Gene Loss and Evolution of the Plastome

Tapan Kumar Mohanta*[1], Adil Khan[1], Abdullatif Khan[1], Elsayed Fathi Abd_Allah[2], Ahmed Al-Harrasi[1]

[1]Natural and Medical Sciences Research Center, University of Nizwa, Nizwa, 616, Oman

[2]Plant Production Department, College of Food and Agricultural Sciences, King Saud University, P.O. Box. 2460 Riyadh 11451, Saudi Arabia.

*Corresponding author

Tapan Kumar Mohanta, e-mail: nostoc.tapan@gmail.com, tapan.mohanta@unizwa.edu.om

Abstract

Chloroplasts are unique organelles within plant cells and are ultimately 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 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 CDS in the chloroplasts of algae are higher relative to other lineages. Approximately 10.31% of the examined species have lost the inverted repeats (IR) that span across the lineages that comprise algae, bryophytes, pteridophytes, gymnosperm, angiosperms, magnoliids, and protists. Genome-wide analyses revealed that the loss of the *RBCL* gene in parasitic and heterotrophic plant species occurred approximately 56 Ma ago. *PsaM, Psb30, ChlB, ChlL, ChlN,* and *RPL21* were found to be characteristic signature genes of algae, bryophytes, pteridophytes, and gymnosperms; while 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 encoding 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 analysed species; suggesting that it is a signature gene of the plastome. Our evolutionary analysis indicated that chloroplast genomes evolved from multiple common ancestors and have undergone significant recombination events across different taxonomic lineages. Additionally, our findings support the hypothesis that these recombination events are the most probable cause behind the dynamic loss of chloroplast genes and inverted repeats in different species.

Key words

Chloroplast genome, Plastome, Evolution, Deletion, Duplication, Recombination, Nucleotide substitution

Introduction

Photosynthesis is a process by which autotrophic plants utilize chlorophyll to transform solar energy into chemical energy. 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. This essential process occurs inside a semi-autonomous organelle, commonly known as a plastid or chloroplast. Current knowledge indicates that the origin and evolution of plastids occurred through the endosymbiosis of ancestral cyanobacteria with non-photosynthesizing cells that dates back to 1.5 to 1.6 billion years ago [1,2]. 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 [2]. 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 [2]. Throughout this evolutionary time scale, the endosymbiont retained its existence inside the cell and its dominant function of photosynthesis without undergoing any evolutionary changes. In addition to photosynthesis, this semi-autonomous organelle also plays an important role in the biosynthesis of amino acids, lipids, carotenoids, and other important biomolecules. Studies indicate that the plastid genome has retained a complete set of protein synthesizing machinery and encodes approximately 100 proteins. 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. The size of the plastid genome of land plants is reported to range from 120 to 190 kb [3–5]. 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 [6]. The size of the IR region ranges from 10-15 kb in non-seed plants to 20-30 kb in angiosperms [6–9]. The IR A and B regions are reported to share a conserved molecular evolutionary pattern [10,11]. 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 [12]. The majority of genes in the chloroplast genome have been either functionally transferred to the nuclear genome or lost during evolution [13,14]. For example, the functional genes *tufA, ftsH, odpB,* and *Rpl5* have been transferred from the plastome to the nucleus [15,16]. Structural rearrangements of the plastid genome have occurred throughout its' evolution; resulting in expansion, contraction, or loss of genetic content [5]. 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 [7,17–20]. Changes in the architecture of the IR regions can affect the entire plastid chromosome and its immediate neighbourhood. For example, several genes associated with the SC region got duplicated, including *Ycf2,* due to the relocation of the IR region [5]. 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 2357 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.

Results

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 (335), austrobaileyales (2), bryophytes (24), chloranthales (2), corals (2), eudicots (1314), Flacourtiaceae (1), gymnosperms (95), magnoliids (67), monocots (570), nymphaeales (14), opisthokonta (1), protists (31), pteridophytes (52), and an unclassified chloroplast genome (1) (Supplementary File 1). A comparison of the analysed genomes indicated that *Haematococcus lacustris* encoded the largest chloroplast genome, comprising 1.352 Mbs; while *Asarum minus* encoded the smallest chloroplast genome, comprising 0.0155 Mbs (Figure 1). 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 it was in 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 found to be 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).

