J. (2018): Whole-Genome Duplication and Plant Macroevolution. − Trends in Plant Science 23: 933-945.

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

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

Cocquyt, E., Verbruggen, H., Leliaert, F., Zechman, F., Sabbe, K., Clerck, O. (2009): Gain and loss of elongation factor genes in green algae. − BMC Evol Biol 9.

Colbath, G.K., Grenfell, H.R. (1995): Review of biological affinities of Paleozoic acid-resistant, organic-walled eukaryotic algal microfossils (including “acritarchs”). − Review of Palaeobotany and Palynology 86: 287-314.

Coneva, V., Chitwood, D.H. (2015): Plant architecture without multicellularity: quandaries over patterning and the soma-germline divide in siphonous algae. − Frontiers in Plant Science 6.

Costa, J.F., Lin, S.-M., Macaya, E.C., Fernández-García, C., Verbruggen, H. (2016): Chloroplast genomes as a tool to resolve red algal phylogenies: a case study in the Nemaliales. − BMC Evolutionary Biology 16: 205.

Cox, C.J., Foster, P.G., Hirt, R.P., Harris, S.R., Embley, T.M. (2008): The archaebacterial origin of eukaryotes. − Proceedings of the National Academy of Sciences 105: 20356.

Cremen, M.C.M., Leliaert, F., Marcelino, V.R., 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.

Crick, F.H. (1958): On protein synthesis. − Symposia of the Society for Experimental Biology 12: 138-163.

De Clerck, O., Kao, S.-M., Bogaert, K.A., Blomme, J., Foflonker, F., Kwantes, M., Vancaester, E., Vanderstraten, E., Aydogdu, E., Boesger, J., Califano, G., Charrier, B., Clewes, R., Del Cortona, A., D'hondt, S., Fernandez-Pozo, N., Gachon, C.M., Hanikenne, M., Latterman, L., Leliaert, F., Liu, X., Maggs, C.A., Popper, Z.A., Raven, J.A., Van Bel, M., Wilhelmsson, P.K.I., Coates, J.C., Rensing, S.A., Van Der Straeten, D., Vardi, A., Sterck, L., Vandepoele, K., Van De Peer, Y., Wichard, T., Bothwell, J.H. (2018): Insights into the Evolution of Multicellularity from the Sea Lettuce Genome. − Current Biology 28: 2921-2933.

De Vries, J., Archibald, J.M. (2018): Plastid genomes. − Current Biology 28: R336-R337.

De Vries, J., Habicht, J., Woehle, C., Huang, C., Christa, G., Wagele, H., Nickelsen, J., Martin, W.F., Gould, S.B. (2013): Is ftsH the key to plastid longevity in sacoglossan slugs? − Genome Biol Evol 5: 2540-2548.

Del Cortona, A., Leliaert, F., Bogaert, K.A., Turmel, M., Boedeker, C., Janouškovec, J., Lopez-Bautista, J.M., 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.e3776.

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

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piégu, B., Ball, S.G., Ral, J.-P., Bouget, F.-Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van De Peer, Y., Moreau, H. (2006): Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. − Proceedings of the National Academy of Sciences 103: 11647.

Díaz-Tapia, P., Maggs Christine, A., West John, A., Verbruggen, H. (2017): Analysis of chloroplast genomes and a supermatrix inform reclassification of the Rhodomelaceae (Rhodophyta). − Journal of Phycology 53: 920-937.

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

Duckett, J.G., Pressel, S. (2018): The evolution of the stomatal apparatus: intercellular spaces and sporophyte water relations in bryophytes—two ignored dimensions. − Philosophical Transactions of the Royal Society B: Biological Sciences 373.

Dunn, M.J., Kinney, G.M., Washington, P.M., Berman, J., Anderson, M.Z. (2018): Functional diversification accompanies gene family expansion of MED2 homologs in Candida albicans. − PLOS Genetics 14: e1007326.

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

Eddy, S.R. (2011): Accelerated profile HMM searches. − PLOS Computational Biology 7.

Edelman, D.B., Mcmenamin, M., Sheesley, P., Pivar, S. (2016): Origin of the vertebrate body plan via mechanically biased conservation of regular geometrical patterns in the structure of the blastula. − Progress in Biophysics and Molecular Biology 121: 212-244.

Ellis, T.H.N., Day, A. (1986): A hairpin plastid genome in barley. − The EMBO Journal 5: 2769-2774.

Emms, D.M., Kelly, S. (2015): OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. − Genome Biology 16.

Enright, A.J., Van Dongen, S., Ouzounis, C.A. (2002): An efficient algorithm for large-scale detection of protein families. − Nucleic Acids Research 30: 1575-1584.

Espelund, M., Minge, M.A., Gabrielsen, T.M., Nederbragt, A.J., Shalchian-Tabrizi, K., Otis, C., Turmel, M., Lemieux, C., Jakobsen, K.S. (2012): Genome fragmentation is not confined to the peridinin plastid in dinoflagellates. − PLoS One 7: e38809.

Falkowski, P.G., Katz, M.E., Knoll, A.H., Quigg, A., Raven, J.A., Schofield, O., Taylor, F.J.R. (2004a): The Evolution of Modern Eukaryotic Phytoplankton. − Science 305: 354.

Falkowski, P.G., Schofield, O., Katz, M.E., Van De Schootbrugge, B., Knoll, A.H. (2004b): Why is the Land Green and the Ocean Red? In *Coccolithophores: From Molecular Processes to Global Impact* (H. R. THIERSTEIN, J. R. YOUNG editors), pp. 429-453. − Springer Berlin Heidelberg, Berlin, Heidelberg.

Fang, L., Leliaert, F., Novis, P.M., 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, Z.-H., Penny, D., Zhong, B.-J. (2017): Evolution of the Chlorophyta: insights from chloroplast phylogenomic analyses. − Journal of Systematics and Evolution.

Featherston, J., Arakaki, Y., Hanschen, E.R., Ferris, P.J., Michod, R.E., Olson, B.J.S.C., Nozaki, H., Durand, P.M. (2018): 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, P.M., Smith, D.R. (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.

Finet, C., Timme, R.E., Delwiche, C.F., Marlétaz, F. (2010): Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. − Current Biology 20: 2217-2222.

Finet, C., Timme, Ruth e., Delwiche, Charles f., Marlétaz, F. (2012): Erratum - Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. − Current Biology 22: 1456-1457.

Floyd, G.L., O'kelly, C.J. (1990): Phylum Chlorophyta. Class Ulvophyceae. In *Handbook of Protoctista. The structure, cultivation, habitats and life histories of the eukaryotic microorganisms and their descendants exclusive of animals, plants and fungi* (L. MARGULIS, J. O. CORLISS, M. MELKONIAN, D. J. CHAPMAN editors), pp. 600-607. − Jones and Bartlett Publishers, Boston.

Fooden, I. (1980): The Macaques. Studies in ecology, Behavior and evolution. Wiley-Blackwell, New York.

Fott, B. (1971): Algenkunde, 2nd edition. VEB Gustav Fischer, Jena.

Friedl, T., O'kelly, C.J. (2002): Phylogenetic relationships of green algae assigned to the genus *Planophila* (Chlorophyta): evidence from 18S rDNA sequence data and ultrastructure. − European Journal of Phycology 37: 373-384.

Fučíková, K., Leliaert, F., Cooper, E.D., Škaloud, P., D'hondt, S., De Clerck, O., Gurgel, F., Lewis, L.A., Lewis, P.O., Lopez-Bautista, J., Delwiche, C.F., Verbruggen, H. (2014): New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. − Frontiers in Ecology and Evolution 2: 63.

Gabaldón, T. (2008): Large-scale assignment of orthology: back to phylogenetics? − Genome Biology 9.

Gabrielsen, T.M., Minge, M.A., Espelund, M., Tooming-Klunderud, A., Patil, V., Nederbragt, A.J., Otis, C., Turmel, M., Shalchian-Tabrizi, K., Lemieux, C. (2011): Genome evolution of a tertiary dinoflagellate plastid. − PLoS One 6: e19132.

Gao, C., Wang, Y., Shen, Y., Yan, D., He, X., Dai, J., Wu, Q. (2014): Oil accumulation mechanisms of the oleaginous microalga Chlorella protothecoides revealed through its genome, transcriptomes, and proteomes. − BMC Genomics 15: 582.

