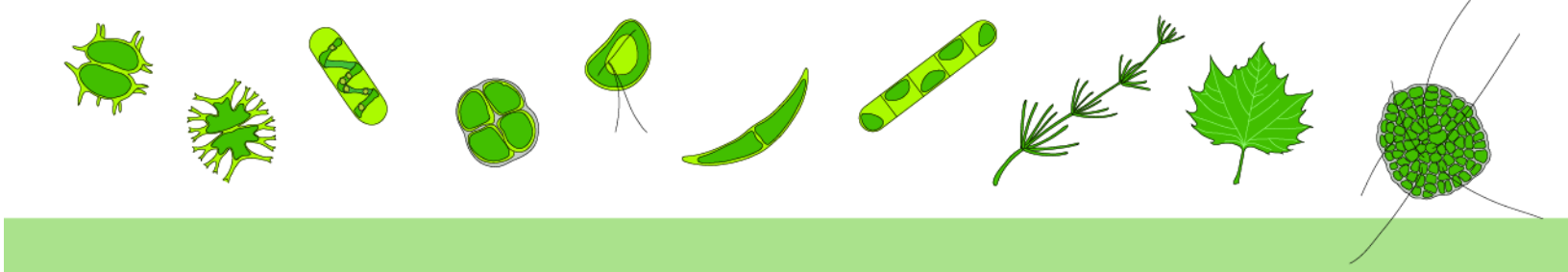


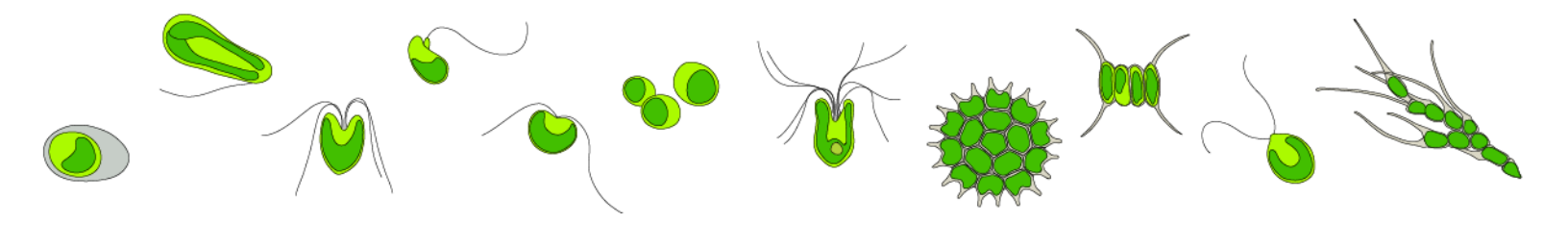
La historia evolutiva de las plantas verdes (Viridiplantae) desde la perspectiva de sus genomas plastídicos

Título de Biólogo

Autor:
Sergio Andrés Muñoz Gómez

Asesores:
Juan Manuel Daza
John Jairo Ramírez

Instituto de Biología
Facultad de Ciencias Exactas y Naturales
Universidad de Antioquia
2013



A la memoria de Juan José, mi amigo.

I think that I shall never see
A thing so awesome as the Tree
That links us all in paths of genes
Down into depths of time unseen;

Whose many branches spreading wide
House wondrous creatures of the tide,
Ocean deep and mountain tall,
Darkened cave and waterfall.

Among the branches we may find
Creatures there of every kind,
From microbe small to redwood vast,
From fungus slow to cheetah fast.

As glaciers move, strikes asteroid
A branch may vanish in the void:
At Permian's end and Tertiary's door,
The Tree was shaken to its core.

The leaves that fall are trapped in time
Beneath cold sheets of sand and lime;
But new leaves sprout as mountains rise,
Breathing life anew 'neath future skies.

On one branch the leaves burst forth:
A jointed limb of firework growth.
With inordinate fondness for splitting lines,
Armored beetles formed myriad kinds.

Wandering there among the leaves,
In awe of variants Time conceived,
We ponder the shape of branching fates,
And elusive origins of their traits.

Three billion years the Tree has grown
From replicators' first seed sown
To branches rich with progeny:
The wonder of phylogeny.

The Tree of Life
David Maddison
Syst Biol (2013) 62 (1): 179.

"THE EVOLUTIONARY HISTORY OF GREEN PLANTS (VIRIDIPLANTAE) AS SEEN FROM THE PERSPECTIVE OF THEIR PLASTID GENOMES"

"LA HISTORIA EVOLUTIVA DE LAS PLANTAS VERDES (VIRIDIPLANTAE) DESDE LA PERSPECTIVA DE SUS GENOMAS PLASTÍDICOS"

ABSTRACT

The green plants comprise the land plants and green algae and together they represent the most diverse group of photosynthetic eukaryotes. Their characteristic green plastids contain genomes that provide valuable phylogenetic information. The present study aims to analyze the evolution of these genomes, as well as to evaluate the historical relationships among green plant lineages based on molecular sequence and structural data derived from plastid genomes. In order to achieve this, diverse methods of phylogenetic analysis were employed together with the ancestral reconstruction of genomic characters. As a result, a consensus Viridiplantae tree that summarizes the relationships among green plants is presented and discussed, and the evolution of green plastid genomes is narrated. Finally, the importance of increasing taxon sampling to solve current phylogenetic uncertainties is emphasized, and the relevance to carefully designing phylogenomic analyses is highlighted.

Keywords: Charophyceae, Chlorophyta, comparative genomics, genome evolution, green algae, phylogenetics, phylogenomics, Prasinophyceae, Streptophyta

RESUMEN

Las plantas verdes (Viridiplantae) comprenden tanto a las plantas terrestres como a las algas verdes y en conjunto representan el grupo más diverso de eucariotas fotosintéticos. Sus característicos plástidos verdes contienen genomas que proveen valiosa información filogenética. El presente estudio pretende analizar la evolución de estos genomas, además de evaluar las relaciones históricas entre los linajes de plantas verdes con base en secuencias moleculares y datos estructurales de sus genomas plastídicos. Para lograr esto, se utilizaron diversos métodos de análisis filogenético, al igual que reconstrucción ancestral de caracteres genómicos. Como resultado se presenta y discute un consenso resumiendo el árbol filogenético de Viridiplantae y se narra la evolución de sus genomas plastídicos. Finalmente, se enfatiza la importancia de aumentar el muestreo taxonómico para solucionar las incertidumbres actuales y se recalca la relevancia del diseño cuidadoso de análisis filogenómicos.

Palabras clave: algas verdes, Charophyceae, Chlorophyta, evolución genómica, filogenética, filogenómica, genómica comparada, Prasinophyceae, Streptophyta

TABLE OF CONTENTS

INTRODUCTION

THE RISE AND SPREAD OF ALGAE IN THE TREE OF EUKARYOTES

Algae as a group has no place in our taxonomic classifications since it is clearly polyphyletic (1,2). Algae have evolved multiples times in the history of life, and its polyphyly is due to a specific kind of horizontal evolution: symbiogenetic mergers of different cell lineages in the tree of life (3,4). The history of algae is thus the history of lateral plastid acquisitions in addition to vertical divergence (Fig. 1).

Algae are more accurately defined as oxygenic photosynthesizers to the exclusion of a derived group within the green algae that adapted to land, the embryophytes. Because the algae are united by their photosynthetic ability and the excretion of oxygen as its byproduct, they play a prominent role as primary producers in ecosystem trophic webs. Algae, therefore, comprises prokaryote cyanobacteria (proalgae) and eukaryote phototrophs (eualgae and meta-algae) (1).

Although algae are said to be polyphyletic, they are only from the perspective of their host cells. Oxygenic photosynthesis originated only once in the entire history of life (5,6) and it has subsequently spread across the eukaryote tree by an original symbiogenesis with a cyanobacterium, and later through higher-order symbiogeneses between two different eukaryotes. This makes that oxygenic photosynthesis and plastids are ultimately monophyletic despite their reticulated history in the evolutionary tree of eukaryotic life.

Of the five major eukaryotic supergroups currently recognized, three of them contain photosynthetic eukaryotes or algae (7–9) (Fig. 1). They are the Archaeplastida, the Chromalveolata, and the Excavata. The Archaeplastida comprises the glaucophytes, the rhodophtes (red algae) and the green plants or viridiplants, which are the focus of this undergraduate thesis. Its origin is the result of a primary symbiogenetic event between a heterotrophic eukaryotic host cell and a cyanobacterium that led to the diversification of primary plastid-containing algae (plants *sensu lato*). The Chromalveolata (10), a good working hypothesis that has received contradictory support from different phylogenetic studies (4,11–14), is also a very diversified supergroup containing several algal lineages. They include the dinoflagellates, the chromerids, the heterokont algae or ochrophytes, the cryptophytes, the haptophytes, and the rhizarian chlorarachniophytes. All chromalveolate algae, except the chlorarachniophytes, have plastids of red algal origin that are hypothesized to have been acquired in a single secondary symbiogenesis between a eukaryotic phagotroph and a red alga. Finally, the Excavata is a eukaryotic supergoup whose members are predominantly heterotrophic, but that contains an important group of algae, the euglenophytes. The secondary plastids of euglenophytes and chlorarachniophytes are of green origin, and each group got its plastids independently from different green algae (15,16).

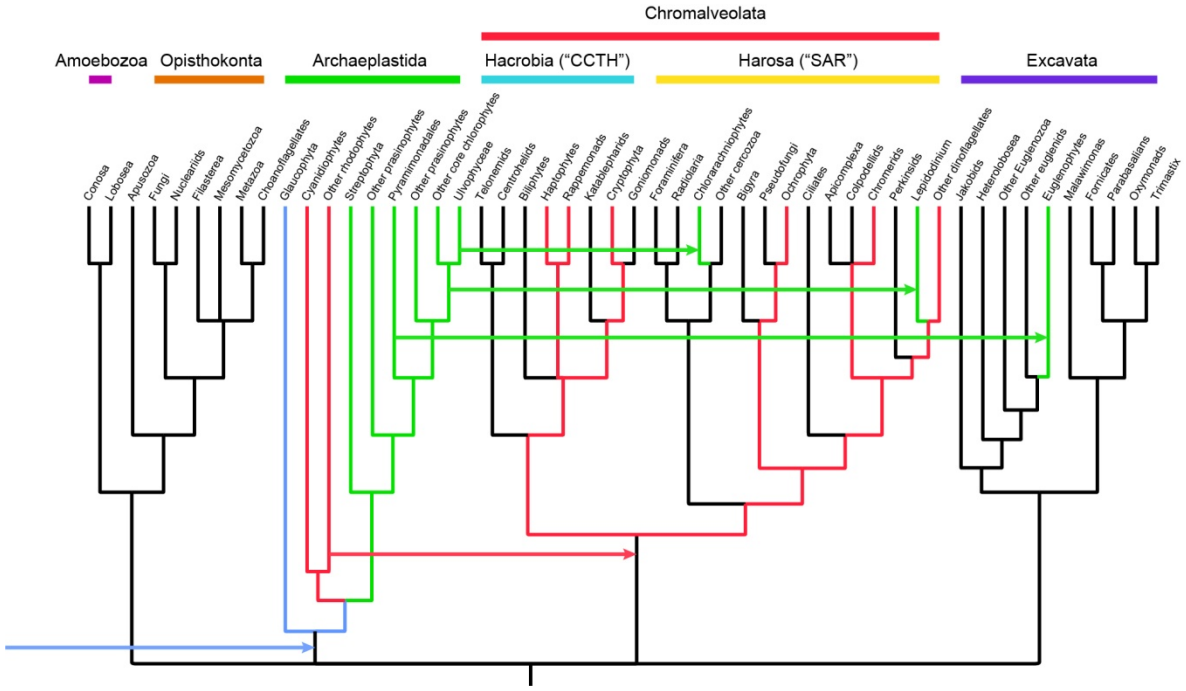


Figure 1. Reticulated phylogenetic diagram showing the evolutionary relationships among major lineages of eukaryotes. A primary symbiogenesis (blue arrow) gave rise to primary plastid-containing eukaryotes (eualgae) that belong to the Archaeplastida supergroup (also known as Plantae). A minimum of one secondary symbiogenesis between a red alga and a eukaryotic phagotroph gave rise to chromalveolate lineages (red arrow). Three additional secondary symbiogenetic events with green algae (green arrows) gave rise to the secondary green plastids of euglenophytes (Excavata), chlorarachniophytes (Rhizaria) and *Lepidodinium* (dinoflagellates). At least one case of tertiary symbiogenesis resulting in haptophyte-derived plastids in dinoflagellates has occurred (e.g. *Karenia* and *Karlodinium*). This is not shown for simplicity.

THE WET AND DRY GREEN ALGAE: VIRIDIPLANTAE

The green color we see in our aquatic and terrestrial surrounding ecosystems is mainly because of Viridiplantae. In land and freshwater our perception of this color is primarily due to organisms belonging to one of the two major Viridiplantae clades, the Streptophyta. In marine waters, however, the green is usually the result of abundant populations of seaweeds in shorelines or oceanic phytoplankton in continental waters; these organisms belong to the other major clade within Viridiplantae, the Chlorophyta. This green wavelength that our eyes perceive is the product of the presence of chlorophylls a and b that are not masked by accessory photosynthetic pigments in green plant's primary plastids. As we saw in the previous section, green primary plastids are the direct descendants of the symbiogenesis that created Archaeplastida, and hence, that established photosynthesis in eukaryotes changing the fate of life's history in our planet since then.

The group comprising the green plants, i.e. green algae and land plants, has received different names including: "Viridiplantae" (17), "Viridaeplantae" (18), "Chlorobiota" (19), "Chlorobionta" (20), and more recently "Chloroplastida" (21,22). Informally, it has also been termed as the "green

plants" or the "green lineage" (23). Several classification schemes have been proposed for the group and future changes to its taxonomy are to be expected (2,24,25). However, major advances have been made in the field and a relatively general and stable consensus of the structure of the Viridiplantae phylogenetic tree is now emerging (see figure 1). The present study largely adopts the classification system and nomenclature proposed by Lewis and McCourt (2004) and recently summarized and refined by Leliaert and colleagues (23,26). Here, in contrast to some previous authors, Chlorophyta is understood as a major monophyletic group that resulted from the earliest split that also gave rise to the Streptophyta within the green plants. Chlorophyta, therefore, does not comprise all the green algae, leaving outside the charophycean green algae that would lead to the origin of land plants.

Different lines of evidence coming from ultrastructural, biochemical and molecular data indicate that early in the evolutionary history of the green linage there was a primary bifurcation leading to the Chlorophyta clade on one side, and the Streptophyta clade on the other side of the tree (26). The Chlorophyta is comprised by a paraphyletic assemblage of lineages that belong to the class "Prasinophyceae" or simply prasinophytes, and by the three main classes named Chlorophyceae, Trebouxiophyceae and Ulvophyceae (core chlorophytes or UTC clade) that contain most of the diversity of the clade. The Streptophyta comprises the paraphyletic streptophyte algae or "charophyceans" and the specialized monophyletic embryophytes (land plants). Our current understanding of the phylogenetic relationships among the major green plant lineages are diagrammed in figure 2.

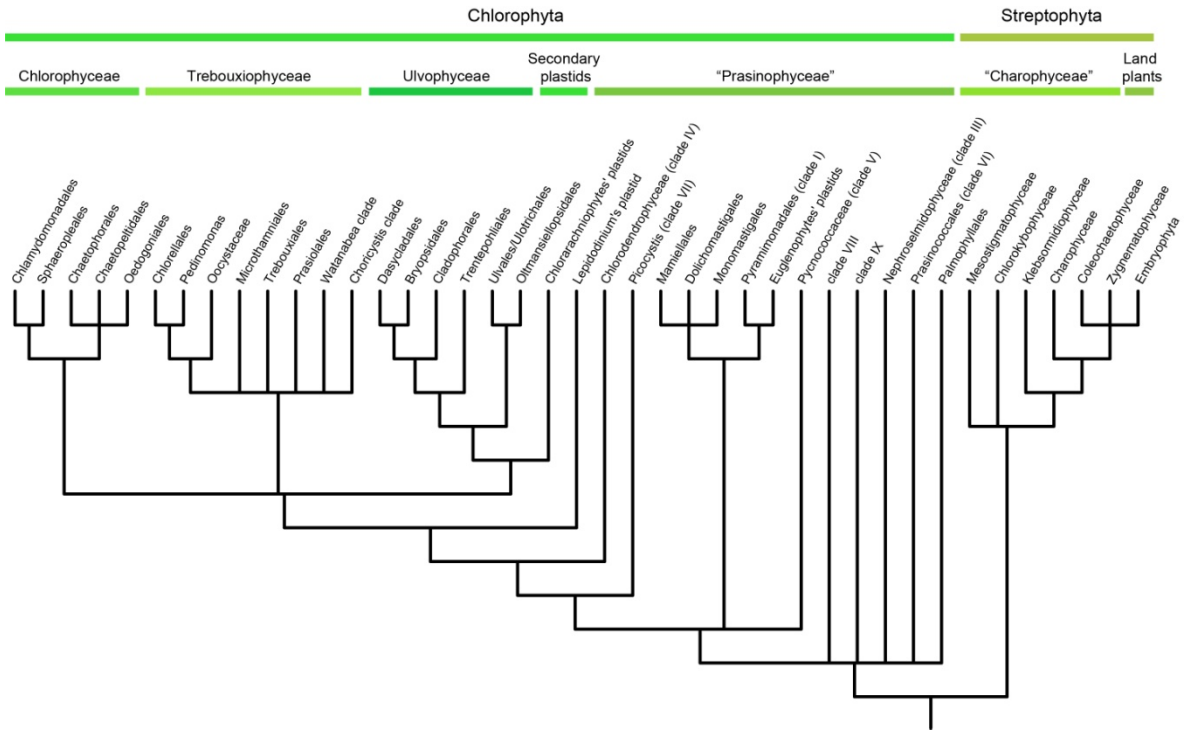


Figure 2. Consensus tree summarizing our current understanding of the relationships among green plants. (After Leliaert *et al*., (2012))

Viridiplantae is an ancient group whose precise origin date has been difficult to determine. Based on molecular clock studies the age of the group has been estimated to be between 1,500 and 700 Ma (27–30). The fossil record of the group dates back to the Proterozoic. Achritarchs, more commonly interpreted as phycomate cysts of prasinophytic chlorophytes, have been found in 1,200 Ma old rocks. Opinions regarding the exact nature of these microfossils differ among authors. The end of the Proterozoic eon and the beginning of the Phanerozoic eon witnessed a great morphological diversification of achritarchs (31,32). These new complexified achritarchs are more easily attributed to cyst stages of planktonic green and other eukaryotic algae. Body fossils associated with complex multicellular green algae have been described from 700 Ma Neo-proterozoic rocks. They have been interpreted as being siphonocladalean chlorophytes such as *Cladophora* or *Cladophoropsis* (33,34).

Because of the ancient nature and great diversification of the group, defining features of the green plants that are present uniquely in all its members is a difficult task. Despite this, some clearly characteristic features are widely distributed, and curiously they are usually associated with their green plastid (2,35). Perhaps the most important synapomorphy of green plants is that they synthesize and store starch in the stroma. No other group of algae do this; they store starch somewhere in the cell outside the plastid. Additionally, the *rbcS* gene, that coding for the small rubisco subunit, is nuclear encoded in green plants, in contrast to plastids of the red lineage and glaucophytes that retain it in their genomes (ptDNA). Moreover, several light-harvesting complex (LHC) proteins and some of their associated accessory pigments evolved in the common ancestor of Viridiplantae after the loss of phycobilisome antenna systems from their plastids (36,37). A host cell character that is thought to be derived from the eukaryotic ancestor of green plants is the ‘stellate structure’ found in the transition zone between flagella and basal bodies which looks like a ninepointed star in cross section and is H-shaped in longitudinal section (35).

Green algae were traditionally classified according to morphological criteria exclusively. This approach led to the rationale that green algae had evolved towards greater complexity, from unicellular flagellates into coccoid and sarcinoid chlorophytes, and then later into colonial, filamentous, coenocytic and siphonous body forms. These criteria mostly based on optical microscopy, were later supplemented with new taxonomic characters at a fine scale, ultrastructural studies of green algae facilitated by the advent of electron microscopy. The ultrastructural study of mitosis, cytokinesis and the flagellar apparatus delineated new groups at higher taxonomic ranks and suggested numerous instances of parallel evolution of similar body plans (26). The ‘Age of Molecules’ that arrived in the 1990’s complemented the Age of Ultrastructure, initially with traditional molecular markers such as SSU, 5.8S, and LSU including ITS-1 and ITS-2, as well as actin, several chloroplast genes (*rbc*L, *atp*B, and others), and mitochondrial genes (*nad*5) (38). Subsequently, complete mitochondrial and chloroplast genomes of green algae and land plants began to be sequenced. To date, 31 green algal and more than 240 land plant chloroplast genomes (cpDNA) have been fully sequenced.

Despite being the group of photosynthetic eukaryotes more widely sampled with regards to complete plastid genomes, the incredible diversity of Viridiplantae is still undersampled. Major lineages of green algae are not represented in databases, and most sequencing efforts have concentrated in embryophytes. In fact, it is surprising that sequencing almost identical flowering plant ptDNAs appears to be more important than sequencing those lineages of green algae that represent the deepest branches of the green tree. Because of this, there is currently limited information in sequence databases that can be used to resolve problematic nodes of the Viridiplantae phylogenetic tree. Moreover, land plant cpDNAs show little diversity in comparison to green algal cpDNAs, and this seriously constraints our broader understanding of the evolution of cpDNA architecture (39).

Plastid genomes represent excellent models to study molecular evolution. For example, group I and II introns have proliferated in some lineages of green plastid genomes, repetitive elements have considerably increased the size of some ptDNAs, the typical quadripartite structure has been lost in some lineages, and some others have experience accelerated rates of gene rearrangements (23,39). Furthermore, they are ideal sources of phylogenetic information for several reasons. First, they are gene-dense and provide a relatively large number of different molecular markers (>120 genes) in a small circular mapping molecule of approximately 200 Kbp. Second, they provide structural genomic data such as gene content, gene order, intron number and insertion sites, etc. Third, they do not suffer from the problem of gene paralogy so common in nuclear genomes. And fourth, being small in size their sequencing is cheap and relatively straightforward with the available standardized protocols.

Leliaert and colleagues recently presented a review in which they synthesized the state of the art of the last decades of research in green algal systematic (23). It is clear from their paper, that after much effort, there are still many uncertainties and problematic relationships among green plants. Moreover, in addition to sequencing most green algae ptDNAs, the group of Monique Turmel and Claude Lemieux has studied plastid genomics and phylogenetics using structural genomic characters to validate their multi-gene sequence trees (40,41). Although their research has been invaluable to the field there appears that an updated synthesis on green plant cpDNA evolution is desirable. Furthermore, most phylogenetic studies on green plants usually do not incorporate all the available data and they usually lack rigor in exploring possible biases in their analyses.

The present study aims to review and critically analyze the supporting evidence for current phylogenetic hypotheses of the group using genome structural and sequence data from the databases. It also intends to analyze the evolution of diverse genomic features of ptDNAs in the history of Viridiplantae.