The number of coding sequences (CDS) in the analyzed chloroplast genomes ranged from 273 (*Pinus koraiensis*) to 4 (*Monoraphidium neglectum*) (Figure 1). The overall average of number of CDS in the chloroplast genome was found to be 91.67 per genome. Some other species, however, were found to contain 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 1). Similarly, some of the species were found to encode a lower number of CDS in the chloroplast genome; including *Asarum minus* (7), *Cytinus hypocistis* (15), *Sciaphila densiflora* (18), *Gastrodia elata* (20), *Burmannia oblonga* (22), *Orobanche gracilis* (24), and others (Supplementary File 1). 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 GC content of the analysed chloroplast genomes ranged from a high of 57.66% (*Trebouxiophyceae* sp. MX-AZ01) to a low of 23.25% (*Bulboplastis apyrenoidosa*) (Figure 4, Supplementary File 1). The average GC content in the chloroplast genome was 36.87%. Some species contained a higher percentage of GC content, including *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 1). 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 also grouped together (Figure 5). The percentage GC content in bryophytes and pteridophytes did not group with the algae and protists or the eudicots, gymnosperms, magnoliids, monocots, or Nymphaeales (Figure 5).

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, 94 were found to possess the *PsaM* gene. All of the species those found to possess the *PsaM* gene were algae, bryophytes, pteridophytes, or gymnosperms (Supplementary File 2). Notably, no species in the angiosperm lineage possessed the *PsaM* gene; clearly indicating that the *PsaM* gene was lost in the angiosperm lineage. The size of the PsaM deduced proteins ranged from 22 to 33 amino acids. The deduced molecular weight (MW) of PsaM protein ranged from 3.568 kDa in *Mesotaenium endlicherianum* to 1.59 kDa in *P. bungeana*, while the isoelectric point ranged from 3.37 in *Taxodium distichum* to 9.95 in *P. bungeana*. The MW of PsaM proteins in species of *Picea* were smaller than the *Cycas, Ginkgo*, and species from bryophytes, pteridophytes, and algae. 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 1). 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 1). Phylogenetic analysis revealed that PsaM genes grouped into five independent clusters, suggesting that they have evolved independently from multiple common ancestors (Supplementary Figure 2A). Duplication and deletion analysis of *PsaM* genes revealed that deletion events were more predominant than the duplication or co-divergence events (Table 1). Among the 84 analysed *PsaM* genes, 12 had undergone duplication and 34 had undergone deletions, while 34 genes had undergone co-divergence (Table 1, Supplementary Figure 2B). The upper and lower boundaries of the time of each duplication and co-divergence revealed that the Jugermanniopsida, Pinaceae, and Streptophyta were in the upper boundary while the Eukaryota, Pinaceae, Streptophyta, Cycadales, Podocarpaceae, *Apopellia endiviifolia*, and Zygnematophyceae were in the lower boundary (Supplementary Figure 2B). The lower boundary represents the oldest species where duplications must have occurred and the upper boundary represents the most recent species where duplication events are not present.

Psb30 encodes the Ycf12 protein, which is essential for the functioning of the photosystem II reaction centre. The size of the translated protein in the analysed chloroplast genomes ranged from 24 (*Pinus nelsonii*) to 34 amino acids (*Isoetes flaccida*). The MW of Ycf12 ranged from 3.75 kDa in *Schizaea pectinata* to 2.46 kDa in *P. nelsonii* and the predicted isoelectric point ranged from 3.13 in *P. nelsonii* to 10.613 in *Cylindrocystis brebissonii*. A 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 3). A total of 164 species were found to possess *psb30* gene and all of the species were belonged to algae, bryophytes, pteridophytes, or gymnosperms (Supplementary File 2). *Psb30* was absent in the chloroplast genome of angiosperms. Phylogenetic analysis of *Psb30* genes resulted in the designation of two major clusters and six minor clusters, suggesting that it evolved from multiple common ancestors (Supplementary Figure 4A). Deletion/duplication analysis indicated that 39 *Psb30* genes had undergone a duplication event and 120 had undergone a deletion event, while 49 were found to be co-diverged (Table 1, Supplementary Figure 4B). The upper boundary species, where duplication events were absent, belonged to the Streptophyta, Pinaceae, Polypodiopsida, Mesotaeniaceae, Zygnematophyceae, Zamiaceae, and Eukaryota. Species in the lower boundary group, where duplication events must have occurred, belonged to the Eukaryota, Streptophyta, Pinaceae, *Cathaya argyrophylla*, Cycadales, *Dioon spinulosum*, *Anthoceros formosae*, Pteridaceae, Aspleniaceae, Polypodiales, Zygnematales, *C. brebissonii*, and Viridiplantae.