Gao, Z., Li, Y., Wu, G., Li, G., Sun, H., Deng, S., Shen, Y., Chen, G., Zhang, R., Meng, C., Zhang, X. (2015): Transcriptome Analysis in *Haematococcus pluvialis*: Astaxanthin Induction by Salicylic Acid (SA) and Jasmonic Acid (JA). − PLOS ONE 10: e0140609.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M.A., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., Mcfadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B. (2002): Genome sequence of the human malaria parasite Plasmodium falciparum. − Nature 419: 498.

Gile, G.H., Novis, P.M., Cragg, D.S., Zuccarello, G.C., Keeling, P.J. (2009): The distribution of Elongation Factor-1 alpha (EF-1a), Elongation Factor-Like (EFL), and a non-canonical genetic code in the Ulvophyceae: discrete genetic characters support a consistent phylogenetic framework. − J Eukaryot Microbiol 56: 367-372.

Glenn Travis, C. (2011): Field guide to next-generation DNA sequencers. − Molecular Ecology Resources 11: 759-769.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S. (2012): Phytozome: a comparative platform for green plant genomics. − Nucleic Acids Research 40: D1178-D1186.

Gouzy, J., Carrere, S., Schiex, T. (2009): FrameDP: sensitive peptide detection on noisy matured sequences. − Bioinformatics 25: 670-671.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A. (2011): Full-length transcriptome assembly from RNA-Seq data without a reference genome. − Nat Biotechnol 29: 644-652.

Graham, L.E., Mcbride, G.E. (1975): The ultrastructure of multilayered structures associated with flagellar bases in motile cells of *Trentepohlia aurea*. − Journal of Phycology 11: 86-96.

Grant, J.R., Katz, L.A. (2014): Building a phylogenomic pipeline for the eukaryotic Tree of Life - Addressing deep phylogenies with genome-scale data. − PLOS Currents Tree of Life 1.

Grantham, P.J., Wakefield, L.L. (1988): Variations in the sterane carbon number distributions of marine source rock derived crude oils through geological time. − Organic Geochemistry 12: 61-73.

Gray, M.W., Boer, P.H. (1988): Organization and expression of algal *Chlamydomonas reinhardtii* mitochondrial DNA. − Philosophical Transactions of the Royal Society of London. B, Biological Sciences 319: 135.

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

Green, B.R. (2011): Chloroplast genomes of photosynthetic eukaryotes. − The Plant Journal 66: 34-44.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G. (2016): Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. − European Journal of Epidemiology 31: 337-350.

Griffith, M., Walker, J.R., Spies, N.C., Ainscough, B.J., Griffith, O.L. (2015): Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. − PLOS Computational Biology 11: e1004393.

Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., Shabalov, I. (2014): MycoCosm portal: gearing up for 1000 fungal genomes. − Nucleic Acids Research 42: D699-D704.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010): New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. − Systematic Biology 59: 307-321.

Guiry, M.D. (2012): How many species of algae are there? − J. Phycol. 48: 1057-1063.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S. (2007): Quantifying similarity between motifs. − Genome Biology 8: R24.

Haber, A.H. (1962): Nonessentiality of concurrent cell divisions for degree of polarization of leaf growth. I. Studies with radiation-induced mitotic inhibition. − American Journal of Botany 49: 583-589.

Haber, A.H., Foard, D.E. (1963): Nonessentiality of concurrent cell divisions for degree of polarization of leaf growth. II. Evidence from untreated plants and from chemically induced changes of the degree of polarization. − American Journal of Botany 50: 937-944.

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

Hamaji, T., Kawai-Toyooka, H., Toyoda, A., Minakuchi, Y., Suzuki, M., Fujiyama, A., Nozaki, H., Smith, D.R. (2017): Multiple Independent Changes in Mitochondrial Genome Conformation in Chlamydomonadalean Algae. − Genome Biology and Evolution 9: 993-999.

Hamant, O., Heisler, M.G., Jönsson, H., Krupinski, P., Uyttewaal, M., Bokov, P., Corson, F., Sahlin, P., Boudaoud, A., Meyerowitz, E.M., Couder, Y., Traas, J. (2008): Developmental patterning by mechanical signals in *Arabidopsis*. − Science 322: 1650-1655.

Hämmerling, J. (1953): Nucleo-cytoplasmic relationships in the development of *Acetabularia*. − International Review of Cytology 2: 475-498.

Hanschen, E.R., Marriage, T.N., Ferris, P.J., Hamaji, T., Toyoda, A., Fujiyama, A., Neme, R., Noguchi, H., Minakuchi, Y., Suzuki, M., Kawai-Toyooka, H., Smith, D.R., Sparks, H., Anderson, J., Bakarić, R., Luria, V., Karger, A., Kirschner, M.W., Durand, P.M., Michod, R.E., Nozaki, H., Olson, B.J.S.C. (2016): The Gonium pectorale genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. − Nature Communications 7: 11370.

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

Herrmann, R.G., Possingham, J.V. (1980): Plastid DNA — The Plastome. In *Chloroplasts* (J. REINERT editor), pp. 45-96. − Springer Berlin Heidelberg, Berlin, Heidelberg.

Herron, M.D., Hackett, J.D., Aylward, F.O., Michod, R.E. (2009): Triassic origin and early radiation of multicellular volvocine algae. − Proceedings of the National Academy of Sciences 106: 3254.

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

Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., Vinh, L.S. (2018): UFBoot2: Improving the Ultrafast Bootstrap Approximation. − Molecular Biology and Evolution 35: 518-522.

Hoelzer, G.A., Meinick, D.J. (1994): Patterns of speciation and limits to phylogenetic resolution. − Trends in Ecology & Evolution 9: 104-107.

Hoffman, P.F., Kaufman, A.J., Halverson, G.P., Schrag, D.P. (1998): A Neoproterozoic Snowball Earth. − Science 281: 1342.

Hollants, J., Decleyre, H., Leliaert, F., De Clerck, O., Willems, A. (2011): Life without a cell membrane: Challenging the specificity of bacterial endophytes within Bryopsis (Bryopsidales, Chlorophyta). − BMC Microbiology 11: 255.

Hollants, J., Leliaert, F., Verbruggen, H., Willems, A., De Clerck, O. (2013): Permanent residents or temporary lodgers: characterizing intracellular bacterial communities in the siphonous green alga *Bryopsis*. − Proceedings of the Royal Society B: Biological Sciences 280.

Hori, T., Enomoto, S. (1978): Developmental cytology of *Dictyosphaeria cavernosa*. I. Light and electron microscope observations on cytoplasmic cleavage in zooid formation. − Botanica Marina 21: 401-408.

Howe, C.J., Nisbet, R.E., Barbrook, A.C. (2008): The remarkable chloroplast genome of dinoflagellates. − Journal of Experimental Botany 59: 1035-1045.

Hozza, M., Vinař, T., Brejová, B. (2015): How big is that genome? Estimating genome size and coverage from k-mer abundance spectra. In *String Processing and Information Retrieval: 22nd International Symposium, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings* (C. ILIOPOULOS, S. PUGLISI, E. YILMAZ editors), pp. 199-209. − Springer International Publishing, Cham.

Huang, X., Madan, A. (1999): Cap3: A DNA sequence assembly program.

Hudson, P.R., Waaland, J.R. (1974): Ultrastructure of mitosis and cytokinesis in the multinucleate green alga *Acrosiphonia*. − The Journal of Cell Biology 62: 274-294.

Hudson, R.R. (1990): Gene genealogies and the coalescent process. − Oxford Surveys in Evolutionary Biology 7: 1-44.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., Bork, P. (2017): Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. − Molecular Biology and Evolution 34: 2115-2122.

Hunt, B.E., Mandoli, D.F. (1996): A new, artificial sea water that facilitated groeth of large numbers of cells of *Acetabularia acetanulum* (Chlorophyta) and reduces the labor inherent in cell culture. − Journal of Phycology 32: 483-495.

I5k-Consortium. (2013): The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. − Journal of Heredity 104: 595-600.

International Rice Genome Sequencing Project. (2005): The map-based sequence of the rice genome. − Nature 436: 793.

Jackson, C., Knoll, A.H., Chan, C.X., 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.

Janouškovec, J., Gavelis, G.S., Burki, F., Dinh, D., Bachvaroff, T.R., Gornik, S.G., Bright, K.J., Imanian, B., Strom, S.L., Delwiche, C.F., Waller, R.F., Fensome, R.A., Leander, B.S., Rohwer, F.L., Saldarriaga, J.F. (2017): Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. − Proc Natl Acad Sci 114: E171-E180.