METHODS

PHYLOGENETIC RECONSTRUCTION BASED ON MOLECULAR SEQUENCES

DATA RETRIEVAL AND SUPERMATRICES' ASSEMBLY

The totality of sequenced plastid genomes for green algae available at the Organelle Genome Resources of GenBank (42) as of September 2011 was chosen for the present study. These include 36 organisms belonging to the group of interest, Viridiplantae, while the remaining ones constitute the outgroup, four rhodophytes and the glaucophyte *Cyanophora paradoxa*. Within the ingroup the taxonomic distribution is the following: four embryophytes and six charophyceans from the Streptophyta clade, and eight prasinophyceans, eight chlorophyceans, seven trebouxiophyceans and three ulvophyceans from the Chlorophyta clade. Moreover, the secondary plastids of the euglenophyte *Euglena gracilis* and the chlorarachniophyte *Bigelowiella natans* were included. In total, they represent 43 taxa. Due to the ancient evolutionary relationships among the organisms being studied, only protein-coding genes were considered. Furthermore, most of the analyses focused on the amino acid rather than the nucleotide sequence of protein-coding genes because of higher conservation and stronger phylogenetic signal at this level.

Amino acid and nucleotide sequence data from each plastid genome-encoded gene were retrieved from NCBI GenBank. All genes and their functional assignments are shown in table S1. The genes were chosen based on their distribution in the selected taxa. Protein-coding genes that are at least present in 10 taxa were considered. Ninety-four genes met this criterion. Single genes in amino acid sequence were aligned using MUSCLE 3.7 (43,44) as implemented in the CIPRES Science Gateway v3.3 (45). The 94 genes were also aligned in their nucleotide sequence using TranslatorX server and the Muscle method (46). This server implements a variety of methods and performs multiple alignments of nucleotide sequences guided by amino acid translations and based on a specified genetic code. The result is a nucleotide alignment of codons. Additionally, each single gene alignment was "cleaned" by GBlocks 0.91b (47) that removes poorly o ambiguously aligned sites. GBlocks from the Castresana Lab server and TranslatorX server was implemented allowing options for less stringent selection.

Finally, single gene alignments were concatenated using SequenceMatrix 1.7.8 (48) and the following supermatrices were assembled:

(i) 42x79-aa. Dataset consisting of 42 taxa and 79 genes in their amino acid sequence (15,029 aa). It contains 79 ptDNA-encoded proteins that are present in at least 15 of the 42 taxa. This supermatrix is 85.29% filled at the gene level;

(ii) 42x94-aa. Dataset consisting of 42 taxa and 94 genes in their amino acid sequence (18,977 aa). It contains 94 ptDNA-encoded proteins that are present in at least ten of the 42 taxa. This supermatrix is 75.96% filled at the gene level;

(iii) 43x94-aa. Dataset consisting of 43 taxa and 94 genes in their amino acid sequence (18,977 aa). It contains 94 ptDNA-encoded proteins that are present in at least ten of

the 43 taxa and includes the fast-evolving plastid genome of the parasite *Helicosporidium* sp. This supermatrix is 74.83% filled at the gene level;

(iv) 43x12-aa. Dataset consisting of 43 taxa and 12 genes in their amino acid sequence (5,628 aa) including *Tetraselmis* sp. for which only these genes are currently available. This supermatrix is 96.12% filled at the gene level;

(v) 51x12-aa. Dataset consisting of 52 taxa and 12 genes in their amino acid sequence (5,683 aa) including *Tetraselmis* sp. and eight additional chlorophytes for which only these genes are currently available. The eight newly added chlorophyta taxa are: *Pedinophyceae* sp. YPF701, *Picocystis salinarum, Prasinococcus capsulatus, Prasinophyceae* sp. CCMP1205, *Pseudoscourfieldia marina, Pterosperma cristatum* and *Ulva arasakii.* Genes from the serial green secondary plastid of the dinoflagellate *Lepidodinium chlorophorum* were included. This supermatrix is 93.58% filled at the gene level;

(vi) 41x79-nt. Dataset consisting of 41 taxa and 79 genes in their nucleotide sequence (43,752 nt). It excludes the bryophyte *Anthoceros formosae* due to massive editing in most of its genes which made it difficult to align their codons. This supermatrix is 85.27% filled at the gene level;

(vii) 41x79-nt-cp12. Dataset consisting of 41 taxa and 79 genes in their nucleotide sequence (29,168 nt) with the exclusion of the third codon position. It excludes the bryophyte *Anthoceros formosae* due to massive editing in most of its genes which made it difficult to align their codons. This supermatrix is 85.27% filled at the gene level;

(viii) 41x94-nt. Dataset consisting of 41 taxa and 94 genes in their nucleotide sequence (55,440 nt). It excludes the bryophyte *Anthoceros formosae* due to massive editing in most of its genes which made it difficult to align their codons. This supermatrix is 75.71% filled at the gene level;

(ix) 41x94-nt-cp12. Dataset consisting of 41 taxa and 94 genes in their nucleotide sequence (36,960 nt) with the exclusion of the third codon position. It excludes the bryophyte *Anthoceros formosae* due to massive editing in most of its genes which made it difficult to align their codons. This supermatrix is 75.71% filled at the gene level.

All necessary sequence format interconversions in the process were carried out with Mesquite 2.75 (49). The removal of third codon positions was done with PAUP4.0b10 (50).

PHYLOGENETIC RECONSTRUCTION BASED ON AMINO ACID SEQUENCES

For the purpose of inferring phylogenetic trees based on amino acid sequences the five previously described supermatrices were used. The phylogenetic analyses relied on statistical/probabilistic methods that are based on explicit models of sequence evolution. For Bayesian inference (BI) both MrBayes and PhyloBayes were used, whereas for maximum likelihood (ML) PhyML and RAxML were the preferred options. MrBayes v3.1.2 (51–53) and RAxML v7.3.2 (54,55) were used as implemented in the CIPRES Science gateway v3.3 (45) that provides probably the fastest hybrid

codes available for each one. PhyloBayes v3.3e (56,57) was used from the web-based portal for phylogenomic analysis Bioportal at the University of Oslo (58). Additionally, computational resources from Centro Nacional de Secuenciación Genómica (CNSG) of the Universidad de Antioquia allowed us to carry out heavy and time-consuming phylogenetic inferences using PhyloBayes MPI 1.1b.

Bayesian analyses performed with MrBayes were done under the mixture model of protein evolution. Two independent runs, each starting from a random tree for Markov chain Monte Carlo (MCMC) chains, were run for 10,000,000 generations and sampled every 1,000 generations. Posterior probabilities and average branch lengths were calculated from the consensus of trees sampled after burn-in set to 25% or 250,000 generations.

Bayesian analyses performed with PhyloBayes used the –ratecat or –dgam 4 option to model rates across sites by a Dirichlet process or by a discrete gamma distribution with eight categories (Γ4) respectively, the -gtr option to model exchange rates and the –cat option to model specific profile mixtures by a Dirichlet process. In each analysis, MCMC chains were run for approximately 2,000 cycles and the first 100 cycles were discarded as burn-in to calculate the posterior consensus from the remaining points. This limited number of cycles was due to the extreme computational burden and time cost associated with the analyses. PhyloBayes is a Bayesian Monte Carlo Markov Chain (MCMC) sampler for phylogenetic reconstruction using protein alignments. Compared to other phylogenetic MCMC samplers (e.g. MrBayes), the main distinguishing feature of PhyloBayes is the underlying probabilistic model, CAT, and the use of nonparametric methods for modelling site-specific features of sequence evolution. It is particularly well suited for large multigene alignments, such as those used in phylogenomics (56,57,59).

Furthermore, datasets 42x79-aa and 42x94-aa were analyzed in PhyloBayes by recoding its amino acids into the 6 Dayhoff classes. This was done in order to evaluate possible compositional biases in the amino acid sequence of the plastid proteins used to build the matrices. These biases can cause phylogenetic artifacts such as the grouping of distantly related taxa due to similar amino acid composition in their proteins. The matrices were run for 5,000 cycles and the posterior probability consensus tree was calculated from the last 4,750 points. Relative exchange rates and across-site variation were modeled using GTR and a Dirichlet process, respectively. Each dataset was run with and without the –dc option that removes constant sites.

ML analyses performed with RAxML used both the PROTCATGTR and PROTGAMMALGF models of protein sequence evolution (60). Each analysis was carried out starting from 20 distinct randomized maximum parsimony (MP) trees. Non-parametric bootstrapping was performed with the same models used in the most likely tree search and 100 iterations (55).

PHYLOGENETIC RECONSTRUCTION BASED ON NUCLEOTIDE SEQUENCES

Phylogenetic analyses of nucleotide datasets were carried out using BI and ML methods. MrBayes v3.1.2 and RAxML v7.3.2 as implemented in the CIPRES Science gateway v3.3 were used.

Bayesian analyses performed with MrBayes were done under the general time reversible (GTR) model, a discrete gamma distribution with eight categories (Γ8) and a proportion of invariable sites (I); a GTR+ Γ8+I model of sequence evolution. Two independent runs, each starting from a random tree for Markov chain Monte Carlo (MCMC) chains, were run for 10,000,000 generations and sampled every 1,000 generations. Posterior probabilities and average branch lengths were calculated from the consensus of trees sampled after burn-in set to 25% or 250,000 generations.

The optimal models GTRGAMMA and GTRCAT available in RAxML v.7.2.6 were chosen to infer ML trees and for the boostrap analysis with 100 pseudoreplicates.

SLOW-FAST ANALYSIS

In order to investigate the impact of fast-evolving sites in the phylogenetic reconstruction of the Viridiplantae tree the SlowFaster software (61) was implemented. Slow-fast analysis is a method to reduce the influence of substitution saturation, one of the causes of phylogenetic noise and long-branch attraction (LBA) artifacts. In several steps of increasing stringency, the slow-fast analysis omits the fastest substituting alignment positions from the analysed dataset and thus increases its signal/noise ratio. Twenty-seven new datasets were generated, each with a lower proportion of fast-evolving positions identified by SlowFaster. The largest dataset 42x94-aa (18,977 aa) was the supermatrix chosen for the slow-fast analysis. The software uses parsimony to calculate the number of changes in each position based on the monophyletic groups in the given topology.

Each one of the above trimmed datasets was subjected to BI analysis with MrBayes v3.1.2. These were done under the mixture model of protein evolution. Only one run per dataset was carried out starting from a random tree for 10,000,000 generations and sampled every 1,000 generations. Posterior probabilities and average branch lengths were calculated from the consensus of trees sampled after burn-in set to 25%. Similarly, each dataset was analysed under ML with RAxML v.7.2.6 using both the PROTCATGTR and PROTGAMMALGF models of protein sequence evolution. Each analysis was carried out starting from 20 distinct randomized maximum parsimony (MP) trees and the boostrap support was calculated from 100 pseudoreplicates analyzed with the same model used for the best tree inference.

Bayesian posterior probabilities and ML bootstrap values obtained were tabulated for the major nodes in order to evaluate branch support variation across datasets with decreasing amount of fast-evolving sites and thus less prone to phylogenetic artifacts such as long-branch attraction (LBA).

SUPERTREE CONSTRUCTION

A supertree combines the topological information of already available trees estimated from different data (source trees). This method, along with the clann software, allows to investigate the underlying phylogenomic information and to evaluate the congruence among the phylogenetic signal contained in different gene trees.

Individual-gene phylogenetic trees for each of the 94 ptDNA potein-coding genes were estimated using the PhyML 3.0 server (62). The selection of the best-fit models of amino acid replacement used during the phylogenetic inferences was done with ProtTest HPC 3.1 (63). Table S1 shows amino acid best-fit models for each gene alignment. Tree searching was done starting from a BioNJ tree and performing SPR topological alterations. Statistical support was evaluated with 100 bootstrap pseudoreplicates.

The construction of supertrees from partially overlapping trees derived from 94 plastid protein-coding genes was carried out with Clann v 3.2.2 (64). The supertree construction method chosen was Most Similar Supertree Algorithm (MSSA) and a heuristic search was carried out with 50 repetitions, a neighbour-joining starting tree, a SPR swapping algorithm allowing 10 steps branch swapping-regrafting, and source trees normalized to avoid tree biases in the scoring system. Boostrapping was done with 100 pseudoreplicates, and same conditions as the best supertree search except for 10 repetitions per heuristic search, and branch swapping-regrafting allowing only 3 steps away from the original position.

PHYLOGENETIC RECONSTRUCTION BASED ON DISCRETE GENOMIC CHARACTERS

CLADISTIC ANALYSIS OF VIRIDIPLANTAE PLASTIDS BASED ON GENE CONTENT

I considerably expanded the data matrix built by Martin *et al*. (2002) (65) and subsequently modified by Nozaki *et al*. (2003) (66). We included 31 new taxa belonging to Viridiplantae, whose whole plastid genomes have been sequenced since then. Twenty one of these new taxa correspond to chlorophyte algae, while five correspond to streptophyte algae. Additionally, the red alga *Gracilaria tenuistipitata*, the secondary green plastid of *Bigelowiella natans* and the two embryophytes *Arabidopsis thaliana* and *Anthoceros formosease* were included in the analysis. The final size of the data matrix was 42 terminal taxa x 274 characters (genes) (43x274-gc).

The cladistic analysis of Viridiplantae plastid gene content was carried out in a modified PAUP4.0b10 (50) version called PAUP* Ratchet that incorporates Kevin Nixon's Parsimony Ratchet algorithm. The algorithm is available at the CIPRES Science gateway v3.3. An irreversible Camin and Sokal model of character type was implemented for 269 of 274 characters. Similarly to Nozaki *et al*. (2003), the remaining five characters corresponding to genes *cuvI*, *matK*, *ycf13*, *ycf68* and *ycf74* were designated as unordered characters. These genes are harboured within group I and II introns, and because the selfish mobile nature of these sequences they can spread horizontally within and between genomes (67). Character state 0 corresponds to the ancestral state, i.e., gene present, whereas character state 1 corresponds to the derive state, i.e., gene absent.

A consensus tree was calculated from all the most parsimonious trees found by the ratchet tree searching algorithm with the program consense of the PHYLIP package (68).

CONSERVED GENE CLUSTER ANALYSIS

In order to analyze the evolution of genome rearrangements a gene order dataset was built and subsequently used as input for the GeneSyn1.0 software (69). GeneSyn algorithm allows studying gene contiguity on a chromosome by detecting gene clusters or strings conserved in a given fraction of genomes.

Gene order data was extracted from the genbank file directly or through the NCBI Graphical Sequence Viewer using the Search option. Thirty-six (36) complete Viridiplantae ptDNAs were used to assemble the gene order dataset. Only protein-coding and rRNA-specifying genes were considered for the final dataset. tRNAs and ORFs were discarded due to poor conservation, duplications and ambiguous annotation that made orthology assessment difficult. Furthermore, where present, inverted repeats (IRs) were excluded to facilitate posterior analyses. Finally, a dataset that would work as the GeneSyn input was built in which genes on the positive strand were positive integers while genes in the opposite orientation (negative strand) were preceded by a minus sign.

The search for subsequences (conserved gene clusters) in at least 4 sequences (genomes) out of 36 was carried out considering only strings of a minimum of 3 contiguous genes. The -r option which says the program to consider also the reverse strand of the input sequences and the –c option that specifies the circularity of the genomes were invoked. GeneSyn outputs a file describing the presense-absense of each conserved gene cluster found. The output file was subsequently assembled in new data matrix (36x440-go) and subjected to parsimony analysis in PAUP* Ratchet treating each character type as unordered. The 36x440-go character matrix consisted of 440 characters.

A consensus tree was calculated from all the most parsimonious trees found by the ratchet tree searching algorithm with the program consense of the PHYLIP package.

MAPPING DISCRETE AND CONTINUOUS GENOMIC CHARACTERS ON THE VIRIDIPLANTAE PHYLOGENETIC TREE

ANCESTRAL GENE CONTENT AND ORDER RECONSTRUCTION

To assess the support from sequence-independent data to the sequence-derived Viridiplantae tree topologies, the standard characters corresponding to presence-absence of genes (gene content) and conserved gene clusters (gene order) were optimized along the branches of topologies 1 and 2. These two topologies were chosen because they better reflect the congruence among most analyses. Topology 1 is primarily derived from phylogenetic analyses based on amino acid data, while topology 2 reflects the relationships derived from nucleotide data. The reconstruction of ancestral states along the two chosen topologies was carried out with the program MacClade 4.08 (70) using the Trace All Changes option.

ANCESTRAL RECONSTRUCTION OF CONTINUOUS GENOME FEATURES

Green plastid genomes are extremely variable in their organization (71). In order to investigate the evolution of different continuous genomic features during the evolution of Viridiplantae, the length of intronic, intergenic, non-coding and genic DNA were determined from whole plastid genomes. This information was extracted from the genbank files loaded on Artemis and using the Overview option in the View menu. Direct visual inspection of individual files was done when necessary. Due to the presence of trans-spliced genes in some of these green plastid genomes corrections for the number of intronic nucleotides were necessary to recalculate intron lengths by excluding the artifactual annotated long lengths of fragmented trans-spliced introns. These may contain numerous other genes, and even most of the plastid genome. This was done by taking the size of the extra long trans-spliced genes of each genome and substracting its coding exons. The amount of genic DNA was determined as the sum of protein-coding, rRNA and tRNA genes excluding introns. The amount representing the non-coding fraction was calculated as the total genome size minus the genic nucleotides. Intergenic DNA was calculated as the total DNA minus the sum of genes including introns. Finally, for calculating intronic DNA the length of each gene category excluding introns was substracted from their respective lengths including introns. These numbers are presented in Table S2.

The length of the proportions of these types of DNA sequences were mapped and optimized to reconstruct their ancestral values along branches of the Viridiplantae phylogenetic tree. This was done using the software Mesquite 2.75 (49) that incorporates the linear cost assumption to reconstruct continuous characters in that the cost of a change from state x to state y is $|x-y|$. Traced genome size values were visually compared to each other traced type of DNA side by side using Mirror Tree Window option available in Mesquite.

RESULTS

VIRIDIPLANTAE RELATIONSHIPS BASED ON PLASTID GENOME-ENCODED PROTEIN SEQUENCES

VIRIDIPLANTAE RELATIONSHIPS BASED ON AMINO ACID SEQUENCES

Phylogenetic analyses of 42x79-aa dataset results in a general topology highly congruent among the different methods employed and with previous studies based on plastid genes (Figs. 3 and S1-5). *Mesostigma viride* appears always as sister to *Chlorokybus atmophyticus* and together basally as sister to the rest of Viridiplantae. This pattern is strongly supported in PhyloBayes analyses (Figs. S2 and S3) but not so well in ML trees (Figs. S4 and S5). Regarding the branching order within Streptophyta, *Chaetosphaeridium globosum* branches as sister to embryophytes, and this position is well supported by BI as well as ML methods. *Chaetosphaeridium* is followed by the zignematophyceans *Staurastrum punctulatum* and *Zygnema circumcarinatum*. *Chara vulgaris* branches unambiguously as sister to the Zygnematophyceae + Coleocahetophyceae +

Embryophyta clade with high support. Within the Embryophyta it is interesting to note that the two bryophytes form a clade where *Anthoceros formosae* appears as sister to *Marchantia polymorpha*.

On the side of the chlorophytes, *Nephroselmis olivacea* is placed as the sister to all chlorophytes with moderate to high posterior probabilities (PP) in PhyloBayes analyses (Figs. S2 and S3), whereas it appears as sister to most prasinophytes in MrBayes and ML analyses (Figs. S1, S4 and S5). Regarding the prasinophytes, a strongly supported clade appears integrated by the Mamiellophyceae and the pyramimonadaleans including the secondary plastid of *Euglena gracilis*. The prasinophyte *Pycnococcus* seems to be more related to core chlorophytes than to the rest of prasinophytes, appearing as its sister with a PP of 1 in PhyloBayes analyses (Figs. S2 and S3) and lower bootstrap support in ML analyses (Figs. S4 and S5). The three classes Chlorophyceae, Trebouxiophyceae and Ulvophyceae are recovered in most analyses of this dataset and a clade formed by the three is found with the highest support. However, their interrelationships could not be resolved. Among these three classes, the Chlorophyceae is strongly supported in all its branches. The Trebouxiophyceae appears with moderate to low support, with *Pedinomonas minor* occurring within it as sister to the Chlorellales. *Oocystis solitaria* is the sister to the clade *Pedinomonas* + Chlorellales. Finally, the ulvophycean representatives *Oltmansiellopsis viridis* and *Pseudendoclonium akinetum* form a consistent clade to which the secondary plastid of the chlorarachniophyte *Bigelowiella natans* is sister with moderate support. The branch leading to *Bryopsis hypnoides* could not be unequivocally placed within the core chlorophyte clade (Fig. 3).

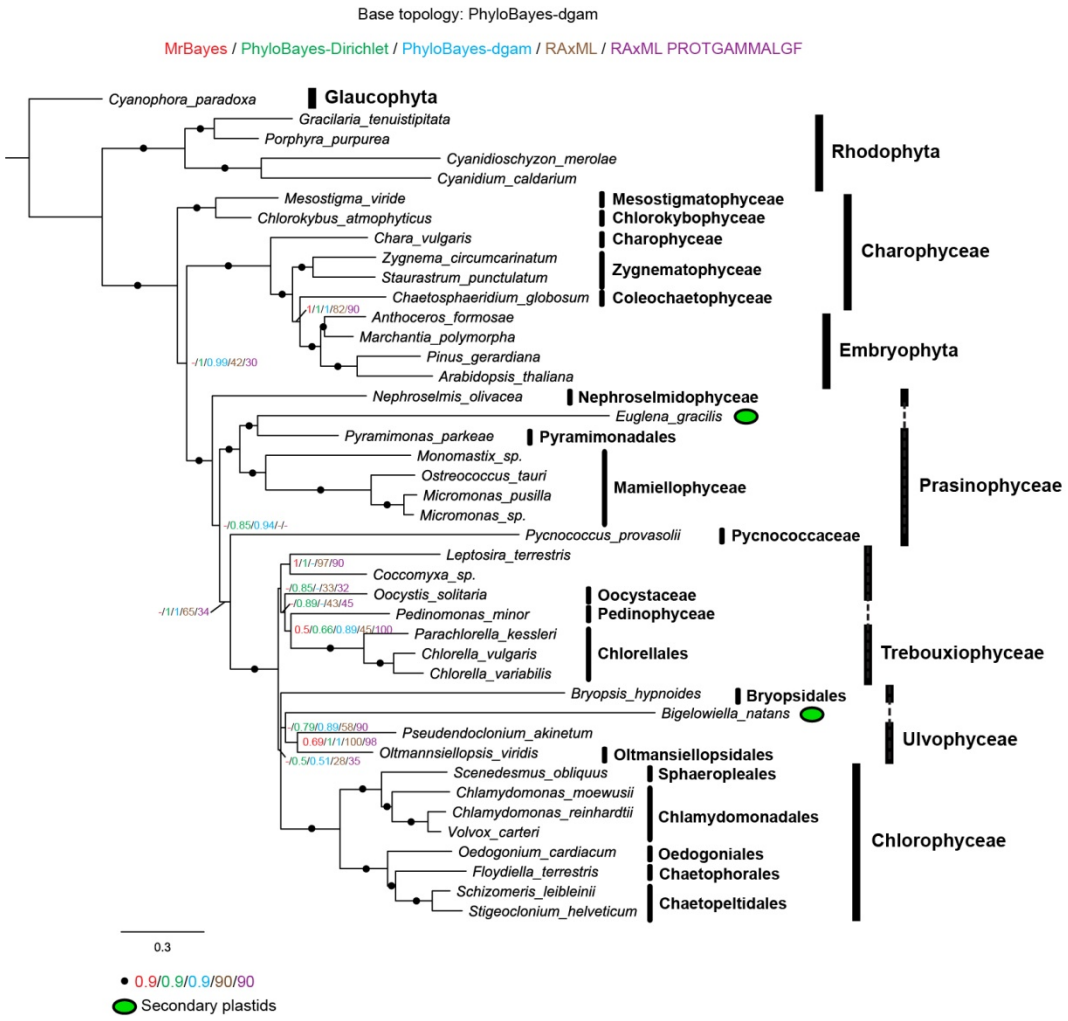


Figure 3. Consensus posterior tree resulting from a PhyloBayes analysis of 42x79-aa dataset using the mixture model CAT and Γ4. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

Relative to 42x79-aa, the phylogenetic trees resulting from analyses of the 42x94-aa dataset presents a very similar topology (Figs. 4 and S6-10). Some exceptions include the uncertainty that starts appearing regarding the sister group of embryophytes. While Coleochaetophyceae remains as their sister in PhyloBayes analyses (Figs. S7 and S8), high support for the grouping of Zygnematophyceae and Embryophyta appears in trees reconstructed with RAxML (Figs. S9 and S10). Support for the basal position of *Nephroselmis olivacea* relative to all Chlorophyta increases in ML trees. The ulvophycean *Bryopsis hypnoides* shows some instability in its placement varying among analyses (Figs. S6-10). Furthermore, the basal placement of the clade *Mesostigma* + *Chlorokybus* loses support in RAxML analyses where it preferentially branches as the first diverging streptophyte lineage (Fig. 4).

