genes

MDPI

Article

Gene Loss and Evolution of the Plastome

Tapan Kumar Mohanta [1,†,*], Awdhesh Kumar Mishra [2,†], Adil Khan [1], Abeer Hashem [3,4], Elsayed Fathi Abd_Allah [5] and Ahmed Al-Harrasi [6,*]

1 Biotech and Omics Laboratory, Natural and Medical Sciences Research Centre, University of Nizwa, Nizwa, 616, Oman; adilsafi122333@gmail.com
2 Department of Biotechnology, Yeungnam University, Gyeongsan, 38541, Republic of Korea; awadhesh.biotech07@gmail.com
3 Botany and Microbiology Department, College of Science, King Saud University, Riyadh, 11451, Saudi Arabia; habeer@ksu.edu.sa
4 Mycology and Plant Disease Survey Department, Plant Pathology Research Institute, 12511, Giza, Egypt
5 Plant Production Department, College of Food and Agricultural Sciences, King Saud University, P.O. Box. 2460 Riyadh 11451, Saudi Arabia; eabdallah@ksu.edu.sa
6 Natural Product Laboratory, Natural and Medical Sciences Research Centre, University of Nizwa, Nizwa, 616, Oman
† These authors contributed equally to this work.
* Correspondence: tapan.mohanta@unizwa.edu.om (T.K.M); aharrasi@unizwa.edu.om (A.A.-H.)

Received: 16 July 2020; Accepted: 14 September 2020; Published: date

Abstract: Chloroplasts are unique organelles within the plant cells and are responsible for sustaining life forms on the earth due to their ability to conduct photosynthesis. Multiple functional genes within the chloroplast are responsible for a variety of metabolic processes that occur in the chloroplast. Considering its fundamental role in sustaining life on the earth, it is important to identify the level of diversity present in the chloroplast genome, what genes and genomic content have been lost, what genes have been transferred to the nuclear genome, duplication events, and the overall origin and evolution of the chloroplast genome. Our analysis of 2511 chloroplast genomes indicated that the genome size and number of coding DNA sequences (CDS) in the chloroplasts genome of algae are higher relative to other lineages. Approximately 10.31% of the examined species have lost the inverted repeats (IR) in the chloroplast genome that span across all the lineages. Genome-wide analyses revealed the loss of the *Rbcl* gene in parasitic and heterotrophic plants occurred approximately 56 Ma ago. *PsaM, Psb30, ChlB, ChlL, ChlN,* and *Rpl21* were found to be characteristic signature genes of the chloroplast genome of algae, bryophytes, pteridophytes, and gymnosperms; however, none of these genes were found in the angiosperm or magnoliid lineage which appeared to have lost them approximately 203–156 Ma ago. A variety of chloroplast-encoded genes were lost across different species lineages throughout the evolutionary process. The *Rpl20* gene, however, was found to be the most stable and intact gene in the chloroplast genome and was not lost in any of the analyzed species, suggesting that it is a signature gene of the plastome. Our evolutionary analysis indicated that chloroplast genomes evolved from multiple common ancestors ~1293 Ma ago and have undergone vivid recombination events across different taxonomic lineages.

Keywords: chloroplast genome; plastome; evolution; deletion; duplication; recombination; nucleotide substitution

1. Introduction

Photosynthesis is a process by which autotrophic plants utilize chlorophyll to transform solar energy into chemical energy [1]. Almost all life forms depend directly or indirectly on this chemical

energy as a source of energy to sustain growth, development, and reproduction of their species [2,3]. This essential process occurs inside a semiautonomous organelle, commonly known as a plastid or chloroplast [4]. Current knowledge indicates that the origin and evolution of plastids occurred through the endosymbiosis of ancestral cyanobacteria with nonphotosynthesizing cells that dates back to 1.5 to 1.6 billion years ago [5,6]. The subsequent divergence of a green plastid lineage occurred prior to 1.2 billion years ago and led to the development of land plants approximately 432 to 476 million years ago, and to seed plants around 355 to 370 million years ago [6]. A subsequent split into gymnosperms and angiosperms occurred approximately 290 to 320 million years ago and the divergence of monocots and eudicots within the angiosperm lineage occurred approximately 90 to 130 million years ago [6]. Throughout this evolutionary time scale, the endosymbiont retained its existence inside the cell and its dominant function of photosynthesis without undergoing any basic evolutionary changes (photosynthesis) [7–10]. In addition to photosynthesis, this semiautonomous organelle also plays an important role in the biosynthesis of amino acids, lipids, carotenoids, and other important biomolecules [11–15]. Studies indicate that the plastid genome has retained a complete set of protein-synthesizing machinery and encodes approximately 100 proteins [16]. All other proteins required by the chloroplast, however, are encoded by the nuclear genome. All of the protein synthesis and photosynthetic machinery used by the plastid is encoded by its own genome, commonly referred to as the plastome, that is arranged in a quadripartite structure [17–20]. The size of the plastid genome of land plants is reported to range from 120 to 190 kb [21–23]. The quadripartite structure consists of four main segments, referred to as the small single-copy region (SSC), large single-copy region (LSC), and the inverted repeat A and B (IR_A and IR_B) regions [24]. The size of the IR region ranges from 10 to 15 kb in nonseed plants to 20–30 kb in angiosperms [24–27]. The IR_A and IR_B regions are reported to share a conserved molecular evolutionary pattern [28,29]. Studies also indicate that the genes in the plastome genome are organized in an operon or operon-like structure that undergoes transcription, producing polycistronic precursors [30]. The majority of genes in the chloroplast genome have been either functionally transferred to the nuclear genome or lost during evolution [31,32]. For example, the functional genes *tufA*, *ftsH*, *odpB*, and *Rpl5* have been transferred from the plastome to the nucleus [33,34]. Structural rearrangements of the plastid genome have occurred throughout its evolution; resulting in expansion, contraction, or loss of genetic content [23]. These events have occurred multiple times during the evolution of the chloroplast and can be specific to a single species, or sometimes to a whole plant order [25,35–38]. Changes in the architecture of the IR regions can affect the entire plastid chromosome and its immediate neighborhood. For example, several genes associated with the SSC region got duplicated, including *Ycf2*, due to the relocation of the IR region [23]. Although several analyses of the plastid genome have been conducted, a comprehensive comparative study of the plastid genome at a large-scale has not yet been reported. Comparative studies have thus far only included a few species of an order or a few species from a few different groups. Therefore, a large-scale analysis of 2511 chloroplast genomes was conducted to better understand the genomics and evolution of the plastid genome. Details of the novel genomic features of the chloroplast genome are reported in the present study.

2. Materials and Methods

2.1. Sequence Retrieval and Annotation

All of the sequenced chloroplast genomes available up until December 2018 were downloaded from the National Center for Biotechnology Information (NCBI) and used in the current study to analyze the genomic details of the chloroplast genome. In total, 2511 full-length complete chloroplast genome sequences were downloaded, including those from algae, bryophytes, pteridophytes, gymnosperms, monocots, dicots, magnoliids, and protist/protozoa (Supplementary File S1). All of the individual genomes were subjected to OGDRAW to check for the presence and absence of inverted repeats in the genome [39]. Genomes that were found to lack inverted repeats (IR), as determined by OGDRAW, were further searched in the NCBI database to cross verify the absence of IR in their genome. The annotated coding DNA sequences (CDS) sequences in each chloroplast

genome were downloaded and the presence or absence of CDS from all chloroplast genomes were searched in each individual genome using Linux programming. Species that were identified as lacking a gene in their chloroplast genome were noted and further rechecked manually in the NCBI database. Each chloroplast genome was newly annotated using the GeSeq-annotation of the organellar genomes pipeline to further extend the study of gene loss in chloroplast genomes [40]. The combined analysis of NCBI and GeSeq-annotation of the organellar genomes were considered in determining the absence of a particular gene in a chloroplast genome.

The CDS of the nuclear genome of 145 plant species were downloaded from the NCBI database. The presence of chloroplast-encoded genes in the nuclear genome was determined using Linux-based commands and collected in a separate file. The chloroplast-encoded genes present in the nuclear genomes were further processed in a Microsoft Excel spreadsheet.

2.2. Multiple Sequence Alignment and Creation of Phylogenetic Trees

Prior to the multiple sequence alignment, the CDS sequences of *PsaM, psb30, ChlB, ChlL, ChlN*, and *RPL21* were converted to amino acid sequences using a sequence manipulation suite (http://www.bioinformatics.org/sms2/translate.html) [41]. The resulting protein sequences were subjected to multiple sequence alignment using the Multalin server to identify conserved amino acid motifs [42]. The CDS sequences of *PsaM, psb30, ChlB, ChlL, ChlN*, and *RPL21* genes were also subjected to multiple sequence alignment using Clustal Omega. The resultant aligned file was downloaded in Clustal format and converted to a MEGA file format using MEGA6 software [43]. The converted MEGA files of *PsaM, psb30, ChlB, ChlL, ChlN*, and *RPL21* were subsequently used for the construction of a phylogenetic tree. Prior to the construction of the phylogenetic tree, a model selection was carried out using MEGA6 software using the following parameters; analysis, model selection; tree to use, automatic (neighbor-joining tree); statistical method, maximum likelihood; substitution type, nucleotide; gaps/missing data treatment, partial deletion; site coverage cut-off (%), 95; branch swap filer, very strong; and codons included, 1st, 2nd, and 3rd. Based on the lowest BIC (Bayesian information criterion) score, the following statistical parameters were used to construct the phylogenetic tree: statistical method, maximum likelihood; test of phylogeny, bootstrap method; number of bootstrap replications, 1000; model/method, general time-reversible model; rates among sites, gamma-distributed with invariant sites (G+I); number. of discrete gamma categories, 5; gaps/missing data treatment, partial deletion; site coverage cut-off (%), 95; ML Heuristic method, nearest-neighbor-interchange (NNI); branch swap filer, very strong; and codons included, 1st, 2nd, and 3rd. The resulting phylogenetic trees were saved as gene trees. Whole-genome sequences of chloroplast genomes were also collectively used to construct a phylogenetic tree to gain insight into the evolution of chloroplast genomes. ClustalW program was used in a Linux-based platform to construct the phylogenetic tree of chloroplast genomes using the neighbor-joining method and 500 bootstrap replicates. The resultant Newick file was uploaded in Archaeopteryx (https://sites.google.com/site/cmzmasek/home/software/archaeopteryx) to view the phylogenetic tree. A separate phylogenetic tree of species with IR-deleted regions was also constructed using the whole sequence of the IR-deleted chloroplast genome using similar parameters as described above. The evolutionary time of plant species used in this study was created using the TimeTree [44]. Cyanobacterial species were used as an outgroup to calibrate the time tree for the other species.

2.3. Analysis of the Deletion and Duplication of Chloroplast-Encoded Genes

A species tree was constructed using the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) prior to the study of deletion and duplication of *PsaM, psb30, ChlB, ChlL, ChlN*, and *RPL21* genes. The gene tree of the individual gene family was uploaded in Notung software v.2.9 followed by uploading the species tree and subsequent reconciliation of the gene tree with the species tree [45–47]. Once reconciled, deletion and duplication events for the genes were visualized and noted.

2.4. Recombination Events and Time Tree Construction of the Chloroplast Genome

The constructed phylogenetic tree of chloroplast genomes was uploaded in IcyTree [48] to analyze the recombination events that occurred in chloroplast genomes. The recombination events in IR-deleted and nondeleted IR species were studied separately. The time tree of the studied tree was constructed using the TimeTree program [44].

2.5. Substitution Rate in Chloroplast Genomes

Chloroplast genomes were grouped into different groups to determine lineage-specific nucleotide substitution rates. The groups were algae, bryophytes, gymnosperms, eudicots, monocots, magnoliids, Nymphaeales, protists, and IR-deleted species. At least 10 chloroplast genomes were included for each lineage when analyzing the rate of nucleotide substitutions. The full-length sequences of chloroplast genomes were subjected to multiple sequence alignment to generate a Clustal file. The MAFT-multiple alignment pipeline was implemented to align the sequences of the different chloroplast genomes. The aligned sequences of individual lineages were downloaded and converted to a MEGA file format using MEGA6 software [43]. The converted files were subsequently uploaded in MEGA6 software to analyze the rate of nucleotide substitution. The following statistical parameters were used to analyze the rate of substitution rate in chloroplast genomes: analysis, estimate transition/transversion bias (MCL); scope, all selected taxa; statistical method, maximum composite likelihood; substitution type, nucleotides; model/method, Tamura–Nei model; and gaps/missing data treatment, complete deletion.

2.6. Statistical Analysis

Principal component analysis and the probability distribution of chloroplast genomes were conducted using Unscrambler software version 7.0 and Venn diagrams were constructed using InteractiVenn (http://www.interactivenn.net/) [49].