Janouškovec, J., Sobotka, R., Lai, D.-H., Flegontov, P., Koník, P., Komenda, J., Ali, S., Prášil, O., Pain, A., Oborník, M. (2013): Split photosystem protein, linear-mapping topology and growth of structural complexity in the plastid genome of *Chromera velia*. − Molecular Biology and Evolution 30: 2447-2462.

Jansen, R.K., Ruhlman, T.A. (2012): Plastid genomes of seed plants. In *Genomics of Chloroplasts and Mitochondria* (R. BOCK, V. KNOOP editors), pp. 103-126. − Springer Netherlands, Dordrecht.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y.W., Faircloth, B.C., Nabholz, B., Howard, J.T., Suh, A., Weber, C.C., Da Fonseca, R.R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M.S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-

Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W.C., Ray, D., Green, R.E., Bruford, M.W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E.P., Bertelsen, M.F., Sheldon, F.H., Brumfield, R.T., Mello, C.V., Lovell, P.V., Wirthlin, M., Schneider, M.P.C., Prosdocimi, F., Samaniego, J.A., Velazquez, A.M.V., Alfaro-Núñez, A., Campos, P.F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D.M., Zhou, Q., Perelman, P., Driskell, A.C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F.E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F.K., Jønsson, K.A., Johnson, W., Koepfli, K.-P., O'brien, S., Haussler, D., Ryder, O.A., Rahbek, C., Willerslev, E., Graves, G.R., Glenn, T.C., Mccormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S.V., Stamatakis, A., Mindell, D.P., Cracraft, J., Braun, E.L., Warnow, T., Jun, W., Gilbert, M.T.P., Zhang, G. (2014): Whole-genome analyses resolve early branches in the tree of life of modern birds. − Science 346: 1320.

Johnson, M.G., Malley, C., Goffinet, B., Shaw, A.J., Wickett, N.J. (2016): A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta). − Molecular Phylogenetics and Evolution 98: 29-40.

Ju, C., Van De Poel, B., Cooper, E.D., Thierer, J.H., Gibbons, T.R., Delwiche, C.F., Chang, C. (2015): Conservation of ethylene as a plant hormone over 450 million years of evolution. − Nature Plants 1: 14004.

Junier, T., Zdobnov, E.M. (2010): The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. − Bioinformatics 26: 1669-1670.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., Jermiin, L.S. (2017): ModelFinder: fast model selection for accurate phylogenetic estimates. − Nature Methods 14: 587.

Kamikawa, R., Brown, M.W., Nishimura, Y., Sako, Y., Heiss, A.A., Yubuki, N., Gawryluk, R., Simpson, A.G.B., Roger, A.J., Hashimoto, T., Inagaki, Y. (2013): Parallel re-modeling of EF-1α function: divergent EF-1α genes co-occur with EFL genes in diverse distantly related eukaryotes. − BMC Evolutionary Biology 13: 131.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. (2014): Data, information, knowledge and principle: back to metabolism in KEGG. − Nucleic Acids Research 42: D199-D205.

Kanehisa, M., Sato, Y., Morishima, K. (2016): BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. − Journal of Molecular Biology 428: 726-731.

Kang, S., Tice, A.K., Spiegel, F.W., Silberman, J.D., Pánek, T., Čepička, I., Kostka, M., Kosakyan, A., Alcântara, D.M., Roger, A.J., Shadwick, L.L., Smirnov, A., Kudryavstev, A., Lahr, D.J.G., Brown, M.W. (2017): Between a pod and a hard test: the deep evolution of amoebae. − Mol Biol Evol 34: 2258–2270.

Kaplan, D.R., Hagemann, W. (1991): The relationship of cell and organism in vascular plants - Are cells the building blocks of plant form? − BioScience 41: 693-703.

Kapraun, D.F. (2007): Nuclear DNA Content Estimates in Green Algal Lineages: Chlorophyta and Streptophyta. − Annals of Botany 99: 677-701.

Karpowicz, S.J., Prochnik, S.E., Grossman, A.R., Merchant, S.S. (2011): The GreenCut2 Resource, a Phylogenomically Derived Inventory of Proteins Specific to the Plant Lineage. − Journal of Biological Chemistry 286: 21427-21439.

Katoh, K., Standley, D.M. (2013): MAFFT multiple sequence alignment software version 7: improvements in performance and usability. − Mol Biol Evol 30: 772-780.

Kawahara, A.Y., Breinholt, J.W. (2014): Phylogenomics provides strong evidence for relationships of butterflies and moths. − Proceedings of the Royal Society B: Biological Sciences 281.

Keeling, P.J. (2010): The endosymbiotic origin, diversification and fate of plastids. − Philosophical Transactions of the Royal Society B-Biological Sciences 365: 729-748.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., Mcmanus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z. (2014): The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. − PLOS Biology 12.

Keeling, P.J., Inagaki, Y. (2004): A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1α. − Proc Natl Acad Sci 101: 15380-15385.

Kennedy, M., Mrofka, D., Von Der Borch, C. (2008): Snowball Earth termination by destabilization of equatorial permafrost methane clathrate. − Nature 453: 642.

Kim, G.H., Klotchkova, T.A., Kang, Y.-M. (2001): Life without a cell membrane: regeneration of protoplasts from disintegrated cells of the marine green alga *Bryopsis plumosa*. − Journal of Cell Science 114: 2009-2014.

Kishino, H., Miyata, T., Hasegawa, M. (1990): Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. − Journal of Molecular Evolution 31: 151-160.

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

Kliman, R.M., Andolfatto, P., Coyne, J.A., Depaulis, F., Kreitman, M., Berry, A.J., Mccarter, J., Wakeley, J., Hey, J. (2000): The Population Genetics of the Origin and Divergence of the &lt;em&gt;Drosophila simulans&lt;/em&gt; Complex Species. − Genetics 156: 1913.

Kloppstech, K., Schweiger, H.G. (1975a): 80 S ribosomes in *Acetabularia* major distribution and transportation within the cell. − Protoplasma 83: 27-40.

Kloppstech, K., Schweiger, H.G. (1975b): Polyadenylated RNA from *Acetabularia*. − Differentiation 4: 115-123.

Knoll, A.H., Javaux, E.J., Hewitt, D., Cohen, P. (2006): Eukaryotic organisms in Proterozoic oceans. − Philosophical Transactions of the Royal Society B: Biological Sciences 361: 1023.

Kocot, K.M., Citarella, M.R., Moroz, L.L., Halanych, K.M. (2013): PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. − Evolutionary Bioinformatics 9: 429-435.

Koepfli, K.-P., Benedict Paten, Scientists, T.G.K.C.O., O'brien, S.J. (2015): The Genome 10K project: a way forward. − Annual Review of Animal Biosciences 3: 57-111.

Koonin, E.V. (2005): Orthologs, Paralogs, and Evolutionary Genomics. − Annual Review of Genetics 39: 309-338.

Koren, S., Harhay, G.P., Smith, T.P., Bono, J.L., Harhay, D.M., Mcvey, S.D., Radune, D., Bergman, N.H., Phillippy, A.M. (2013): Reducing assembly complexity of microbial genomes with single-molecule sequencing. − Genome Biology 14: 1-16.

Kuroa, K., Manabe, E. (1983): Microtubule-associated cytoplasmic Streaming in *Caulerpa*. − Proceedings of the Japan Academy 59: 131-134.

La Claire, J.W., Ii. (1992): Contractile movements in the algae: the Siphonocladales as model systems. In *The Cytoskeleton of the Algae.*, pp. 239-253. − CRC Press.

La Claire, J.W., Loudenslager, C.M., Zuccarello, G.C. (1998): Characterization of novel extrachromosomal DNA from giant-celled marine green algae. − Current Genetics 34: 204-211.

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

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

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

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

Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S. (2012): PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. − Mol Biol Evol 29.

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

Lang, B.F., Nedelcu, A.M. (2012): Plastid genomes of algae. In *Genomics of Chloroplasts and Mitochondria* (R. BOCK, V. KNOOP editors), pp. 59-87. − Springer Netherlands, Dordrecht.

Lartillot, N., Lepage, T., Blanquart, S. (2009): PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. − Bioinformatics 25: 2286-2288.

Laurin-Lemay, S., Brinkmann, H., Philippe, H. (2012): Origin of land plants revisited in the light of sequence contamination and missing data. − Current Biology 22: R593-R594.

Le Bail, A., Dittami, S.M., De Franco, P.-O., Rousvoal, S., Cock, M.J., Tonon, T., Charrier, B. (2008): Normalisation genes for expression analyses in the brown alga model *Ectocarpus siliculosus*. − BMC Molecular Biology 9: 1-9.