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 2) among protists, algae, bryophytes, pteridophytes, and gymnosperms. The *ChlB* gene was absent in species in the Chloranthales, corals, or angiosperm lineage. The predicted size of the ChlB proteins ranged from 724 (*Gonium pectorale*) to 177 (*Welwitschia mirabilis*) amino acids. The MW of ChlB protein ranged from 20.99 (*W. mirabilis*) to 79.89 (*G. pectorale*) kDa, while the isoelectric point ranged from 5.10 (*Sequoia sempervirens*) to 10.74 (*S. pectinata*). A multiple sequence alignment revealed the presence of several

highly conserved amino acid motifs (Supplementary Figure 8). 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 5). A phylogenetic analysis indicated that *ChlB* genes grouped into two major clusters and thirteen minor clusters, reflecting multiple evolutionary ancestor (Supplementary Figure 6A). *ChlB* genes were comprised of a few groups. Specifically, deletion and duplication analysis revealed that 35 *ChlB* genes had undergone duplications and 126 had undergone deletions, while 116 exhibited co-divergence in their evolutionary history (Table 1, Supplementary Figure 6B). The lower time boundary contained species where duplication events must have occurred. These taxa included members of the Viridiplantae, Streptophyta, Cycadales, *D. spinulosum*, *Ampelopteris prolifera*, Polypodiales, *A. endiviifolia*, Zygnematophyceae, Chlorophyta, Trebouxiophyceae, Hydrodictyaceae, Bryopsidales, and *Pseudochloris wilhelmii*. The upper time boundary included taxa where duplication events were not present. These included members of the Streptophyta, Zamiaceae, Thelypteridaceae, Jungermanniopsida, Zygnematophyceae, Chlorophyta, Trebouxiophyceae, Sphaeropleales, and Ulvophyceae.

ChlL is a light-independent protochlorophyllide reductase iron-sulphur ATP-binding protein that functions in the reduction of ferredoxin, reducing the D ring of protochlorophyllide to form chlorophyllide. It plays a role in the light-independent reaction of photosynthesis and the L component serves as the electron donor to the NB-component. This protein is also involved in light-independent biosynthesis of chlorophyll. Analysis of the chloroplast genome sequences identified 303 species that possess *ChlL* genes (Supplementary File 2). All of the identified species those possess *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. The length of the predicted ChlL proteins ranged from 186 amino acids in *Pycnococcus provasolii* to 299 amino acids in *Macrothelypteris torresiana*. The MW of ChlL proteins ranged from 20.19 (*Pycnococcus provasolii*) to 33.08 (*Retrophyllum piresii*) kDa, while the isoelectric point ranged from 11.74 (*Hypodematium crenatum*) to 4.53 (*Leptosira terrestris*). A 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 7). The phylogenetic analysis indicated that *ChlL* genes grouped into one major independent cluster and eleven minor clusters, suggesting that they also evolved independently from different common ancestors (Supplementary Figure 8A). Deletion and duplication analysis indicated that 49 *ChlL* genes had undergone duplication events and 184 had undergone deletions, while 100 *ChlL* genes exhibited co-divergence (Table 1, Supplementary Figure 8B). The lower boundary contains taxa where duplication events must have occurred. These include members of the Viridiplantae, Streptophyta, Cycadales, *D. spinulosum*, Polypodiopsida, Pteridaceae, Diplaziopsidaceae, Zygnematophyceae, Chlorophyta, Hydrodictyaceae, Trebouxiophyceae, and *P. wilhelmii*. The upper boundary includes taxa where deletion events must have occurred. These include members of the Streptophyta, Zamiaceae, Polypodiopsida, Thelypteridaceae, Chlorophyta, and Trebouxiophyceae.