3. Results

3.1. The Genomic Features of Chloroplast Genomes Are Diverse and Dynamic

A study of 2511 chloroplast genomes was conducted to gain insight into the genomic structure and evolution of the chloroplast genome. The analysis included the complete genome sequences of algae, austrobaileyales, bryophytes, chloranthales, corals, eudicots, Flacourtiaceae, gymnosperms, magnoliids, monocots, Nymphaeales, opisthokonta, protists, pteridophytes, and an unclassified chloroplast genome (Supplementary File S1). A comparison of the analyzed genomes indicated that *Haematococcus lacustris* encoded the largest chloroplast genome, comprising 1.352 Mbs; however, *Pilostyles aethiopica* encoded the smallest chloroplast genome, comprising only 0.01134 Mbs (Figure 1) followed by *Pilostyles hamiltoni* (0.01516 Mb), and *Asarum minus* (0.0155 Mb). The overall average size of the chloroplast genome was found to be 0.152 Mbs. The order of the average size (Mbs) of the chloroplast genome in different plant groups was 0.164 (algae), 0.160 (Nymphaeales), 0.154 (eudicot), 0.154 (Magnoliid), 0.149 (pteridophyte), 0.144 (monocot), 0.134 (bryophyte), 0.131 (gymnosperm), and 0.108 (protist). The average chloroplast genome size in algae (0.164 Mbs) and the Nymphaeales (0.160 Mbs) was larger than eudicots (0.154 Mbs), monocots (0.144 Mbs), and gymnosperms (0.131 Mbs). The average size of the protist chloroplast genome (0.108 Mbs) was the smallest. Principal component analysis (PCA) of the chloroplast genome size of algae, bryophytes, eudicots, gymnosperms, magnoliids, monocots, Nymphaeales, protists, and pteridophytes reveals a clear distinction between the different plant groups (Figure 2). The size of the chloroplast genome of gymnosperm and bryophytes grouped together; and eudicots, magnoliids, and pteridophytes grouped together. In contrast, the algae and protists were independently grouped (Figure 2). This shows that the chloroplast genome of algae and protists might have evolved from their respective common ancestors.

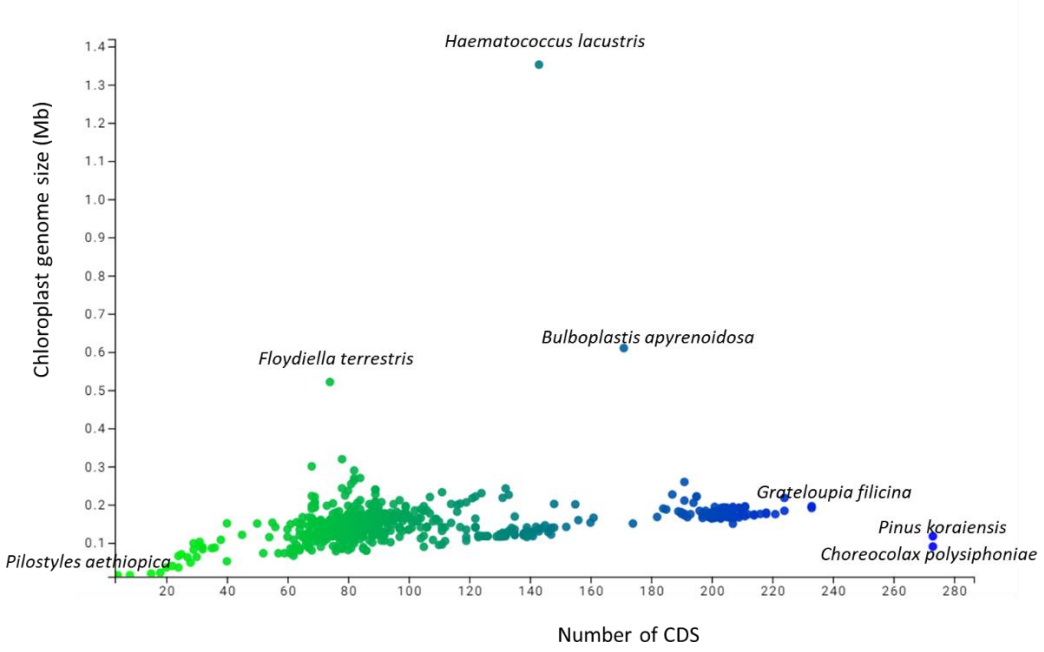

Figure 1. Genome size and number of coding sequences (CDS) in the chloroplast genome. The blue dot present at the right side indicates the genome size of the largest chloroplast genome that encodes 1.35 Mbs in *Haematococcus lacustris* and the green dot present at the top of the figure represents 273 CDS found in *Pinus koraiensis*.

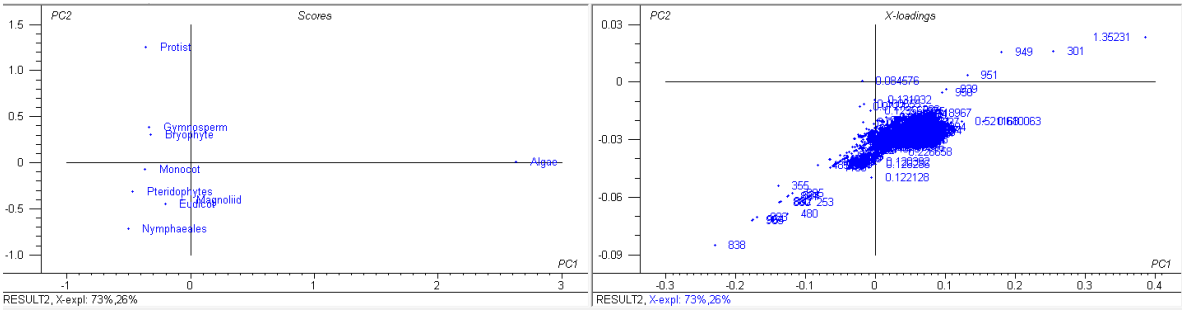

Figure 2. Principal component analysis of chloroplast genome sizes. The genome size of gymnosperms and bryophytes fall in one group and eudicots, magnoliids, monocots, and pteridophytes fall in the other group; however, algae and protists fall distantly.

The number of coding sequences (CDS) in the analyzed chloroplast genomes ranged from 273 (*Pinus koraiensis* and *Choreocolax polysiphoniae*) to 3 (*Pilostyles aethiopica*; Figure 1). The average number of CDS in all the studied chloroplast genome was 89.90 per genome. However, some other species contained a higher number of CDS in the chloroplast genome, including *Grateloupia filicina* (233), *Osmundaria fimbriata* (224), *Porphyridium purpureum* (224), *Lophocladia kuetzingii* (221), *Kuetzingia canaliculata* (218), *Spyridia filamentosa* (218), *Bryothamnion seaforthii* (216) and others (Supplementary File S1). Similarly, some species encoded a lower number of CDS in the chloroplast genome, including *Pilostyles aethiopica* (3), *Pilostyles hamiltoni* (4), *Asarum minus* (8), *Cytinus hypocistis* (15), *Sciaphila densiflora* (18), *Gastrodia elata* (20), *Burmannia oblonga* (22), *Orobanche gracilis* (24), and others (Supplementary File S1). PCA analysis indicated that the number of CDS in bryophytes, eudicots, magnoliids, monocots, and pteridophytes grouped together (Figure 3). The number of CDS in algae, gymnosperms, and protists grouped very distantly from the above-mentioned grouping (Figure 3). The average CDS number in algae (140.93) was quite high compared to magnoliid (84), eudicot (83.55), monocot (82.53), gymnosperm (82.56), and protist (98.97). However, algae and protists encoded a higher number of CDS compared to the magnoliid, eudicot, monocot, gymnosperm, and protist. The larger genome size of algae and protist is associated with a greater number of CDS in the

chloroplast genome and they fall distantly in the PCA plot. This suggests that the evolution of chloroplast genome and CDS number of algae and protist share a slightly similar trend compared to other plant species. However, they might have evolved from their respective ancestors.

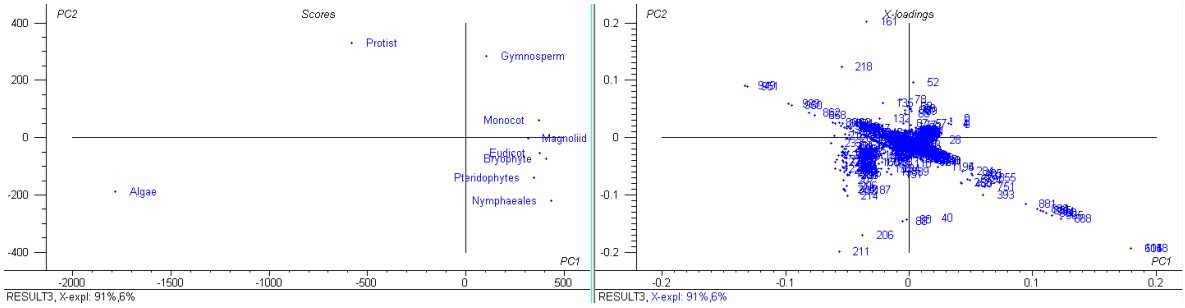

Figure 3. Principal component analysis of CDS numbers of chloroplast genomes. The CDS number of algae, gymnosperms, and protists fall separately; however, bryophytes, eudicots, pteridophytes, and Nymphaeales fall together.

The GC content of the analyzed chloroplast genomes ranged from a high of 57.66% (*Trebouxiophyceae* sp. MX-AZ01) to a low of 20.46% (*Choreocolax polysiphoniae*; Figure 4, Supplementary File S1). The average GC content in the chloroplast genome was 36.82%. Some species contained a higher percentage of GC content, including *Trebouxiophyceae* sp. MX-AZ01 (57.664%), *Coccomyxa subellipsoidea* C-169 (50.73%), *Paradoxia multiseta* (50.58%), *Haematococcus lacustris* (49.88%), *Chromerdia* sp. RM11 (47.74%), *Elliptochloris bilobata* (45.76%), *Choricystis parasitica* (45.44%), and others. On the other hand, some species had a lower percentage of GC content, including *Ulva prolifera* (24.78%), *Ulva linza* (24.78%), *Ulva fasciata* (24.86%), *Ulva flexuosa* (24.97%), and others (Supplementary File S1). PCA analysis revealed that the percentage GC content of eudicots, gymnosperms, magnoliids, monocots, and Nymphaeales grouped together, and the percentage of GC content in algae and protists grouped together (Figure 5). The percentage of GC content in bryophytes and pteridophytes did not group with the algae and protists or the eudicots, gymnosperms, magnoliids, monocots, or Nymphaeales (Figure 5). The GC content of algae and protists showed that they have a common trend of evolution with regard to genome size, CDS number, and GC content. The evolutionary similarity of algae and protist is closer than other lineages.

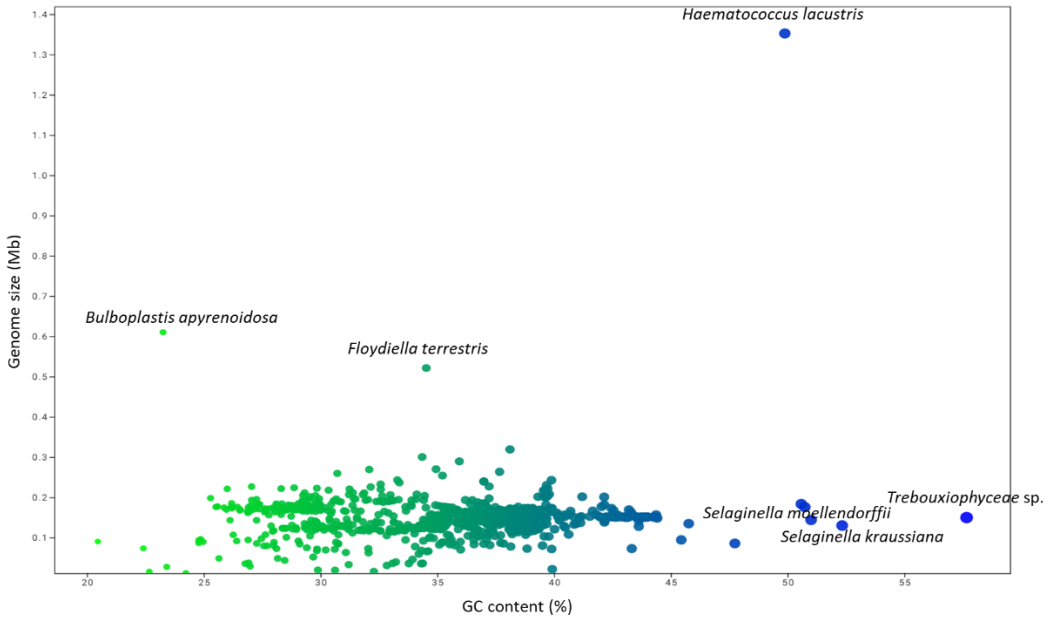

Figure 4. Genome size and GC (%) content in the chloroplast genome. The genome size of *Haematococcus lacustris* was highest (1.352 Mb) present in the upper right side (blue dot). The blue dot

present at the right side of the figure represents the GC content of *Trebouxiophyceae* sp. MX-AZ01 that contain 57.66% GC nucleotides; however, the green dot present at the left upper part of the figure represents the lower GC content (23.25%) of *Bulboplastis apyrenoidosa*.