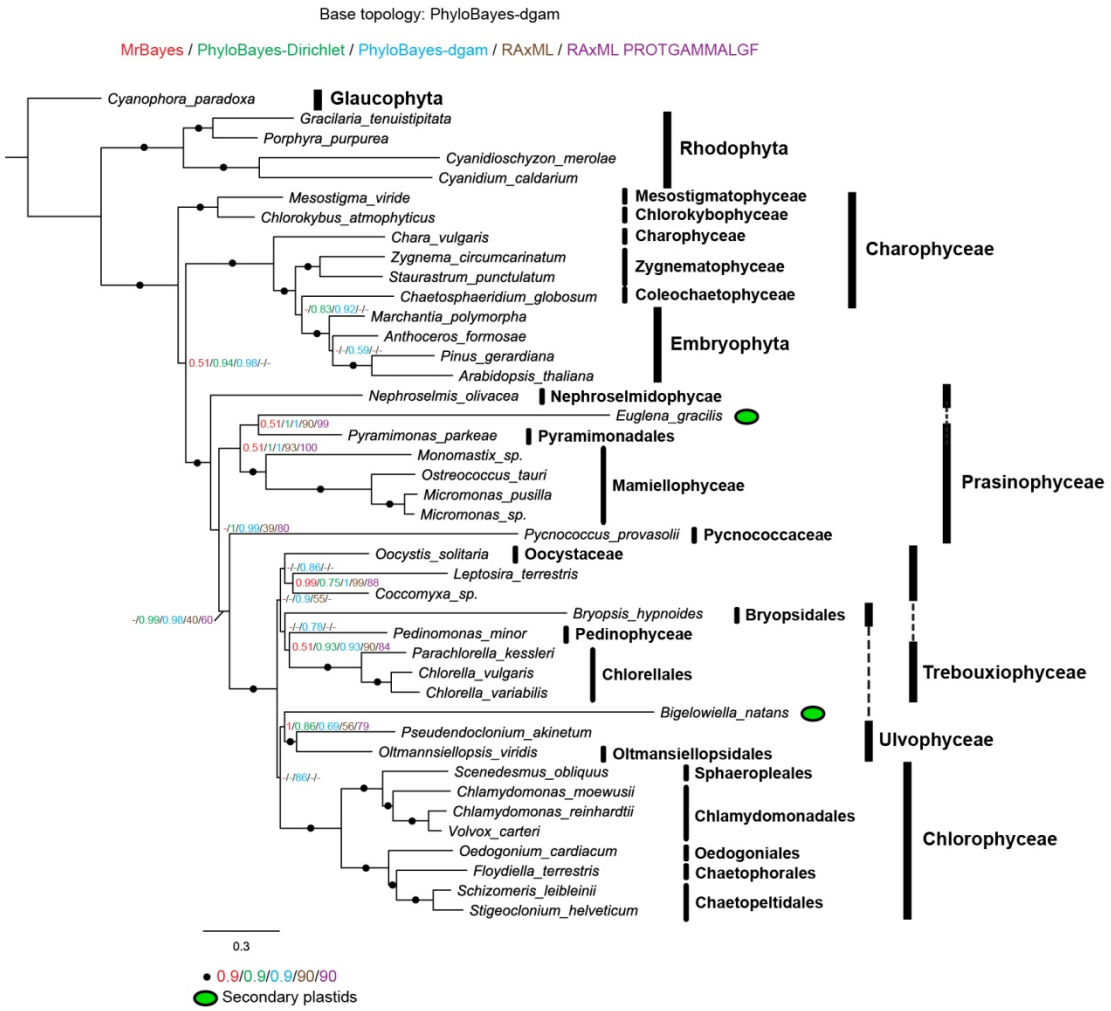


Figure 4. Consensus posterior tree resulting from a PhyloBayes analysis of 42x94-aa dataset using the mixture model CAT and Γ4. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

Phylogenetic analyses of the 43x94-aa dataset, which basically differs from the previous one by the addition of the parasite *Helicosporidium* sp., shows general agreement with the previous results (Figs. 5 and S11-13). Unsurprisingly, this topology resembles more closely the topology obtained from 42x94-aa than from 42x79-aa datasets (Fig. 5). A complete Streptophyta clade including the *Mesostigma* + *Chlorokybus* clade as sister to the Phragmoplastophytina is now relatively better supported by the three methods employed to analyze this dataset (Figs. S11-13). Although *Oocystis solitaria* is confidently placed within the trebouxiophyte clade, its exact position is ambiguous, branching sometimes as sister to Chlorellales, whereas others as sister to the *Leptosira* + *Coccomyxa* clade. Also, similar to previously described results of other datasets no support for a solid position of *Bryopsis hypnoides* is found. The tree inferred is essentially the same tree obtained from the 42x94-aa dataset analysis, with the addition of the consistent placement with maximum support of *Helicospordium* sp. as sister to Chlorellales (Fig. 5).

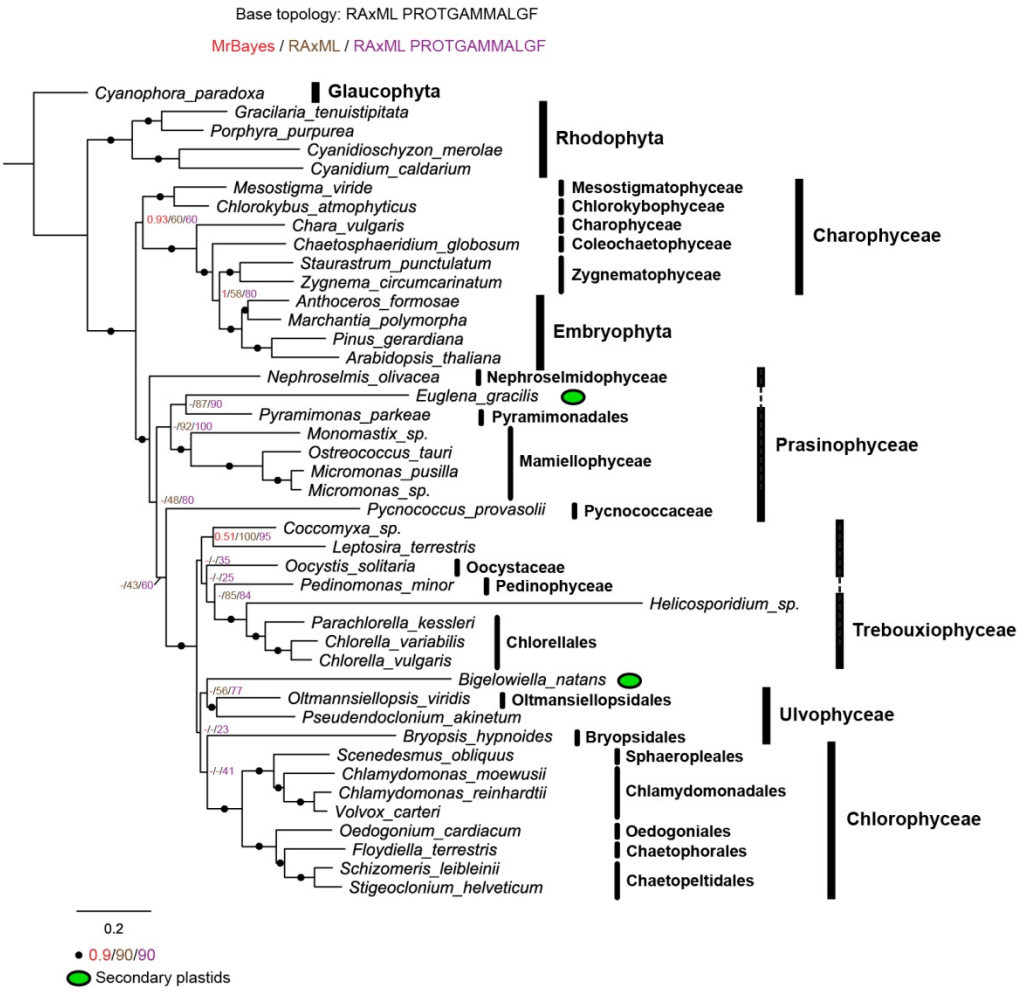


Figure 5. Consensus posterior tree resulting from a RAxML analysis of 43x94-aa dataset using the PROTGAMMALGF. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

Dataset 43x12-aa is a smaller dataset built to assess the position of the chlorodendral *Tetraselmis* sp. It differs from the previous topologies in some respects (Figs. 6 and S14-18). First, *Chaetosphaeridium globosum* branches within the zygnematophyceans. Second, *Pycnococcus provasolii* appears as the sister of the pyramimonadalean clade that includes *Pyramimonas parkeae* and the secondary plastid of *Euglena gracilis*. Third, *Pedinomonas* minor followed by *Bigelowiella natans* branch basally with respect to the core chlorophytes. Finally, contrary to expectations, *Tetraselmis* sp. is placed within the ulvophyceans, challenging its placement as the closer prasinophyte lineage to the core chlorophytes (Fig. 6).

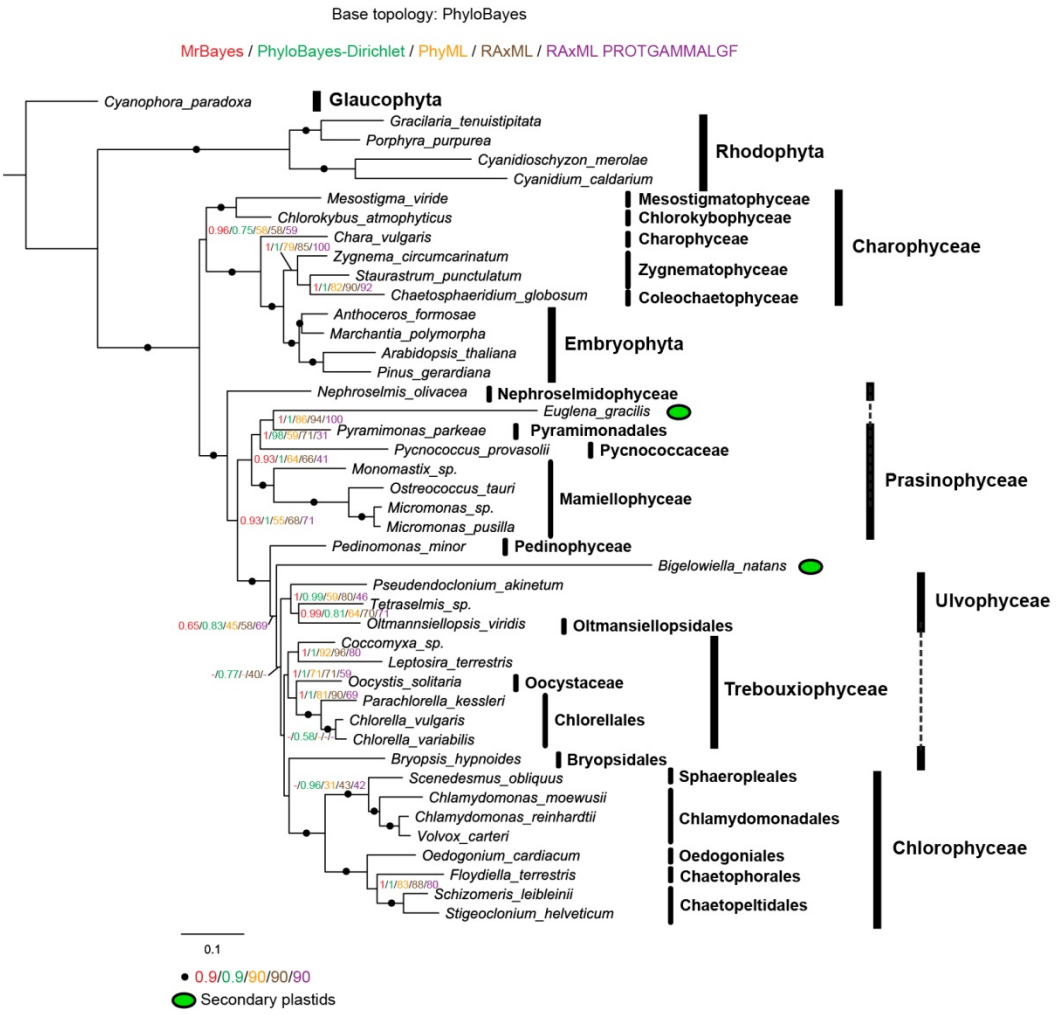


Figure 6. Consensus posterior tree resulting from a PhyloBayes analysis of 43x12-aa dataset using the mixture model CAT and a Dirichlet process. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

To further examine the position of *Tetraselmis* sp. and the other 42 taxa, the 43x12-aa dataset was expanded with the recently produced sequences of eight more chlorophytes. Regarding the streptophyte side of the tree, where taxa were not added, the topology is identical to the previous result (Figs. 7 and S19-23). On the other hand, it is interesting to note that the prasinophyte *Picocystis salinarum* is the sister to all chlorophytes to the exclusion of *Nephroselmis olivacea* which conserves its basal chlorophytan placement. Additionally, a new clade is formed where *Pycnococcus provasolii* is strongly supported as the sister of *Pseudocourfieldia marina*, and plus their sister *Prasinococcus capsulatus* group together with the pyramimonadalean clade. Furthermore, some degree of support exists to group this clade with the Mamiellophyceae class (Fig. 7).

The Chlorophyceae class remains highly supported and the Trebouxiophyceae along with the Ulvophyceae are moderately well supported. *Bryopsis hypnoides* branches as the sister of the Chlorophyceae class. The position of *Tetraselmis* sp. remains constant within ulvophyceans. It is

important to highlight that the Pedinophyceae, now with one more representative, does not branch as sister to the Chlorellales; now its position seems to be more basal without much support, however (Fig. 7). The divergence pattern among major lineages, including the three main classes, of the core chlorophytes is not resolved.

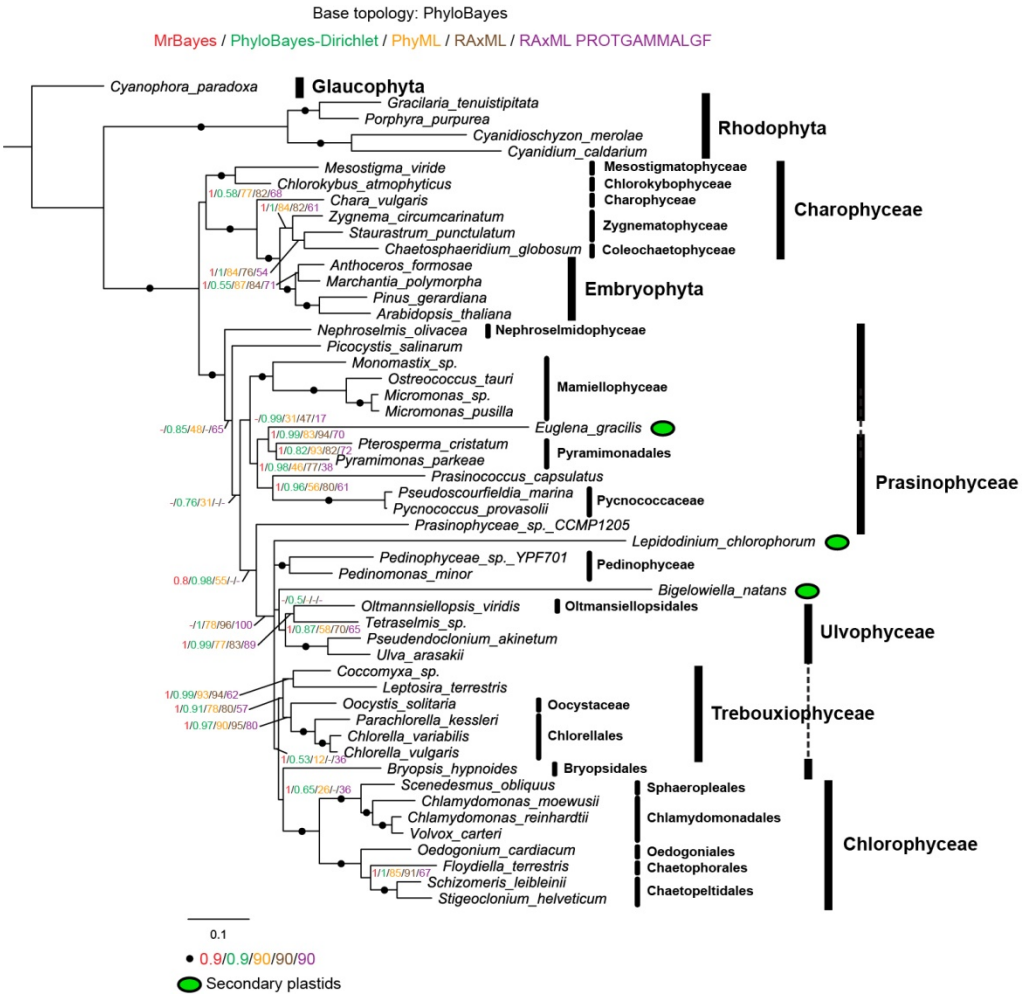


Figure 7. Consensus posterior tree resulting from a PhyloBayes analysis of 51x12-aa dataset using the mixture model CAT and a Dirichlet process. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

It has been suggested that compositional biases at the amino acid level might be introducing false groupings into the results of phylogenetic analyses based on protein data. To examine if previous patterns are artifacts caused by this problem datasets 42x79-aa and 42x94-aa were analyzed in PhyloBayes by recoding its amino acids into the 6 Dayhoff classes. The resulting topologies and associated support values (PPs) are nearly identical among the four analyses of these datasets (Fig. 8).

Figure 8 shows that the vast majority of the nodes are very well supported with the exception of the node that unites *Oocystis solitaria* to the clade formed by *Pedinomonas minor* and

Chlorellales, and the node that groups *Bryopsis hypnoides* as sister to the Chlorophyceae. Nodes to which a posterior probability of less than 0.5 is associated are not shown. The polytomous clade constituted by all core chlorophytes highlights the low support received by specific hypotheses concerning interrelationships among the three core chlorophyte classes.

Under the conditions of this type of analysis, the early diverginig *Mesostigma viride* and *Chlorokybus atmophyticus* group with the rest of streptophytes with high support. The chlorophyte *Nephroselmis olivacea* is placed as sister to all other chlorophytes, and the prasinophyte *Pycnococcus provasolii* as sister to core chlorophytes. A clade formed by the pyramimonadaleans and the mamiellophyceans emerges again with maximum support. The exact placement of the *Bigeowiella natans*' secondary plastid could not be resolved, whereas *Bryopsis hypnoides* appears as the sister of the chlorophyceans with a PP of 0.74. Finally, the Pedinophyceae member *Pedinomonas minor* keeps branching as sister to Chlorellales despite the correction for compositional biases intended with this method, and together with *Oocystis* and the *Coccomyxa* + *Leptosira* clade give support for the Trebouxiophyceae class (Fig. 8).

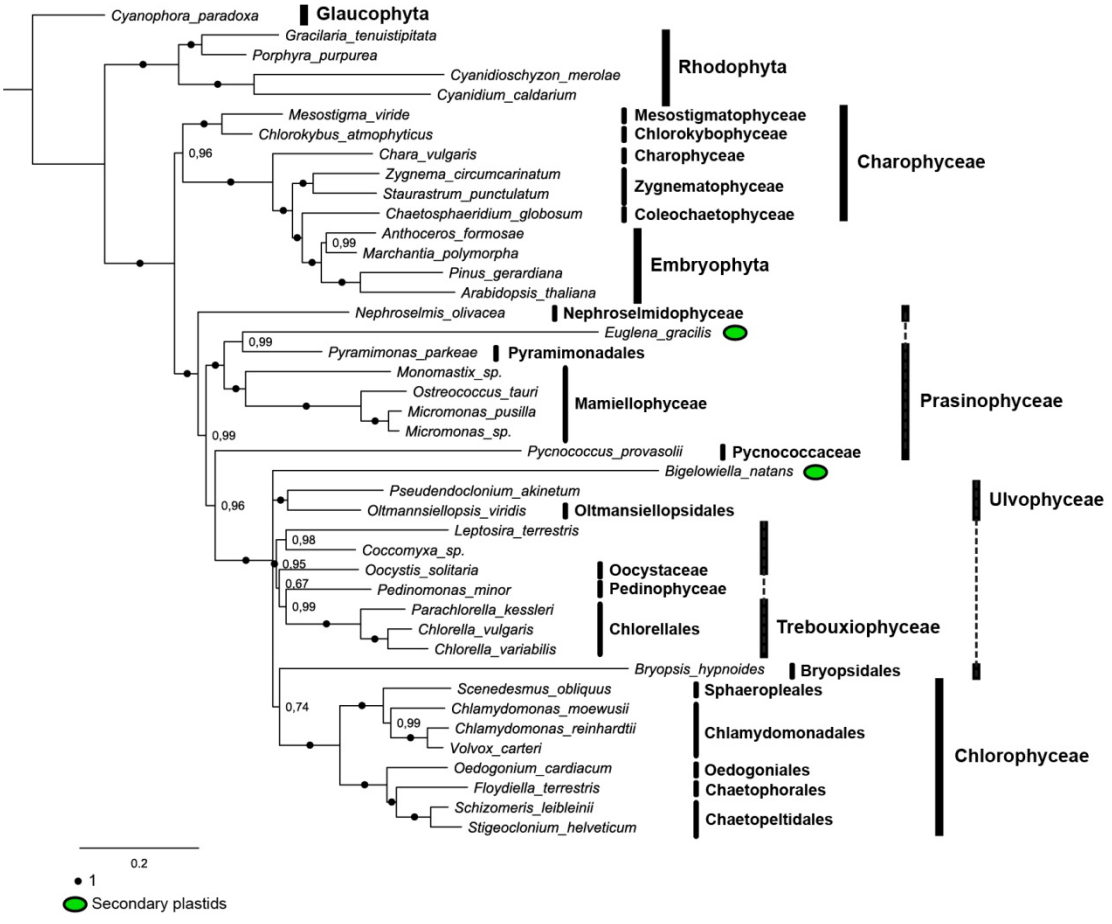


Figure 8. Consensus posterior tree resulting from a PhyloBayes analysis of 42x79-aa dataset using the Dayhoff 6 classes recoding scheme. This result is essentially the same as that obtained for the 42x94-aa dataset under the same conditions.

VIRIDIPLANTAE RELATIONSHIPS BASED ON NUCLEOTIDE SEQUENCES

When 41x79-nt and 41x94-nt datasets, which include the three codon positions that are treated as an unique partition, were subjected to model-based phylogenetic analyses, the most incongruent results were obtained in comparison to the other analyzed datasets (Figs. 9, S24-26 and S27-30). When streptophytes are considered a congruent topology with previous analyses is revealed. With both datasets *Chaetosphaeridium globosum* branches as the sister of embryophytes with maximum support. The clade *Mesostigma* + *Chlorokybus* robustly appears as the earliest-diverging streptophyte.