Leinonen, R., Sugawara, H., Shumway, M. (2011): The Sequence Read Archive. − Nucleic Acids Research 39: D19-D21.

Leister, D. (2005): Genomics-based dissection of the cross-talk of chloroplasts with the nucleus and mitochondria in Arabidopsis. − Gene 354: 110-116.

Leliaert, F., De Clerck, O., Verbruggen, H., Boedeker, C., Coppejans, E. (2007): Molecular phylogeny of the Siphonocladales (Chlorophyta: Cladophorophyceae). − Mol Phylogenet Evol 44: 1237-1256.

Leliaert, F., Lopez-Bautista, J.M. (2015): The chloroplast genomes of *Bryopsis plumosa* and *Tydemania expeditiones* (Bryopsidales, Chlorophyta): compact genomes and genes of bacterial origin. − BMC Genomics 16.

Leliaert, F., Rueness, J., Boedeker, C., Maggs, C.A., Cocquyt, E., Verbruggen, H. (2009): Systematics of the marine microfilamentous green algae *Uronema curvatum* and *Urospora microscopica* (Chlorophyta). − Eur J Phycol 44: 487-496.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., De Clerck, O. (2012): Phylogeny and molecular evolution of the green algae. − Critical Reviews in Plant Sciences 31: 1-46.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., Depriest, M.S., Bhattacharya, D., Karol, K.G., Fredericq, S., Zechman, F.W., Lopez-Bautista, J.M. (2016): Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. − Scientific Reports 6: 25367.

Lemieux, C., Otis, C., Turmel, M. (2014a): Chloroplast phylogenomic analysis resolves deep-level relationships within the green algal class Trebouxiophyceae. − BMC Evolutionary Biology 14: 1-15.

Lemieux, C., Otis, C., Turmel, M. (2014b): Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. − BMC Genomics 15: 1-20.

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

Lemieux, C., Vincent, A.T., Labarre, A., Otis, C., Turmel, M. (2015): Chloroplast phylogenomic analysis of chlorophyte green algae identifies a novel lineage sister to the Sphaeropleales (Chlorophyceae). − BMC Evolutionary Biology 15.

Lewis, L.A., Mccourt, R.M. (2004): Green Algae and the Origin of Land Plants. − American Journal of Botany 91: 1535-1556.

Lewitus, E., Morlon, H. (2016): Characterizing and Comparing Phylogenies from their Laplacian Spectrum. − Systematic Biology 65: 495-507.

Li, L., Stoeckert, C.J., Roos, D.S. (2003): OrthoMCL: identification of ortholog groups for eukaryotic genomes. − Genome Res 13: 2178-2789.

Li, Q., Liu, J., Zhang, L., Liu, Q. (2014): De dovo transcriptome analysis of an aerial microalga *Trentepohlia jolithus*: pathway description and gene discovery for carbon fixation and carotenoid biosynthesis. − PLOS ONE 9.

Li, W., Godzik, A. (2006): Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. − Bioinformatics 22: 1658-1659.

Liddle, L., Berger, S., Schweiger, H.-G. (1976): Ultrastructure during development of the nucleus of *Batophora oerstedii* (Chlorophyta; Dasycladaceae). − Journal of Phycology 12: 261-272.

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

Liu, K., Jia, S., Du, Q., Zhang, C. (2017): NanoAsPipe: A transcriptome analysis and alternative splicing detection pipeline for MinION long-read RNA-seq. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1823-1826.

Lohaus, R., Van De Peer, Y. (2016): Of dups and dinos: evolution at the K/Pg boundary. − Current Opinion in Plant Biology 30: 62-69.

Lokhorst, G.M., Star, W. (1983): Fine structure of mitosis and cytokinesis in *Urospora* (Acrosiphoniales, Chlorophyta). − Protoplasma 117: 142-153.

Lopez-Bautista, J.M., Chapman, R.L. (2003): Phylogenetic affinities of the Trentepohliales inferred from small-subunit rDNA. − Int J Syst Evol Microbiol 53: 2099-2106.

Lü, F., Xü, W., Tian, C., Wang, G., Niu, J., Pan, G. (2011): The *Bryopsis hypnoides* plastid genome: multimeric forms and complete nucleotide sequence. − PLoS One 6.

Luttke, A. (1988): The lack of chloroplast DNA in *Acetabularia mediterranea* (*acetabulum*) (Chlorophyceae): a reinvestigation. − Journal of Phycology 24: 173-180.

Macdonald, F.A., Schmitz, M.D., Crowley, J.L., Roots, C.F., Jones, D.S., Maloof, A.C., Strauss, J.V., Cohen, P.A., Johnston, D.T., Schrag, D.P. (2010): Calibrating the Cryogenian. − Science 327: 1241.

Maddison, W. (1989): Reconstructing character evolution on polytomous cladograms. − Cladistics 5: 365-377.

Maeshima, K., Ide, S., Hibino, K., Sasai, M. (2016): Liquid-like behavior of chromatin. − Current Opinion in Genetics & Development 37: 36-45.

Magallón, S., Hilu Khidir, W., Quandt, D. (2013): Land plant evolutionary timeline: Gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. − American Journal of Botany 100: 556-573.

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

Marcelino, V., Cremen, M.C.M., Jackson, C.J., Larkum, A.a.W., 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.

Martens, C., Vandepoele, K., Van De Peer, Y. (2008): Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. − Proceedings of the National Academy of Sciences 105: 3427.

Massouh, A., Schubert, J., Yaneva-Roder, L., Ulbricht-Jones, E.S., Zupok, A., Johnson, M.T.J., Wright, S.I., Pellizzer, T., Sobanski, J., Bock, R., Greiner, S. (2016): Spontaneous Chloroplast Mutants Mostly Occur by Replication Slippage and Show a Biased Pattern in the Plastome of &lt;em&gt;Oenothera&lt;/em&gt;. − The Plant Cell 28: 911.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J.G., Gitzendanner, M.A., Wafula, E., Der, J.P., Depamphilis, C.W., Roure, B., Philippe, H., Ruhfel, B.R., Miles, N.W., Graham, S.W., Mathews, S., Surek, B., Melkonian, M., Soltis, D.E., Soltis, P.S., Rothfels, C., Pokorny, L., Shaw, J.A., Degironimo, L., Stevenson, D.W., Villarreal, J.C., Chen, T., Kutchan, T.M., Rolf, M., Baucom, R.S., Deyholos, M.K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., Wong, G.K.-S. (2014): Data access for the 1,000 Plants (1KP) project. − GigaScience 3.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.-Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., Wasserman, W.W. (2013): JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. − Nucleic Acids Research 42: D142-147.

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

Matsuo, Y., Imagawa, H., Nishizawa, M., Shizuri, Y. (2005): Isolation of an algal morphogenesis inducer from a marine bacterium. − Science 307: 1598.

Mattox, K.R., Stewart, K.D. (1984): Classification of the green algae: a concept based on comparative cytology. In *Systematics of the green algae* (D. E. G. IRVINE, D. M. JOHN editors), pp. 29-72. − Academic Press, London.

Mazza, A., Casale, A., Sassone-Corsi, P., Bonotto, S. (1980): A minicircular component of *Acetabularia acetabulum* chloroplast DNA replicating by the rolling circle. − Biochemical and Biophysical Research Communications 93: 668-674.

Mccourt, R.M., Delwiche, C.F., Karol, K.G. (2004): Charophyte algae and land plant origins. − Trends Ecol. Evol. 19: 661-666.

Mcnaughton, E.E., Goff, L.J. (1990): The role of microtubules in establishing nuclear spatial patterns in multinucleate green algae. − Protoplasma 157: 19-37.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D.M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., Van helden, J. (2015): RSAT 2015: Regulatory Sequence Analysis Tools. − Nucleic Acids Research 43: W50-56.

Melkonian, M. (1990): Phylum Chlorophyta. Class Prasinophyceae. In *Handbook of Protoctista. The structure, cultivation, habitats and life histories of the eukaryotic microorganisms and their descendants exclusive of animals, plants and fungi* (L. MARGULIS, J. O. CORLISS, M. MELKONIAN, D. J. CHAPMAN editors), pp. 600-607. − Jones and Bartlett Publishers, Boston.

Melnick, D.J., Hoelzer, G.A., Absher, R., Ashley, M.V. (1993): mtDNA diversity in rhesus monkeys reveals overestimates of divergence time and paraphyly with neighboring species. − Molecular Biology and Evolution 10: 282-295.