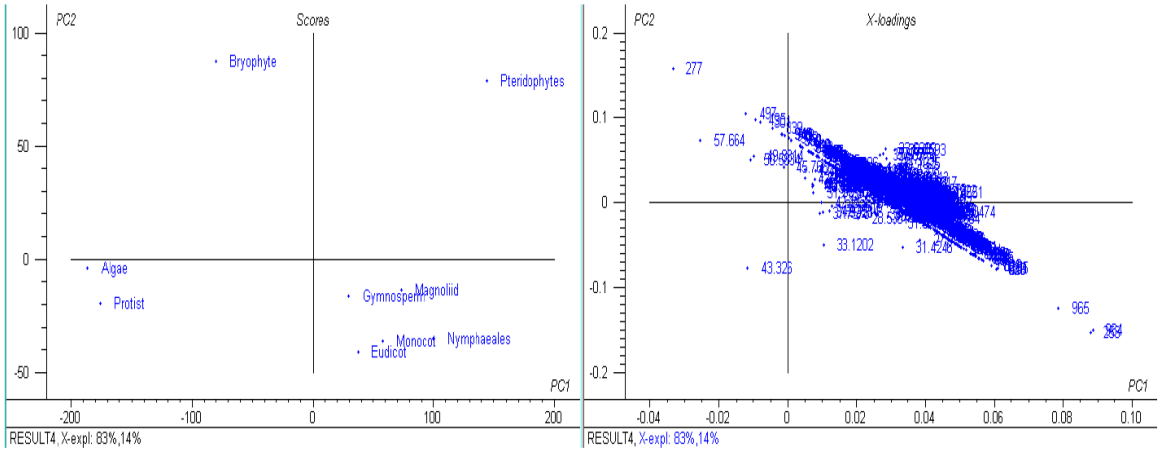

Figure 5. Principal component analysis of GC content of the chloroplast genomes. The GC content of algae and protists and gymnosperms, magnoliids, monocots, eudicots, and Nymphaeales grouped together; however, the GC content of the bryophytes and pteridophytes fall distantly.

3.2. PsaM, Psb30, ChlB, ChlL, ChlN, *and* RPL21 *Are Chloroplast Genes Characteristic of Algae, Bryophytes, Pteridophytes, and Gymnosperms*

The PsaM protein is subunit XII of photosystem I. Among the 2511 studied species, 84 were found to possess the *PsaM* gene. All of the species found to possess the *PsaM* gene belonged to algae, bryophytes, pteridophytes, and gymnosperms (Supplementary File S2). Notably, no the species in the angiosperm lineage possessed the *PsaM* gene; clearly indicating that the *PsaM* gene was lost in the angiosperm lineage. The PsaM protein was found to contain the characteristic conserved amino acid motif Q-x_3-A-x_3-A-F-x_3-I-L-A-x_2-L-G-x_2-L-Y (Supplementary Figure S1). A few species, including *Cephalotaxus, Podocarpus tortara, Retrophyllum piresii, Dacrycarpus imbricatus, Glyptostrobus pensilis, T. distichum, Cryptomeria japonica, Pinus contorta, Pinus taeda*, and *Ptilidium pulcherrimum*, however, did not contain the conserved amino acid motif. Instead, they possessed the conserved motif, F-x-S-x_3-C-F-x_4-F-S-x_2-I (Supplementary Figure S1). Phylogenetic analysis revealed that *PsaM* genes grouped into five independent clusters, suggesting that they have evolved independently from multiple common ancestral nodes (Supplementary Figure S2A). Duplication and deletion analysis of *PsaM* genes revealed that deletion events were more prominent than the duplication or codivergence events (Table 1). Among the 84 analyzed *PsaM* genes, 12 underwent duplication and 34 underwent deletions, while 34 underwent codivergence (Table 1, Supplementary Figure S2B).

Table 1. Deletion and duplication events of *PsaM, Psb30, ChlB, ChlL, ChlN*, and *Rpl21* genes. Analysis revealed gene loss was dominated compared to the duplication and codivergence.

Name of the Gene	Total No. of Sequences Studied	No. of Duplication	No. of Codivergence	No. of Losses	Transfer
PsaM	84	12 (14.28%)	37 (44.04%)	34 (40.47%)	0
Psb30	157	39 (24.84)	49 (31.21%)	120 (76.43%)	0
ChlB	288	35 (12.15%)	116 (40.27%)	126 (43.75%)	0
ChlL	283	49 (17.31%)	100 (35.33%)	184 (65.01%)	0
ChlN	83	8 (9.63%)	34 (40.47%)	46 (55.42%)	0
Rpl21	22	3 (13.63%)	9 (40.90%)	8 (36.36%)	0

A total of 164 species were found to possess *Psb30* gene and all of the species belonged to algae, bryophytes, pteridophytes, or gymnosperms (Supplementary File S2). *Psb30* was absent in the

chloroplast genome of angiosperms. Multiple sequence alignment revealed the presence of a conserved consensus amino acid sequence, N-x-E-x_3-Q-L-x_2-L-x_6-G-P-L-V-I (Supplementary Figure S3). Phylogenetic analysis of *Psb30* genes resulted in the designation of two major clusters and six minor clusters, suggesting that it evolved from multiple common ancestral nodes (Supplementary Figure S4A). Deletion/duplication analysis indicated that 39 of *Psb30* genes underwent a duplication event and 120 underwent a deletion event, while 49 were found to be codiverged (Table 1, Supplementary Figure S4B).

ChlB encodes a light-independent protochlorophyllide reductase. A total of 288 of the examined chloroplast genome sequences were found to possess a *ChlB* gene (Supplementary File S2) among protists, algae, bryophytes, pteridophytes, and gymnosperms. The *ChlB* gene was absent in species in the chloranthales, corals, or angiosperm lineage. Multiple sequence alignment revealed the presence of several highly conserved amino acid motifs (Supplementary Figure S8). At least seven conserved motifs were identified, including A-Y-W-M-Y-A, L-P-K-A-W-F, E-N-Y-I-D-Q-Q, S-Q-A-A-W-F, H-D-A-D-W-F, E-P-x_2-I-F-G-T, E-K-F/Y-A-R-Q-Q, and E-V-M-Y-A-A (Supplementary Figure S5). Phylogenetic analysis indicated that *ChlB* genes grouped into two major clusters and 13 minor clusters, reflecting multiple evolutionary nodes (Supplementary Figure S6A). *ChlB* genes were composed of a few groups. Specifically, deletion and duplication analysis revealed that 35 *ChlB* genes underwent duplications and 126 underwent deletions, while 116 exhibited codivergence in their evolutionary history (Table 1, Supplementary Figure S6B).

Analysis of the chloroplast genome sequences identified 303 species that possess *ChlL* genes (Supplementary File S2). All of the identified species possessing the *ChlL* gene belonged to algae, bryophytes, gymnosperms, protists, and pteridophytes. None of the taxa in the angiosperm or magnoliid lineage were found to possess a *ChlL* gene. Within the protist lineage, only species in the genera *Nannochloropsis*, *Vaucheria*, *Triparma*, and *Alveolata* encode a *ChlL* gene. Multiple sequence alignment revealed the presence of several highly conserved amino acid motifs, including K-S-T-T-S-C-N-x-S, W-P-E-D-V-I-Y-Q, K-Y-V-E-A-C-P-M-P, C-D-F-Y-L-N, Q-P-E-G-V-V/I, and S-D-F-Y-L-N (Supplementary Figure S7). The phylogenetic analysis indicated that *ChlL* genes grouped into one major independent cluster and 11 minor clusters, suggesting that they also evolved independently from different common ancestors (Supplementary Figure S8A). Deletion and duplication analysis indicated that 49 *ChlL* genes underwent duplication events and 184 underwent deletions, while 100 *ChlL* genes exhibited codivergence (Table 1, Supplementary Figure S8B).

The analysis revealed that at least 289 species possess *ChlN* genes. These genomes were from taxa within the protists, algae, bryophytes, pteridophytes, and gymnosperms (Supplementary File S2). Multiple sequence alignment revealed the presence of highly conserved amino acid motifs, including N-Y-H-T-F, A-E-L-Y-Q-K-I-E-D-S, M-A-H-R-C-P, and Q-I-H-G-F (Supplementary Figure S9). Phylogenetic analysis revealed that *ChlN* genes group into two independent clusters (Supplementary Figure S10A). No lineage-specific grouping, however, was identified in the phylogenetic tree. Deletion and duplication analysis indicated that eight *ChlN* genes underwent duplication events, 46 underwent deletion events and 34 genes exhibited codivergence (Table 1, Supplementary Figure S10B).

The chloroplast genomes of at least 137 of the examined species were found to possess an *RpL21* gene which belonged to algae, bryophytes, pteridophytes, and gymnosperms (Supplementary File S2). In the majority of cases, full-length CDS was not found. Instead, the CDS of the *Rpl21* genes were found to be truncated. Therefore, only 22 full-length CDS were used to identify deletion and duplication events. Rpl21 proteins were found to contain the conserved amino acid motifs, Y-A-I-I-D-x-G-G-x-Q-L-R-E-V-x-G-R-F, R-V-L-M-I, G-x-P-W-L, R-I-L-H, and K-x_2-I/V-x_5-K-K (Supplementary Figure S11). Phylogenetic analysis shows the presence of three clusters, reflecting their origin from multiple common ancestral nodes (Supplementary Figure S12A). Deletion/duplication analysis indicated that three *RpL21* genes underwent duplication events, eight underwent deletion events, and nine exhibited codivergence (Table 1, Supplementary Figure S12B).

3.3. The Rbcl *Gene Has Been Lost in Parasitic and Heterotrophic Plant Species*

The analysis found, at least 19 species have lost the *Rbcl* gene in their chloroplast genome. The species lacking an *Rbcl* gene were *Pilostyles aethiopica*, *Pilostyles hamiltoni*, *Alveolata* sp. CCMP3155, *A. minus, Bathycoccus prasinos* (picoplankton), *Burmannia oblonga* (orchid), *Codonopsis lanceolate* (eudicot), *Cytinus hypocistis* (parasite), *Gastrodia elata* (saprophyte), *Monotropa hypopitys* (mycoheterotroph), *Orobanche austrohispanica* (parasite), *Orobanche densiflora* (parasite), *Orobanche gracilis* (parasite), *Orobanche pancicii* (parasite), *Phelipanche purpurea* (parasite), *Phelipanche ramosa* (parasite), *Prototheca cutis* (parasitic algae), *Prototheca stagnorum* (parasitic algae), and *Sciaphila densiflora* (mycoheterotroph).

3.4. Deletion of Inverted Repeats (IRs) Has Occurred Across All Plastid Lineages

Inverted repeats (IR) are one of the major characteristic features of the chloroplast genomes. The analysis conducted in the present study revealed the deletion of inverted repeats in the chloroplast genome of 259 (10.31%) species from the 2511 species examined (Supplementary File S3). IR deletion events were identified in protists (14), protozoans (one), algae (126), bryophytes (one), gymnosperms (64), magnoliids (one) monocots (nine), and eudicots (43). The average size of the deleted IR region in algae was 0.177 Mb, which is larger than the overall size of the chloroplast genome in the respective taxa. The average size of the deleted IR region in eudicots, monocots, and gymnosperms was 0.124, 0.131, and 0.127 Mb, respectively, which is smaller than the overall size of the chloroplast genome in the respective lineages.