Regarding the chlorophytes, relationships are more scrambled relative to previous topologies. Clades that remain well supported include the Chlorphyceae class, the Chlorellales, and a mamiellophycean clade formed by the two *Micromonas* species and *Ostreococcus tauri*. *Pycnococcus provasoli* is the sister of the latter clade, and *Nephroselmis olivacea* sister to this.

Additional inconsistent groupings that appear in the analyses of these two datasets include a clade composed of *Oocystis solitaria*, *Bryopsis hypnoides*, *Leptosira terrestris* and *Bigelowiella natans*' secondary plastid. All other lineages present ambiguous positions (Fig. 9).

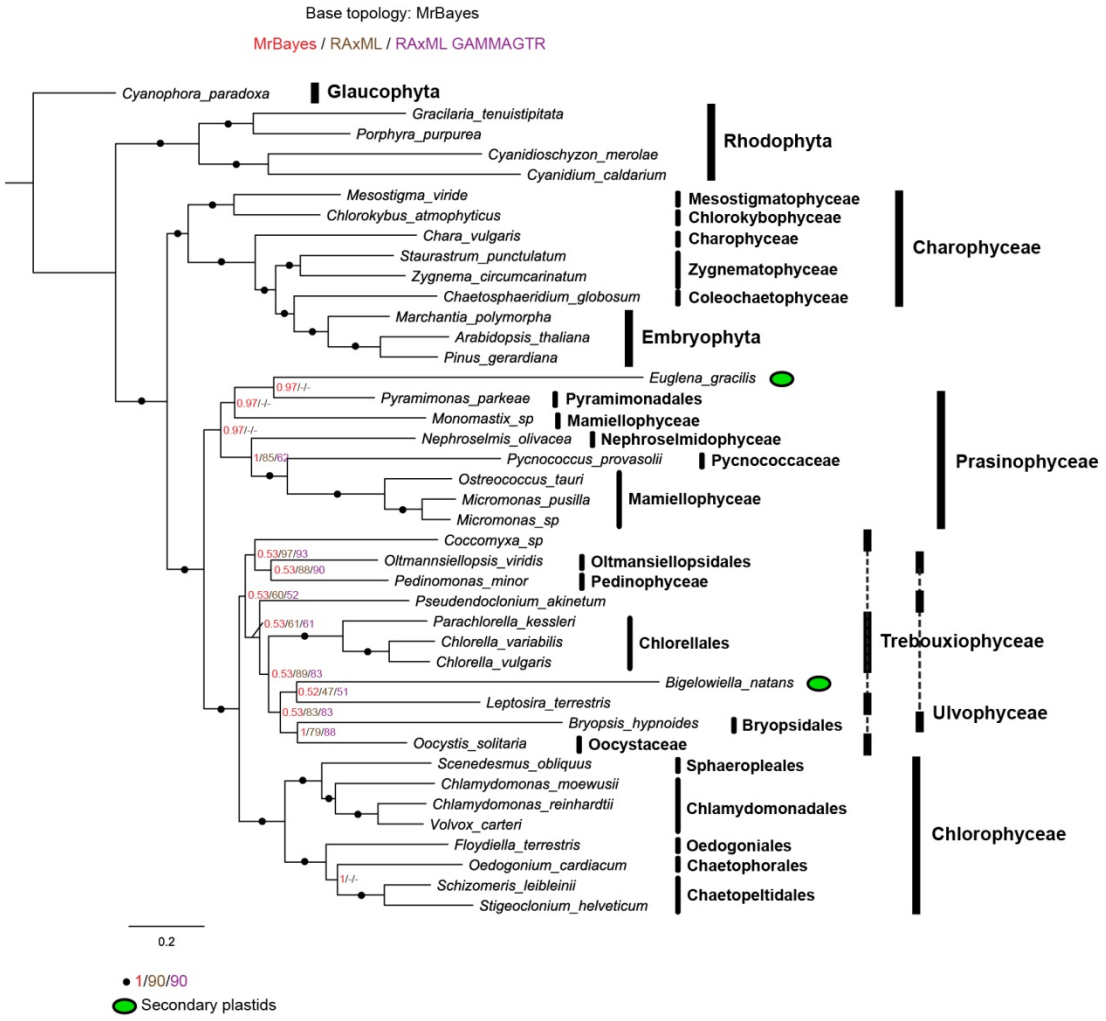


Figure 9. Consensus posterior tree resulting from a MrBayes analysis of 41x79-nt dataset using the GTR+Γ8+I model. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

Topologies resulting from the analyses of nucleotide datasets that exclude the third codon position (41x72-nt-cp12 and 41x94-nt-cp12) show more congruent topologies with those derived from amino acid sequence analyses (Figs. 10 and S31-39). However, some interesting but contrasting patterns are well defined. First, *Chaetosphaeridium globosum* branches with embryophytes and this relationship is relatively well supported in both Bayesian and ML analyses. A streptophyte clade uniting the *Mesostigma* + *Chlorokybus* clade with the Phragmoplastophytina receives the highest support (Fig. 10).

With respect to the chlorophytan branch of the Viridiplantae tree, all prasinophytes formed a strongly supported clade. Pyramimonadaleans are the sisters of mamiellophyceans, and these are sisters to a relatively well supported clade consisting of *Nephroselmis olivacea* and *Pycnococcus provasolii*.

Pedinomas minor no longer is placed within trebouxiophyceans as sister to Chlorellales. Instead, it appears basally as sister to all core chlorophytes, occupying the position that *Pycnococcus* did in amino acid based trees. The three main classes are recovered: Chlorophyceae with strong support as in previously described cases, Ulvophyceae (including the chlorarachnean secondary plastid) to the exclusion of the unstable *Bryopsis* is also well supported, and Trebouxiophyceae lacks good support. Moreover, Ulvophyceae branches as the sister of the weakly supported *Bryopsis* + Chlorophyceae clade, however, this relationship is not well supported. Relationships among the three classes tend to remain uncertain (Fig. 10).

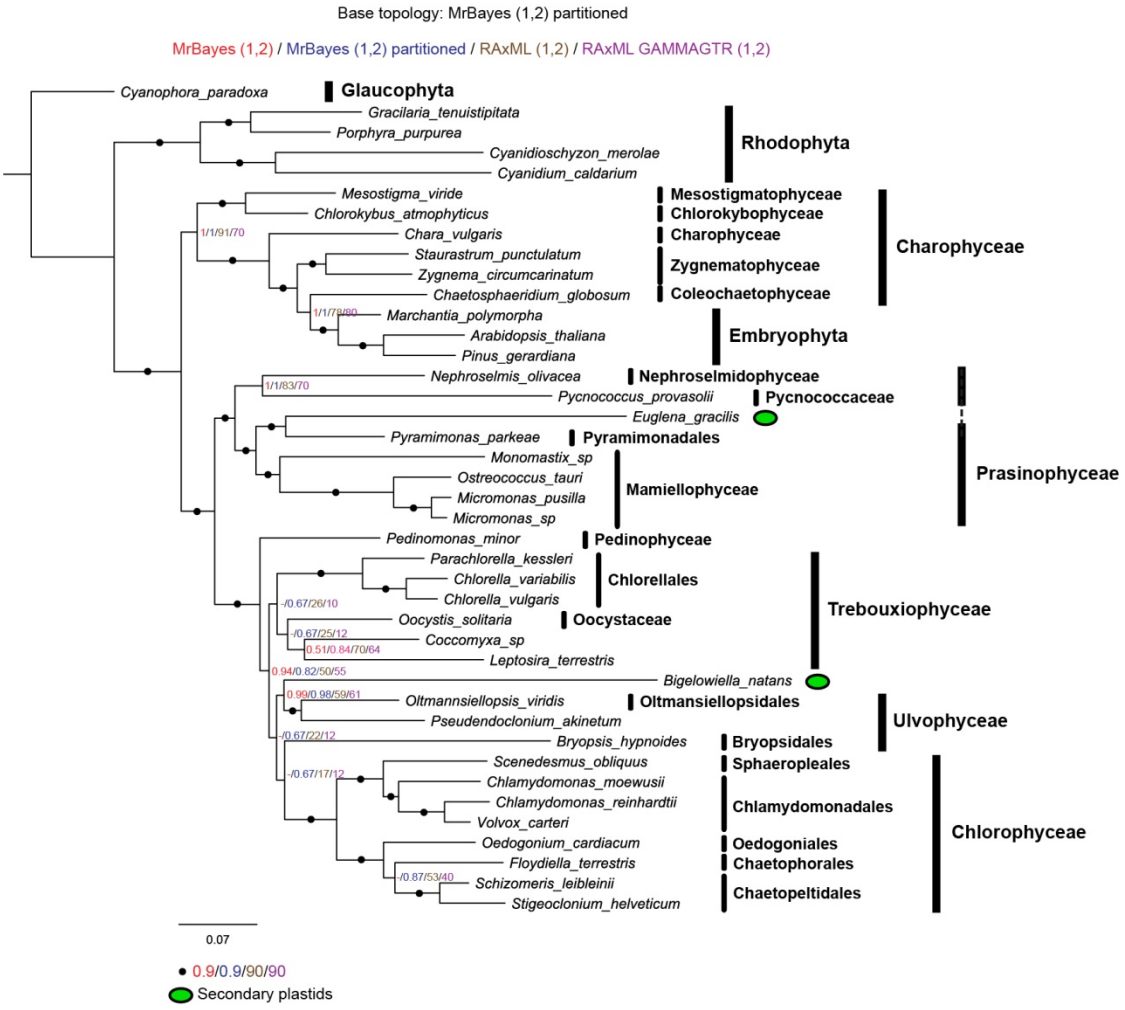


Figure 10. Consensus posterior tree resulting from a MrBayes analysis of 41x79-nt-cp12 dataset using the GTR+Γ8+I model and unliked parameters for the two partitions. It is taken as a base topology on which support values derived from supplementary analyses are superimposed.

SLOW-FAST ANALYSIS

Impact of the removal of fast-evolving positions on the 42x94-aa dataset was assessed by performing the slow-fast method. After analyzing each dataset with a decreasing number of variable sites, the support values for major nodes that turned out to be problematic in previous analyses showed trends in some cases (Tabs. 1 and S3-4). Results from the ML analyses using RAxML and the PROTGAMMALGF and PROTCATGTR models are less variable and show clear patterns in a more definite manner where present (Tabs. 1 and S4). In contrast, MrBayes results in which the mixture model of protein evolution was used present more variable support values (PPs) across datasets from which patterns are more difficult to abstract (Tab. S3).

Table 1. RAxML-PROTGAMMALGF bootstrap values supporting Viridiplantae selected clades across 27 datasets with decreasing number of fast-evolving sites.

Clade	Number of positions																										
	18975	18971	18961	18948	18923	18889	18851	18792	18704	18604	18472	18308	18097	17859	17561	17237	16889	16492	16039	15469	14818	13999	13022	11915	10630	9080	6574
	S26	**S25**	**S24**	**S23**	**S22**	**S21**	**S20**	**S19**	**S18**	**S17**	**S16**	**S15**	**S14**	**S13**	**S12**	**S11**	**S10**	**S9**	**S8**	**S7**	**S6**	**S5**	**S4**	**S3**	**S2**	**S1**	**S0**
Streptophyta	62	63	59	55	66	71	74	68	66	71	73	79	78	83	85	82	87	85	80	84	92	84	81	93	56	-	-
(*Anthoceros*(Spermatophyta))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Anthoceros*, *Marchantia*)	90	93	90	90	92	91	88	91	90	90	94	86	86	86	77	64	76	55	69	58	65	76	92	96	95	100	-
(*Chaetosphaeridium*(Embryophyta))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	52	60	66	-
(Zygnematophyceae(Embryophyta))	72	68	68	67	76	73	75	73	75	79	86	78	80	76	88	73	81	71	45	62	81	62	58	-	-	-	-
(*Nephroselmis*(prasinophytes))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	23	-	73	-	-
(Chlorophyta)-*Nephroselmis*	54	46	51	54	58	44	51	52	45	38	49	42	43	57	60	59	64	59	63	56	69	49	-	71	-	-	-
(*Nephroselmis*, *Pycnococcus*)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	69	-
(Mamiellophyceae(Pyramimonadales))	97	99	97	100	99	99	99	95	98	96	97	98	98	96	99	97	99	99	99	99	100	98	94	97	87	70	-
(*Pycnococcus*, *Euglena*)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Pycnococcus*(core chlorophtes+*Bigelowiella*))	71	75	66	72	70	61	64	59	55	53	51	49	55	66	72	67	71	73	62	70	78	83	51	-	-	-	-
(*Pycnococcus*(core chlorophytes-*Bigelowiella*))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Pedinomonas*(core chlorophytes+*Bigelowiella*))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(core chlorophytes)-*Bigelowiella*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Trebouxiophyceae (including *Pedinomonas*)	33	32	37	35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Pedinomonas*(Chlorellales))	92	83	93	85	87	90	96	83	85	74	71	81	77	80	53	76	74	76	63	-	-	-	-	56	52	-	-
(*Pedinomonas*(*Oocystis*(Chlorellales)))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51	60	61	31	-	-	-	-
(*Oocystis*(*Pedinomonas*(Chlorellales)))	47	41	45	44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	39	-	-	-	-	46	-	-	-
(*Coccomyxa*, *Leptosira*)	94	95	93	93	96	91	96	98	95	100	100	98	100	100	99	96	96	100	95	97	97	88	62	47	59	92	-
(*Oocystis*(*Coccomyxa*, *Leptosira*))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Oocystis*(Chlorellales))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	70	62	65	58	-	-	-	-
(*Bigelowiella*(*Pseudendoclonium*, *Oltmansiellopsis*))	73	75	82	81	72	72	84	81	76	84	62	62	58	64	60	61	65	83	72	81	46	-	-	-	21	-	-
(*Bryopsis*(ulvophyceans+*Bigelowiella*))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Bryopsis*(ulvophyceans))	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Bigelowiella*, *Bryopsis*)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Bryopsis*(Chlorophyceae))	48	55	51	64	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46	70	55	-	-	-	-	28	-
(*Pedinomonas*, *Bryopsis*)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
(*Floydiella*(*Schizomeris*, *Stigeoclonium*))	97	93	91	84	92	92	92	93	92	92	89	85	86	85	91	80	90	88	90	95	94	94	95	99	93	99	-

Groups that remain well supported (generally above 80% BV) in ML analyses across the whole spectrum include the chlorophycean clade (*Floydiella*(*Schizomeris*, *Stigeoclonium*)), the trebouxiophycean clade (*Coccomyxa*, *Leptosira*), the bryophytes (*Anthoceros*, *Marchantia*) and the prasinophyte clade (Mamiellophyceae(Pyramimonadales)).

The complete Streptophyta, i.e., that including the *Mesostigma* + *Chlorokybus* clade, receives moderate support (55-78%) across the spectrum, generally increasing its support towards the second half of the spectrum (80-93%) in the S14-S3 interval. In respect to the sister group of embryophytes, zygnematophyceans appear with moderate support as their sisters (62-81%), but in the last three datasets lose their sister position to Coleochaetophyceae, which starts appearing with moderate support (52-66%) as the sister of the land plants (Tab. 1).

The early-diverging prasinophyte *Nephroselmis olivacea* is excluded from the clade formed by all other chlorophytes, appearing thus as their sister. This placement receives low to moderate support along most of the spectrum, except in the last four datasets (S4-S1) where its position alternates between the two following topologies: (*Nephroselmis*(prasinophytes)) and (*Nephroselmis*, *Pycnococcus*).

Pycnococcus is usually placed as sister to all chlorophytes including the secondary plastid of the chlorarachniophyte *Bigelowiella natans*. This occurs throughout most of the spectrum, but support disappears in the last three datasets. Support for the placement of *Bigelowiella natans* is moderate (62-84%) until S5 dataset when it drastically decreases.

The affiliation of *Pedinomonas* minor to the Chlorellales within Trebouxiophyceae is strongly supported when slow-fast datasets are larger, but it decreases with smaller datasets accordingly. *Oocystis* exhibits a similar pattern where topology (*Oocystis*(*Coccomyxa*, *Leptosira*)) is favoured at the beginning, but topologies (*Oocystis*(*Pedinomonas*(Chlorellales))) and *(Pedinomonas*(*Oocystis*(Chlorellales))) receive better support at the end of the spectrum.

When patterns could be derived from the MrBayes slow-fast analysis they are a congruent subset from those found in the RAxML slow-fast analysis (Tabs. 1 and S3-4).

In general when patterns could be extracted from the table these describe a decrease in support from initially well supported clades towards less supported ones at the end of the spectrum, starting from datasets S8-S6. Cases where clades only receive support in very trimmed datasets are the (*Chaetosphaeridium*(Embryophyta)) clade and to some extent the (*Oocystis*(Chlorellales)) trebouxiophycean clade (Tab. 1).

VIRIDIPLANTAE RELATIONSHIPS BASED ON SUPERTREES

The supertree derived from the combination of the 94 sources trees allows the examination of the congruence among single-gene topologies. The resultant supertree is largely compatible with previous topologies produced in this study with the exception of the following taxa that are placed differently: *Pedinomonas minor*, *Helicosporidium* sp., *Bryopsis hypnoides*, *Bigelowiella natans*,

Leptosira terrestris, *Oocystis solitaria* and *Coccomyxa* sp (Fig. 11). Incongruence relates mainly to the monophyly of trebouxiophyceans and ulvophyceans. Despite this, the trebouxiophycean subclades Chlorellales and (*Lepstosira*, *Coccomyxa*, *Oocystis*) and the ulvophycean clade *Pseudendoclonium* + *Oltmansiellopsis* are recovered.

The ulvophycean *Bryopsis hypnoides* and the chlorarachnean secondary plastid (*Bigelowiella natans*) are taxa that have shown to be difficult to place. The supertree reconstruction is no exception to this pattern where they do not group with the two other ulvophyceans. *Pedinomonas minor* branches neither robustly with Chlorellales as in amino acid sequence-based trees (Fig. 3), nor sister to core chlorophytes as in nucleotide sequence-based trees (Fig. 10). This new fragile placement of *Pedinomonas* might indicate lack of congruence among individual gene trees.

The best supertree found through a heuristic search is very similar to a bootstrap consensus from the 94 source trees. Figure 11 shows a comparison between these two topologies and the minor differences among them. It is notable the basal position of *Nephroselmis olivacea* relative to all other viridiplants. The Streptophyta clade, including the basal *Mesostigma* and *Chlorokybus*, is recovered with a moderate bootstrap value, and *Chaetosphaeridium globosum* appears as the sister to land plants. The presumed clade formed by the pyramimonadaleans and the Mamiellophyceae is also recovered. *Pycnococcus* is the sister to the core chlorophytes and Chlorophyceae class is strongly supported as in all previous analyses.

Most clades, however, tend to have moderate to low bootstrap percentages associated. This could be an indication of some incongruence among the phylogenetic signal of different ptDNA protein-coding gene trees and a result of the relatively low number of them. Clades that receive high bootstrap values are less inclusive with the exception of the Chlorophyceae class and the Phragmoplastophytina within streptophytes. The Streptophyta and the core chlorophytes (including *Pedinomonas*) receive moderate support (>70%).

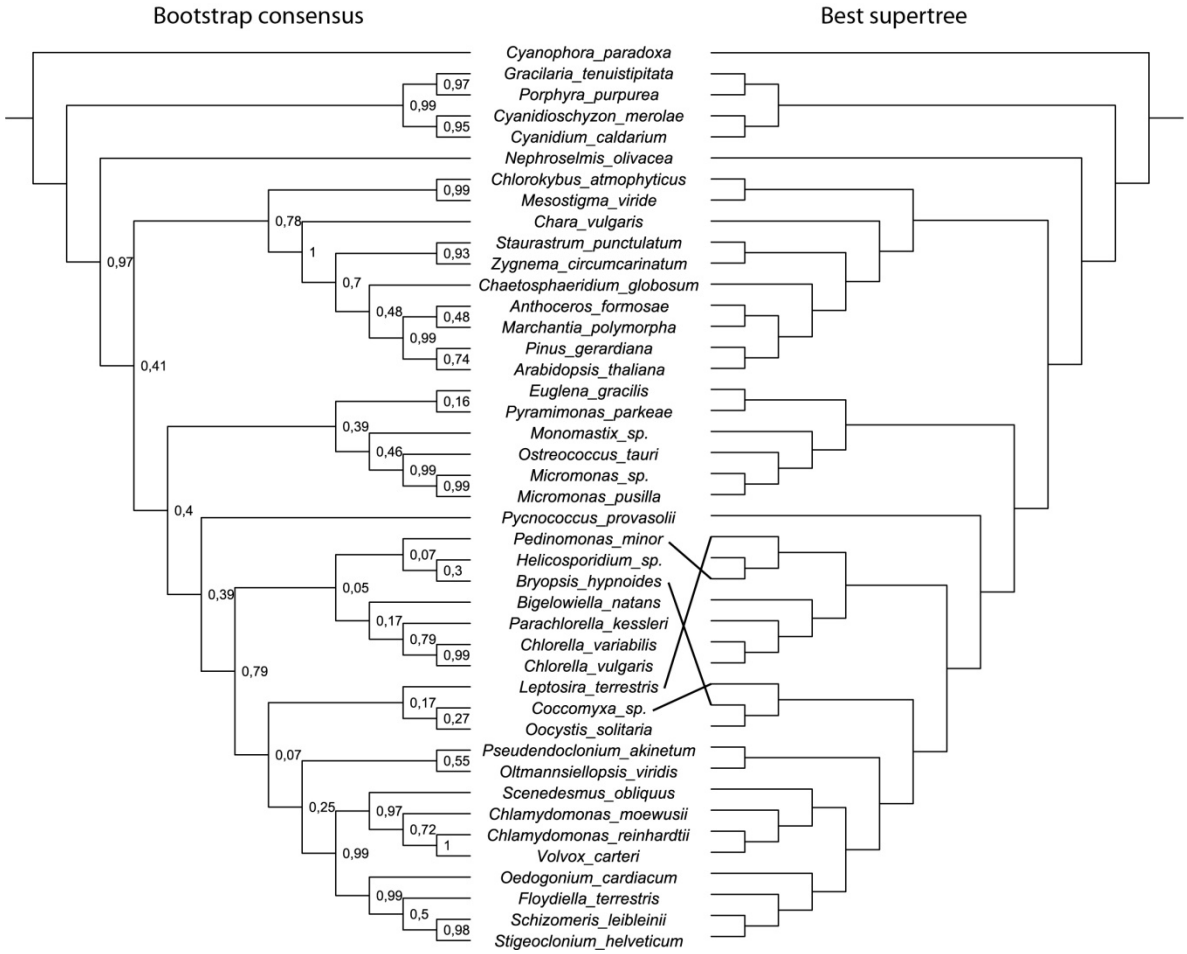


Figure 11. Comparison between the best supertree and the supertree bootstrap consensus derived from analyses in the Clann 3.2.2. software.