ChlN protein is a dark-operative, light-independent, protochlorophyllide reductase. It utilizes Mg^{2+}-ATP mediated reduction of ferredoxin to reduce the D ring of protochlorophyllide that is subsequently converted to chlorophyllide. At least 289 of the analyzed chloroplast genomes possess *ChlN* genes. These genomes were from taxa within the protists, algae, bryophytes, pteridophytes, and gymnosperms (Supplementary File 2). The length of the predicted ChlN proteins range from 373 (*Toxarium undulatum*) to 523 (*Chlorella mirabilis*) amino acids and have a predicted MW ranging from 44.93 (*P. provasolii*) to 58.86 (*C. mirabilis*) kDa, while the isoelectric point ranges from 4.92 (*Ginkgo biloba*) to 9.83 (*R. piresii*). A 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 9). Phylogenetic analysis revealed that *ChlN* genes group into two independent clusters (Supplementary Figure 10A). No lineage specific grouping, however, was identified in the phylogenetic tree. Deletion and duplication analysis indicated that 8 *ChlN* genes had undergone duplication events, 46 had undergone deletion events and 34 genes exhibited co-divergence (Table 1, Supplementary Figure 10B). The lower boundary, which indicates where duplication events must have occurred, contained members of the

Chlorophyta, Polypodiales, and *A. prolifera*. The upper boundary contains taxa where deletion events must have occurred and included members of the Thelypteridaceae, Polypodiales, and Chlorophyta.

Rpl21 protein is a component of the 60S ribosomal subunit. The chloroplast genomes of at least 137 of the examined species were found to possess a *RpL21* gene and included taxa within the algae, bryophytes, pteridophytes, and gymnosperms (Supplementary File 2). In the majority of cases, however, the CDS of the *Rpl21* genes were 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 11). Phylogenetic analysis shows the presence of three clusters, reflecting their origin from multiple common ancestors (Supplementary Figure 12A). Deletion/duplication analysis indicted that 3 *RpL21* genes had undergone duplication events, 8 had undergone deletion events, and 9 exhibited co-divergence (Table1, Supplementary Figure 12B).

The *RBCL* gene has been lost in parasitic and heterotrophic plant species

Ribulose-1,5-biphosphate carboxylase (RBCL), the most abundant enzyme of the earth, functions in the process of carbon fixation during the dark reaction of photosynthesis to produce carbohydrate. The presence of the *RBCL* gene and its role in carbon assimilation contributes to the role of plants as producers in natural ecosystems. Therefore, a characteristic feature of almost all photosynthetic plants is that their chloroplast genome encodes an *RBCL* gene to modulate photosynthesis. However, not all the plants with chloroplasts/plastids, possess an *RBCL* gene. Among the plant taxa analysed, we identified at least 17 species that did not encode an *RBCL* gene in their chloroplast genome. The species lacking an *RBCL* gene were *Alveolata* sp. CCMP3155, *A. minus*, *Bathycoccus prasinos* (picoplankton), *Burmannia oblonga* (orchid), *Codonopsis lanceolate* (eudicot), *Cytinus hypocistis* (parasite), *Gastrodia elata* (saprophyte), *Monotropa hypopitys* (myco-heterotroph), *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* (myco-heterotroph). The length of the RBCL protein in taxa that did possess an *RBCL* gene ranged from 334 (*Asterionellopsis glacialis*) to 493 (*Gossypium australe*) amino acids and the MW of the RBCL protein ranged from 36.809 (*Nannochloropsis oceanica*) to 54.87 (*G. australe*) kDa; while the isoelectric point ranged from 4.47 (*Primula sinensis*) to 7.79 (*Oogamochlamys gigantea*) (Supplementary Figure 13). An analysis of the composition of the RBCL protein revealed that Ala and Gly were the most abundant amino acids and that Trp was the least abundant.

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 3). IR deletion events were identified in protists (14), protozoans (1), algae (126), bryophytes (1), gymnosperms (64), magnoliids (1) monocots (9), 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 (Figure 7). 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 (Figure 6). The middle cluster contained at least four major phylogenetic groups (Figure 6). 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 the different groups of the phylogenetic tree, the majority of the algal species were present in a single group (yellow) (Figure 6). A phylogenetic tree of taxa with an IR-deleted chloroplast genome and taxa with chloroplast genomes that did not lose the IR region (*Floydiella terrestris*, *Carteria cerasiformis*, *B. apyrenoidosa*, *E.*

grandis, O. 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 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 had undergone vivid recombination (Supplementary Figure 14A and 14B). In addition, the IR deleted chloroplast genomes were also undergone vivid recombination (Supplementary Figure 15).