Phylogenetic analysis of chloroplast genomes containing deleted IR regions produced three major clusters (Supplementary Figure S16). Gymnosperms were in the upper cluster (cyan) while the lower cluster (red) comprised the algae, bryophytes, eudicots, gymnosperms, and pteridophytes. No chloroplast genomes from monocot plants were present in the lower cluster (Supplementary Figure S16). The middle cluster contained at least four major phylogenetic groups (Supplementary Figure S16). Monocot plants were present in two groups (pink) in the middle cluster. Gymnosperm (cyan) and eudicot (green) chloroplast genomes were also present in two of the groups in the middle cluster. Although there was some sporadic distribution of algae in different groups of the phylogenetic tree, the majority of the algal species were present in a single group (yellow; Supplementary Figure S16). A phylogenetic tree of taxa with deleted IR and taxa with chloroplast genomes that did not lose the IR region (*Floydiella terrestris*, *Carteria cerasiformis, B. apyrenoidosa, Eucalyptus grandis, Oryza sativa,* and others) did not reveal any specific difference in their clades. Instead, they also grouped with the genomes in which the IR region was deleted. Inverted repeats stabilize the chloroplast genome [50,51] and the loss of a region of inverted repeats most likely leads to a genetic rearrangement in the chloroplast genome. The lower cluster (red) contained the oldest group. Genomic recombination analysis revealed that the chloroplast genomes across different lineages also underwent vivid recombination (Supplementary Figure S14A,B). In addition, the IR-deleted chloroplast genomes also underwent vivid recombination (Supplementary Figure S15).

3.5. Several Genes in the Chloroplast Genome Have Been Lost

The chloroplast genome encodes genes for photosynthesis, amino acid biosynthesis, transcription, protein translation, and other important metabolic processes. The major genes involved in such events are *AccD* (acetyl-coenzyme A carboxylase carboxyl transferase), *AtpA, AtpB, AtpE, AtpF, AtpH, AtpI, CcsA* (cytochrome C biogenesis protein), *CemA* (chloroplast envelope membrane), *ChlB* (light-independent protochlorophyllide reductase), *ChlL, ChlN, ClpP* (ATP-dependent Clp protease), *MatK* (maturase K), *NdhA* (NADPH-quinone oxidoreductase), *NdhB, NdhC, NdhD, NdhE, NdhF, NdhG, NdhH, NdhI, NdhJ, NdhK, Pbf1* (photosystem biogenesis factor 1), *PetA* (cytochrome precursor), *PetB, PetD, PetG, PetL, PetN, PsaA* (photosystem I protein), *PsaB, PsaC, PsaI, PsaJ, PsaM, Psb30, PsbA* (photosystem II protein), *PsbB, PsbC, PsbD, PsbE, PsbF, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM, PsbT, PsbZ, Rbcl* (ribulose 1,5-bisphosphate carboxylase), *Rpl2* (60S ribosomal protein), *Rpl14, Rpl16, Rpl20, Rpl21, Rpl22, Rpl23, Rpl32, Rpl33, Rpl36, RpoA* (DNA-directed RNA polymerase), *RpoB, RpoC1, RpoC2, Rps2 (40S ribosomal protein), Rps3, Rps4, Rps7, Rps8, Rps11, Rps12, Rps14, Rps15, Rps16, Rps18,*

Rps19, Ycf1, Ycf2, Ycf3, and Ycf4. Our analysis revealed that a number of these genes were lost in one or other species in a dynamic manner (Table 2). The analysis indicated that the ribosomal proteins Rpl and Rpo were lost less frequently than the other chloroplast genes (Table 2). *Ndh* genes were lost in a number of different species. Several other genes had been deleted in a considerable number of species across different lineages. These included *AccD* (402), *AtpF* (217), *Clp* (194), *Ycf2* (226), *Ycf4* (111), *PetL* (248), *PetN* (125), *PsaI* (129), *PsbM* (166), *PsbZ* (145), *Rpl22* (137), *Rpl23* (221), *Rpl32* (182), *Rpl33* (163), *Rps15* (263), and *Rps16* (372), where the number in parentheses indicates the number of taxa in which the gene has been deleted from the chloroplast genome (Table 2). Detailed about the loss of all the chloroplast genes across can be found in Supplementary Data S1.

Table 2. Deletion of different genes in the chloroplast genomes. Almost all of the genes have been deleted in the chloroplast genome of one or another species. However, *Rpl20* was found to be the most intact gene and found in all the species studied so far.

rpoA	rpoB	rpoC1	rpoC2												
26	19	21	13												
atpA	atpB	atpE	atpF	atpH	atpI										
8	8	12	14	13	12										
accD	ccsA	cemA	chlB	chlL	ChlN										
387	29	29	2054	2062	2066										
ClpP	Rbcl	Ycf1	Ycf2	Ycf3	Ycf4										
142	19	161	219	30	39										
ndhA	ndhB	ndhC	ndhD	ndhE	ndhF	ndhG	NdhH	ndhI	ndhJ	ndhK					
339	258	339	293	322	346	335	322	378	340	331					
petA	petB	PetD	petG	petL	petN										
33	15	36	13	71	135										
psaA	psaB	psaC	psaI	psaJ	psaM										
16	10	19	72	24	2214										
psbA	psbB	psbC	psbD	psbE	psbF	psbH	psbI	psbJ	psbK	psbL	psbM	psbN	psbT	psbZ	Psb30
12	18	16	17	21	21	20	18	21	13	22	157	23	22	31	2126
Rpl2	Rpl14	Rpl16	Rpl20	Rpl22	Rpl23	Rpl32	Rpl33	Rpl36							
2	4	3	0	127	24	114	133	5							
Rps2	Rps3	Rps4	Rps7	Rps8	Rps11	Rps12	Rps14	Rps15	Rps16	Rps18	Rps19				
3	3	4	3	3	2	2	7	249	284	5	5				

3.6. The loss of Genes in Chloroplast Genomes is Dynamic

When the collection of all the lost genes were grouped, it was evident that a large number of genes had been found to be lost in algae, eudicots, magnoliids, and monocots (Supplementary Table S1). Only a small number of genes were lost in bryophytes, gymnosperms, protists, and pteridophytes (Supplementary Table S1). When the species of algae, gymnosperms, monocots, eudicots, magnoliids, and bryophytes were grouped together, *NdhA, NdhC, NdhD, NdhE, NdhF, NdhG, NdhH, NdhI, NdhJ,* and *NdhK* genes were found to be lost in all six lineages; however, *AtpB, AtpE, AtpH, AtpI, CemA, PetA, PetB, PetD, PetG, PetL, PsaA, PsaB, PsaC, PsaI, PsbA, PsbB, PsbC, PsbD, PsbE, PsbF, PsbH, PsbJ, PsbL, PsbZ, Psbf1, Rpl22, Rpl33, RpoB,* and *RpoC2* had been lost in algae, eudicots, magnoliids, and monocots (Supplementary Figure S17, Supplementary Table S1). *AccD, NdhB, PsaJ, Rpl23*, and *Rpl32* genes were only absent in species of algae, eudicots, gymnosperms, magnoliids, and monocots. When species of algae, bryophytes, gymnosperms, angiosperms (monocot and dicot), pteridophytes, and protists were grouped together, at least 11 genes were found to be lost in all of the lineages (Supplementary Table S1, Supplementary Figure S18). The most commonly lost genes were *NdhA, NdhC, NdhD, NdhE, NdhF, NdhG, NdhH, NdhI, NdhJ, NdhK*, and *Rps16*. The *NdhB* gene, however, was lost in algae, angiosperms, gymnosperms, protists, and pteridophytes; however, it was present in all species of bryophytes. When the higher groupings of plant lineages (gymnosperms, magnoliids, and monocots) were grouped together, it was found that *AccD, NdhA, NdhB, NdhC, NdhD, NdhE, NdhF, NdhG, NdhH, NdhI, NdhJ, NdhK, PsaJ, Rpl23*, and *Rpl32* had been lost in all four lineages (Supplementary Figure S19, Supplementary Table S1). *AtpB, AtpE, AtpH, AtpI, CcsA, CemA, PetA, PetB, PetD, PetG, PetL, PetN, PsaA, PsaB, PsaC, PsaI, PsbA, PsbB, PsbC, PsbD, PsbE, PsbF, PsbH, PsbJ, PsbL, PsbZ, Psbf1, Rpl22, Rpl33, RpoB, RpoC1, RpoC2*, and *Rps19* were found to be lost in eudicots, magnoliids, and monocots. *ClpP* was found to be lost in eudicots, gymnosperms, and magnoliids. A comparative analysis of gene loss in eudicot and monocot plants revealed that gene loss was more frequent in eudicots (69 genes) than in monocots (59 genes). Eudicots and monocots share the loss of 59 genes in their chloroplast genomes. The loss of *ClpP, Rpl2, Rpl14, Rpl36, RpoA, Rps2, Rps8, Rps11, Rps14*, and *Rps18* occurred only in eudicots and not in monocots. A comparative analysis of gene loss in eudicots, gymnosperms, and monocots indicated that the loss of *Rps7* was unique to the gymnosperms. The loss of at least 17 genes (*accD, ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK, psaJ, rpl23, rpl32, rps15,* and *rps16*) were found to be common in between eudicots, gymnosperms, and monocots.

3.7. Chloroplast-Derived Genes are Present in the Nuclear Genome

It has been speculated that genes lost from chloroplast genomes may have moved to the nuclear genome and are regulated as a nuclear-encoded gene [52,53]. Therefore, a genome-wide analysis of fully sequenced and annotated genomes of 145 plant species was analyzed to explore this question. Results indicated a maximum presence of the chloroplast-encoding genes in the nuclear genome. We found the presence of 189,381 putative nuclear encoding chloroplast gene from the study of 145 plant species (Supplementary File S5). Some of the chloroplast-derived genes that were found in the nuclear genome were: Rubisco accumulation factor, 30S ribosomal 30S ribosomal proteins (1, 2, 3, S1, S2, S3, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S18, S19, S20, S21, and S31) 50S ribosomal proteins (5, 6, L1, L2, L3, L4, L5, L6, L9, L10, L11, L12, L13, L14, L15, L16, L17, L18, L19, L20, L21, L22, L23, L24, L27, L28, L29, L31, and L32), *Psa* (A, B, C, I, and J), *Psb* (A, B, D, E, F, H, I, J, K, L, M, N, P, Q, T, and Z), *Rpl* (12 and 23), *RpoA, RpoB, RpoC1, RpoC2, Rps7, Rps12, Ycf* (1, 2, and 15), *YlmG* homolog, Ribulose bisphosphate carboxylase small chain (1A, 1B, 2A, 3A, 3B, 4, F1, PW9, PWS4, and S4SSU11A), Ribulose bisphosphate carboxylase/oxygenase activase A and B, (-)-beta-pinene synthase, (-)-camphene/tricyclene synthase, (+)-larreatricin hydroxylase, (3S,6E)-nerolidol synthase, (E)-beta-ocimene synthase, 1,4-alpha-glucan-branching enzyme, 10 kDa chaperonin, 1,8-cineole synthase, 2-carboxy-1,4-naphthoquinone phytyltransferase, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase, ABC transporter B family, AccD, acyl-carrier-protein, adenylate kinase, ALBINO protein, allene oxide

cyclase, anion transporter, anthranilate synthase, APO protein, aspartokinase, ATP synthase, Atp (A, B, E, F, H, I), ATP-dependent Clp protease, beta carbonic anhydrase, calcium-transporting ATPase, Calvin cycle protein CP12, carbonic anhydrase, cation/H(+) antiporter, chaperone protein Clp (B, C, and D), DnaJ, chaperonin 60 subunit, chlorophyll a-b binding protein (1, 2, 3, 4, 6, 7, 8, 13, 15, 16, 21, 24, 26, 29, 36, 37, 40, 50, 80, M9, LHCII, and P4), chlorophyll(ide) b reductase (NOL and NYC), chloroplastic acetyl-coenzyme A carboxylase, chloroplastic group IIA intron splicing facilitator CRS (S1, A, and B), chorismate mutase, cytochrome b6/f complex subunit (1, 2, IV, V, VI, and VIII), cytochrome c biogenesis protein CCS1, DEAD-box ATP-dependent RNA helicase, DNA gyrase A and B, DNA polymerase A and B, DNA repair protein recA homolog, DNA-(apurinic or apyrimidinic site) lyase, DNA-damage-repair/toleration protein, DNA-directed RNA polymerase, early light-induced protein, fatty acid desaturase, ferredoxin--NADP reductase, fructokinase, gamma-terpinene synthase, geraniol synthase, geranylgeranyl pyrophosphate synthase, glucose-1-phosphate adenylyltransferase small and large subunit, glutathione S-transferase, GTP diphosphokinase CRSH, inactive ATP-dependent zinc metalloprotease FTSHI, inactive shikimate kinase, kinesin protein KIN (D, E, K, L, and M), L-ascorbate peroxidase, light-harvesting complex protein, light-induced protein, light-regulated protein, lipoxygenase, magnesium transporter, magnesium-chelatase, MATE efflux family protein, multiple organellar RNA editing factor, N-(5'-phosphoribosyl)anthranilate isomerase, NAD Kinase, NAD(P)H-quinone oxidoreductase subunits (1, 2, 3, 4, 5, 6, H, I, J, K, L, M, N, O, S, T, and U), NADH dehydrogenase subunits (1, 2, 3, 4, 5, 6, 7, I, J, and K), NADH-plastoquinone oxidoreductase subunits (1, 2, 3, 4, 5, 6, 7, I, J, and K), NADPH-dependent aldehyde reductase, nifU protein, nudix hydrolases, outer envelope pore proteins, oxygen-evolving enhancer proteins, pentatricopeptide repeat-containing protein (CRP1, DOT4, DWY1, ELI1, MRL1, OTP51, PPR5), peptide chain release factor, peptide methionine sulfoxide reductase, peptidyl-prolyl cis-trans isomerases, Pet (A, B, G, and L), phospholipase, photosynthetic NDH subunit of lumenal location, photosynthetic NDH subunit of subcomplex B, protochlorophyllide reductase subunits (B, L, and N), phytol kinase, plastid-lipid-associated proteins, protease Do 1, protein cofactor assembly of complex c subunits, protein CutA, DCL, pyruvate dehydrogenase E1 component subunits, sodium/metabolite cotransporter BASS, soluble starch synthase, stearoyl-[acyl-carrier-protein] 9-desaturase, thioredoxins, thylakoid luminal proteins, translation initiation factor, transcription factor GTE3, transcription termination factor MTERF, translocase of chloroplast, zinc metalloprotease EGY, and others (Supplementary File S6).