VIRIDIPLANTAE RELATIONSHIPS BASED ON PLASTID GENE CONTENT

Figure 12 presents the result from a gene presence-absence analysis of 274 plastid genes, most of them treated as irreversible characters due to the high improbability of gaining a lost ptDNA-encoded gene. The extended majority rule consensus summarized from the 184 most parsimonious trees found reveals some interesting aspects regarding the groups recovered and the nature of the characters used. Some groupings coincide with monophyletic groups found in sequence-based analyses. These include (1) the close relationship between *Mesostigma* and *Chlorokybus*, (2) the clade that unites *Chara* with *Chaetosphaeridium*, the zygnematophyceans and the embryophytes (Phragmoplastophytina), (3) a group composed of the trebouxiophyceans *Coccomyxa* sp., *Chlorella variabilis*, *Parachlorella kessleri* and the incertae sedis *Pedinomonas minor*, (4) the ulvophycean subgroup formed by *Pseudendoclonium* and *Oltmansiellopsis*, and (5) the Chlorophyceae class.

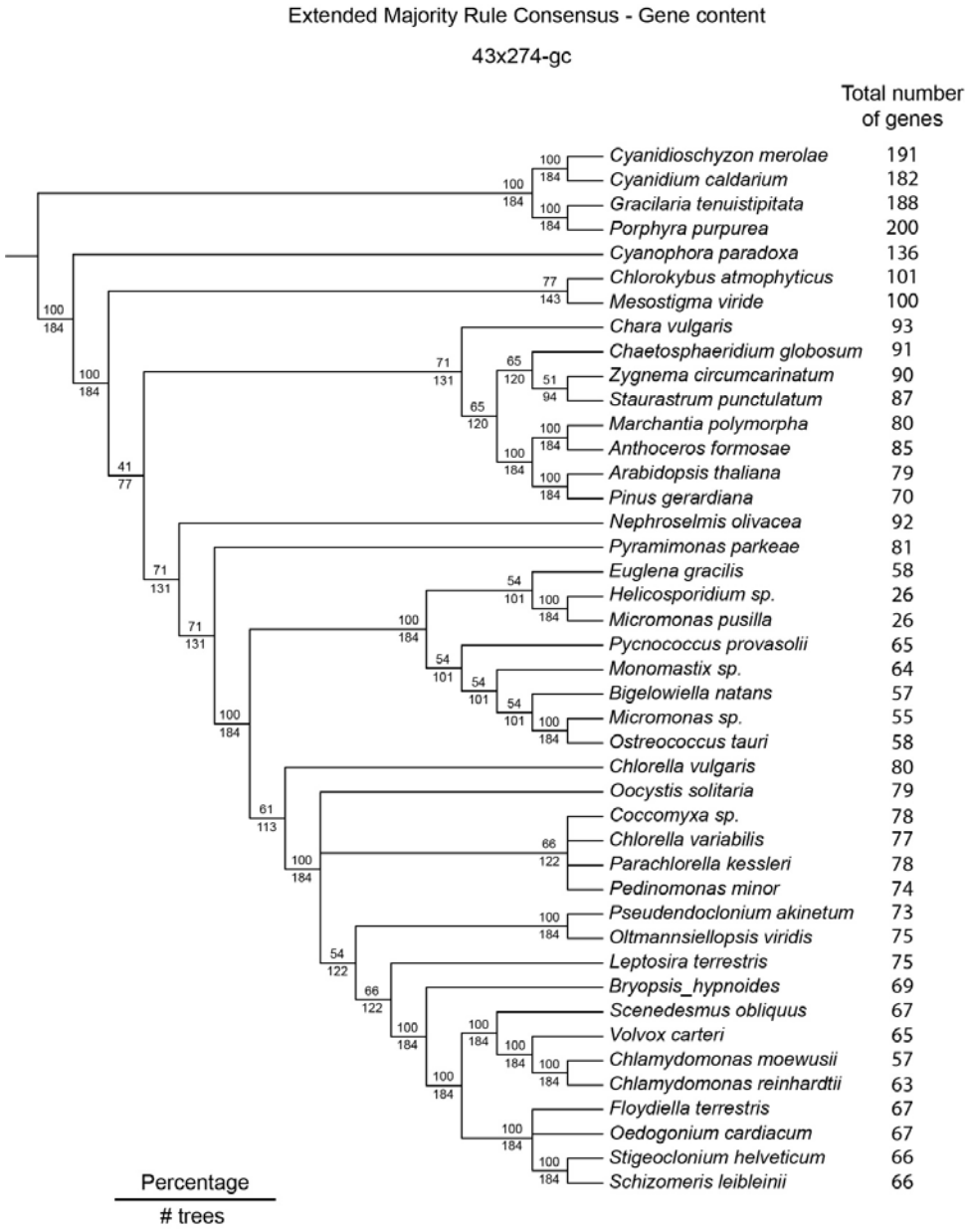


Figure 12. Extended majority rule consensus of most parsimonious trees after a PAUPRatchet parsimony analysis using the gene content dataset.

The distribution of the number of genes present out of the 274 genes selected for the parsimony analysis might imply that some of the unexpected groupings could be the result of homoplasy. A pattern is evident in some cases such as the affiliation of *Helicosporidium* sp. and *Micromonas pusilla*. Both of them appear distantly related in previous analyses, where the prasinophytes *Micromonas pusilla* is placed robustly within the Mamiellophyceae while the trebouxiophycean *Helicosporidium* sp. groups confidently with the Chlorellales. However, in Figure 12 they are sisters to each other, and this pattern was found in all of the 184 most parsimonious trees. Both species have drastically reduced their genomes converging into the same small gene set of 26 genes.

Another artifactual group, probably the result of similar gene content (i.e., homoplasy), is the one that goes from *Euglena gracilis* to *Ostreococcus tauri*, all of which composing taxa have the smallest plastid gene repertoire of all, from 26 to 58 genes. The basal placement of (1) the *Bryopsis* lineage relative to the Chlorophyceae class, (2) *Nephroselmis* and *Pyraminomnas* to all chlorophytes, and (3) *Mesostigma* + *Chlorokybus* to all other green plants, could be a pattern produced by the same cause, because they tend to have bigger gene complements than their sister groups.

VIRIDIPLANTAE RELATIONSHIPS BASED ON PLASTID GENE ORDER DATA

The gene order analysis tends to reveal more incongruence and less resolution in comparison to sequence-based and gene-content derived topologies (Fig. 13). The 36x440-go extended majority rule consensus cladogram, which was built from a dataset containing contiguous-gene characters of at least 3 genes, recovers a Phragmoplastophytina group to the exclusion of the zygnematophycean *Zygnema circumcarinatum*. However, these streptophytes are grouped together with unexpected interrelationships. The suspected streptophyte grouping of *Mesostigma* and *Chlorokybus* is found as sister to all other green plants.

On the chlorophytan side, the trebouxiophycean clade formed by the Chlorellales and *Oocystis* is recovered, as well as the Chlorophyceae class that is usually found very well supported in analyses with different data. Other taxa occupy conflicting positions in regard to sequence-based topologies.

Figure 13. Extended majority rule consensus cladogram derived from a PAUPRatchet parsimony analysis of the conserved gene cluster matrix 36x440-go.

MAPPING OF CONTINUOUS GENOMIC FEATURES (DNA SIZES)

Optimizing sequence-independent genomic characters on robust topologies derived from sequence-based phylogenetic analyses allows analyzing their evolution along the branches of the Viridiplantae tree. Genome size in plastid organelles is a function of the number and length of genes, the number and length of gene-interrupting introns, and the length of intergenic spacers. Comparing the amount of DNA dedicated to each type of DNA relative to genome size allows us to have a better understanding of the major tendencies in plastid genome evolution in Viridiplantae.

Based on the amino acid topology 1 (Fig. 14) and the nucleotide topology 2 (Fig. S40), genome sizes have been relatively homogeneous during streptophyte history, whether the *Mesostigma* +

Chlorokybus clade is placed among them or not. It ranges from 117,618 bp in the gymnosperm *Pinus gerardiana* to 184,933 bp in *Chara vulgaris* (Fig. 14A and Tab. S2). Similarly to the pattern found in streptophytes regarding plastid genome size, the genic DNA size of their genomes shows no major variation during its evolution (Fig. 14A). This size range for genic DNA or one a little bigger seems to be the ancestral one to all viridiplants. When the non-genic DNA size is considered it is also observed that no major changes have occurred in this fraction, except in the charophyte *Chara vulgaris* in which an increase in the intronic DNA size is observed (Fig. 14B). A less drastic increase in the amount of intronic DNA has also occurred in the clade formed by the embryophytes and *Chaetosphaeridium globosum* (Fig. 14C). Intergenic spacers have not suffered expansions or reduction during streptophyte evolution (Fig. 14D).

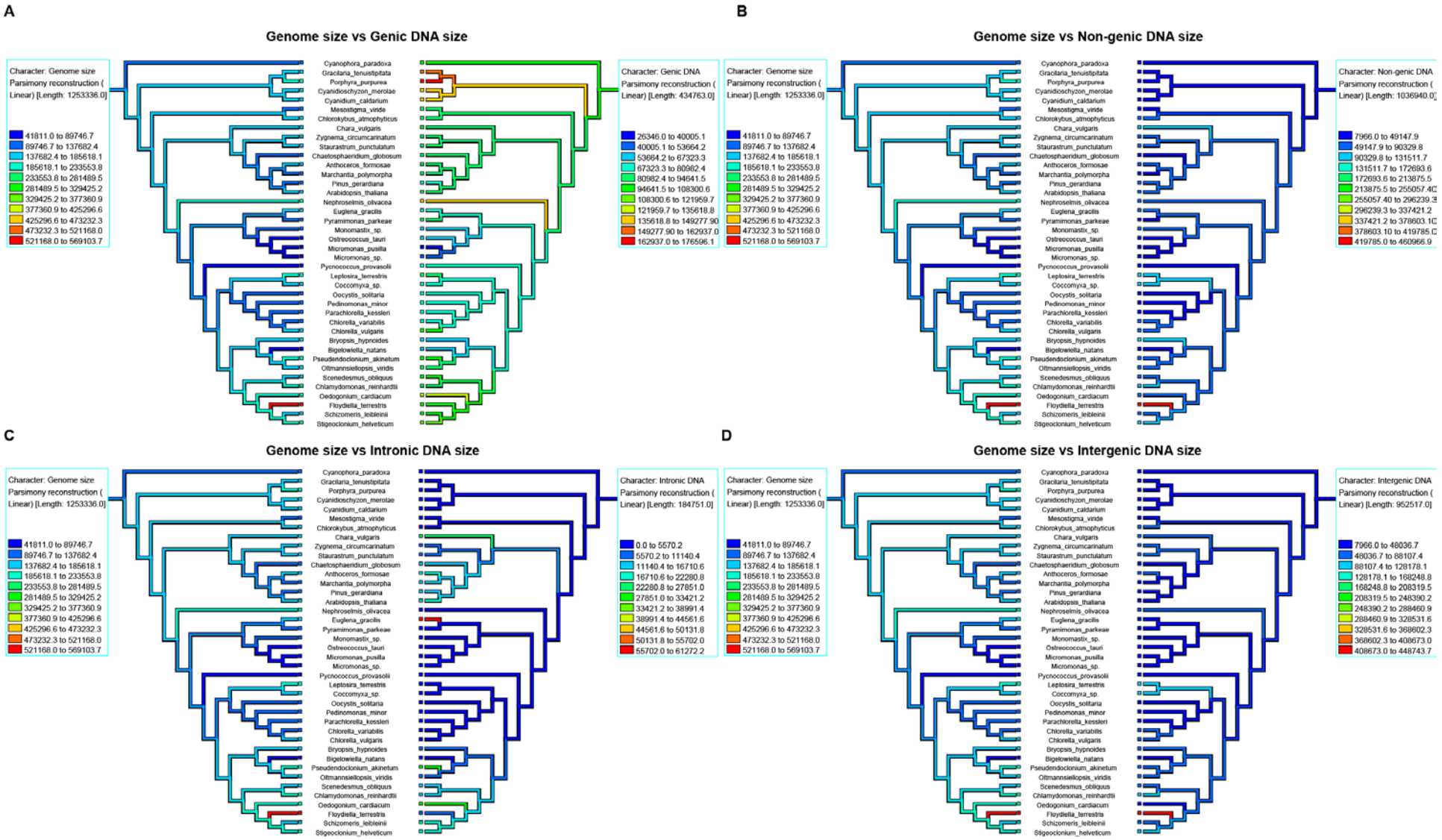


Figure 14. Mirror trees (topology 1) comparing the ancestral reconstruction of continuous values for plastid genome size vs (A) genic DNA size, (B) non-genic DNA size, (C) intronic DNA size, and (D) intergenic DNA size.

Relative to Chlorophyta, *Nephroselmis olivacea* has a relatively large genome size (200,799 bp) in contrast to most chlorophytes, with the exception of chlorophyceans. The size of its genome appears to have increased and not gained much non-genic DNA in the form of intronic or intergenic DNA (Fig. 14).

The Mamiellales (e.g., *Micromonas* spp., *Ostreococcus*) posses the smallest plastid genome sizes of all photosynthetic green plants sequenced so far. The bigger clade of which the Mamiellales is part, the Mamiellophyceae plus Pyramimonadales, is also characterized by small genome sizes no bigger than 114,528 bp (*Monomastix* sp.), with the exception of the secondary plastid genome of *Euglena gracilis* that has a size of 143,171 bp (Tab. S2). The expansion of the plastid genome of *Euglena gracilis* is caused by a dramatic increase in the size of intronic DNA, as it is easily observable in red in Figure 14C.

After the divergence of most chlorophytes from *Nephroselmis*, a reduction of the genic DNA is observed in the reconstructed states. In a manner congruent with the pattern described for plastid genome size, the genic DNA size of mamiellaleans has been reduced drastically, especially in *Micromonas pusilla* where it has reached a value between 26,346 and 40,005 bp. In addition, the mamiellalean plastid genome is extremely streamlined in regard to non-genic DNA (Fig. 14B).

This pattern indicates that the extreme genome reduction in mamiellaleans began from an already reduced genome, and that the secondary plastid genome of *Euglena* has experienced a secondary increase in its size by intron proliferation.

In contrast to the plastid genome of the euglenophyte, the secondary plastid of the chlorarachniophyte *Bigelowiella natans* has followed the opposite path, reducing its genome to a size of 679,166 bp. This reduction has not only been in the loss of genic DNA, but probably also by losing non-genic DNA (Fig. 14B).

The *Pycnococcus* taxon, whether it is placed sister to *Nephroselmis* or to all core chlorophytes, has a small total plastid DNA length (80,211 bp) in comparison to its sister group, and has not gained practically any non-genic DNA conserving its relatively ancestral streamlined nature. The trebouxiophyceans have moderate to large genome sizes that go from 96,287 bp in *Oocystis solitaria* to 175,731 bp in *Coccomyxa* sp. and 195,081 bp in *Leptosira terrestris* (Tab. S2). Interestingly, the two latter plastid genomes have increased their size by accumulating non-genic DNA relative to its sister group, not in the form of intronic DNA but only in intergenic spacers (Fig. 14D).

The plastid genome of *Bryopsis hypnoides* has followed a similar evolutionary pattern of reduction to that observed in *Pycnococcus* and *Bigelowiella*'s plastid. In them, the loss of genic DNA has been accompanied by a loss or not gain of secondary non-genic DNA (Figs. 14A and B).

Plastid genome sizes among the green plants that are close to the upper limit of their size range (41,811-223,902 bp) include the genomes of the ulvophycean *Pseudendoclonium akinetum* (195,867 bp), and the chlorophyceans *Chlamydomonas reinhardtii* (203,828 bp), *Oedogonium*

cardiacum (196,547 bp) and *Stigeoclonium helveticum* (223,902 bp) (Tab. S2). All of these taxa have larger amounts of genic DNA compared to other chlorophytes, and similar to the genic DNA sizes observed in streptophytes (Fig. 14A). With respect to intronic DNA, the ulvophycean *Psudendoclonium alinetum* and the chlorophycean *Oedogonium cardiacum* have gained in the number or length of their introns. A similar pattern is observed in *Schizomeris leibleinii* and *Stigeoclonium helveticum*, although to a lesser extent.

An exceptional outlier is the plastid genome of the chaetopeltidalean *Floydiella terrestris* (Chloropyceae) that has a size of 521,168 bp. The extreme expansion of the plastid genome size of the chaetopeltidalean is the result of massive increases of intergenic spacers, while intronic DNA has probably even been reduced according to the ancestral continuous character state reconstruction (Figs. 14C and D).

On the whole, the classes Ulvophyceae and Chlorophyceae seem to present more variability in plastid organization related to these three types of DNA sizes (Fig. 14). Most green plant plastid genomes have reduced their size relative to the ancestral streptophyte-like ptDNA genome size. Exceptions to this rule are observed among the ulvophyceans mentioned above, and the chlorophyceans in which there seems to have been a trend in their plastid genome sizes to increase.

GENE CONTENT CHARACTER MAPPING

Gene content evolution in green plastid genomes has been highly homoplasic. Because of this, unambiguous synapomorphies are not common (Fig. 12) and unique changes traced to a single branch in the ancestral character state reconstruction of gene content are few (Figs. S41 and S42). Several genes have been lost independently in parallel fashion in more than one lineage. Despite this clear case for convergent evolution, homoplasic synapomorphies are still distinguishable and allow identifying some groups. Supplemental Figures S41 and S42 show which characters are more homoplasic than others based on their consistency index (CI).

Due to the fact that topologies 1 and 2 are robust inferences from large molecular datasets, and disagreements between them are not significant, reconstructed changes along the branches could be interpreted as support for them. In addition, Figure 15 allows us to trace the evolution of gene loss along the ancestral branches of the Viridiplantae tree. Some groups have suffered several plastid gene losses during their origin, whereas others have retained a more ancestral inferred gene repertoire. Both figures show the number of gene losses that have occurred along the branches. To see the specific genes please refer to Figures S41 and S42.

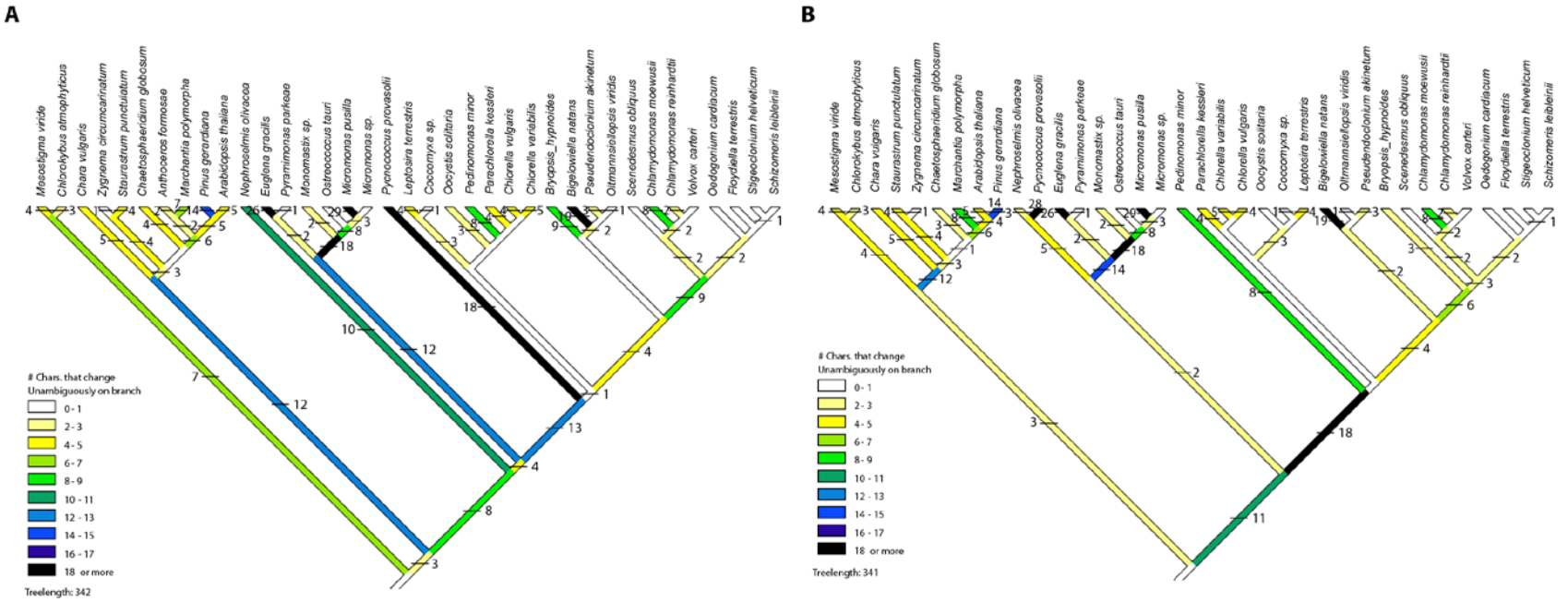


Figure 15. Gene presence-absence characters mapped and optimized along the branches of the amino acid-derived topology 1 (A) and the nucleotide derived topology 2 (B).

Topologies 1 (Fig. 15A) and 2 (Fig. 15B) differ in the positions of the *Mesostigma* + *Chlorokybus* clade, the prasinophytes *Nephroselmis olivacea* and *Pycnococcus provasolii*, *Pedinomonas minor* and the ulvophycean *Bryopsis hypnoides*. Given the taxon sampling and the topologies analyzed at least twelve changes have occurred during the separation of the Phragmoplastophytina from other green plants. This number remains irrespective of whether the *Mesostigma* + *Chlorokybus* clade is sister to them or to all viridiplants. Overall, streptophytes have retained more genes than their chlorophytan sisters, therefore resembling more the hypothetical ancestral green plastid regarding gene content. This pattern is also seen from the low numbers existing along streptophyte branches in comparison to chlorophyte branches. The gymnosperm land plant *Pinus geradiana* is the streptophyte that has lost more genes.

Several gene content changes have also occurred during the origin of chlorophytes. In contrast to all other chlorophytes and similarly to streptophytes, the plastid genomes of the prasinophytes *Nephroselmis olivacea* and *Pyramimonas parkeae* have retained most of the complete set of *ndh* genes involved in electron transfer during photosynthesis (Figs. S41 and S42). The evolutionary emergence of the clade composed by the pyramimonadaleans and the mamiellaleans involved a considerable amount of gene loss (14 genes). Subsequently, mamiellaleans reduced their plastid genomes by losing about the same number of genes (18 genes). Furthermore, *Micromonas pussilla* lost at least 29 more genes in a drastic reduction of its plastid genome. Another lineage that shows a significant number of gene losses in both topologies 1 and 2 is *Pycnococcus provasolii* (Fig. 15).

Core chlorophytes are relatively homogenous in their gene composition with the exception of the secondary plastid of presumptive ulvophycean origin that lost 19 gene in comparison to its sister group. A bigger number of changes differentiate chlorophycean genomes from the rest, and stasis in gene loss during chlorophytan evolution is easily seen in the chlorophycean CCW clade where practically no changes have occurred (Fig. 15).

These results show good support for the Phragmoplastiphytina clade, the Pyramimonadales-Mamiellophyceae clade, the Chlorophyta, the core chlorophytes, and the Chlorophyceae class. Groups that are not well supported by changes in the gene content of their plastids are the Trebouxiophyceae and the Ulvophyceae.

GENE ORDER CHARACTER MAPPING

As seen in Figure 13 the parsimony reconstruction based on gene order data did not recover many relationships congruent with other analyses. Thus, similarly to gene presence-absence data, gene order data is prone to convergence.

Mapping the gene order dataset 36x440-go allows observing the evolution of gene rearrangements on the Viridiplantae tree. The number of changes listed on each branch corresponds to the sum of gene cluster losses and gains. From Figure 16 it is only possible to say that some ptDNAs have undergone more changes relative to others. To see the specific gains or losses or gene cluster please refer to Figures S43 and S44. According to the argument presented

above for gene content, character state changes could be interpreted as branch support in these reconstructions.