Melton, J.T., Iii, Leliaert, F., Tronholm, A., Lopez-Bautista, J.M. (2015): The complete chloroplast and mitochondrial genomes of the green macroalga *Ulva* sp. UNA00071828 (Ulvophyceae, Chlorophyta). − PLoS ONE 10: e0121020.

Menzel, D. (1987): The cytoskeleton of the giant coenocytic green alga *Caulerpa* visualized by immunocytochemistry. − Protoplasma 139: 71-76.

Menzel, D. (1988): How do giant plant cells cope with injury?—The wound response in siphonous green algae. − Protoplasma 144: 73-91.

Menzel, D. (1994): Cell differentiation and the cytoskeleton in *Acetabularia*. − New Phytologist 128: 369-393.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., Marshall, W.F., Qu, L.-H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.-L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.A., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S., Ral, J.-P., Riaño-Pachón, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.-J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y.W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W.C.A., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I.V., Rokhsar, D.S., Grossman, A.R. (2007): The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. − Science 318: 245.

Micheels, A.M., Michael. (2008): A snowball Earth versus a slushball Earth: Results from Neoproterozoic climate modeling sensitivity experiments. − Geosphere 4: 401-410.

Mikhailov, K.V., Janouškovec, J., Tikhonenkov, D.V., Mirzaeva, G.S., Diakin, A.Y., Simdyanov, T.G., Mylnikov, A.P., Keeling, P.J., Aleoshin, V.V. (2014): A Complex Distribution of Elongation Family GTPases EF1A and EFL in Basal Alveolate Lineages. − Genome Biology and Evolution 6: 2361-2367.

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

Miller, M.A., Pfeiffer, W., Schwartz, T. (2011): The CIPRES science gateway: a community resource for phylogenetic analyses. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, pp. 1-8. − ACM, Salt Lake City, Utah.

Mine, I., Anota, Y., Menzel, D., Okuda, K. (2005): Poly(A)+ RNA and cytoskeleton during cyst formation in the cap ray of *Acetabularia peniculus*. − Protoplasma 226: 199-206.

Mine, I., Menzel, D., Okuda, K. (2008): Morphogenesis in giant-celled algae. In *International Review of Cell and Molecular Biology*, pp. 37-83.

Mirarab, S., Bayzid, M.S., Warnow, T. (2016): Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. − Systematic Biology 65: 366-380.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T. (2014): ASTRAL: genome-scale coalescent-based species tree estimation. − Bioinformatics 30: i541-i548.

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., Mckenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., Von Reumont, B.M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M., Zhou, X. (2014): Phylogenomics resolves the timing and pattern of insect evolution. − Science 346: 763.

Molloy, E.K., Warnow, T. (2018): To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. − Systematic Biology 67: 285-303.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M.F., Piganeau, G., Rouzé, P., Da Silva, C., Wincker, P., Van De Peer, Y., Vandepoele, K. (2012): Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. − Genome Biology 13: R74.

Moreira, D., Philippe, H. (2000): Molecular phylogeny: pitfalls and progress. − International Microbiology 3: 9-16.

Morlon, H., Lewitus, E., Condamine Fabien, L., Manceau, M., Clavel, J., Drury, J. (2015): RPANDA: an R package for macroevolutionary analyses on phylogenetic trees. − Methods in Ecology and Evolution 7: 589-597.

Motomura, T. (1996): Cell cycle analysis in a multinucleate green alga, *Boergesenia forbesii* (Siphonocladales, Chlorophyta). − Phycological Research 44: 11-17.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J., Gerstein, M. (2016): The real cost of sequencing: scaling computation to keep pace with data generation. − Genome Biology 17.

Mungpakdee, S., Shinzato, C., Takeuchi, T., Kawashima, T., Koyanagi, R., Hisata, K., Tanaka, M., Goto, H., Fujie, M., Lin, S., Satoh, N., Shoguchi, E. (2014): Massive Gene Transfer and Extensive RNA Editing of a Symbiotic Dinoflagellate Plastid Genome. − Genome Biology and Evolution 6: 1408-1422.

Muñoz-Gómez, S.A., Mejía-Franco, F.G., Durnin, K., Colp, M., Grisdale, C.J., Archibald, J.M., Slamovits, C.H. (2017): The new red algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. − Current Biology 27: 1677-1684. e1674.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., Kanaya, S. (2011): Sequence-specific error profile of Illumina sequencers. − Nucleic Acids Research 39: e90-e90.

Nakayama, T., Watanabe, S., Inouye, I. (1996): Phylogeny of wall-less green flagellates inferred from 18S rDNA sequence data. − Phycological Research 44: 151-161.

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

Nedelcu, A.M., Lee, R.W., Lemieux, C., Gray, M.W., Burger, G. (2000): The Complete Mitochondrial DNA Sequence of Scenedesmus obliquus Reflects an Intermediate Stage in the Evolution of the Green Algal Mitochondrial Genome. − Genome Research 10: 819-831.

Negrutskii, B.S., El'skaya, A.V. (1998): Eukaryotic translation elongation factor 1 alpha: structure, expression, functions, and possible role in aminoacyl-tRNA channeling. − progress in Nucleic Acid Research and Molecular Biology 60: 47-78.

Nelson, D.R., Khraiwesh, B., Fu, W., Alseekh, S., Jaiswal, A., Chaiboonchoe, A., Hazzouri, K.M., O'connor, M.J., Butterfoss, G.L., Drou, N., Rowe, J.D., Harb, J., Fernie, A.R., Gunsalus, K.C., Salehi-Ashtiani, K. (2017): The genome and phenome of the green alga Chloroidium sp. UTEX 3007 reveal adaptive traits for desert acclimatization. − eLife 6: e25783.

Nelson, M.J., Green, B.R. (2005): Double hairpin elements and tandem repeats in the non-coding region of *Adenoides eludens* chloroplast gene minicircles. − Gene 358: 102-110.

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q. (2015): IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. − Molecular Biology and Evolution 32: 268-274.

Niklas, K.J. (2014): The evolutionary-developmental origins of multicellularity. − Am J Bot 101: 6-25.

Noble, G.P., Rogers, M.B., Keeling, P.J. (2007): Complex distribution of EFL and EF-1α proteins in the green algal lineage. − BMC Evolutionary Biology 7.

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

O'kelly, C.J., Floyd, G.L. (1984): Flagellar apparatus absolute orientations and the phylogeny of the green algae. − BioSystems 16: 227-251.

Oakley, T.H., Wolfe, J.M., Lindgren, A.R., Zaharoff, A.K. (2013): Phylotranscriptomics to bring the understudied into the fold: monophyletic Ostracoda, fossil placement, and pancrustacean phylogeny. − Molecular Biology and Evolution 30: 215-233.

Ochsenreiter, T., Hajduk, S. (2008): The Function of RNA Editing in Trypanosomes. In *RNA Editing* (H. U. GÖRINGER editor), pp. 181-197. − Springer Berlin Heidelberg, Berlin, Heidelberg.

Okuda, K., Mine, I., Morinaga, T., Kuwaki, N. (1997): Cytomorphogenesis in cenocytic green algae. V. Segregative cell division and cortical microtubules in *Dictyosphaeria cavernosa* (Siphonocladales, Chlorophyceae). − Phycological Research 45: 189-196.

Oldenburg, D.J., Bendich, A.J. (2016): The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. − Current genetics 62: 431-442.

Padmanabhan, U., Green, B.R. (1978): The kinetic complexity of *Acetabularia* chloroplast DNA. − Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis 521: 67-73.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen, R., Derelle, E., Rombauts, S., Zhou, K., Otillar, R., Merchant, S.S., Podell, S., Gaasterland, T., Napoli, C., Gendler, K., Manuell, A., Tai, V., Vallon, O., Piganeau, G., Jancek, S., Heijde, M., Jabbari, K., Bowler, C., Lohr, M., Robbens, S., Werner, G., Dubchak, I., Pazour, G.J., Ren, Q., Paulsen, I., Delwiche, C., Schmutz, J., Rokhsar, D., Van De Peer, Y., Moreau, H., Grigoriev, I.V. (2007): The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. − Proceedings of the National Academy of Sciences 104: 7705.

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

Parfrey, L.W., Lahr, D.J.G., Knoll, A.H., Katz, L.A. (2011): Estimating the timing of early eukaryotic diversification with multigene molecular clocks. − Proceedings of the National Academy of Sciences 108: 13624.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., Searson, S. (2015): Open science resources for the discovery and analysis of Tara Oceans data. − Scientific Data 2.