3.8. The Ratio of Nucleotide Substitution is Highest in Pteridophytes and Lowest in Nymphaeales

Determining the rate of nucleotide substitution in the chloroplast genome can be an important parameter that needs to be more precisely understood to further elucidate the evolution of the chloroplast genome. Single base substitutions, and insertion and deletion (indels) events play an important role in shaping the genome. Therefore, an analysis was conducted to determine the rate of substitution in the chloroplast genome by grouping them according to their respective lineages. Results indicated that the transition/transversion substitution ratio was highest in pteridophytes (k1 = 4.798 and k2 = 4.043) and lowest in Nymphaeales (k1 = 2.799 and k2 = 2.713; Supplementary Table S2). The ratio of nucleotide substitution in species with deleted IR regions was 2.951 (k1) and 3.42 (k2; Supplementary Table S2). The rate of transition of A > G substitution was highest in pteridophytes (15.08) and lowest in protists (8.51) and the rate of G > A substitution was highest in protists (22.15) and lowest in species with deleted IR regions (16.8). The rate of substitution of T > C was highest in pteridophytes (14.01) and lowest in protists (8.95; Supplementary Table S2). The rate of substitution of C > T was highest in protists (22.34) and lowest in Nymphaeales. The rate of transversion is two-times less frequent than the rate of transition. The rate of transversion of A > T was highest in protists (6.80) and lowest in pteridophytes (4.64), while the rate of transversion of T > A was highest in algae (6.98) and lowest in pteridophytes (Supplementary Table S2). The rate of substitution of G > C was highest in Nymphaeales (4.31) and lowest in protists (2.46), while the rate of substitution of C > G was highest in Nymphaeales (4.14) and lowest in protists (2.64; Supplementary Table S2). Based on these results, it is concluded that the highest rates of transition and transversion were more frequent

in lower eukaryotic species, including algae, protists, Nymphaeales, and pteridophytes; however, high rates of transition/transversion were not observed in bryophytes, gymnosperms, monocots, and dicots (Supplementary Table S2). Notably, G > A transitions were more prominent in chloroplast genomes with deleted IR regions (Supplementary Table S2).

3.9. Chloroplast Genomes Have Evolved from Multiple Common Ancestral Nodes

A phylogenetic tree was constructed to obtain an evolutionary perspective of chloroplast genomes (Figure 6). All of the 2511 studied species were used to construct a phylogenetic tree (Figure 6). The phylogenetic analysis produced four distinct clusters, indicating that chloroplast genomes evolved independently from multiple common ancestral nodes. Lineage-specific groupings of chloroplast genomes were not present in the phylogenetic tree. The genomes of algae, bryophytes, gymnosperms, eudicots, magnoliids, monocots, and protists grouped dynamically in different clusters. Although the size of the chloroplast genome in protists was far smaller than other lineages and still, they were distributed sporadically throughout the phylogenetic tree. Time tree analysis indicated that the origin of the cyanobacterial species (used as outgroup) date back to ~2180 Ma and that the endosymbiosis of the cyanobacterial genome occurred ~1768 Ma ago and was incorporated into the algal lineage ~1293-686 Ma ago (Supplementary Figure S20); which then further evolved into the Viridiplantae ~1160 Ma, Streptophyta ~1150 Ma, Embryophyta ~532 Ma, Tracheophyte ~431 Ma, Euphyllophyte 402 Ma, and Spermatophyta 313 Ma (Supplementary Figure S20). The molecular signature genes *PsaM, ChlB, ChlL, ChlN, Psb30*, and *Rpl21* in algae, bryophytes, pteridophytes, and gymnosperms were lost ~203 (Cycadales) and -156 (Gnetidae) Ma ago, and as a result, are not found in the subsequently evolved angiosperm lineage (Supplementary Figure S20).

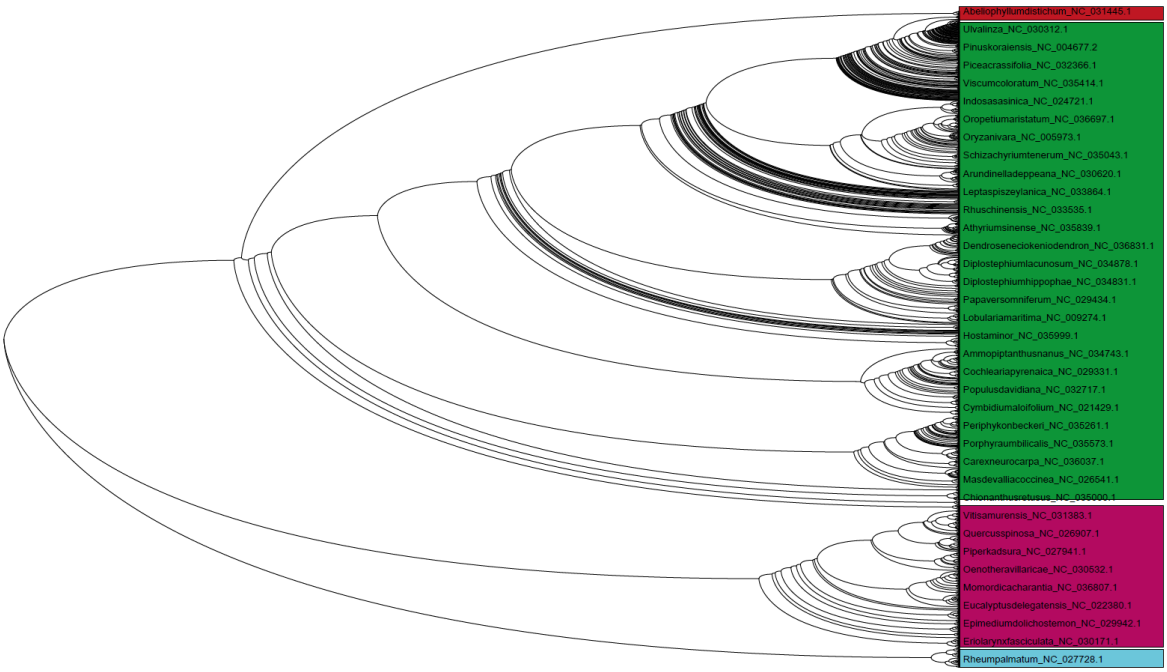

Figure 6. Phylogenetic tree of chloroplast genomes. The phylogenetic tree showed the presence of four major clusters in the chloroplast genomes, suggesting their evolution from multiple common ancestral nodes. The phylogenetic tree considered all of the genomes used during the study and was constructed by a Neighbor-joining program with 500 bootstrap replicates and ClustalW.

4. Discussion

Chloroplasts are an indispensable part of plant cells function as semiautonomous organelles due to the presence of their own genetic material, potential to self-replicate, and capability to modulate cell metabolism [4,54–56]. The size of the chloroplast genome is highly variable and does not correlate

to the size of the corresponding nuclear genome of the species. The average size of the chloroplast genome is 0.152 Mb and encodes an average of 91.67 CDS per genome. The deletion of IR regions in the chloroplast genome is supposed to drastically reduce the genetic content of the chloroplast genome and also the number of CDS. However, the current analysis does not support this premise. The average number of CDS in algae (140.93) was higher than protists (98.97), pteridophytes (86.54), eudicots (83.55), bryophytes (83.38), gymnosperms (82.54), and monocots (82.53). The larger genome size (0.177 Mb) of the chloroplast genome in algae with deleted IR regions, and the higher number of CDS (172.16 per genome) in IR-deleted taxa of algae indicates that the loss of IR regions in algae led to a genetic rearrangement and an enlargement in the chloroplast genome. However, the average CDS number of other lineages in IR-deleted genomes was quite lower than their average CDS count (86.28 for protist, 63 for monocot, 81.42 for gymnosperm, and 71.88 for eudicot). The average size of IR-deleted chloroplast genomes in eudicots, monocots, protists, and gymnosperms was smaller than the average size of chloroplast genomes of taxa where IR regions have not been deleted. Thus, the lower number of CDS in these taxa may be related to the deletion of IR regions. This suggests that the deletion of IR regions in the chloroplast genome of algae is directly proportional to the increase in the genome size and concomitant increase in the CDS number; however, this was not true in the other plant lineages where the relationship was inversely proportional. The deletion of IR regions has been previously reported in a few species of algae, magnoliids, and other genomes [57–61]. The present study, however, provided clear evidence regarding the loss of IR regions across all plant and protist lineages. The deletion of IR repeats and an increase in the genome size in algae has largely been attributed to the duplication of the chloroplast genome. The evolutionary age of IR-deleted species of algae dates back to ~965-850 Ma. This provides strong evidence that the deletion of IR repeats and duplications of the chloroplast genome has been a continuous process since the initial evolution of the chloroplast genome in algae. Zhu et al. also suggested a role for duplication in the evolution of IR-deleted chloroplast genomes [60]. Characterizing the pattern and frequency of neutral mutations (substitution, insertions, and deletion) is important for deciphering the molecular basis of the evolution of genes and genomes. Turmel et al. reported that a differential loss of genes from the chloroplast genome resulted in the loss of IR regions in the chloroplast genome for all the lineages, except algae and protists [57]. The transition/transversion ratio of purine substitutions in all IR-deleted species (k1 = 2.951) was much lower than in non-IR-deleted species, except for species in the Nymphaeales, and the substitution of pyrimidines in all IR-deleted species was higher (k2 = 3.42), except pteridophytes (Supplementary Table S2). These data suggest that, in addition to a duplication event, a lower rate of purine substitution and a higher rate of pyrimidine substitution are closely associated with the deletion of IR regions.

In addition to the loss of IR regions, the loss of genes from chloroplast genomes was also analyzed. The loss of important genes from the chloroplast genome has been previously reported in some species of green algae, bryophytes, and magnoliids (Supplementary Data S1) [62–65]. The results of the present study indicate the loss of the *Rbcl* gene in at least 19 species among parasitic, mycoparasitic, and saprophytic plant species across different lineages, including algae, eudicots, magnoliids, monocots, and protists. The parasitic plant *Conopholis* of Orobanchacea lost the photosynthetic gene *Rbcl*; however, it was present in other parasitic plants in Orobanchacea [66,67]. The loss of *Rbcl*, however, was not observed in any species of bryophytes, pteridophytes, or gymnosperms. The number of CDS in the *Rbcl*-deleted chloroplast genome was much lower (27 per genome) relative to the average number of CDS found in the chloroplast genomes; except for *Alveolata* sp. CCMP3155 which possessed 81 CDS. The loss of the *Rbcl* gene in the chloroplast genome is associated with a drastic reduction in the number of other protein-coding genes. The reduction in the genome size is associated with the massive loss of ancestral protein-coding genes [68]. Interestingly, the parasitic genus, *Cuscuta*, possesses an *Rbcl* gene which suggests that the parasitic nature of a species is not always associated with the deletion of the *Rbcl* gene and vice versa, the loss of the *Rbcl* gene is not a prerequisite of becoming a parasitic plant as well. However, it is quite clear that parasitism is getting more prone towards the loss of chloroplast-encoding genes. Although a few contain the *Rbcl* gene, they cannot sustain themselves for their own photosynthesis. The losses of

these molecular features are providing an important platform to understand the plant–parasite interactions and evolution of parasitic plants. The loss of genes is most possibly associated with a high level of contraction of the nuclear genome as well. Most possibly, the autotrophic plant evolved parasitic characters through neofunctionalization and transcriptional reprogramming of its older lineage. The study reported that transition from the autotrophic plants to parasitic plants relaxes the functional constraints in a stepwise manner for plastid genes [69].