Several genes in the chloroplast genome have been subject to deletion events.

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- biphosphate 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 of 2511 chloroplast genomes revealed that a number of these genes were lost in one or more species (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).

Chloroplast genomes possess genes that encode at least six different ATP molecules, namely, AtpA, AtpB, AtpE, AtpF, AtpH, and AtpI. Among the 2511 analysed chloroplast genomes, *AtpA*, *AtpB*, *AtpE*, *AtpF*, *AtpH*, and *AtpI* were found to be lost in 8, 8, 12, 14, 13, and 12 species, respectively (Table 2, Supplementary File 4). The loss of *Atp* genes occurred in algae, eudicots, magnoliids, and monocots, while no loss of Atp genes occurred in any species of bryophytes, pteridophytes, and gymnosperms (Supplementary File 4). The *AccD* gene in chloroplasts, which encodes the beta-carboxyl transferase subunit of acetyl Co-A carboxylase (ACC) complex, was found to be lost from 387 plant species (Supplementary File 4). The *AccD* gene was lost in taxa belonging to algae (92 species), eudicots (32 species), gymnosperms (7 species), magnoliids (1 species), monocots (227 species), and protists (27 species), while the *AccD* gene was found to be present in all bryophytes and pteridophytes.

The *CcsA* gene in the chloroplast genome encodes a cytochrome C biogenesis protein. *CcsA* genes were found to be lost in at least 29 species (Table 2). It was lost in taxa belonging to members of algae (8 species), bryophyte (3 species), eudicots (6 species), magnoliids (1 species), monocots (5 species), and protists (6 species), while no evidence of a loss was observed in members of the pteridophytes and gymnosperms. The *CemA* gene encodes a chloroplast envelope membrane protein and was found to be lost in 29 species (Table 2). The loss of the *CemA* gene was found in algae, eudicots, magnoliids, monocots, and protists, while no evidence of deletion was observed in taxa of bryophytes, pteridophytes, and gymnosperms. The *ClpP* gene encodes an ATP-dependent Clp protease proteolytic subunit that is necessary for ATP hydrolysis. It was observed to be lost in at least 142 species (Supplementary File 4) belonging to the algae (108 species), eudicots (2 species), gymnosperms (3 species), magnoliids (1 species), and protists (28 species). Loss of the *ClpP* gene was not found in members of bryophytes, pteridophytes, and monocots (Supplementary File 4). The chloroplast genome possesses at least six different *Psa* genes, *PsaA*, *PsaB*, *PsaC*, *PsaI*, *PsaJ*, and *PsaM* (Table). The *PsaA* gene was absent in 16 species (3 algae, 8 eudicots, 1 magnoliid, and 4 monocots). *PsaB* was lost in 10 species (3 algae, 2

eudicots, 1 magnoliid, and 4 monocots). *PsaC* was lost in 19 species (2 algae, 10 eudicots, 1 magnoliid, and 6 monocots). *PsaI* was absent in 72 species (42 algae, 18 eudicots, 1 magnoliid, 6 monocots, and 5 protists), and *PsaJ* was lost in 24 species (6 algae, 12 eudicots, 1 gymnosperm, 1 magnoliid, 3 monocots, and 1 protist). The *PsaM* gene was lost in 2214 species distributed amongst all of the angiosperm lineages examined in our analysis. The chloroplast encodes 16 different *Psb* genes (*PsbA*, *PsbB*, *PsbC*, *PsbD*, *PsbE*, *PsbF*, *PsbH*, *PsbI*, *PsbJ*, *PsbK*, *PsbL*, *PsbM*, *PsbN*, *PsbT*, *PsbZ*, and *Psb30*). The analysis indicated that PsbA, PsbB, PsbC, PsbD, PsbE, PsbF, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM, PsbN, PsbT, PsbZ, and *Psb30* were lost in a variety of species. Evidence for the deletion of these genes was observed as follows: *PsbA* was lost in 12 species (2 algae, 6 eudicots, and 4 monocots); *PsbB* was lost in 18 species (2 algae, 11 eudicots, 1 magnoliid, and 4 monocots); *PsbC* was lost in 16 species (2 algae, 7 eudicots, 1 magnoliid, and 6 monocots); *PsbD* was lost in 17 species (2 algae, 9 eudicots, 1 magnoliid, and 5 monocots); *PsbE* and *PsbF* were lost in 21 species (2 algae, 12 eudicots, 1 magnoliid, and 6 monocots in both cases); *PsbH* was lost in 20 species (2 algae, 14 eudicots, 1 magnoliid and 3 monocots); *PsbI* was lost in 18 species (3 algae, 10 eudicots, and 5 monocots); *PsbJ* was lost in 21 species (2 algae, 12 eudicots, 1 magnoliid, 6 monocots); *PsbK* was lost in 13 species (2 algae, 5 eudicots, and 6 monocots); *PsbL* was lost in 22 species (2 algae, 12 eudicots, 1 magnoliid, 6 monocots, and 1 protist); *PsbM* was lost in 158 species (115 algae, 9 eudicots, 6 monocots, and 27 protists); *PsbN* was lost in 22 species (2 algae, 14 eudicots, 1 magnoliid, and 5 monocots); *PsbT* was lost in 23 species (3 algae, 14 eudicots, and 6 monocots); and *PsbZ* was lost in 31 species (14 algae, 7 eudicots, 1 magnoliid, 4 monocots, and 5 protists) (Supplementary File 4).