Philippe, H., Vienne, D.M.D., Ranwez, V., Roure, B., Baurain, D., Delsuc, F. (2017): Pitfalls in supermatrix phylogenomics. − European Journal of Taxonomy 283: 1-25.

Pombert, J.-F., Beauchamp, P., Otis, C., Lemieux, C., Turmel, M. (2006a): The complete mitochondrial DNA sequence of the green alga Oltmannsiellopsis viridis: evolutionary trends of the mitochondrial genome in the Ulvophyceae. − Current Genetics 50: 137-147.

Pombert, J.-F., Blouin, N.A., Lane, C., Boucias, D., Keeling, P.J. (2014): A Lack of Parasitic Reduction in the Obligate Parasitic Green Alga Helicosporidium. − PLOS Genetics 10: e1004355.

Pombert, J.-F., Lemieux, C., Turmel, M. (2006b): The complete chloroplast DNA sequence of the green alga *Oltmannsiellopsis viridis* reveals a distinctive quadripartite architecture in the chloroplast genome of early diverging ulvophytes. − BMC Biology 4.

Pombert, J.F., Otis, C., Lemieux, C., Turmel, M. (2005): The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages. − Mol Biol Evol 22: 1903-1918.

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

Porter, S.M. (2004): The fossil record of early eukaryotic diversification. − The Paleontological Society Papers 10: 35-50.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldon, T., Rattei, T., Creevey, C., Kuhn, M., Jensen, L.J., Von Mering, C., Bork, P. (2014): eggNOG v4.0: nested orthology inference across 3686 organisms. − Nucleic Acids Res 42: D231–D239.

Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K., Hellsten, U., Chapman, J., Simakov, O., Rensing, S.A., Terry, A., Pangilinan, J., Kapitonov, V., Jurka, J., Salamov, A., Shapiro, H., Schmutz, J., Grimwood, J., Lindquist, E., Lucas, S., Grigoriev, I.V., Schmitt, R., Kirk, D., Rokhsar, D.S. (2010): Genomic Analysis of Organismal Complexity in the Multicellular Green Alga *Volvox carteri*. − Science 329: 223.

Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van De Peer, Y., Vandepoele, K. (2009): PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants. − The Plant Cell 21: 3718.

Puttick, M.N., Morris, J.L., Williams, T.A., Cox, C.J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Schneider, H., Pisani, D., Donoghue, P.C.J. (2018): The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. − Current Biology 28: 733-745.e732.

Ranjan, A., Townsley, B.T., Ichihashi, Y., Sinha, N.R., Chitwood, D.H. (2015): An intracellular transcriptomic atlas of the giant coenocyte *Caulerpa taxifolia*. − PLOS Genetics 11: e1004900.

Raup, D.M., Sepkoski, J.J. (1982): Mass Extinctions in the Marine Fossil Record. − Science 215: 1501.

Raup, D.M., Sepkoski, J.J. (1984): Periodicity of extinctions in the geologic past. − Proceedings of the National Academy of Sciences 81: 801.

Rensing, S.A. (2018): Plant Evolution: Phylogenetic Relationships between the Earliest Land Plants. − Current Biology 28: R210-R213.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E.A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S.-I., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W.B., Barker, E., Bennetzen, J.L., Blankenship, R., Cho, S.H., Dutcher, S.K., Estelle, M., Fawcett, J.A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K.A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D.R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P.J., Sanderfoot, A., Schween, G., Shiu, S.-H., Stueber, K., Theodoulou, F.L., Tu, H., Van De Peer, Y., Verrier, P.J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A.C., Hasebe, M., Lucas, S., Mishler, B.D., Reski, R., Grigoriev, I.V., Quatrano, R.S., Boore, J.L. (2008): The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. − Science 319: 64.

Revell, L.J. (2011): phytools: an R package for phylogenetic comparative biology (and other things). − Methods in Ecology and Evolution 3: 217-223.

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

Rindi, F., Lam, D.W., López-Bautista, J.M. (2009): Phylogenetic relationships and species circumscription in *Trentepohlia* and *Printzina* (Trentepohliales, Chlorophyta). − Molecular Phylogenetics and Evolution 52: 329-339.

Roch, S., Steel, M. (2015): Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. − Theoretical Population Biology 100: 56-62.

Roth, M.S., Cokus, S.J., Gallaher, S.D., Walter, A., Lopez, D., Erickson, E., Endelman, B., Westcott, D., Larabell, C.A., Merchant, S.S., Pellegrini, M., Niyogi, K.K. (2017): Chromosome-level genome assembly and transcriptome of the green alga &lt;em&gt;Chromochloris zofingiensis&lt;/em&gt; illuminates astaxanthin production. − Proceedings of the National Academy of Sciences 114: E4296.

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

Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., Burleigh, J.G. (2014): From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. − BMC Evolutionary Biology 14: 23.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., Barrell, B. (2000): Artemis: sequence visualization and annotation. − Bioinformatics 16: 944-945.

Sabnis, D.D., Jacobs, W.P. (1967): Cytoplasmic Streaming and Microtubules in the Coenocytic Marine Alga *Caulerpa Prolifera* −Journal of Cell Science 2: 465-472.

Sakofsky, C.J., Malkova, A. (2017): Break induced replication in eukaryotes: mechanisms, functions, and consequences. − Critical Reviews in Biochemistry and Molecular Biology 52: 395-413.

Salazar-Ciudad, I. (2010): Morphological evolution and embryonic developmental diversity in metazoa. − Development 137: 531-539.

Sánchez-Baracaldo, P., Raven, J.A., Pisani, D., Knoll, A.H. (2017): Early photosynthetic eukaryotes inhabited low-salinity habitats. − Proceedings of the National Academy of Sciences 114: E7737-E7745.

Santos, M.a.S., Moura, G., Massey, S.E., Tuite, M.F. (2004): Driving change: the evolution of alternative genetic codes. − Trends in Genetics 20: 95-102.

Sayyari, E., Mirarab, S. (2018): Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. − Genes 9.

Schatz, M.C., Witkowski, J., Mccombie, W.R. (2012): Current challenges in de novo plant genome sequencing and assembly. − Genome Biology 13.

Schirmer, M., D’amore, R., Ijaz, U.Z., Hall, N., Quince, C. (2016): Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. − BMC Bioinformatics 17: 125.

Schirmer, M., Ijaz, U.Z., D'amore, R., Hall, N., Sloan, W.T., Quince, C. (2015): Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. − Nucleic Acids Research 43: e37-e37.

Schodde, R., Mason, I.J. (1999): Directory of Australian Birds: Passerines. Csiro Publishing, Australia.

Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E. (2012): Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. − Bioinformatics 28: 1086-1092.

Schvartzman, M.S., Corso, M., Fataftah, N., Scheepers, M., Nouet, C., Bosman, B., Carnol, M., Motte, P., Verbruggen, N., Hanikenne, M. (2018): Adaptation to high zinc depends on distinct mechanisms in metallicolous populations of Arabidopsis halleri. − New Phytologist 218: 269-282.

Sears, J.R. (1967): Mitotic waves in the green alga *Blastophysa rizhopus* as related to coenocyte form. − Journal of Phycology 3: 136-139.

Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, Juan j., Sabidó, E., Gómez-Skarmeta, José l., Di croce, L., Ruiz-Trillo, I. (2016): The dynamic regulatory genome of *Capsaspora* and the origin of animal multicellularity. − Cell 165: 1224-1237.

Serikawa, K.A., Porterfield, D.M., Mandoli, D.F. (2001): Asymmetric subcellular mRNA distribution correlates with carbonic anhydrase activity in *Acetabularia acetabulum*. − Plant Physiology 125: 900-911.

Sharp, L.W. (1926): An introduction to cytology. McGraw-Hill Book Company, Inc., New York.

Shen, B., Dong, L., Xiao, S., Kowalewski, M. (2008): The Avalon Explosion: Evolution of Ediacara Morphospace. − Science 319: 81.

Shen, X.-X., Hittinger, C.T., Rokas, A. (2017): Contentious relationships in phylogenomic studies can be driven by a handful of genes. − Nature Ecology & Evolution 1.

Shields, G.A. (2008): Marinoan meltdown. − Nature Geoscience 1: 351.