The deletions of one or more important genes of the chloroplast genome observed in numerous species (Supplementary Data S1). It is difficult to decipher the exact reason for the loss of these individual genes in different chloroplast genomes. *NdhA, NdhC, NdhD, NdhE, NdhF, NdhG, NdhH, NdhI, NdhJ, NdhK,* and *Rps16* were genes that were most commonly lost across the analyzed chloroplast genomes. The *NdhB* gene, however, was found to be intact in all species of bryophytes, suggesting that it could serve as a signature gene for the bryophyte chloroplast genome. *Ndh* genes encode a component of the thylakoid Ndh-complex involved in photosynthetic electron transport. The loss of specific *Ndh* genes in different species suggests that not all *Ndh* genes are involved in or needed for functional photosynthetic electron transport. The loss of one *Ndh* gene may be compensated for by other *Ndh* genes or by nuclear-encoded genes. The functional role of the *Ndh* gene was previously reported to be closely related to the adaptation of land plants and photosynthesis [70]. The loss of *Ndh* genes in species across all the plant lineages, including algae, suggests that *Ndh* genes are not associated with the adaptation of photosynthesis to terrestrial ecosystems. Previous studies have reported the loss of *Ndh* genes in the Orchidaceae, where the deletion was reported to occur independently after the orchid family split into different subfamilies [71]. These data suggest that the loss of *Ndh* genes in the parental lineage of orchids led to the loss of *Ndh* genes in the subfamilies in the downstream lineages of orchids.

A comparison of gene loss in monocots and dicots revealed that species in the eudicots are more prone to gene loss than monocot species. Monocots and dicots chloroplast genome shared a common loss of 59 genes, while eudicots have lost 10 more genes (*ClpP, Rpl14, Rpl2, Rpl36, RpoA, Rps2, Rps8, Rps11, Rps14,* and *Rps18*) than monocots, suggesting that these genes represent the molecular signature of the chloroplast genomes of monocot species. *Ycf* (*Ycf1, Ycf2, Ycf3,* and *Ycf4*) genes were found to be intact in all species of bryophytes, gymnosperms, and pteridophytes, suggesting that they represent a common molecular signature for these lineages. Various genes, including *MatK, Rbcl, Ndh,* and *Ycf*, are commonly used as universal molecular markers in DNA barcoding studies for determining the genus and species of the plants. The loss of these genes in the chloroplast genome of various lineages makes their use as universal markers questionable in future studies for DNA barcoding [72–76].

The loss of *RpoA* from the chloroplast genome of mosses was previously reported and it was suggested that *RpoA* had relocated to the nuclear genome [63,77]. The loss of *Psa* and *Psb* genes were quite prominent in algae, eudicot, magnoliid, monocot, and protist lineages. *Psa* and *Psb* genes were always found in species of bryophytes, pteridophytes, and gymnosperms, suggesting that these genes could serve as a common molecular signature for these lineages. *PsaM, Psb30, ChlB, ChlL, ChlN,* and *Rpl21* are characteristic molecular signature genes for lower eukaryotic plants, including algae, bryophytes, pteridophytes, and gymnosperms. Additionally, these genes are completely absent in the eudicots, magnoliids, monocots, and protists. The absence of these genes in angiosperm and magnoliid lineages reflect their potential role in the origin of flowering plants. Duplication events for *PsaM, Psb30, ChlB, ChlL, ChlN*, and *Rpl21* genes were much lower than deletion and codivergence events (Table 1). In fact, codivergence was the dominant event for all of these genes (Table 1). The recombination events that occurred in the chloroplast genome directly reflect the potential possibility of codivergent and divergent evolution in these genes. The presence of *PsaM, Psb30, ChlB, ChlL,* and *ChlN* genes in their respective lineages support the premise that these genes are orthologous and resulted from a speciation event [78–81]. *Chl* genes are involved in photosynthesis in cyanobacteria, algae, pteridophytes, and conifers [82–87]; indicating that the *Chl* genes were originated at least ~2180 Ma ago and remained intact up to the divergence of the angiosperms at ~156 Ma. The loss of *Psa* and *Psb* genes in different species also suggests that they are not essential for a complete and functional

photosynthetic process. The loss of a *Psa* or *Psb* gene in a species might be compensated for by other *Psa* or *Psb* genes or by a nuclear-encoded gene. The loss of *Psa* and *Psb* genes in species across all plant lineages has not been previously reported. Thus, this study is the first to report the loss of *Psa* and *Psb* genes in the chloroplast genome of species across all plant lineages, as well as protists. The loss of *Rpl22*, *Rpl32*, and *Rpl33* genes was more prominent than the loss of *Rpl2*, *Rpl14*, *Rpl16*, *Rpl20*, *Rpl23*, and *Rpl36*, suggesting the conserved nature of *Rpl2*, *Rpl14*, *Rpl16*, *Rpl20*, *Rpl23*, and *Rpl36* genes and the conserved transfer of these genes to subsequent downstream lineages as intact genes. *Rpl20* was found to be an intact gene in all 2511 of the studied species, suggesting that *Rpl20* is the most evolutionary conserved gene in the chloroplast genome of the plants and protists. Therefore, *Rpl20* can be considered as the molecular signature gene of the chloroplast genome. Similarly, the loss of *Rps15* and *Rps16* was more frequently relative to the loss of *Rps2*, *Rps3*, *Rps4*, *Rps7*, *Rps8*, *Rps11*, *Rps12*, *Rps14*, *Rps18*, and *Rps19*.

There are several reports regarding the transfer of genes from the chloroplast to the nucleus [4,31,88–90]. In the present study, almost all of the genes encoded by the chloroplast genomes were also found in the nuclear genome. The presence of the chloroplast-encoded genes in the nuclear genome, however, was quite dynamic. If a specific chloroplast-encoded gene was found in the nuclear genome of one species, it may not have been present in the nuclear genome of the other species. One report also indicated that genes transferred to the nuclear genome may not provide a one to one correspondence function [90]. The question also arises as to how almost all of the chloroplast-encoded genes can be found in the nuclear genome and how were they transferred? If the transfers and correspondence are real, it is plausible that almost all chloroplast-encoded genes have been transferred to the nuclear genome in one or more species and that the transfer of chloroplast genes to the nuclear genome is a common process in the plant kingdom and exchange of chloroplast genes with nuclear genome have already completed.

5. Conclusions

The underlying exact mechanism regarding the deletion of IR regions from the chloroplast genome is still unknown and the loss of specific chloroplast-encoded genes and IR regions in diverse lineages makes it more problematic to decipher the mechanism or selective advantage behind the loss of the genes and IR regions. It is likely that nucleotide substitutions and the dynamic recombination of chloroplast genomes are the factors that are most responsible for the loss of genes and IR regions. Although the evolution of parasitic plants can, to some extent, be attributed to the loss of important chloroplast genes (including *Rbcl*); still it is not possible to draw any definitive conclusions regarding the loss of genes and IR regions. The presence of all chloroplast-encoded genes in the nuclear genome in one or another species is quite intriguing. A question arises, however: do the chloroplast genomes complete the transfer of different chloroplast-encoding genes in different species based on some adaptive requirement? The presence of a completely intact *Rpl20* gene without any deletions in the chloroplast genome of all the species indicates that the *Rpl20* gene can be considered as a molecular signature gene of the chloroplast genome.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Supplementary Data S1: Loss of various chloroplast genes across the plant lineage, Supplementary File S1: File showing the name and genomic details of the species whose chloroplast genome was used during this study, Supplementary File S2: File showing the presence of *Psb30*, *PsaM*, *ChlL*, *ChlN*, *ChlB*, and *Rpl21* genes in the chloroplast genome of species belonged to algae, bryophyte, pteridophyte, and gymnosperm. These genes were not found in the chloroplast genome of angiosperm lineage, Supplementary File S3: File showing the loss of IR region in the chloroplast genome of different species, Supplementary File S4: File showing the loss of different chloroplast-encoding genes in different species, Supplementary File S5: Complete list of putative nuclear encoding chloroplast genes studied from the 145 fully annotated nuclear genomes, Supplementary File S6: List of the chloroplast-encoding genes found in the nuclear genomes, Supplementary Table S1: Loss of chloroplast-encoding genes in different species of respective lineages, Supplementary Table S2: Maximum composite likelihood substitution of nucleotides. The entry reflects the probability of substitution (r) from one base (row) to another base (column). The rates of

transitions are highlighted in bold and rates of transversion are highlighted in italics. The nucleotide frequencies (%) of A, T/U, G, and C for the respective study are mentioned in the rows. The transition/transversion ratios are mentioned as K1 (purine) and K2 (pyrimidine). The transition/transversion bias R = [A*G*k1 + T*C*k2]/[(A+G) * (T+C)]. The codon position included were 1st + 2nd + 3rd + noncoding. All the positions with less than 95% site coverage were eliminated. That is fewer than 5% alignment gaps. Missing data and ambiguous bases were allowed at any position. The C > T substitution is more frequent than T > C substitution and G > A substitution more frequently than A > G. The major mechanism mutation is deamination of 5′-methyl cytosine to uracil (thiamine) producing C > T or on the complementary strand G > A, Supplementary Figure S1: Conserved amino acid sequences of PsaM proteins. Blue mark indicates conservation of amino acids below 90%, Supplementary Figure S2: (A) Phylogenetic tree of *PsaM* genes showing five clusters. (B) Deletion and duplication event of *PsaM* genes. Duplications: 12, codivergences: 37, transfers: 0, Losses: 34; number of temporally feasible Optimal Solutions: 1; tree without Losses, total nodes: 171, internal nodes: 85, leaf nodes: 86; polytomies: 0, size of largest polytomy: 0, height: 18; tree with losses, total nodes: 239, internal nodes: 119, leaf nodes: 120, size of largest polytomy: 0 and height: 22, Supplementary Figure S3: Conserved amino acid sequences of Psb30 proteins. Red mark indicates conservation of amino acids of 90% or more, Supplementary Figure S4: (A) phylogenetic tree of *Psb30* genes. (B) deletion and duplication event of *Psb30* genes. Duplications: 39, codivergences: 49, transfers: 0, losses: 120; number of temporally feasible optimal solutions: 1; tree without losses, total nodes: 313, internal nodes: 156, leaf nodes: 157; polytomies: 0, size of largest polytomy: 0; height: 24, tree with losses; total nodes: 553, internal nodes: 276, leaf nodes: 277, size of largest polytomy: 0, and height: 34, Supplementary Figure S5: Conserved amino acid sequences of ChlB proteins. Red mark indicate conservation of 90% or more, Supplementary Figure S6: (A) phylogenetic tree of *ChlB* genes. (B) deletion and duplication event of *ChlB* genes. Duplications: 35, codivergences: 116, transfers: 0, losses: 126, number of temporally feasible optimal solutions: 1; tree without losses, total nodes: 575, internal nodes: 287, leaf nodes: 288, polytomies: 0, size of largest polytomy: 0, height: 34; tree with losses, total nodes: 827, internal nodes: 413, leaf nodes: 414, size of largest polytomy: 0, and height: 37, Supplementary Figure S7: Conserved amino acid sequences of ChlL proteins. Red mark indicate conservation of 90% or more, Supplementary Figure S8: (A) Phylogenetic tree of *ChlL* genes. (B) Deletion and duplication event of *ChlL* genes. Duplications: 49, codivergences: 100, transfers: 0, losses: 184, number of temporally feasible optimal solutions: 1; tree without losses, total nodes: 565, internal nodes: 282, leaf nodes: 283, polytomies: 0, size of largest polytomy: 0, height: 35; tree with losses, total nodes: 933, internal nodes: 466, leaf nodes: 467, size of largest polytomy: 0 and height: 39, Supplementary Figure S9: Conserved amino acid sequences of ChlN proteins. Red mark indicate conservation of 90% or more, Supplementary Figure S10: (A) Phylogenetic tree of *ChlN* genes. (B) Deletion and duplication event of *ChlN* genes. Duplications: 8, codivergences: 34, transfers: 0, losses: 46, number of temporally feasible optimal solutions: 1; tree without losses, total nodes: 161, internal nodes: 80, leaf nodes: 81, polytomies: 0, size of largest polytomy: 0 height: 17; tree with losses; total nodes: 253, internal nodes: 126, leaf nodes: 127, size of largest polytomy: 0, and height: 23, Supplementary Figure S11: Conserved amino acid sequences of Rpl21 proteins. Red mark indicate conservation of 90% or more, Supplementary Figure S12: Phylogenetic tree of *Rpl21* genes. (B) Deletion and duplication event of *Rpl21* genes. Duplications: 3, codivergences: 9, transfers: 0, losses: 8, number of temporally feasible optimal solutions: 1; tree without losses, total nodes: 43, internal nodes: 21, leaf nodes: 22, polytomies: 0, size of largest polytomy: 0, height: 10; tree with losses, total nodes: 59, internal nodes: 29, leaf nodes: 30, size of largest polytomy: 0 and height: 11, Supplementary Figure S13: Molecular weight and isoelectric point (pI) of RBCL proteins, Supplementary Figure S14: Recombination events of chloroplast genomes (A) unresolved (B) resolved. Chloroplast genomes were found to undergo vivid genomic recombination; which might be one of the possible reasons regarding the loss of the IR region in the chloroplast genomes. The color represents their link of recombination events in different taxon/groups. The genomic recombination of chloroplast genomes was studied using the IcyTree viewer (https://icytree.org/) server, Supplementary Figure S15: Recombination event of inverted repeat deleted chloroplast genomes. Each color indicates a locus and their distribution in different clusters indicates they underwent vivid recombination, Supplementary Figure S16. Phylogenetic tree of inverted repeat (IR)-deleted chloroplast genomes. The phylogenetic tree of chloroplast genomes was constructed with ClustalW using a neighbor-joining approach with 1000 bootstrap replicates and three major clusters were identified. The phylogenetic tree was constructed in combination with the species containing the inverted repeats (Floydiella terrestris, Carteria cerasiformis, Bulboplastis apyrenoidosa, Eucalyptus grandis, Oryza sativa, and others) to decipher the differences. Deletion of inverted repeats did not have a considerable impact on the phylogeny, Supplementary Figure S17: Venn diagram showing group specific loss of chloroplast-