The *Ndh* gene encodes a NAD(P)H-quinone oxidoreductase that shuttles electrons from plastoquinone to quinone in the photosynthetic chain reaction. The analysis indicated that the chloroplast genome encodes at least 11 *Ndh* genes, *NdhA*, *NdhB*, *NdhC*, *NdhD*, *NdhE*, *NdhF*, *NdhG*, *NdhH*, *NdhI*, *NdhJ*, and *NdhK* (Table 2). Evidence for the deletion of these genes was observed as follows: *NdhA* was lost in: 339 species (207 algae, 1 bryophyte, 35 eudicots, 37 gymnosperms, 2 magnoliids, 27 monocots, 28 protists, and 2 pteridophytes); *NdhB* was lost in 258 species (211 algae, 5 eudicots, 7 gymnosperms, 1 magnoliid, 3 monocots, 29 protists, and 2 pteridophytes); *NdhC* was lost in 339 species (212 algae, 38 eudicots, 7 gymnosperms, 2 magnoliids, 49 monocots, 29 protists, and 2 pteridophytes); *NdhD* was lost in 293 species (214 algae, 28 eudicots, 9 gymnosperms, 1 magnoliid, 9 monocots, 30 protists, and 2 pteridophytes); *NdhE* was lost in 322 species (218 algae, 30 eudicots, 19 gymnosperms, 1 magnoliid, 20 monocots, 30 protists, 4 pteridophytes); *NdhF* was lost in 346 species (207 algae, 37 eudicots, 37 gymnosperms, 1 magnoliid, 33 monocots, 30 protists, and 2 pteridophytes); *NdhG* was lost in 335 species (213 algae, 1 bryophyte, 35 eudicots, 37 gymnosperms, 2 magnoliids, 16 monocots, 30 protists, and 1 pteridophyte); *NdhH* was lost in 322 species (213 algae, 1 bryophyte, 34 eudicots, 15 gymnosperms, 1 magnoliid, 26 monocots, 30 protists, and 2 pteridophytes); *NdhI* was lost in 378 species (213 algae, 1 bryophyte, 43 eudicots, 37 gymnosperms, 2 magnoliids, 50 monocots, 30 protists, and 2 pteridophytes); *NdhJ* was lost in 340 species (215 algae, 40 eudicots, 37 gymnosperms, 2 magnoliids, 15 monocots, 30 protists, and 2 pteridophytes); and *NdhK* was lost in 331 species (204 algae, 1 bryophyte, 39 eudicots, 7 gymnosperms, 2 magnoliids, 46 monocots, 30 protists, and 2 pteridophytes) (Supplementary File 4). The loss of *Ndh* genes was found to occur in members of algae, bryophytes, pteridophytes, gymnosperms, monocots, eudicots, magnoliids, and protists.