Shimodaira, H. (2002): An Approximately Unbiased Test of Phylogenetic Tree Selection. − Systematic Biology 51: 492-508.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi, T., Hisata, K., Tanaka, M., Fujiwara, M., Hamada, M., Seidi, A., Fujie, M., Usami, T., Goto, H., Yamasaki, S., Arakaki, N., Suzuki, Y., Sugano, S., Toyoda, A., Kuroki, Y., Fujiyama, A., Medina, M., Coffroth, Mary a., Bhattacharya, D., Satoh, N. (2013): Draft Assembly of the *Symbiodinium minutum* Nuclear Genome Reveals Dinoflagellate Gene Structure. − Current Biology 23: 1399-1408.

Si Quang, L., Gascuel, O., Lartillot, N. (2008): Empirical profile mixture models for phylogenetic reconstruction. − Bioinformatics 24: 2317-2323.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. (2015): BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. − Bioinformatics 31: 3210-3212.

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J.C., Manuel, M., Philippe, H., Telford, M.J. (2018): A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. − BMC Biology 16: 28.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., Manuel, M. (2017): A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. − Current Biology 27: 958-967.

Simpson, C.L., Stern, D.B. (2002): The treasure trove of algal chloroplast genomes. Surprises in architecture and gene content, and their functional implications. − Plant Physiology 129: 957-966.

Škaloud, P., Kalina, T., Nemjová, K., De Clerck, O., Leliaert, L. (2013): Morphology and phylogenetic position of the freshwater green microalgae *Chlorochytrium* (Chlorophyceae) and *Scotinosphaera* (Scotinosphaerales, ord. nov., Ulvophyceae). − J Phycol 49: 115-129.

Škaloud, P., Rindi, F., Boedeker, C., Leliaert, F. (2018): Chlorophyta: Ulvophyceae. In *Süßwasserflora von Mitteleuropa, Freshwater Flora of Central Europe, Vol. 13* (B. BÜDEL, GÄRTNER, G., KRIENITZ, L., PREISIG, H.-R., SCHAGERL, M. (EDS.) editor), − Heidelberg, Springer Spektrum, Berlin, Heidelberg.

Slamovits, C.H., Saldarriaga, J.F., Larocque, A., Keeling, P.J. (2007): The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. − Journal of molecular biology 372: 356-368.

Slowinski, J.B. (2001): Molecular Polytomies. − Molecular Phylogenetics and Evolution 19: 114-120.

Sluiman, H.J. (1989a): The green algal class Ulvophyceae. An ultrastructural survey and classification. − Crypt. Bot. 1: 83-94.

Sluiman, H.J. (1989b): The green algal class Ulvophyceae: An ultrastructural survey and classification. − Crypt. Bot. 1: 83-94.

Smith, D.R. (2016): The mutational hazard hypothesis of organelle genome evolution: 10 years on. − Molecular Ecology 25: 3769-3775.

Smith, D.R. (2017): Evolution: In Chloroplast Genomes, Anything Goes. − Current Biology 27: R1305-R1307.

Smith, D.R., Burki, F., Yamada, T., Grimwood, J., Grigoriev, I.V., Van Etten, J.L., Keeling, P.J. (2011): The GC-Rich Mitochondrial and Plastid Genomes of the Green Alga Coccomyxa Give Insight into the Evolution of Organelle DNA Nucleotide Landscape. − PLOS ONE 6: e23624.

Smith, D.R., Hua, J., Archibald, J.M., Lee, R.W. (2013): Palindromic Genes in the Linear Mitochondrial Genome of the Nonphotosynthetic Green Alga Polytomella magna. − Genome Biology and Evolution 5: 1661-1667.

Smith, D.R., Keeling, P.J. (2015): Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. − Proceedings of the National Academy of Sciences of the United States of America 112: 10177-10184.

Smith, D.R., Lee, R.W. (2009): The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA. − BMC Genomics 10.

Smith, D.R., Lee, R.W., Cushman, J.C., Magnuson, J.K., Tran, D., Polle, J.E.W. (2010): The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. − BMC Plant Biology 10: 83.

Smith, S.A., Brown, J.W., Walker, J.F. (2018): So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. − PLOS ONE 13: e0197433.

Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W., Cooper, A. (2012): The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. − Molecular Biology and Evolution 29: 3345-3358.

Spoerner, M., Wichard, T., Bachhuber, T., Stratmann, J., Oertel, W. (2012): Growth and thallus morphogenesis of *Ulva mutabilis* (Chlorophyta) depends on a combination of two bacterial species excreting regulatory factors. − Journal of Phycology 48: 1433-1447.

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

Staves, M.P., La Claire, J.W. (1985): Nuclear synchrony in *Valonia macrophysa* (Chlorophyta): light microscopy and flow cytometry. − Journal of Phycology 21: 68-71.

Steijger, T., Abril, J.F., Engström, P.G., Kokocinski, F., The, R.C., Hubbard, T.J., Guigó, R., Harrow, J., Bertone, P. (2013): Assessment of transcript reconstruction methods for RNA-seq. − Nature Methods 10: 1177.

Sturmbauer, C., Meyer, A. (1993): Mitochondrial phylogeny of the endemic mouthbrooding lineages of cichlid fishes from Lake Tanganyika in eastern Africa. − Molecular Biology and Evolution 10: 751-768.

Suh, A. (2016): The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. − Zoologica Scripta 45: 50-62.

Sulston, J.E., Schierenberg, E., White, J.G., Thomson, J.N. (1983): The embryonic cell lineage of the nematode *Caenorhabditis elegans*. − Developmental Biology 100: 64-119.

Sun, L., Fang, L., Zhang, Z., Chang, X., Penny, D., Zhong, B. (2016): Chloroplast phylogenomic inference of green algae relationships. − Sci Rep 6.

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

Takahashi, K., Terai, Y., Nishida, M., Okada, N. (2001): Phylogenetic Relationships and Ancient Incomplete Lineage Sorting Among Cichlid Fishes in Lake Tanganyika as Revealed by Analysis of the Insertion of Retroposons. − Molecular Biology and Evolution 18: 2057-2066.

Talavera, G., Castresana, J. (2007): Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. − Syst Biol 56: 564-577.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S. (2011): MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. − Molecular Biology and Evolution 28: 2731-2739.

The Arabidopsis Genome Initiative. (2000): Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. − Nature 408: 796.

Timme, R.E., Bachvaroff, T.R., Delwiche, C.F. (2012): Broad phylogenomic sampling and the sister lineage of land plants. − PLOS ONE 7.

Torruella, G., De mendoza, A., Grau-Bové, X., Antó, M., Chaplin, Mark a., Del campo, J., Eme, L., Pérez-Cordón, G., Whipps, Christopher m., Nichols, Krista m., Paley, R., Roger, Andrew j., Sitjà-Bobadilla, A., Donachie, S., Ruiz-Trillo, I. (2015): Phylogenomics reveals convergent evolution of lifestyles in close Relatives of animals and fungi. − Current Biology 25: 2404-2410.

Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K., Ruiz-Trillo, I. (2012): Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. − Molecular Biology and Evolution 29: 531-544.

Turmel, M., Bellemare, G., Lee, R.W., Lemieux, C. (1986): A linear DNA molecule of 5.9 kilobase-pairs is highly homologous to the chloroplast DNA in the green alga *Chlamydomonas moewusii*. − Plant Molecular Biology 6: 313-319.

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

Turmel, M., Lemieux, C. (2018): Chapter Six - Evolution of the Plastid Genome in Green Algae. In *Advances in Botanical Research* (S.-M. CHAW, R. K. JANSEN editors), pp. 157-193. − Academic Press.

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

Turmel, M., Otis, C., Lemieux, C. (2016b): Mitochondrion-to-chloroplast DNA transfers and intragenomic proliferation of chloroplast Group II introns in *Gloeotilopsis* green algae (Ulotrichales, Ulvophyceae). − Genome Biology and Evolution 8: 2789-2805.

Turmel, M., Otis, C., Lemieux, C. (2017): Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophycean green algae. − Scientific Reports 7.

Tymms, M.J., Schweiger, H.G. (1985): Tandemly repeated nonribosomal DNA sequences in the chloroplast genome of an *Acetabularia mediterranea* strain. − Proc Natl Acad Sci 82: 1706-1710.

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van De Peer, Y., Vandepoele, K. (2013): TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. − Genome Biology 14.

Van Den Hoek, H., Mann, D., Jahns, H.M. (1995): Algae: An Introduction to Phycology. Cambridge University Press.

Van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C. (2014): Ten years of next-generation sequencing technology. − Trends in Genetics 30: 418-426.