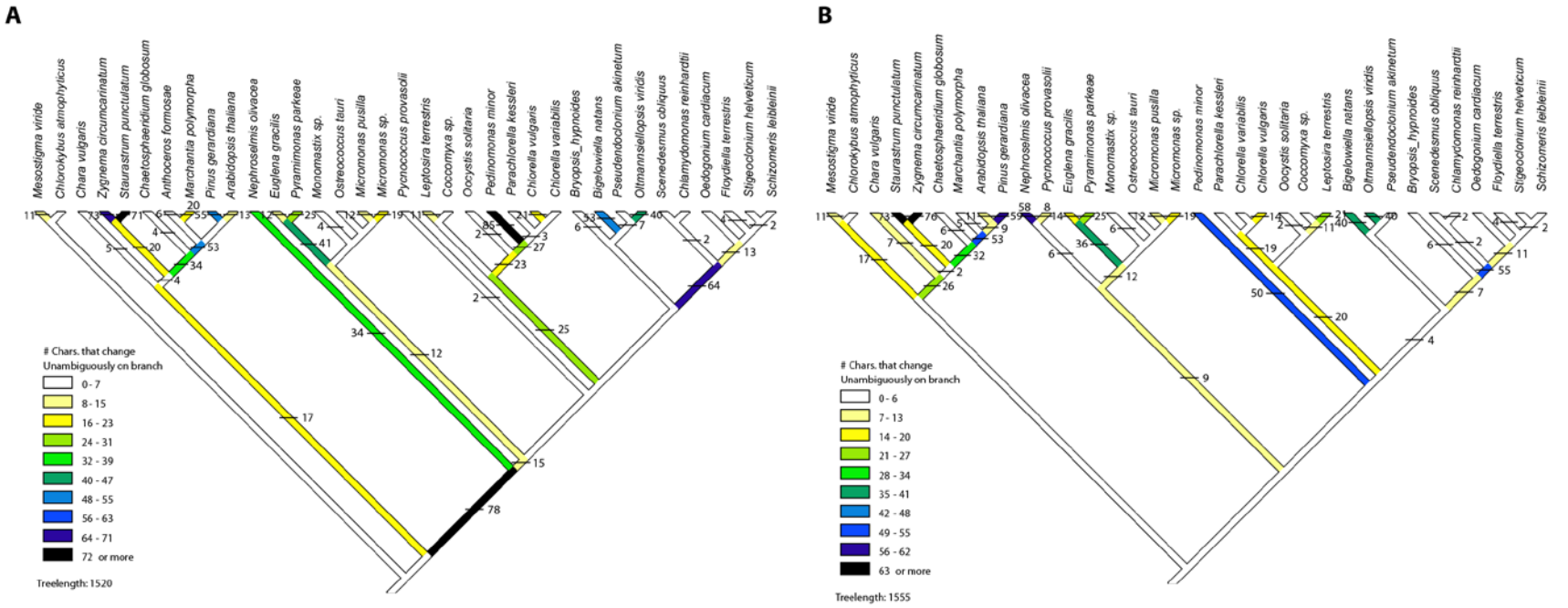


Figure 16. Conserved gene cluster characters mapped and optimized along the branches of the amino acid-derided topology 1 (A) and the nucleotide derived topology 2 (B).

Among streptophytes, the zygnematophyceans *Zygnema circumcarinatum* and *Chaetosphaeridium globosum* have suffered a significant amount of genome rearrangements in comparison to other streptophytes. The trend exhibited by these zygnematophyceans is also followed by the embryophyte *Pinus geradiana* to a lesser extent. The Phramoplastophytina is relatively well supported by at least 17 changes in gene clusters (Fig. 16A).

Among chlorophytes, *Pedinomonas minor*, *Nephroselmis oliveacea*, *Bigelowiella natans* plastid and the ulvophycean *Oltmansiellopsis viridis* have undergone more than 30 changes during the evolution of their lineages. This is true in both ancestral reconstructions. The Pyramimonadales clade comprised of *Pyramimonas parkeae* and *Euglena gracilis* secondary plastid is also supported by at least 36 changes (according to topology 2 (Fig. 16B)) that occurred in their genome prior to their divergence. The Pyramimonadales + Mamiellophyceae clade is also moderately supported by 12 changes in its branch.

Interestingly, several changes in gene clusters appear to have occurred prior to the divergence of trebouxiophcyean lineages. This support from gene order data comes unexpectedly after the gene content ancestral reconstruction from which no specific gene losses were seen to support the class. Also, when comparing topologies 1 (Fig. 16A) and 2 (Fig. 16B) it is observed that more than 20 changes suggest that *Oocystis* is the sister to Chlorellales rather than to the *Leptosira* + *Coccomyxa* clade.

Gene order adds additional support to some clades that lack it from gene content data, such as in the cases of Pyramimonadales and Trebouxiophyceae. It further corroborates the strong support existing from different sets of independent data to the Chlorophyceae class.

DISCUSSION

RELATIONSHIPS AMONG GREEN PLANT LINEAGES

In addition to trace the evolution of diverse genomic features, the present study also aimed to evaluate the available phylogenetic evidence in support of specific hypotheses of historical relationships among green plants, based on plastid genomes. The results presented by this study are in agreement with previous phylogenomic studies based on whole plastid genome sequences. Moreover, because of the employment of different datasets and methods to evaluate these hypotheses some interesting contrasting patterns emerged that are worthy of discussion.

PROBLEMATIC NODES AND UNSTABLE LINEAGES

While some lineages are well resolved and their placements do not vary in most analyses, others exhibit unstable positions and jump from one place to another in the green tree depending on the conditions of analysis. The most unstable lineage of all taxa considered for the present study is the ulvophycean *Bryopsis hypnoides,* which is sometimes placed as sister to all chlorophyceans (e.g., Fig. 5). This giant, macroscopic unicell, which exhibits a siphonous organization and is highly

differentiated in its cytomorphology, is confidently classified within the Bryopsidales (or Caulerpales) in taxonomic schemes (2,24). Ecological, morphological, cytological, and biochemical evidence relates this green alga with other siphonous green algae within the class Ulvophyceae (72,73). Furthermore, small subunit ribosomal RNA (SSU rRNA) (74) and more recently a dataset composed of seven nuclear genes, nrDNA and the plastid genes *rbc*L and *atp*B consistently placed *Bryopsis* with other Ulvophyceans, and recovered the Ulvophyceae class as a well supported monophyletic group (72). In all these analyses, the branch leading to *Bryopsis*, and in general those of the BD clade (Bryopsidales + Dasycladales), are the longest of the tree inferred. This is clearly an indication of fast-evolving lineages whose rates of molecular evolution have been greatly accelerated during their evolution. The Bryopsidales not only display long branches in sequence trees but they also have divergent translation machineries, in addition to derived siphonous body plans and an old fossil record dating to the Neoproterozoic (35).

A similar incongruent pattern to that found here is observed in the study of Zuccarello *et al.* (2009), where the seaweed *Caulerpa filiformis* is found closer to the trebouxiophycean *Chlorella* than to the ulvophyceans *Oltmansiellopsis* and *Pseudendoclonium.* This finding led the authors, very probably incorrectly, to conclude that the core chlorophyte classes Ulvophyceae and Trebouxiophcyeae are non-monophyletic (75). Similarly, the authors of the *Bryopsis* genome paper arrive to the same conclusion, although acknowledging the possible impact of insufficient taxon sampling of their analysis (76).

A new, better supported position for *Bryopsis hypnoides* was not found after successively removing fast-evolving sites in the slow-fast analysis (Tabs. 1 and S3-4). Under the MrBayes slow-fast analysis, it branches as sister to different taxa including the Chlorophyceae class, *Bigelowiella natans* and *Pedinomonas minor* (Tab. S3). Under the RAxML slow-fast analysis, the ulvophycean tends to branch as sister to the Chlorophyceae in the largest datasets, but with boostrap values generally below 50% (Tabs. 1 and S4). This would suggest that the bias is stronger in the larger datasets. Also, after recoding the amino acid datasets into the 6 Dayhoff classes the ulvophycean kept branching with the Chlorophyceae with a low posterior probability of 0.74 (Fig. 8). This is the most common placement for *Bryopsis* in my analyses, although it usually receives poor supporting values from both ML bootstrap and Bayesian PPs.

It is well known that lineages whose genetic molecular sequences evolve very rapidly are a source of phylogenetic artifacts in tree reconstruction such as long-branch attraction (77). Moreover, the insufficient taxon sampling of the Ulvophyceae class at the whole plastid genome level could increase the effect of long-branch attraction and produce other artefacts result of systematic biases (78). Despite employing the slow-fast and the Dayhoff recoding methods, these problems could not be alleviated and a more congruent pattern with independent evidence (e.g. morphology, nrRNA markers) could not be recovered. The variable and low-supported placement of *Bryopsis hypnoides* in the present study could be the result of a combination of these phenomena.

Other taxa show alternative positions in the trees inferred, one associated with amino acid data and the other resulting from nucleotide analyses. The *Mesostigma* + *Chlorokybus* clade is an example of this. It is usually sister to all green plants in amino acid trees (e.g., Fig. 3), whereas it occupies a position as the earliest streptophytes in nucleotide trees (e.g., Fig. 10). The pedinophycean *Pedinomonas minor* branches robustly as sister to Chlorellaceae in amino acid trees but basal to core chlorophytes in nucleotide trees. The pseudocourfiedialean *Pycnococcus provasolii* is the sister of core chlorophytes in amino acid trees, but it is the sister of nephrosemidophyceans in nucleotide trees. Linked to this is the pattern observed for *Nephroselmis olivacea*, in which the nephroselmidophycean appears as the sister to all chlorophytes in amino acid trees, but sister to *Pycnococcus* in nucleotide trees in a well supported and more inclusive monophyletic prasinophyte group.

Finally, the branching order among the three phycoplast-containing core chlorophytes could be resolved. The possible causes of these patterns and the probable true organismal relationships among green plants despite conflicting topologies are discussed in the next sections.

IS STREPTOPHYTA A CLADE?

The Streptophyta is the clade that contains the land plants (embryophytes) and their closest algal relatives, the charophycean algae. Among the streptophyte algae six monophyletic groups are usually considered. These have recently been raised to the class rank and include the Charophyceae, the Coleochaetophyceae, the Zygnematophyceae, the Klebsormidiophyceae, the Chlorokybophyceae and the Mesostigmatophyceae (23). There is currently certain confidence among green algae systematists that the latter two classes belong to the Streptophyta clade. This confidence has emerged during the last years in which the debate regarding the position of the scaly biflagellate unicell *Mesostigma viride* has been partially resolved to this side (23,40,79,80). More recently, independent evidence based on the presence of specific gene families in the genome of *Mesostigma viride* strongly suggests that this formerly considered prasinophyte unicell is in fact a member of the streptophytes (81,82). Another important point of discussion is the branching pattern of early streptophytes. Are *Mesostigma* and *Chlorokybus* sisters? Or did *Mesostigma* diverge first followed by *Chlorokybus* which is closer to all other streptophytes?

First of all, it has to be mentioned that all trees that were reconstructed here robustly grouped the Mesostigmatophyceae with the Chlorokybophyceae (e.g., Figs. 3 and 10). This alone constitutes evidence that suggests the clade is within the streptophytes and not sister to all other viridiplants, because *Chlorokybus atmophyticus* is with no doubt considered as a streptophyte based on its cellular organization (40). The clade received high support values from practically all analyses. In support of this, a careful analysis of chloroplast structural genomic features of several streptophytes performed by Lemieux and colleagues corroborated this affiliation (40). In my own analyses, gene order, as well as gene content consensus cladograms associate these two taxa (Figs. 12 and 13). Furthermore, the Dayhoff trees (Fig. 8) and those derived from the slow-fast analyses unambiguously recovered this clade (Tab. 1).

Some previous studies have placed *Mesostigma* basally followed by *Chlorokybus* in the divergence order of Streptophyta (83). However, these usually lacked sufficient gene sampling. An exception to this deficiency is the study of Finet *et al*. 2011 that found the same topology starting from a dataset of 77 nuclear genes (84). In contrast, whole-chloroplast phylogenomic investigations (40) as well as a recent analysis that used 160 proteins and 14 taxa (85) are in clear agreement with the results here presented. At present, evidence tends to favour a grouping of *Mesostigma* and *Chlorokybus* in a single clade, however in the face of some conflicting data further investigation is recommendable.

In the present study, the nucleotide datasets, even including the third codon position, grouped the *Mesostigma* + *Chlorokybus* clade always as sister to other streptophytes (e.g., Figs. 9 and 10), whereas the amino acid dataset was ambiguous regarding its position, sometimes coinciding with the nucleotide position and others appearing as sister to all green plants (e.g., Figs., 3-7). The basal placement of the clade in question could have been an artefact of long-branch attraction to the red algae outgroup. A good way to test this is to perform the same phylogenetic analyses with the exclusion of the outgroup and see how this affects the Viridiplantae tree topology (77,78). Because of time limitations, it was not feasible during this undergraduate project.

Other analyses were performed to overcome the possible artifacts encountered. In clear agreement with the current consensus, when the two larger amino acid datasets (42x79-aa and 42x94-aa) were recoded into the 6 Dayhoff classes and submitted to phylogenetic analyses using PhyloBayes, a Streptophyta group including the *Mesostigma* + *Chlorokybus* clade as its earliest branch is recovered unambiguously with a PP of 0.96 (Fig. 8). This is also the case in the supertree analysis where a Streptophyta is found with a relatively good bootstrap support, which suggests that there is a common phylogenetic signal among most plastid genes for the whole clade and the *Mesostigma* + *Chlorokybus* clade (Fig. 11). In the slow-fast analysis, support for the complete Streptophyta is found in most (MrBayes) or all (RAxML) analyses with an increasing support when fast-evolving sites are removed (Tab. 1). In summary, the evidence suggests that Mesostigmatophyceae is sister to Chlorokybophyceae and that they are part of the Streptophyta.

Increased taxon sampling among streptophytes, and the addition of klebsormidiophcyean sequences will very probably help to unambiguously recover a complete streptophyte clade in the future, by attracting the *Mesostigma* + *Chlorokybus* clade to the rest of streptophytes. However, the present evidence and that thrown by this study make us confident that the relationships discussed are indeed the true organismal relationships.

THE SISTER LINEAGE OF EMBRYOPHITIC LAND PLANTS

The nature of the sister group to the more than 500,000 species of embryophitic land plants has been a topic of intense debate for long time. Several lineages of green algae have been proposed to occupy this place during the history of the field, however, with the advent of ultrastructural data it became clear that charophycean algae were the most closely related of all green algae to the land plants. Currently, three streptophyte classes are regarded as the probable sister of embryophytes. They are the Charophyceae, the Coleochaetophyceae, and the Zygnematophyceae.

The genera *Chara* and *Nitella* belong to the Charophyceae class, and because these informally called stoneworts exhibit big body sizes and complex branched filamentous morphologies they have traditionally been allied to the embryophytes. This group also display other important characters that are considered homologous to those of land plants. The study of Karol *et al*. (2001) used combined sequences of four genes from the nucleus (18S rRNA genes), chloroplast (*atpB* and *rbcL*), and mitochondria (*nad5*) of 25 charophycean algae, eight land plants, and five chlorophytes, and found a sister relationship between the stoneworts and the land plants; however, moderate bootstrap support was observed for the positions of the other charophycean groups (83,86).

More recent analyses based on chloroplast genomes or multiple nuclear genes tend to favour a closer relationship between embryophytes and coleochaetophyceans or zygnematophyceans (40,85,87,88). While coleochaeophyceans have complex body plans (thalli) as branched filaments with oogamus reproduction, matrotrophy, zygotic esporopollenin and lignin precursors, among other characteristics, zygnematophyceans are simpler morphologically being composed of unicells and non-branched filaments with conjugating sexual reproduction. Plastid multigene trees have generally favour the close affiliation of zygnematophyceans to embryophytes (40,87,89). In 2010, Finet and colleagues (84) argued to have offered the first multigene (77 nuclear genes; 12,149 amino acid positions) phylogenetic evidence that Coleocahetophyceae represent the closest living relatives of land plants. Posterior phylogenomic investigations using larger and non-overlapping sets of 129 (88) and 160 nuclear-encoded proteins (85) suggest that Zygnematophcyeae is indeed the living sister-group of Embryophyta. All these studies agree in that the charophyceans *sensu stricto* are not the sisters of embryophytes.

The present results are compatible with the latest studies previously mentioned in that Charophyceae is excluded from a clade formed by embryophyes, coleochaetophcyeans and zygnematophyceas (e.g., Figs. 3-10). Nonetheless, results are ambiguous regarding the exact nature of the embryophytes' sister. The coleochatophycean *Chaetospharidium* is preferentially placed as sister to land plants in most analyses, especially in those using large, either amino acid or nucleotide, datasets and Bayesian inference (e.g., Figs. 3 and 10). The Dayhoff tree also favours this pattern very confidently (Fig. 8). The same occurs in the supertree, although the bootstrap value received for the association is only of 48% (Fig. 11). The slow-fast analysis (both MrBayes and RAxML), however, showed that Zygnematophyceae is sister to land plants in fuller datasets, and that *Chaetospharidium* is favoured in the last three datasets with the smallest number of fast-evolving sites (Tab. 1). This could be because of a loss of informative sites that reveals an artefactual relationship, or the loss of noisy signal that is artefactually grouping *Zygnema* and *Staurastrum* with land plants, revealing thus the true relationship.

The question of the sister group of embryophytes remains open. Deciphering the true organismal relationships among streptophyte algae and uncovering the closest living relative of land plants will provide to be useful to understand character evolution in the group and the colonization of terrestrial ecosystems by the ancestral green alga that gave rise to embryo-bearing plants (90–94). Increased gene and taxon sampling could eventually help answer it, but it is also possible that the enormous phylogenetic gaps existing due to extinction will never allows us to do it so correctly.

THE BRANCHING ORDER OF PRASINOPHYTE LINEAGES

Prasinophytes are a paraphyletic assemblage of unicellular green algae that are found as flagellates and non-flagellates (coccoid and palmelloid stages). They exhibit a wide range of variation at the molecular and cellular level attesting their paraphyletic nature. The paraphyletic class Prasinophyceae used to be considered in formal classifications, but with recent progress in the study of several prasinophyte lineages, some of these independent monophyletic groups have been raised to the class rank (23,95).

In the last two decades, at least ten independent monophyletic groups of prasinophytes have been elucidated (80,96–101). Of these, evidence for clades VIII and IX comes from environmental sequences. Clades I to VII have cultivable representatives deposited in culture collections. In order, each clade is assigned to the following taxa: Pyramimonadales (clade I), Mamiellophyceae (clade II), Nephroselmydophyceae (clade III), Chlorodendrophyceae (clade IV), Pycnococcaceae (claed V), Prasinococcales (clade VI), and *Picocystis* (clade VII) (101). Recently, a new deep-branching Chlorophyta clade has been reported whose members have palmelloid thalli, the Palmophyllales (102,103).

Because only some prasinophyte lineages have been sampled at the level of their whole chloroplast genome (40,104–108), phylogenetic relationships among all prasinophytes could not be investigated in the present study. Only representatives of Pyramimonadales, Mamiellophyceae, Nephroselmidophyceae and Pycnococcaceae were available. Members of the remaining prasinophyte clades, for which some individual plastid genes have been sequenced, were used for the analyses involving the small 52x12-aa dataset (109). Considering the prasinophyte taxa employed in this study, a well supported and inclusive clade emerged that grouped the Pyramimonadales with the new Mamiellophyceae class (e.g., Figs. 3 and 10). This is true for all sequence trees inferred here. This grouping is consistent with previous studies based on plastid protein-coding genes (105). On the contrary, trees inferred from rRNAs usually do not group these two major groups (95,101). A common pattern observed for rRNA trees is that they usually do not unite major prasinophyte lineages into clades; instead, they display an unbalanced comb-like tree in which each major prasinophyte lineage diverges after another (96,101). It is possible that the rRNA sequences lack the sufficient phylogenetic information two recover the true historical relationships among prasinophytes lineages, especially at deep nodes. On the other hand, the grouping that emerges from complete cpDNA trees could be an artifact due to poor taxon sampling among prasinophytes.

Conflicting results between nucleotide and amino acid trees present two alternative topologies regarding *Nephroselmis* and *Pycnococcus*. In amino acid topologies, including the Dayhoff tree, *Nephroselmis* is always placed as sister to all chlorophytes (e.g., Fig. 3); in nucleotide topologies, *Nephroselmis* is placed as sister to *Pycnoccocus* in a clade sister to the Pyramimonadales + Mamiellophyceae clade (e.g., Fig. 10). The slow-fast results based on MrBayes trees indicate that there is some phylogenetic signal for the *Nephroselmis* + *Pycnococcus* clade (Tab. S3). In constrast, the slow-fast results based on RAxML trees show that *Nephroselmis* branches at the base of the

Chlorophyta with moderate to low bootstrap support (Tabs. 1 and S4). This indicates that there is conflicting signal in the data, and based on current knowledge it is difficult to say which topology is more probably reflecting the correct relationships. The basal *Nephroselmis* position is congruent with the ancestral features of its plastid genome such as having the largest chlorophyte gene repertoire and conserving more ancestral gene clusters (41). The affiliation of Nephroselmydophyceae with Pycnococcaceae is in agreement with some trees derived from rRNA sequences that put them together, previously in the Pseudocourfieldiales order (101). However, as explained above, rRNA trees tend to be very comb-like and considerably variable at the deepest nodes depending on the taxa and method employed.

The chlorodendrophycean genus *Tetraselmis*, wich is present in the two trees derived from the 42x12-aa and 52x12-aa datasets, was considered in the present study because its position as the closest prasinophyte relative to the core chlorophytes and the feasibility of gathering data from its plastid genome. Unfortuntely, technical inconvenicencies and time limitations did not allow me to generate new sequence data from this green alga. Despite this, prelimnary analyses were carried out to investigate the placement of the genus based on plastid proteins. The results challenge the traditional position of *Tetraselmis* in the green tree based on rRNA genes. The 12-protein dataset suggests that *Tetraselmis* branches within early-diverging ulvophyceans, as sister to the unicell *Oltmansiellopsis viridis* (Figs. 6 and 7). The placement of *Tetraselmis* within the Ulvophyceae is not contradicted by ultrastructural data. The disagreement between protein and rRNA genes presents an interesting case that is worthy of future analyses. It is important to consider that rRNA trees might be biased and that these markers could not have enough phylogenetic information to resolve the branching order of ancient divergences such as those among prasinphyte lineages and the UTC clade. Because the rapid nature of the evolution of some regions of rRNA molecules, it is also possible that their trees are prone to long-branch attraction artefacts. These results also illustrate the importance of sequencing disparate prasinophyte lineages not sampled to date, and to increase the number of taxa included in phylogenetic analyses.

PEDINOPHYCEAE: AN INDEPENDENT CLASS OR A TREBOUXIOPHYCEAN LINEAGE?

The Pedinophyceae constitutes an enigmatic lineage of naked and uniflagellate single-celled green algae that has been considered “primitive” along with the prasinophyceans. In 2009, Turmel and colleagues reported the complete cpDNA of *Pedinomonas minor* and in their analyses they found that the pedinophycean is robustly nested within Chlorellales, as sister to the Chlorellaceae family (106). They also provided genomic structural evidence in support of their multigene protein trees. More recently, in 2011, Birger Marin sequenced complete nuclear and plastid-encoded rRNA operons for 37 taxa of green algae including 6 members of the Pedinophyceae to test the hypothesis of Turmel and colleagues (110). His trees rejected any affiliation between Chlorellales and the Pedinophyceae. He then made a strong case based on sequence, structural genomic, cellular and ultrastructual evidence for the high imprability that Pedinophyceae is somewhere close to a derived lineage such as the Chlorellales. He finally argued based on his re-examination of plastid protein data that the signal among plastid protein-coding genes was very heterogenous, some genes supporting one topology and others supporting the alternative topology. He

concluded that there is a biased phylogenetic signal among different protein partitions, indicating the pusblished *Pedinomonas* + Chlorellales association as likely artificial.