encoding genes in algae, gymnosperm, bryophyte, monocot, eudicot, and magnoliid, Supplementary Figure S18: Venn diagram showing group specific loss of chloroplast-encoding genes in algae, bryophyte, gymnosperm, angiosperm, pteridophyte and protist, Supplementary Figure S19. Venn diagram showing group specific loss of chloroplast-encoding genes in eudicot, gymnosperm, monocot, and magnoliid, Supplementary Figure S20. Time tree of chloroplast genomes. An evolutionary time tree was constructed using the species used in this study. Time tree study revealed, cyanobacteria were evolved ~2180 Ma ago and subsequently transferred to rhodophyta ~1333 Ma, algae ~1293 Ma, viridiplantae ~1160 Ma, streptophyta ~1150 Ma, embryophyte and ~491 Ma. The time tree uses the impact of earth, oxygen, carbon dioxide, and solar luminosity in the evolution.

Author Contributions: Conceptualization, T.K.M.; formal analysis, T.K.M., A.K.M. and A.K.; investigation, T.K.M.; software, T.K.M.; validation, T.K.M.; writing—original draft preparation, T.K.M. and A.K.M.; writing—review and editing, T.K.M., A.H., E.F.A.A. and A.A.-H.

Funding: This research received no external funding.

Acknowledgement: Authors would like to extend their sincere thanks to Natural and Medical Sciences Research Center, University of Nizwa, Oman for extending its support and facility to conduct the research. Authors would also like to extend their sincere appreciation to the Researchers Supporting Project Number (RSP-2020/134), King Saud University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barber, J.; Andersson, B. Revealing the blueprint of photosynthesis. *Nat.* **1994**, *370*, 31–34, doi:10.1038/370031a0.
2. Mishra, B.S.; Singh, M.; Aggrawal, P.; Laxmi, A. Glucose and Auxin Signaling Interaction in Controlling Arabidopsis thaliana Seedlings Root Growth and Development. *PLoS ONE* **2009**, *4*, e4502.
3. Moore, B.; Zhou, L.; Rolland, F.; Hall, Q.; Cheng, W.-H.; Liu, Y.-X.; Hwang, I.; Jones, T.; Sheen, J. Role of the Arabidopsis Glucose Sensor HXK1 in Nutrient, Light, and Hormonal Signaling. *Science* **2003**, *300*, 332–336, doi:10.1126/science.1080585.
4. Osteryoung, K.W.; Weber, A.P.M. Plastid Biology: Focus on the Defining Organelle of Plants. *Plant Physiol.* **2011**, *155*, 1475–1476, doi:10.1104/pp.111.900408.
5. Hedges, S.B.; Blair, J.E.; Venturi, M.L.; Shoe, J.L. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evolut. Biol.* **2004**, *4*, 2, doi:10.1186/1471-2148-4-2.
6. Yoon, H.S.; Hackett, J.D.; Ciniglia, C.; Pinto, G.; Bhattacharya, D. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Mol. Biol. Evolut.***2004**, *21*, 809–818.
7. Gibbs, S.P. The Chloroplasts of Some Algal Groups May Have Evolved from Endosymbiotic Eukaryotic Algae. *Ann. N. Y. Acad. Sci.* **1981**, *361*, 193–208, doi:10.1111/j.1749-6632.1981.tb54365.x.
8. Mohanta, T.K.; Pudake, R.N.; Bae, H. Genome-wide identification of major protein families of cyanobacteria and genomic insight into the circadian rhythm. *Eur. J. Phycol.* **2017**, *52*, doi:10.1080/09670262.2016.1251619.
9. Raven, J.A.; Allen, J.F. Genomics and chloroplast evolution: What did cyanobacteria do for plants? *Gen. Biol.* **2003**, *4*, 209, doi:10.1186/gb-2003-4-3-209.
10. Cavalier-Smith, T. Membrane heredity and early chloroplast evolution. *Trends Plant Sci.* **2000**, *5*, 174–182, doi:10.1016/S1360-1385(00)01598-3.
11. Kirk, P.R.; Leech, R.M. Amino Acid Biosynthesis by Isolated Chloroplasts during Photosynthesis. *Plant Physiol.* **1972**, *50*, 228–234, doi:10.1104/pp.50.2.228.
12. Noctor, G.; Arisi, A.-C.M.; Jouanin, L.; Foyer, C.H. Manipulation of Glutathione and Amino Acid Biosynthesis in the Chloroplast. *Plant Physiol.* **1998**, *118*, 471–482, doi:10.1104/pp.118.2.471.
13. Hawke, J.C.; Rumsby, M.G.; Leech, R.M. Lipid biosynthesis by chloroplasts isolated from developing Zea mays. *Phytochemistry* **1974**, *13*, 403–413, doi:10.1016/S0031-9422(00)91224-X.
14. Britton, G. Biosynthesis of Chloroplast Carotenoids. In *Current Research in Photosynthesis*; Baltscheffsky, M., Ed.; Springer: Dordrecht, The Netherlands, 1990; pp. 2733–2740. ISBN 978-94-009-0511-5.
15. García-Cerdán, J.G.; Schmid, E.M.; Takeuchi, T.; McRae, I.; McDonald, K.L.; Yordduangjun, N.; Hassan, A.M.; Grob, P.; Xu, C.S.; Hess, H.F.; et al. Chloroplast Sec14-like 1 (CPSFL1) is essential for normal chloroplast development and affects carotenoid accumulation in Chlamydomonas. *Proc. Natl. Acad. Sci.*

USA **2020**, *117*, 12452–12463, doi:10.1073/pnas.1916948117.
16. Ries, F.; Herkt, C.; Willmund, F. Co-translational protein folding and sorting in chloroplasts. *Plants* **2020**, *9*, 214, doi:10.3390/plants9020214.
17. Guo, X.; Liu, J.; Hao, G.; Zhang, L.; Mao, K.; Wang, X.; Zhang, D.; Ma, T.; Hu, Q.; Al-Shehbaz, I.A.; et al. Plastome phylogeny and early diversification of Brassicaceae. *BMC Genomics* **2017**, *18*, 176, doi:10.1186/s12864-017-3555-3.
18. Graham, S.W.; Lam, V.K.Y.; Merckx, V.S.F.T. Plastomes on the edge: The evolutionary breakdown of mycoheterotroph plastid genomes. *N. Phytol.* **2017**, *214*, 48–55, doi:10.1111/nph.14398.
19. Gruenstaeudl, M.; Jenke, N. PACVr: Plastome assembly coverage visualization in R. *BMC Bioinform.* **2020**, *21*, 1–21, doi:10.1186/s12859-020-3475-0.
20. Darshetkar, A.M.; Datar, M.N.; Tamhankar, S.; Li, P.; Choudhary, R.K. Understanding evolution in Poales: Insights from Eriocaulaceae plastome. *PLoS ONE* **2019**, *14*, e0221423.
21. Sugiura, M. The chloroplast genome. In *10 Years Plant Molecular Biology*; Schilperoort, R.A., Dure, L., Eds.; Springer: Dordrecht, The Netherlands, 1992; pp. 149–168, ISBN 978-94-011-2656-4.
22. Yu, Q.-B.; Huang, C.; Yang, Z.-N. Nuclear-encoded factors associated with the chloroplast transcription machinery of higher plants. *Front. Plant Sci.* **2014**, *5*, 316, doi:10.3389/fpls.2014.00316.
23. Wicke, S.; Schneeweiss, G.M.; dePamphilis, C.W.; Müller, K.F.; Quandt, D. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol. Biol.* **2011**, *76*, 273–297, doi:10.1007/s11103-011-9762-4.
24. Kolodner, R.; Tewari, K.K. Inverted repeats in chloroplast DNA from higher plants. *Proc. Natl. Acad. Sci. USA* **1979**, *76*, 41–45.
25. Wolf, P.G.; Der, J.P.; Duffy, A.M.; Davidson, J.B.; Grusz, A.L.; Pryer, K.M. The evolution of chloroplast genes and genomes in ferns. *Plant Mol. Biol.* **2011**, *76*, 251–261, doi:10.1007/s11103-010-9706-4.
26. Raubeson, L.; Jasen, R. Chloroplast genomes of plants. In *Diversity and Evolution of Plants—Genotypic and Phenotypic Variation in Higher Plants*; Henry, R., Ed.; CABI Publishing: Wallingford, UK, 2005; pp. 45–68.
27. Wu, C.S.; Lai, Y.T.; Lin, C.P.; Wang, Y.N.; Chaw, S.M. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: Selection toward a lower-cost strategy. *Mol. Phylogenet. Evol.* **2009**, *52*, doi:10.1016/j.ympev.2008.12.026.
28. Daniell, H.; Lin, C.-S.; Yu, M.; Chang, W.-J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **2016**, *17*, 134, doi:10.1186/s13059-016-1004-2.
29. Rivas, J.D.L.; Lozano, J.J.; Ortiz, A.R. Comparative Analysis of Chloroplast Genomes: Functional Annotation, Genome-Based Phylogeny, and Deduced Evolutionary Patterns. *Genome Res.* **2002**, *12*, 567–583, doi:10.1101/gr.209402.
30. Stern, D.B.; Goldschmidt-Clermont, M.; Hanson, M.R. Chloroplast RNA Metabolism. *Ann. Rev. Plant Biol.* **2010**, *61*, 125–155, doi:10.1146/annurev-arplant-042809-112242.
31. Cullis, C.A.; Vorster, B.J.; Van Der Vyver, C.; Kunert, K.J. Transfer of genetic material between the chloroplast and nucleus: How is it related to stress in plants? *Ann. Bot.* **2009**, *103*, 625–633, doi:10.1093/aob/mcn173.
32. Eckardt, N.A. Genomic Hopscotch: Gene Transfer from Plastid to Nucleus. *Plant Cell* **2006**, *18*, 2865–2867, doi:10.1105/tpc.106.049031.
33. 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**, *23*, 1324–1338.
34. Gao, L.; Su, Y.-J.; Wang, T. Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. *J. Syst. Evol.* **2010**, *48*, 77–93, doi:10.1111/j.1759-6831.2010.00071.x.
35. Downie, S.R.; Palmer, J.D. Restriction Site Mapping of the Chloroplast DNA Inverted Repeat: A Molecular Phylogeny of the Asteridae. *Ann. Mo. Bot. Gard.* **1992**, *79*, 266–283, doi:10.2307/2399769.
36. Goulding, S.E.; Wolfe, K.H.; Olmstead, R.G.; Morden, C.W. Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet. MGG* **1996**, *252*, 195–206, doi:10.1007/BF02173220.
37. Plunkett, G.M.; Downie, S.R. Expansion and Contraction of the Chloroplast Inverted Repeat in Apiaceae Subfamily Apioideae. *Syst. Bot.* **2000**, *25*, 648–667, doi:10.2307/2666726.
38. Guisinger, M.M.; Kuehl, J.V.; Boore, J.L.; Jansen, R.K. Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage. *Mol. Biol. Evol.* **2011**, *28*, 583–600.
39. Greiner, S.; Lehwark, P.; Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit

for the graphical visualization of organellar genomes. *Nucleic Acid. Res.* **2019**, doi:10.1101/545509.
40. Tillich, M.; Lehwark, P.; Pellizzer, T.; Ulbricht-Jones, E.S.; Fischer, A.; Bock, R.; Greiner, S. GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acid. Res.* **2017**, *45*, W6–W11, doi:10.1093/nar/gkx391.
41. Stothard, P. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. *BioTechniques* **2000**, *28*, 1102–1104, doi:10.2144/00286ir01.
42. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acid. Res.* **1988**, *16*, 10881–10890.
43. Tamura, K.; Filipski, A.; Peterson, D.; Stecher, G.; Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **2013**, *30*, 2725–2729, doi:10.1093/molbev/mst197.
44. Kumar, S.; Stecher, G.; Suleski, M.; Hedges, S.B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **2017**, *34*, 1812–1819, doi:10.1093/molbev/msx116.
45. Stolzer, M.; Lai, H.; Xu, M.; Sathaye, D.; Vernot, B.; Durand, D. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **2012**, *28*, i409–i415, doi:10.1093/bioinformatics/bts386.
46. Darby, C.A.; Stolzer, M.; Ropp, P.J.; Barker, D.; Durand, D. Xenolog classification. *Bioinformatics* **2017**, *33*, 640–649, doi:10.1093/bioinformatics/btw686.
47. Chen, K.; Durand, D.; Farach-Colton, M. NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. *J. Comput. Biol.* **2000**, *7*, 429–447, doi:10.1089/106652700750050871.
48. Vaughan, T.G. IcyTree: Rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* **2017**, *33*, 2392–2394, doi:10.1093/bioinformatics/btx155.
49. Heberle, H.; Meirelles, G.V.; da Silva, F.R.; Telles, G.P.; Minghim, R. InteractiVenn: A web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinform.* **2015**, *16*, 169, doi:10.1186/s12859-015-0611-3.
50. Zhang, R.; Ge, F.; Li, H.; Chen, Y.; Zhao, Y.; Gao, Y.; Liu, Z.; Yang, L. PCIR: a database of Plant Chloroplast Inverted Repeats. *Database: J. Biol. Databases Curation* **2019**, *2019*, baz127, doi:10.1093/database/baz127.
51. Ma, J.; Yang, B.; Zhu, W.; Sun, L.; Tian, J.; Wang, X. The complete chloroplast genome sequence of Mahonia bealei (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. *Gene* **2013**, *528*, 120–131, doi:10.1016/j.gene.2013.07.037.
52. Bryant, N.; Lloyd, J.; Sweeney, C.; Myouga, F.; Meinke, D. Identification of Nuclear Genes Encoding Chloroplast-Localized Proteins Required for Embryo Development in Arabidopsis. *Plant Physiol.* **2011**, *155*, 1678, doi:10.1104/pp.110.168120.
53. Savage, L.J.; Imre, K.M.; Hall, D.A.; Last, R.L. Analysis of Essential Arabidopsis Nuclear Genes Encoding Plastid-Targeted Proteins. *PLoS ONE* **2013**, *8*, e73291.
54. Taylor, D.L. Chloroplasts as Symbiotic Organelles. In *International Review of Cytology*; Bourne, G.H., Danlelli, J.F., JeonAcademic Press New York and London: 1970; 27, pp. 29–64. ISBN 0074-7696.
55. Trench, R.K. Chloroplasts: presumptive and de facto organelles. *Ann. N. Y. Acad. Sci.* **1981**, *361*, 341–355, doi:10.1111/j.1749-6632.1981.tb54376.x.
56. Stern, D.S.; Higgs, D.C.; Yang, J. Transcription and translation in chloroplasts. *Trends Plant Sci.* **1997**, *2*, 308–315, doi:10.1016/S1360-1385(97)89953-0.
57. Turmel, M.; Otis, C.; Lemieux, C. Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophycean green algae. *Sci. Rep.* **2017**, *7*, 994, doi:10.1038/s41598-017-01144-1.
58. Wolfe, K.H. The site of deletion of the inverted repeat in pea chloroplast DNA contains duplicated gene fragments. *Curr. Genet.* **1988**, *13*, 97–99, doi:10.1007/BF00365763.
59. Strauss, S.H.; Palmer, J.D.; Howe, G.T.; Doerksen, A.H. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 3898–3902.
60. Zhu, A.; Guo, W.; Gupta, S.; Fan, W.; Mower, J.P. Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. *N. Phytol.* **2016**, *209*, 1747–1756, doi:10.1111/nph.13743.
61. Palmer, J.; Osorio, B.; Aldrich, J.; Thompson, W. Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. *Curr. Genet.* **1987**, *11*, 275–286.
62. Lemieux, C.; Otis, C.; Turmel, M. Comparative Chloroplast Genome Analyses of Streptophyte Green Algae Uncover Major Structural Alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front. Plant Sci.* **2016**, *7*, 697, doi:10.3389/fpls.2016.00697.
63. Sugiura, C.; Kobayashi, Y.; Aoki, S.; Sugita, C.; Sugita, M. Complete chloroplast DNA sequence of the moss Physcomitrella patens: Evidence for the loss and relocation of rpoA from the chloroplast to the nucleus.

Nucleic Acid. Res. **2003**, *31*, 5324–5331, doi:10.1093/nar/gkg726.

64. Sinn, B.T.; Sedmak, D.D.; Kelly, L.M.; Freudenstein, J. V. Total duplication of the small single copy region in the angiosperm plastome: Rearrangement and inverted repeat instability in Asarum. *Am. J. Bot.* **2018**, *105*, 71–84, doi:10.1002/ajb2.1001.
65. Alverson, A.J.; Ruck, E.C.; Theriot, E.C.; Nakov, T.; Jansen, R.K. Serial Gene Losses and Foreign DNA Underlie Size and Sequence Variation in the Plastid Genomes of Diatoms. *Genome Biol. Evolut.* **2014**, *6*, 644–654, doi:10.1093/gbe/evu039.
66. Wolfe, A.D.; dePamphilis, C.W. The effect of relaxed functional constraints on the photosynthetic gene rbcL in photosynthetic and nonphotosynthetic parasitic plants. *Mol. Biol. Evolut.* **1998**, *15*, 1243–1258, doi:10.1093/oxfordjournals.molbev.a025853.
67. Wicke, S.; Müller, K.F.; de Pamphilis, C.W.; Quandt, D.; Wickett, N.J.; Zhang, Y.; Renner, S.S.; Schneeweiss, G.M. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *Plant Cell* **2013**, *25*, 3711–3725, doi:10.1105/tpc.113.113373.
68. Hughes, A.L.; Friedman, R. Genome Size Reduction in the Chicken Has Involved Massive Loss of Ancestral Protein-Coding Genes. *Mol. Biol. Evolut.* **2008**, *25*, 2681–2688, doi:10.1093/molbev/msn207.
69. Sun, G.; Xu, Y.; Liu, H.; Sun, T.; Zhang, J.; Hettenhausen, C.; Shen, G.; Qi, J.; Qin, Y.; Li, J.; et al. Large-scale gene losses underlie the genome evolution of parasitic plant Cuscuta australis. *Nat. Commun.* **2018**, *9*, 2683, doi:10.1038/s41467-018-04721-8.
70. Martín, M.; Sabater, B. Plastid ndh genes in plant evolution. *Plant Physiol. Biochem.* **2010**, *48*, 636–645, doi:10.1016/j.plaphy.2010.04.009.
71. Lin, C.-S.; Chen, J.J.W.; Huang, Y.-T.; Chan, M.-T.; Daniell, H.; Chang, W.-J.; Hsu, C.-T.; Liao, D.-C.; Wu, F.-H.; Lin, S.-Y.; et al. The location and translocation of ndh genes of chloroplast origin in the Orchidaceae family. *Sci. Rep.* **2015**, *5*, 9040.
72. Heckenhauer, J.; Barfuss, M.H.J.; Samuel, R. Universal multiplexable matK primers for DNA barcoding of angiosperms. *Appl. Plant Sci.* **2016**, *4*, apps.1500137, doi:10.3732/apps.1500137.
73. Yu, J.; Xue, J.H.; Zhou, S.L. New universal matK primers for DNA barcoding angiosperms. *J. Syst. Evol.* **2011**, *49*, 176–181, doi:10.1111/j.1759-6831.2011.00134.x.
74. Hollingsworth, P.; Forrest, L.L.; Spouge, J.L.; Hajibabaei, M.; Ratnasingham, S.; van der Bank, M.; Chase, M.W.; Cowan, R.S.; Erickson, D.L.; Fazekas, A.J.; et al. A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12794–12797, doi:10.1073/pnas.0905845106.
75. Li, F.-W.; Kuo, L.-Y.; Rothfels, C.J.; Ebihara, A.; Chiou, W.-L.; Windham, M.D.; Pryer, K.M. rbcL and matK Earn Two Thumbs Up as the Core DNA Barcode for Ferns. *PLoS ONE* **2011**, *6*, e26597.
76. Dong, W.; Xu, C.; Li, C.; Sun, J.; Zuo, Y.; Shi, S.; Cheng, T.; Guo, J.; Zhou, S. ycf1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* **2015**, *5*, 8348, doi:10.1038/srep08348.
77. Goffinet, B.; Wickett, N.J.; Shaw, A.J.; Cox, C.J. Phylogenetic significance of the rpoA loss in the chloroplast genome of mosses. *Taxon* **2005**, *54*, 353–360, doi:10.2307/25065363.
78. Gabaldón, T.; Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **2013**, *14*, 360–366, doi:10.1038/nrg3456.
79. Jensen, R.A. Orthologs and paralogs—We need to get it right. *Genome Biol.* **2001**, *2*, interactions 1002.1-1002.3 doi: 10.1186/gb-2001-2-8-interactions1002.
80. Sonnhammer, E.L.L.; Koonin, E. V. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **2002**, *18*, 619–620, doi:10.1016/S0168-9525(02)02793-2.
81. Palenik, B.; Grimwood, J.; Aerts, A.; Rouzé, P.; Salamov, A.; Putnam, N.; Dupont, C.; Jorgensen, R.; Derelle, E.; Rombauts, S.; et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7705–7710, doi:10.1073/pnas.0611046104.
82. Suzuki, J.Y.; Bauer, C.E. Light-Independent Chlorophyll Biosynthesis: Involvement of the Chloroplast Gene chlL (frxC). *Plant Cell* **2007**, *4*, 929, doi:10.2307/3869460.
83. Fujita, Y.; Takagi, H.; Hase, T. Identification of the chlB gene and the gene product essential for the light-independent chlorophyll biosynthesis in the cyanobacterium Plectonema boryanum. *Plant Cell Physiol.* **1996**, *37*, 313–323, doi:10.1093/oxfordjournals.pcp.a028948.
84. Wu, Q.; Yu, J.; Zhao, N. Partial recovery of light-independent chlorophyll biosynthesis in the chlL-deletion mutant of Synechocystis sp. PCC 6803. *IUBMB Life* **2001**, *51*, 289–293, doi:10.1080/152165401317190789.
85. Burke, D.H.; Raubeson, L.A.; Alberti, M.; Hearst, J.E.; Jordan, E.T.; Kirch, S.A.; Valinski, A.E.C.; Conant, D.S.; Stein, D.B. The chlL (frxC) gene: Phylogenetic distribution in vascular plants and DNA sequence from

Polystichum acrostichoides (Pteridophyta) and *Synechococcus* sp. 7002 (Cyanobacteria). *Plant Syst. Evol.* **1993**, *187*, 89–102, doi:10.1007/BF00994092.

86. Kapoor, M.; Wakasugi, T.; Yoshinaga, K.; Sugiura, M. The chloroplast chlL gene of the green alga *Chlorella vulgaris* C-27 contains a self-splicing group I intron. *Mol. Gener. Gen.* **1996**, *250*, 655–664, doi:10.1007/BF02172976.
87. Karpinska, B.; Karpinski, S.; Hällgren, J.E. The chlB gene encoding a subunit of light-independent protochlorophyllide reductase is edited in chloroplasts of conifers. *Curr. Gen.* **1997**, *31*, 343–347, doi:10.1007/s002940050214.
88. Stegemann, S.; Hartmann, S.; Ruf, S.; Bock, R. High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 8828–8833, doi:10.1073/pnas.1430924100.
89. Baldauf, S.L.; Palmer, J.D. Evolutionary transfer of the chloroplast tufA gene to the nucleus. *Nature* **1990**, *344*, 262–265, doi:10.1038/344262a0.
90. Martin, W.; Herrmann, R.G. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? *Plant Physiol.* **1998**, *118*, 9–17, doi:10.1104/pp.118.1.9.