Van Leeuwe , M., Tedesco, L., Arrigo, K.R., Assmy, P., Campbell, K., Meiners, K.M., Rintala, J.-M.S., Virginia, Thomas, D.N., Stefels, J. (2018): Microalgal community structure and primary production in Arctic and Antarctic sea ice: A synthesis. − Elem Sci Anth. 6.

Van bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de peer, Y., Coppens, F., Vandepoele, K. (2018): PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. − Nucleic Acids Research 46: D1190-D1196.

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.

Vanneste, K., Baele, G., Maere, S., Van De Peer, Y. (2014): Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. − Genome Research 24: 1334-1347.

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

Verbruggen, H., Ashworth, M., Loduca, S.T., Vlaeminck, C., Cocquyt, E., Sauvage, T. (2009): A multi-locus time-calibrated phylogeny of the siphonous green algae. − Mol Phylogenet Evol 50: 642-653.

Verbruggen, H., Marcelino, V.R., Guiry, M.D., Cremen, M.C.M., Jackson, C.J. (2017): Phylogenetic position of the coral symbiont *Ostreobium* (Ulvophyceae) inferred from chloroplast genome data. − Journal of Phycology 53: 790-803.

Verbruggen, H., Theriot, E.C. (2008): Building trees of algae: some advances in phylogenetic and evolutionary analysis. − European Journal of Phycology 43: 229-252.

Verbruggen, H., Tribollet, A. (2011): Boring algae. − Current Biology 21: R876-R877.

Vogel, H., Grieninger, G.E., Zetsche, K.H. (2002): Differential messenger RNA gradients in the unicellular alga *Acetabularia acetabulum*. Role of the cytoskeleton. − Plant Physiology 129: 1407-1416.

Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., Barakat, A., Wolcott, E., Liang, H., Landherr, L., Tomsho, L.P., Hu, Y., Carlson, J.E., Ma, H., Schuster, S.C., Soltis, D.E., Soltis, P.S., Altman, N., Depamphilis, C.W. (2009): Comparison of next generation sequencing technologies for transcriptome characterization. − BMC Genomics 10: 347.

Waller, R.F., Jackson, C.J. (2009): Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. − Bioessays 31: 237-245.

Walsh, H.E., Kidd, M.G., Moum, T., Friesen, V.L. (1999): Polytomies and the Power of Phylogenetic Inference. − Evolution 53: 932-937.

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C., Ware, D. (2016): Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. − Nature Communications 7: 11708.

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

Watanabe, S., Kuroda, N., Maiwa, F. (2001): Phylogenetic status of *Helicodictyon planctonicum* and *Desmochloris halophila* gen. et comb. nov. and the definition of the class Ulvophyceae (Chlorophyta). − Phycologia 40: 421-434.

Watanabe, S., Nakayama, T. (2007): Ultrastructure and phylogenetic relationships of the unicellular green algae *Ignatius tetrasporus* and *Pseudocharacium americanum* (Chlorophyta). − Phycol Res 55: 1-16.

Wichard, T. (2015): Exploring bacteria-induced growth and morphogenesis in the green macroalga order Ulvales (Chlorophyta). − Frontiers in Plant Science 6: 86.

Wicke, S., Schneeweiss, G.M., Depamphilis, C.W., Müller, K.F., Quandt, D. (2011): The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. − Plant Molecular Biology 76: 273-297.

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., Ruhfel, B.R., Wafula, E., Der, J.P., Graham, S.W., Mathews, S., Melkonian, M., Soltis, D.E., Soltis, P.S., Miles, N.W., Rothfels, C.J., Pokorny, L., Shaw, A.J., Degironimo, L., Stevenson, D.W., Surek, B., Villarreal, J.C., Roure, B., Philippe, H., Depamphilis, C.W., Chen, T., Deyholos, M.K., Baucom, R.S., Kutchan, T.M., Augustin, M.M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G.K.-S., Leebens-Mack, J. (2014): Phylotranscriptomic analysis of the origin and early diversification of land plants. − Proc Natl Acad Sci 111: E4859-E4868.

Wodniok, S., Brinkmann, H., Glockner, G., Heidel, A.J., Philippe, H., Melkonian, M., Becker, B. (2011): Origin of land plants: do conjugating green algae hold the key? − BMC Evol Biol 11.

Woodcock, C.L.F., Bogorad, L. (1970): Evidence for variation in the quantity of DNA amonf plastids of *Acetabularia*. − the Journal of Cell Biology 44: 361-375.

Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., Foulon, E., Grimwood, J., Gundlach, H., Henrissat, B., Napoli, C., Mcdonald, S.M., Parker, M.S., Rombauts, S., Salamov, A., Von Dassow, P., Badger, J.H., Coutinho, P.M., Demir, E., Dubchak, I., Gentemann, C., Eikrem, W., Gready, J.E., John, U., Lanier, W., Lindquist, E.A., Lucas, S., Mayer, K.F.X., Moreau, H., Not, F., Otillar, R., Panaud, O., Pangilinan, J., Paulsen, I., Piegu, B., Poliakov, A., Robbens, S., Schmutz, J., Toulza, E., Wyss, T., Zelensky, A., Zhou, K., Armbrust, E.V., Bhattacharya, D., Goodenough, U.W., Van De Peer, Y., Grigoriev, I.V. (2009): Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes *Micromonas*. − Science 324: 268.

Wu, T.D., Reeder, J., Lawrence, M., Becker, G., Brauer, M.J. (2016): GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In *Statistical Genomics: Methods and Protocols* (E. MATHÉ, S. DAVIS editors), pp. 283-334. – Springer New York, New York, NY.

Wysoker, A., Tibbets, K., Fennell, T. (2013): Picard: a set of command line tools for manipulating high-thorughput sequencing data. – Available online at: http://broadinstitute.github.io/picard/.

Xu, J., Fan, X., Zhang, X., Xu, D., Mou, S. (2012): Evidence of coexistence of C3 and C4 photosynthetic pathways in a green-tide-forming alga, *Ulva prolifera*. – PLoS One 7.

Yang, Y., Smith, S.A. (2014): Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. – Molecular Biology and Evolution 31: 3081-3092.

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

Ye, Q., Tong, J., Xiao, S., Zhu, S., An, Z., Tian, L., Hu, J. (2015): The survival of benthic macroscopic phototrophs on a Neoproterozoic snowball Earth. – Geology 43: 507-510.

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

Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simão, F.A., Ioannidis, P., Seppey, M., Loetscher, A., Kriventseva, E.V. (2017): OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. – Nucleic Acids Research 45: D744-D749.

Zechman, F.W., Theriot, E.C., Zimmer, E.A., Chapman, R.L. (1990): Phylogeny of the Ulvophyceae (Chlorophyta): cladistic analysis of nuclear-encoded rRNA sequence data. – J Phycol 26: 700-710.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S. (2018): ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. – BMC Bioinformatics 19: 153.

Zhang, J., Hao, Q., Bai, L., Xu, J., Yin, W., Song, L., Xu, L., Guo, X., Fan, C., Chen, Y., Ruan, J., Hao, S., Li, Y., Wang, R.R.C., Hu, Z. (2014): Overexpression of the soybean transcription factor GmDof4 significantly enhances the lipid content of *Chlorella ellipsoidea*. – Biotechnology for Biofuels 7: 128.

Zhang, X., Ye, N., Liang, C., Mou, S., Fan, X., Xu, J., Xu, D., Zhuang, Z. (2012): *De novo* sequencing and analysis of the *Ulva linza* transcriptome to discover putative mechanisms associated with its successful colonization of coastal ecosystems. – BMC Genomics 13.

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

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

Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., Hao, P. (2011): Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. – BMC Bioinformatics 12: S2.

Zhong, B., Liu, L., Yan, Z., Penny, D. (2013): Origin of land plants using the multispecies coalescent model. – Trends in Plant Science 18: 492-495.

Zhong, B., Xi, Z., Goremykin, V.V., Fong, R., Mclenachan, P.A., Novis, P.M., Davis, C.C., Penny, D. (2014): Streptophyte Algae and the Origin of Land Plants Revisited Using Heterogeneous Models with Three New Algal Chloroplast Genomes. − Molecular Biology and Evolution 31: 177-183.

Zimorski, V., Ku, C., Martin, W.F., Gould, S.B. (2014): Endosymbiotic theory for organelle origins. − Current Opinion in Microbiology 22: 38-48.

Zuccarello, G.C., Price, N., Verbruggen, H., Leliaert, F. (2009): Analysis of a plastid multigene data set and the phylogenetic position of the marine macroalga *Caulerpa filiformis* (Chlorophyta). − J. Phycol. 45: 1206-1212.

ONE
LOVE