The results presented here exemplified perfectly the same incongruence between nucleotide and amino acid data. Interestingly, nucleotide datasets of plastid protein-coding genes produce the same topology as that observed for nuclear and plastid rRNA operons (e.g., Fig. 10). Amino acid derived topologies agree with those published by Turmel and colleagues (e.g., Fig. 3). Two methods were implemented to try to unravel the true phylogenetic signal from the amino acid data. Recoding the amino acid dataset into the 6 Dayhoff classes did not alter the standard *Pedinomonas* + Chlorellaes topology (Fig. 8). Similarly, the slow-fast analyses recovered the *Pedinomonas* + Chlorellales association (Tab. 1). In the RAxML slow-fast analysis, support for the association decreased when more rapid sites were removed; however, the association could not be disrupted and no topology congruent with those derived from nucleotide data was reached. These results indicate that there is a strong bias in the amino acid dataset that places *Pedinomonas* within the Chlorellales. To see if this is the case, a principal component analysis (PCA) was performed in order to test the hypothesis that an amino acid compositional bias in the plastid-encoded proteins of *Pedinomonas* is causing its presumably wrong position. The PCA for the 42x94-aa dataset put *Pedinomonas* and Chlorellales representatives closely together, thereby confirming the suspicion (Fig. S45).

As Marin (2011) puts it, if the Pedinophyceae is indeed nested within the Chlorellales, it would seriously question the utility of ultrastructural characters for the higher-level classification of green algae. Members of the Pedinophyceae are flagellates in their vegetative stage, which means that they divide being flagellated, in contrast to chlorellaleans and trebouxiophyceans that have lost flagella in their vegetative stages to only have them in their reproductive cells (zoospores). The position of the eyespot and the presence of flagellar roots with system I fibers in pedinophyceans also argue against its placement within trebouxiophcyeans. All characters that are shared between pedinophyceans and trebouxiophyceans are probably plesiomorphic, i.e. characters that have been retained conservatively from their ancestors. Furthermore, other important ultrastructural characters that clearly differentiate the Pedinophyceae from the Chlorellales (or Trebouxiophyceae) are the absence of a phycoplast and the presence of a telophase collapsing mitotic spindle in the former, and the existence of a derived mitotic metacentric spindle in trebouxiphyceans. The secondary loss of the phycoplast and the metacentric spindle, and the gaining of a persistent post-metaphase spindle apparatus seem to be improbable events. In conclusion, an evolutionary scenario for character transformation to explain the chlorellalean origin of pedinophyceans appears very unlikely.

UTC, TUC OR CTU CLADE?

The phycoplast-containing core chlorophytes represent the most morphologically and ecophysiologically diverse of all green algae (23). The Chlorophyceae and Trebouxiophyceae classes have adapted to freshwater and terrestrial environments, whereas the Ulvophyceae class has predominantly diversified as seaweeds in coastal ecosystems (90). Deciphering the branching

order among these three classes will provide important insights into the evolution of body plan complexity from prasinophyte-like ancestors, and the ecological diversification and specialization of green algae in most ecosystems.

The task of resolving the interrelationships among the UTC classes has proven to be difficult. All three possible topologies (UTC, TUC and CTU) have been proposed based on the interpretation of different types of data. In addition to the type of data considered, gene and taxon sampling, as well as the phylogenetic methods used, have also produced ambivalent results in molecular analyses (23,26,74,111,112). Ultrastructural data has been interpreted as providing evidence for a relationship between Chlorophyceae and Trebouxiophyceae (e.g. non-persitent mitotic spindle) (113) or Trebouxiophyceae and Ulvophyceae (e.g. counter-clockwise flagellar apparatus) (114). Molecular data generally support an early diverging Trebouxiophyceae (106,115,116). This topology has been supported by mitochondrial multi-gene analysis, a phylogenetic analysis of eight nuclear and two plastid genes, and structural chloroplast genome data (72). Furthermore, there remains some skepticism regarding the monophyletic nature of the Trebouxiophyceae and the Ulvophyceae.

The results reported here based on complete chloroplast sequences tend to support the notion that Trebouxiophyceae and Chlorophyceae are sister groups (e.g., Figs. 3-5). Even so, confidence for this hypothesis is low because the node uniting these two classes generally received poor supporting values. The present study could not provide convincing evidence in support of any of the three possible hypotheses of interrelationships among the three core chlorophyte classes. This difficulty probably stems from the antiquity of the groups (i.e. Proterozoic origin), and a rapid radiation with short-time intervals that gave rise to them.

THE PROGENITORS OF SECONDARY GREEN PLASTIDS

Numerous evidences suggest that the secondary green plastids of *Bigelowiella natans* and *Euglena gracilis* were acquired independently in two separate symbiogenetic events (117,118). *Euglena* is a euglenophyte, a group that is nested within in euglenids; *Bigelowiella* is a chlorarachniphyte, a group that is nested within cercozoans (7). These groups are distantly related to each other and belong to different supergroups, which imply that a common origin of their green secondary plastids would represent the unparsimonious scenario of several independent plastid losses in their heterotrophic and aplastidic sister lineages.

Phylogenetic studies have shown conclusively that their plastids branch separately within the Chlorophyta side of the Viridiplantae tree. Within this green algae lineage, the plastids of euglenophytes have been confidently placed as sister of the pyramimonadalean plastids (105). In contrast, the placement of chlorarachniophyte plastids among chlorophytes has been more elusive. Some studies suggest an ulvophycean origin, whereas others, more generally, an early core chlorophyte origin (15,109,119).

The phylogenetic analyses performed corroborate the close relationship between euglenophyte secondary plastids and pyramimonadalean primary plastids (e.g., Figs. 3 and 10). Unfortunately,

the same cannot be said for the chlorarachniophyte secondary plastids. Some analyses tended to place *Bigelowiella* as sister to the ulvophycean clade formed by *Oltmasiellopsis* and *Pseudendoclonium* (e.g., Fig. 3), but this was not always the case and support for it was not significant.

An increase in the rate of molecular evolution is associated with the origin of secondary plastids from primary green plastids in these two lineages. This makes of their branches the longest of the trees inferred. Long branches are the result of considerable divergence, usually associated with rapid evolution, which in consequence produce multiple nucleotide substitutions and amino acid replacements that make the taxa affected difficult to place and the analysis more prone to systematic errors. Improving taxon sampling to 'cut' the long branches, and using appropriate models of sequence evolution that account for evolutionary heterogeneities (e.g. mixture models) will be very valuable to resolve the phylogenetic origin of secondary green plastids (e.g. chlorarachniophytes' and *Lepidodinium*'s).

A CONSENSUS VIRIDIPLANTAE TREE

The following diagram represents a subjective consensus cladogram that is derived from the discussion of the patterns found in this study and the current evidence from the literature.

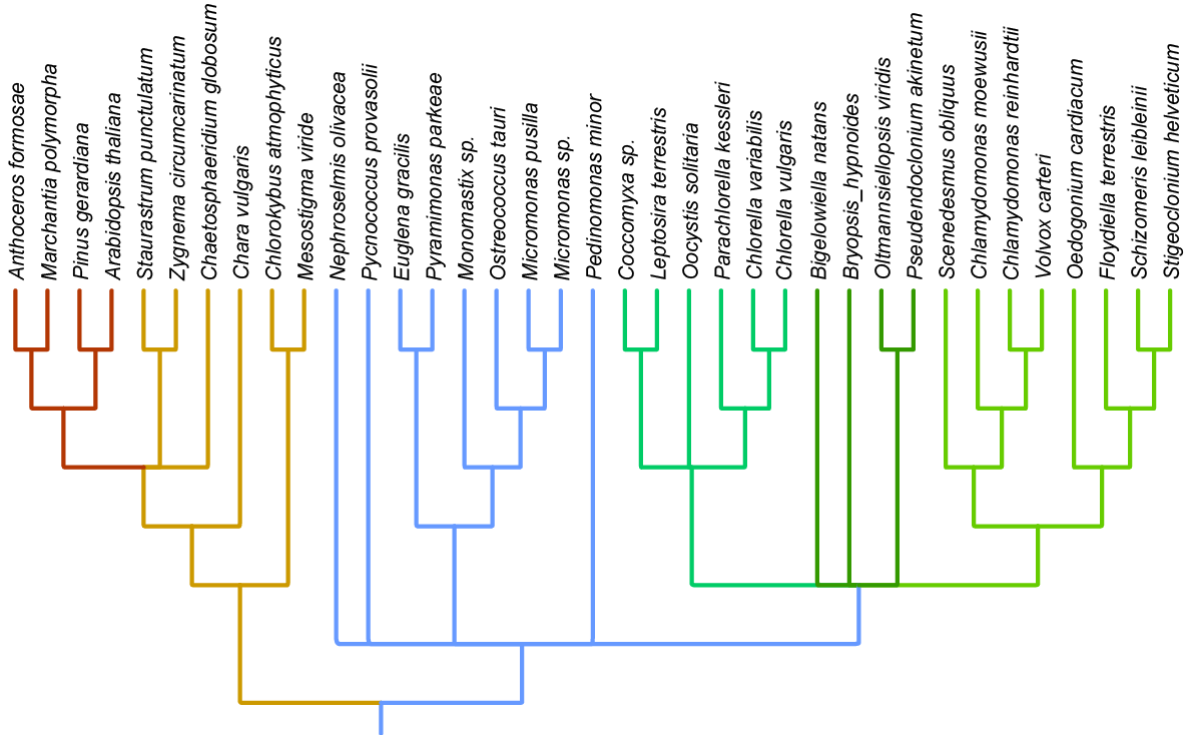


Figure 17. Subjective consensus cladogram derived from the discussion of the patterns found in this study and the current evidence from the literature. It is intended to show uncertainties in relationships as polytomies among the taxa for which whole plastid genomes have been sequenced.

EVOLUTION OF DIVERSE PLASTID GENOMIC FEATURES IN VIRIDIPLANTAE

The genomes of the endosymbiotic organelles named plastids display an astonishing diversity in terms of their size, structure and organization (120,71). In relation to plastid genome size the major factors contributing to its evolution are the increase in non-genic DNA and the loss of genes by permanent loss or endosymbiotic gene transfer (EGT) to the nucleus (117,121). Other factors that contribute to plastid genome size evolution to a lesser degree are gene duplications (e.g. expansion of inverted repeats) followed by some minor cases of lateral gene transfer (HGT) that have been documented (41,118). Proliferation of introns (group I and II) and repetitive elements (122,123), expansion of intergenic spacers and introns, gene rearrangements through inversions and transpositions, and the loss both genic and non-genic DNA through deletions constitute major mechanisms by which the organization of plastid genomes evolve (124). Another important factor that has remodelled plastid genome architecture during the evolutionary history of plastids is the loss of the ancestral quadripartite structure (i.e., an rRNA operon-containing inverted repeat), which has been hypothesized to favour an increase in the rate of genome rearrangements (125).

The green plastid lineage has suffered a less conservative evolutionary pattern than its red and glaucophyte counterparts (23,39). Gene complements of green plastids are in the range of 88-138 genes (not counting leucoplasts) (126). The unique sequenced glaucophyte plastid has a slightly larger number of genes (191 genes), and the red plastids of rhodphytes possess the largest gene repertoire of all eukaryotic plastids (230-259 genes), with the ptDNA of *Porphyra yezoensis* having 259 genes (118). As it can be seen from the distribution of gene contents among the three main lineages of primary plastids, most genes were lost outright or tranferred to the nucleus before the first diversification of Archaeplastida. This is apparent from the fact that when one considers the smallest cyanobacterial genome (1,604 Kbp), that of the prochlorophyte *Prochlorococcus marinus* MIT 9301, it encodes approximately 8 times more genes than the most gene-rich plastid DNA (2,005 vs. 259 genes). The remaining RNA-specifying and protein-coding genes retained by chloroplast genomes are primarily involved in transcription, translation and photosynthesis. A fewer number of genes are dedicated to the processes of protein quality control, membrane biogenesis and metabolism (118).

With respect to genome sizes, viridiplants also exhibit a wide range of variation in their plastid genomes. The mamiellalean *Ostrococcus tauri* has the smallest ptDNA size for a photosynthetic plastid, whereas the chlamydomonadalean *Volvox carteri* has a genome size of ~525 Kbp, and some data indicate that the certain species belonging to the dasycladalean (Ulvophyceae) genus *Acetabularia* have ptDNAs in excess of 2 Mbp (23,71). This is in sharp contrast to the plastid genome size ranges found in other groups of photosynthetic eukaryotes. Similar patterns that illustrate the greater diversity of green ptDNAs are observed for other characteristics like nucleotide composition (%GC content), length of intergenic spacers, and number and length of introns. The latter is especially notorious in green algae and land plants where selfish group I and II introns have proliferated in their plastid genomes in contrast to the situation present in glaucophyte and rhodophyte plastids where there is a paucity of them.

Discussing patterns of plastid genome evolution among the green plants is not an easy task since variation tends to be greater within than between groups. Nevertheless, some patterns can be described based on the complete cpDNAs available and their evolutionary analysis on the Viridiplantae tree.

It can be inferred that the last common ancestor of all green plastid genomes had a gene repertoire of more than 138 genes (40), and that it was gene-rich being depauperate in non-genic DNA (Fig. 14). This ancestral genome was also probably in the genome size range of 137 to 185 Kbp. The primary bifurcation among green plants that led to streptophytes and chlorophytes was not accompanied by major changes of the ancestral genome as can be inferred from the distribution of DNA sizes in early representatives of each group (Fig. 14). However, the evolution of chlorophyte plastid genomes is generally characterized by an increased number of gene losses (Fig. 15).

Streptophytes have followed a more conservative mode of evolution regarding general ptDNA organization relative to their chlorophyte sisters. In general they have retained more genes and their plastid genomes have not experienced major changes in intergenic DNA, but introns have proliferated in them especially since the divergence of Phragmoplastophytina from the *Mesostigma* + *Chlorokybus* clade (Figs. 14-16) (124,127). Group I introns have remained fairly constant, whereas intron expansion has been due to the activity of group II introns (23,118). The charophyte *Chara vulgaris* was more affected by the intron expansion than their relatives in the Phragmoplastophytina (Fig. 14C). Sixteen or more introns are generally found among the cpDNAs of land plants whose cpDNAs have conserved the ancestral quadripartite structure composed of two copies of a large inverted repeat (IR) and two sections of unique DNA, which are referred to as the large and small single copy regions (LSC and SSC, respectively) (120,126). The only two streptophytes genomes that have lost the IR, with the exception of few derived clades within embryophytes, are the zygnemtaophyceans in which this has occurred in association with a larger number of gene rearrangements (Fig. 16). Overall, among streptophyte green algae, the ptDNAs of the charophytes *Mesostigma* and *Chlorokybus* exhibit the most ancestral features (including the largest gene complement among Viridiplantae; 137-138 genes), while the genomes of *Chara* and *Chaetosphaeridium* resemble most their land plants counterparts (Fig. 14-16) (87,128,129).

Among chlorophytes, the paraphyletic “Prasinophyceae” exhibits considerable variation in genome sizes, ranging from the smallest ptDNA described of 71.6 Kbp in size to the larger genome size of *Nephroselmis* (200.8 Kbp) (Fig. 14). They are also very variable in gene complements and intergenic spacer lengths, but not in the number of the introns they carry (Figs. 14A, B and C). *Nephroselmis* has a plastid genome characterized by ancestral features. Its lineage has retained the largest gene complement (128 genes) among prasinophytes and it has not accumulated much non-genic DNA in comparison to other chlorophytes (Figs. 14 and 15). The apparent increase in the genic DNA of its genome is the result of an expanded IR with more duplicated genes (41). Together with the plastid genome of *Pyramimonas parkeae*, *Nephroselmis* and all streptophytes are the only photosynthetic eukaryotes that have retained a set of *ndh* genes (105). *Nephroselmis* ptDNA also conserves more ancestral gene clusters than any other chlorophyte cpDNA (41). These ancestral

features at the cpDNA level appear in conjuction with some ancestral phenotypic traits (e.g. sex, eyespot, five types of scales, flagella) that are congruent with the inferred ancestral green flagellate (AGF) from which green plants probably evolved (23,103). This, in consequence, suggests ceratin degree of stasis in the *Nephroselmis* lineage and tends to favour its early position in amino acid sequence trees.

The Mamiellales (e.g., *Micromonas* spp., *Ostreococcus*) posses the smallest plastid genome sizes of all photosynthetic green plants sequenced so far (107). The bigger clade of which the Mamiellales is part, the Mamiellophyceae plus Pyramimonadales, is also characterized by small genome sizes no bigger that the 114 Kbp of the monomastigalean *Monomastix* sp (Fig. 14). These prasinophyte clades do not exhibit major increases in non-genic DNA; they exhibit a trend towards the loss of genic and non-genic DNA (Figs. 14A and B). Although the ptDNA of *Ostreococcus* is the smallest of all, *Micromonas pusilla* has lost more genes, thereby having less genic DNA in comparison. *Pyramimonas parkeae* also has a relatively small plastid genome (101 Kbp) but it conserves a more ancestral genome organization overall, with a larger gene repertoire (110 genes) than their mamiellalean relatives, and also more ancestral conserved gene clusters (105). Although *Pyramimonas* has a plastid genome that preserves less ancestral features than that of *Nephroselmis*, it shows greater stasis than their closest relatives, coinciding with exceptional ancestral traits exhibited at the morphological level such as a food-uptake apparatus, ancestral flagellar root system and phycoma cysts.

It appears that the common ancestor of *Monomastix* and *Ostreococcus* had already experienced multiple chloroplast gene losses (Figs. 14 and 15), implying that these events might have accompanied the simplification of cell organization that presumably coincided with the emergence of the Mamiellales (105). Moreover, as indicated by the higher frequency of genes losses in the *Ostreococcus* lineage compared with the *Monomastix* lineage, part of the gene losses in the former lineage were likely connected with the evolution of the coccoid cell organization and the reduction in cell size. *Pycnococcus* represents an independent coccoid lineage that sustained considerable reduction of the chloroplast genome, and as observed for *Ostreoccocus*, there was strong pressure to maintain a compact genome organization (107,130,131). A similar pattern is observed for the ptDNA of *Pedinomonas minor*, a tiny naked (no scaly) uniflagallete cell that has been simplified during evolution (Fig. 14), but not to the extreme of the coccoid cell organization displayed by some mamiellaleans.

The evolution of plastid genomes among core chlorophytes have produced a broad range of ptDNAs that differ in several respects (Fig. 14) (39). Fully characterized trebouxiophycean ptDNAs have similar gene contents (106-115 genes) and their differences are mainly because of expanded intergenic regions in the plastid genomes of *Leptosira* and *Coccomyxa* (Fig. 14D). All of them are deficient in introns (Fig. 14C). The ancestral trebouxiophycean ptDNA can be inferred as having low amounts of non-genic DNA and to have diverged considerably at the gene order level from other chlorophyte ptDNAs (Fig. 16). The two available ulvophycean plastid genomes are similar in size and contain several group I introns but no group II introns (23,39). They contain less genes

(104 and 105 genes) that trebouxiophycean plastid genomes but more than those of the chlorophyceans. The ptDNA of *Bryopsis hypnoides* differs significantly from its two ulvophycean relatives in several respects (Figs. 14-16). This is also true in its sequence that is highly divergent and difficult to place in phylogenetic trees. The chlorophyceans show tremendous variation in terms of genome size, intergenic spacers and intron numbers. At the same time, the number of genes encoded in these genomes has been kept remarkably constant (91-99 genes), within the range of derived prasinophyte plastid genomes (e.g. *Pycnococcus*, *Monomastix*). In terms of general genome organization, both types – with or without inverted repeats – are found among chlorophcyean ptDNAs. They usually have a high number of introns, but most of them are group I rather than group II introns (39).

CONCLUSION AND FUTURE DIRECTIONS

The evolution of green plastid genomes has shown to be extremely plastic. Their diversity, resulting from an evolutionary process that started at least since the Ediacaran, is bigger than that exhibited by any other major plastid lineage. Green plastids have revealed to us an astonishing array of patterns that have improved our understanding of how organellar genomes evolve. Furthermore, they have provided pivotal sequence data that have served to uncover previously unknown relationships among green plants and to reconstruct the Viridiplantae phylogenetic tree.

The present work provides a synthesis of the evolutionary relationships among green plants from the perspective of their green plastid genomes. I have emphasized current uncertainties and evaluated different phylogenetic hypotheses based on whole plastid genome sequence and structural data. Finally, I summarized the support for different hypotheses of relationships among green plants and described an evolutionary narrative of the evolution of the organization of green plastid genomes.

Future studies will have to carefully design their phylogenomic analyses. Important steps in the experimental design of phylogenetic analyses include (1) the careful selection of ingroup and outgroup taxa, as well as compatible orthologous gene markers that share a homogenous phylogenetic signal, (2) establishment of positional homology through deliberate multiple alignments and removal of ambiguously aligned regions, (3) the appropriate division of the dataset and assignment of the fittest evolutionary models to each partition, (4) the implementation of diverse phylogenetic methods that employ different algorithms to assess the behaviour of the dataset under different conditions, and (5) the performance of different methods to detect possible artifacts resulting from systematic biases.

There is still a lot to learn about green plastids and the historical relationships among green algae. The sequencing of chloroplast genomes will continue to be a relatively easy, cheap and fast way to gather sequences for phylogenetics. These genomes are small, gene dense and do not commonly suffer of the problem of gene paralogy unlike nuclear genomes. Similarly, the genomic structural data that they provide will continue to be important as an independent source of data to test and

validate sequence derived tree hypotheses. The existence of several hypotheses is the result of conflicting sequence trees due to inherent errors in sequence tree reconstruction procedures. Together with the utilization of most complex models that better describe the complexities of sequence evolution, increasing taxon sampling constitutes the main strategy to overcome phylogenetic artefacts.

Sequencing more plastid genomes will undoubtedly help to determine the branching order among prasinophyte clades, resolve the relationships within the TUC clade, confidently assign a sister group to the secondary plastids of chlorarachniophytes and green dinoflagellates, and to probably decide between the two possible sister lineages of land plants. Plastid genomes that prove to be critical to resolve deep phylogenetic relationships in the Viridiplantae tree are those of the major prasinophytes lineages (e.g. Chlorodendrales, Picocystis clade, Prasinococcales), major ulvophyte clades (e.g. Chalophorales, Trentepohliales, Dasycladales) and phragmoplast-containing charophytes (e.g. Coleochaetiophyceae, Zygnematophyceae and Charophyceae). Targeting disparate lineages to sequence should be our primary focus in order to try to cover most of the wide phylogenetic spectrum present in the green lineage.

AGRADECIMIENTOS

Quiero agradecer primeramente a mis padres, sin los cuales ninguno de mis logros hasta ahora hubiera sido posible. Su apoyo incondicional a lo largo de mi vida, y especialmente durante mi carrera académica ha sido indispensable. A mi hermana Sarita también le doy gracias por su apoyo, al igual que a mi abuelo, quien ha sido paciente y me ha apoyado notoriamente durante mi pregrado. Dentro de las personas que también me han brindado un gran apoyo emocional y académico esta mi novia, Valeria.

De manera especial deseo agradecer a mis dos asesores y maestros, John y Juan, quienes han sido pacientes conmigo y de gran ayuda y acompañamiento durante mi trabajo de grado. Aprecio enormemente que me hayan motivado con respecto a mis intereses particulares de investigación que difieren de los suyos. De la misma forma agradezco a mi maestro Ricardo de quien he tenido la oportunidad de aprender enormemente. También a Juan Saldarriaga por darme su confianza, asesorarme y guiarme, especialmente en esta última etapa del pregrado y en los inicios de lo que espero sea una gratificante carrera en el campo de la biología evolutiva y protistología.

Los profesores Edwin Patiño, Zulma Monsalve, Lucía Atehortúa, Gabriel Bedoya, Carlos Muskus y Juan Fernando Alzate también encuentran un espacio especial en estos agradecimientos, especialmente por prestarme su colaboración durante mi trabajo de grado. A Giovany Olaya, Catalina Lugo, Andrés Lara, Jose Usme y todos aquellos otros que estuvieron ahí, también manifiesto mis agradecimientos por su colaboración y asesoría.

Finalmente, me siento muy agradecido hacia mis compañeros de carrera Camilo Calderón, Laura Toro, Maicol Ospina, Juan Pablo Narváez, Fabián Mejía y Laura González, quienes me han brindado su amistad y han sido de gran compañía durante todo este tiempo en la universidad.

REFERENCES

1. Cavalier-Smith T. Evolution and relationships of algae: major branches of the tree of life. In: Brodie J, Lewis J, editors. Unravelling the algae: the past, present, and future of algal systematics. CRC Press; 2007. p. 402.

2. Graham JE, Wilcox LW, Graham LE. Algae. 2nd ed. Benjamin Cummings; 2008.

3. Martin W, Kowallik KV. Annotated English Translation of Mereschkowsky's 1905 Paper. European Journal of Phycology. 1999;34(03):287–95.

4. Cavalier-Smith T. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 2003 Jan 29;358(1429):109 –134.

5. Blankenship RE. Early Evolution of Photosynthesis. Plant Physiol. 2010 Oct 1;154(2):434–8.

6. Hohmann-Marriott MF, Blankenship RE. Evolution of Photosynthesis. Annual Review of Plant Biology. 2011;62(1):515–48.

7. Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, et al. The tree of eukaryotes. Trends in Ecology & Evolution. 2005 Dec;20(12):670–6.

8. Keeling PJ. Chromalveolates and the evolution of plastids by secondary endosymbiosis. J. Eukaryot. Microbiol. 2009 Feb;56(1):1–8.

9. Gould SB, Waller RF, McFadden GI. Plastid evolution. Annu Rev Plant Biol. 2008;59:491–517.

10. Cavalier-Smith T. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Eukaryot. Microbiol. 1999 Aug;46(4):347–66.

11. Sanchez-Puerta MV, Delwiche CF. A HYPOTHESIS FOR PLASTID EVOLUTION IN CHROMALVEOLATES1. Journal of Phycology. 2008;44(5):1097–107.

12. Bodył A, Stiller JW, Mackiewicz P. Chromalveolate plastids: direct descent or multiple endosymbioses? Trends in Ecology & Evolution. 2009 Mar;24(3):119–21.

13. Baurain D, Brinkmann H, Petersen J, Rodriguez-Ezpeleta N, Stechmann A, Demoulin V, et al. Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. Molecular Biology and Evolution. 2010 Mar 1;27(7):1698–709.

14. Dorrell RG, Smith AG. Do Red and Green Make Brown?: Perspectives on Plastid Acquisitions within Chromalveolates. Eukaryotic Cell. 2011 Jul 1;10(7):856 –868.

15. Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol. Biol. Evol. 2007 Jan;24(1):54–62.

16. Ishida K, Cao Y, Hasegawa M, Okada N, Hara Y. The origin of chlorarachniophyte plastids, as inferred from phylogenetic comparisons of amino acid sequences of EF-Tu. J. Mol. Evol. 1997 Dec;45(6):682–7.

17. Cavalier-Smith T. Eukaryote kingdoms: Seven or nine? Biosystems. 1981;14(3–4):461–81.

18. Cavalier-Smith T. A revised six-kingdom system of life. Biological Reviews. 1998;73(3):203–66.

19. Jeffrey C. Thallophytes and Kingdoms: A Critique. Kew Bulletin. 1971;25(2):291.

20. Jeffrey C. Kingdoms, Codes and Classification. Kew Bulletin. 1982;37(3):403.

21. Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, et al. The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists. Journal of Eukaryotic Microbiology. 2005;52(5):399–451.

22. Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, et al. The Revised Classification of Eukaryotes. Journal of Eukaryotic Microbiology. 2012;59(5):429–514.

23. Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF, et al. Phylogeny and Molecular Evolution of the Green Algae. Critical Reviews in Plant Sciences. 2012;31(1):1–46.

24. Hoek C van den, Mann D, Jahns HM. Algae: An Introduction to Phycology. Cambridge University Press; 1996.

25. Lee RE. Phycology. 4th ed. Cambridge University Press; 2008.

26. Lewis LA, McCourt RM. Green algae and the origin of land plants. American Journal of Botany. 2004 Oct 1;91(10):1535 –1556.

27. Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. Mol Biol Evol. 2004 May 1;21(5):809–18.

28. Berney C, Pawlowski J. A Molecular Time-Scale for Eukaryote Evolution Recalibrated with the Continuous Microfossil Record. Proc. R. Soc. B. 2006 Aug 7;273(1596):1867–72.

29. Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H. The Timing of Eukaryotic Evolution: Does a Relaxed Molecular Clock Reconcile Proteins and Fossils? PNAS. 2004 Oct 26;101(43):15386–91.

30. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the Timing of Early Eukaryotic Diversification with Multigene Molecular Clocks. PNAS. 2011 Aug 16;108(33):13624–9.

31. Huntley JW, Xiao S, Kowalewski M. 1.3 Billion years of acritarch history: An empirical morphospace approach. Precambrian Research. 2006 Jan 20;144(1–2):52–68.

32. Gaucher C, Sprechmann P. Chapter 9.1 Neoproterozoic Acritarch Evolution. Neoproterozoic-Cambrian Tectonics, Global Change And Evolution: A Focus On South Western Gondwana [Internet]. Elsevier; 2009. p. 319–26. Available from: http://www.sciencedirect.com/science/article/pii/S0166263509016223

33. Butterfield NJ. MACROEVOLUTION AND MACROECOLOGY THROUGH DEEP TIME. Palaeontology. 2007 Jan;50(1):41–55.

34. Butterfield NJ. Modes of pre-Ediacaran multicellularity. Precambrian Research. 2009 Sep;173(1-4):201–11.

35. J. O'kelly C. Chapter 13 - The Origin and Early Evolution of Green Plants. In: Paul G. Falkowski, Andrew H. Knoll, editors. Evolution of Primary Producers in the Sea [Internet]. Burlington: Academic Press;

2007 [cited 2013 Jan 25]. p. 287–XII. Available from: http://www.sciencedirect.com/science/article/pii/B978012370518150014X

36. Neilson JAD, Durnford DG. Structural and functional diversification of the light-harvesting complexes in photosynthetic eukaryotes. Photosynthesis Research. 2010 Jul 2;106(1-2):57–71.

37. Koziol AG, Borza T, Ishida KI, Keeling P, Lee RW, Durnford DG. Tracing the evolution of the light-harvesting antennae in chlorophyll a/b-containing organisms. Plant physiology. 2007;143(4):1802–16.

38. Proschold T, Leliaert F. Systematics of the green algae: conflict of classic and modern approaches. In: Brodie J, Lewis J, editors. Unravelling the algae: the past, present, and future of algal systematics. 1st edition. CRC Press; 2007. p. 402.

39. Lang BF, Nedelcu AM. Plastid genomes of algae. In: Bock R, Knoop V, editors. "Genomics of Chloroplasts and Mitochondria". Advances in Photosynthesis and Respiration Series. Springer; p. 59–87.

40. Lemieux C, Otis C, Turmel M. A clade uniting the green algae Mesostigma viride and Chlorokybus atmophyticus represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. BMC Biol. 2007;5:2.

41. Turmel M, Otis C, Lemieux C. The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: Insights into the architecture of ancestral chloroplast genomes. PNAS. 1999 Aug 31;96(18):10248–53.

42. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D16–D20.

43. Edgar R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004 Mar 8;32(5):1792–7.

44. Edgar R. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004 Aug 19;5(1):113.

45. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE), 2010. IEEE; 2010. p. 1–8.

46. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Research. 2010 Apr 30;38(Web Server):W7–W13.

47. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 2007 Aug;56(4):564–77.

48. Vaidya G, Lohman DJ, Meier R. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics. 2011 Apr 1;27(2):171–80.

49. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis [Internet]. 2011. Available from: http://mesquiteproject.org

50. Swofford DL. PAUP*. Sinauer Associates, Sunderland, Massachussetts; 2003.

51. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug;17(8):754–5.

52. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003 Aug 11;19(12):1572–4.

53. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics. 2004 Jan 22;20(3):407–15.

54. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006 Aug 23;22(21):2688–90.

55. Stamatakis A, Hoover P, Rougemont J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. Systematic Biology. 2008 Oct 16;57(5):758–71.

56. Lartillot N. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Molecular Biology and Evolution. 2004 Feb 12;21(6):1095–109.

57. Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evolutionary Biology. 2007;7(Suppl 1):S4.

58. Kumar S, Skjæveland Å, Orr RJ, Enger P, Ruden T, Mevik B-H, et al. AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. BMC Bioinformatics. 2009;10(1):357.

59. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009 Sep 1;25(17):2286–8.

60. Stamatakis A. Phylogenetic models of rate heterogeneity: a high performance computing perspective. IEEE; 2006 [cited 2012 Nov 4]. p. 8 pp. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1639535

61. Kostka M, Uzlikova M, Cepicka I, Flegr J. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics. 2008;9(1):341.

62. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010 May 1;59(3):307 –321.

63. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011 Feb 17;27(8):1164–5.

64. Creevey CJ, McInerney JO. Clann: Investigating Phylogenetic Information Through Supertree Analyses. Bioinformatics. 2005 Feb 1;21(3):390–2.

65. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U.S.A. 2002 Sep 17;99(19):12246–51.

66. Nozaki H, Ohta N, Matsuzaki M, Misumi O, Kuroiwa T. Phylogeny of Plastids Based on Cladistic Analysis of Gene Loss Inferred from Complete Plastid Genome Sequences. Journal of Molecular Evolution. 2003 Oct 1;57(4):377–82.

67. Lang BF, Laforest M-J, Burger G. Mitochondrial introns: a critical view. Trends in Genetics. 2007 Mar 1;23(3):119–25.

68. Felsenstein J. PHYLIP (Phylogeny Inference Package) [Internet]. Department of Genome Sciences, University of Washington, Seattle.: Distributed by the author; 2005. Available from: http://evolution.genetics.washington.edu/phylip.html

69. Pavesi G, Mauri G, Iannelli F, Gissi C, Pesole G. GeneSyn: a tool for detecting conserved gene order across genomes. Bioinformatics. 2004 Jun 12;20(9):1472–4.

70. Maddison DR, Maddison WP. MacClade 4: Analysis of phylogeny and character evolution. Sunderland, Massachusetts: Sinauer Associates; 2011.

71. Palmer JD. Comparative Organization of Chloroplast Genomes. Annual Review of Genetics. 1985;19(1):325–54.

72. Cocquyt E, Verbruggen H, Leliaert F, De Clerck O. Evolution and Cytological Diversification of the Green Seaweeds (Ulvophyceae). Molecular Biology and Evolution. 2010;27(9):2052 –2061.

73. Woolcott GW, Knöller K, King RJ. Phylogeny of the Bryopsidaceae (Bryopsidales, Chlorophyta): cladistic analyses of morphological and molecular data. Phycologia. 2000 Nov;39(6):471–81.

74. Zechman FW, Theriot EC, Zimmer EA, Chapman RL. PHYLOGENY OF THE ULVOPHYCEAE (CHLOROPHYTA): CLADISTIC ANALYSIS OF NUCLEAR-ENCODED rRNA SEQUENCE DATA1. Journal of Phycology. 1990;26(4):700–10.

75. Zuccarello GC, Price N, Verbruggen H, Leliaert F. Analysis of a Plastid Multigene Data Set and the Phylogenetic Position of the Marine Macroalga Caulerpa Filiformis (chlorophyta)1. Journal of Phycology. 2009;45(5):1206–12.

76. Lü F, Xü W, Tian C, Wang G, Niu J, Pan G, et al. The Bryopsis hypnoides Plastid Genome: Multimeric Forms and Complete Nucleotide Sequence. PLoS ONE. 2011 Feb 14;6(2):e14663.

77. Bergsten J. A review of long-branch attraction. Cladistics. 2005;21(2):163–93.

78. Verbruggen H, Theriot EC. Building trees of algae: some advances in phylogenetic and evolutionary analysis. European Journal of Phycology. 2008;43(3):229–52.

79. Rodríguez-Ezpeleta N, Philippe H, Brinkmann H, Becker B, Melkonian M. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. Mol. Biol. Evol. 2007 Mar;24(3):723–31.

80. Marin B, Melkonian M. Mesostigmatophyceae, a New Class of Streptophyte Green Algae Revealed by SSU rRNA Sequence Comparisons. Protist. 1999 Dec;150(4):399–417.

81. Nedelcu AM, Borza T, Lee RW. A Land Plant–Specific Multigene Family in the Unicellular Mesostigma Argues for Its Close Relationship to Streptophyta. Mol Biol Evol. 2006 May 1;23(5):1011–5.

82. Grauvogel C, Petersen J. Isoprenoid biosynthesis authenticates the classification of the green alga Mesostigma viride as an ancient streptophyte. Gene. 2007 Jul;396(1):125–33.

83. Karol KG, McCourt RM, Cimino MT, Delwiche CF. The closest living relatives of land plants. Science. 2001 Dec 14;294(5550):2351–3.

84. Finet C, Timme RE, Delwiche CF, Marlétaz F. Multigene Phylogeny of the Green Lineage Reveals the Origin and Diversification of Land Plants. Current Biology. 2010 Dec;20(24):2217–22.

85. Timme RE, Bachvaroff TR, Delwiche CF. Broad Phylogenomic Sampling and the Sister Lineage of Land Plants. Joly S, editor. PLoS ONE. 2012 Jan 13;7(1):e29696.

86. McCourt RM, Delwiche CF, Karol KG. Charophyte algae and land plant origins. Trends in Ecology & Evolution. 2004 Dec 1;19(12):661–6.

87. Turmel M, Otis C, Lemieux C. The chloroplast genome sequence of Chara vulgaris sheds new light into the closest green algal relatives of land plants. Mol. Biol. Evol. 2006 Jun;23(6):1324–38.

88. Wodniok S, Brinkmann H, Glöckner G, Heidel AJ, Philippe H, Melkonian M, et al. Origin of land plants: Do conjugating green algae hold the key? BMC Evolutionary Biology. 2011;11(1):104.

89. Turmel M, Ehara M, Otis C, Lemieux C. PHYLOGENETIC RELATIONSHIPS AMONG STREPTOPHYTES AS INFERRED FROM CHLOROPLAST SMALL AND LARGE SUBUNIT rRNA GENE SEQUENCES1. Journal of Phycology. 2002;38(2):364–75.

90. Becker B, Marin B. Streptophyte algae and the origin of embryophytes. Ann Bot. 2009 May;103(7):999–1004.

91. Graham LE, Cook ME, Busse JS. The origin of plants: Body plan changes contributing to a major evolutionary radiation. Proceedings of the National Academy of Sciences. 2000 Apr 25;97(9):4535–40.

92. Graham LE. Origin of Land Plants. 1st ed. Wiley; 1993.

93. Delwiche CF, Graham LE, Thomson N. Lignin-Like Compounds and Sporopollenin Coleochaete, an Algal Model for Land Plant Ancestry. Science. 1989 Jul 28;245(4916):399–401.

94. Graham LE. Coleochaete and the Origin of Land Plants. American Journal of Botany. 1984 Apr;71(4):603.

95. Marin B, Melkonian M. Molecular Phylogeny and Classification of the Mamiellophyceae class. nov. (Chlorophyta) based on Sequence Comparisons of the Nuclear- and Plastid-encoded rRNA Operons. Protist. 2010 Apr;161(2):304–36.

96. Steinkotter J, Bhattacharya D, Semmelroth I, Bibeau C, Melkonian M. PRASINOPHYTES FORM INDEPENDENT LINEAGES WITHIN THE CHLOROPHYTA: EVIDENCE FROM RIBOSOMAL RNA SEQUENCE COMPARISONS1. Journal of Phycology. 1994 Apr;30(2):340–5.

97. Nakayama T, Marin B, Kranz HD, Surek B, Huss VAR, Inouye I, et al. The Basal Position of Scaly Green Flagellates among the Green Algae (Chlorophyta) is Revealed by Analyses of Nuclear-Encoded SSU rRNA Sequences. Protist. 1998 Dec;149(4):367–80.

98. Nakayama T, Watanabe S, Inouye I. Phylogeny of wall-less green flagellates inferred from 18SrDNA sequence data. Phycological Research. 1996;44(3):151–61.

99. Daugbjerg N, Moestrup Ø, Arctander P. Phylogeny of genera of Prasinophyceae and Pedinophyceae (Chlorophyta) deduced from molecular analysis of the rbcL gene. Phycological Research. 1995;43(4):203–13.

100. Fawley MW, Yun Y, Qin M. Phylogenetic analyses of 18s rdna sequences reveal a new coccoid lineage of the prasinophyceae (Chlorophyta). Journal of Phycology. 2001 Dec 25;36(2):387–93.

101. Guillou L, Eikrem W, Chrétiennot-Dinet M-J, Le Gall F, Massana R, Romari K, et al. Diversity of Picoplanktonic Prasinophytes Assessed by Direct Nuclear SSU rDNA Sequencing of Environmental Samples and Novel Isolates Retrieved from Oceanic and Coastal Marine Ecosystems. Protist. 2004 Jun;155(2):193–214.

102. Zechman FW, Verbruggen H, Leliaert F, Ashworth M, Buchheim MA, Fawley MW, et al. An Unrecognized Ancient Lineage of Green Plants Persists in Deep Marine Waters1. Journal of Phycology. 2010;46(6):1288–95.

103. Leliaert F, Verbruggen H, Zechman FW. Into the deep: New discoveries at the base of the green plant phylogeny. BioEssays. 2011;33(9):683–92.

104. Lemieux C, Otis C, Turmel M. Ancestral chloroplast genome in Mesostigma viride reveals an early branch of green plant evolution. Nature. 2000 Feb 10;403(6770):649–52.

105. Turmel M, Gagnon M-C, O'Kelly CJ, Otis C, Lemieux C. The chloroplast genomes of the green algae Pyramimonas, Monomastix, and Pycnococcus shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol. Biol. Evol. 2009 Mar;26(3):631–48.

106. Turmel M, Otis C, Lemieux C. The chloroplast genomes of the green algae Pedinomonas minor, Parachlorella kessleri, and Oocystis solitaria reveal a shared ancestry between the Pedinomonadales and Chlorellales. Mol. Biol. Evol. 2009 Oct;26(10):2317–31.

107. Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Peer YV de. The Complete Chloroplast and Mitochondrial DNA Sequence of Ostreococcus tauri: Organelle Genomes of the Smallest Eukaryote Are Examples of Compaction. Mol Biol Evol. 2007 Apr 1;24(4):956–68.

108. Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. Science. 2009 Apr 10;324(5924):268–72.

109. Matsumoto T, Shinozaki F, Chikuni T, Yabuki A, Takishita K, Kawachi M, et al. Green-colored plastids in the dinoflagellate genus Lepidodinium are of core chlorophyte origin. Protist. 2011 Apr;162(2):268–76.

110. Marin B. 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. 2012 Sep;163(5):778–805.

111. Watanabe S, Nakayama T. Ultrastructure and phylogenetic relationships of the unicellular green algae Ignatius tetrasporus and Pseudocharacium americanum (Chlorophyta). Phycological Research. 2007;55(1):1–16.

112. López-Bautista JM, Chapman RL. Phylogenetic affinities of the Trentepohliales inferred from small-subunit rDNA. Int J Syst Evol Microbiol. 2003 Nov 1;53(6):2099–106.

113. Mattox KR, Stewart KD. Classification of the green algae: a concept based on comparative cytology. Systematics of the green algae. 1984;29–72.

114. Sluiman HJ. The green algal class Ulvophyceae. An ultrastructural survey and classification. Crypt. Bot [Internet]. 1989 [cited 2013 Jan 31]; Available from: http://europepmc.org/abstract/AGR/IND91012197/reload=0

115. Turmel M, Brouard J, Gagnon C, Otis C, Lemieux C. Deep Division in the Chlorophyceae (chlorophyta) Revealed by Chloroplast Phylogenomic Analyses1. Journal of Phycology. 2008 Jun 1;44(3):739–50.

116. Brouard J-S, Otis C, Lemieux C, Turmel M. Chloroplast DNA sequence of the green alga Oedogonium cardiacum (Chlorophyceae): unique genome architecture, derived characters shared with the Chaetophorales and novel genes acquired through horizontal transfer. BMC Genomics. 2008;9:290.

117. Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010 Mar 12;365(1541):729 –748.

118. Kim E, Archibald JM. Diversity and Evolution of Plastids and Their Genomes. In: Sandelius AS, Aronsson H, editors. The Chloroplast [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009 [cited 2012 Feb 12]. p. 1–39. Available from: http://www.springerlink.com/content/v45925432166hn72/

119. Takahashi F, Okabe Y, Nakada T, Sekimoto H, Ito M, Kataoka H, et al. Origins of the secondary plastids of Euglenophyta and Chlorarachniophyta as revealed by an analysis of the plastid-targeting, nuclear-encoded gene psbO1. Journal of Phycology. 2007;43(6):1302–9.

120. Simpson CL, Stern DB. The Treasure Trove of Algal Chloroplast Genomes. Surprises in Architecture and Gene Content, and Their Functional Implications. Plant Physiol. 2002 Jul 1;129(3):957–66.

121. Copertino DW, Hallick RB. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. Trends Biochem. Sci. 1993 Dec;18(12):467–71.

122. Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, et al. Complete sequence of Euglena gracilis chloroplast DNA. Nucleic Acids Research. 1993 Jul 25;21(15):3537 –3544.

123. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, et al. The Chlamydomonas reinhardtii Plastid Chromosome Islands of Genes in a Sea of Repeats. Plant Cell. 2002 Nov 1;14(11):2659–79.

124. Jansen RK, Raubeson LA, Boore JL, Depamphilis CW, Chumley TW, Haberle RC, et al. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods in enzymology. 2005;395:348–84.

125. Pombert J-F, Otis C, Lemieux C, Turmel M. 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. 2005 Sep;22(9):1903–18.

126. Green BR. Chloroplast genomes of photosynthetic eukaryotes. The Plant Journal. 2011;66(1):34–44.

127. Qiu Y-L, Palmer JD. Phylogeny of early land plants: insights from genes and genomes. Trends in Plant Science. 1999 Jan;4(1):26–30.

128. Turmel M, Otis C, Lemieux C. The chloroplast and mitochondrial genome sequences of the charophyte Chaetosphaeridium globosum: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc. Natl. Acad. Sci. U.S.A. 2002 Aug 20;99(17):11275–80.

129. Turmel M, Otis C, Lemieux C. The complete chloroplast DNA sequences of the charophycean green algae Staurastrum and Zygnema reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biol. 2005;3:22.

130. Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S, et al. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proc. Natl. Acad. Sci. U.S.A. 2006 Aug 1;103(31):11647–52.

131. Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N, et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. Proc. Natl. Acad. Sci. U.S.A. 2007 May 1;104(18):7705–10.