The chloroplast genome encodes *PetA* (cytochrome f precursor), *PetB* (cytochrome b6), *PetD* (cytochrome b6-f complex subunit 4), *PetG* (cytochrome b6-f complex subunit 5), *PetL* (cytochrome b6-f complex subunit 6), and *PetN* (cytochrome b6-f complex subunit 8). genes. Evidence for deletion of these genes was observed as follows: *PetA* was lost in 33 species (8 algae, 10 eudicot, 1 magnoliid, 6 monocots, and 8 protists); *PetB* was lost in 15 species (2 algae, 8 eudicots, 1 magnoliid, and 4 monocots); *PetD* was lost in 36 species (7 algae, 13 eudicots, 1 magnoliid, 6 monocots, and 9 protists); *PetL* was lost in 71 species (39 algae, 11 eudicots, 1 magnoliid, 4 monocots, and 16 protists); and *PetN* gene was lost in135 species (106 algae, 5 bryophytes, 11 eudicots, 1 magnoliid, 6 monocots, and 6 protists) (Supplementary File 4). *PetA* was lost in taxa of members of algae, eudicots, magnoliids, monocots, and protists, while *PetA* was found to be present in bryophytes, pteridophytes, and gymnosperms. *PetB* gene was found to be lost in taxa of members of algae, eudicots, magnoliids, and monocots, while it was found to be present in bryophytes, pteridophytes, and gymnosperms (Supplementary File 4). *PetD* was found to

be lost in taxa of members of algae, eudicots, magnoliids, monocots, and protists, while it was found to be intact in bryophytes, pteridophytes, and gymnosperms. *PetL* was found to be lost in taxa of members of algae, eudicots, magnoliids, monocots, and protists, while it was found to be intact in bryophytes, pteridophytes, and gymnosperms. *PetN* genes were found to be lost in taxa of members of algae, bryophytes, eudicots, magnoliids, monocots, and protists, while it was intact in pteridophytes and gymnosperms.

The chloroplast genome encodes at least nine *Rpl* genes, *Rpl2, Rpl14, Rpl16, Rpl20, Rpl22, Rpl23, Rpl32, Rpl33,* and *Rpl36* (Table 2). Deletion of these genes was found in taxa of different lineages (Table 2). *Rpl2* was lost in two species (1 eudicot and 1 magnoliid). *Rpl14* was lost in four species (2 algae, 1 eudicot, and 1 magnoliid). *Rpl16* was lost in three species (2 algae and 1 magnoliid). *Rpl22* was lost in 127 species (107 algae, 12 eudicots, 3 magnoliids, 2 monocots, and 3 protists). *Rpl32* was lost in 114 species (21 algae, 73 eudicots, 5 gymnosperms, 1 magnoliid, 6 monocots, 8 protists). *Rpl33* was lost in 133 species (111 algae, 7 eudicots, 1 magnoliid, 4 monocots, and 10 protists). *Rpl23* was lost in 24 species (8 algae, 4 eudicots, 6 gymnosperms, 1 magnoliid, and 5 monocots) (Supplementary File 4).

The chloroplast genome encodes 12 *Rps* genes, *Rps2, Rps3, Rps4, Rps7, Rps8, Rps11, Rps12, Rps14, Rps15, Rps16, Rps18,* and *Rps19* (Table). Our analysis indicated that different *Rps* genes were lost from a variety of species (Supplementary Table). Specifically, *Rps2* was lost in three species (1 algae, 1 eudicot, and 1 magnoliid); *Rps3* was lost in three species (2 algae and 1 magnoliid); *Rps4* was lost in four species (3 algae and 1 magnoliid); *Rps7* was lost in three species (1 algae, 1 gymnosperm, and 1 magnoliid); *Rps8* was lost in three species (1 eudicot, 1 magnoliid, and 1 protist); *Rps11* was lost in two species (1 eudicot and 1 magnoliid) and *Rps12* was lost in two species (1 algae and 1 magnoliid). The chloroplast genome encodes four *Rpo* genes, *RpoA, RpoB, RpoC1* and *RpoC2* (Table 2). *RpoA* and *RpoC1* encode for the alpha-subunit and *RpoB* and *RpoC2* encode the beta-subunit of DNA-dependent RNA polymerase. The analysis revealed the loss of *RpoA, RpoB, RpoC1,* and *RpoC2* genes from the chloroplast genome of several taxa (Supplementary File 4). Specifically, *RpoA1* was lost in 26 species (5 algae, 6 bryophytes, 7 eudicots, 1 magnoliid, 4 monocots, and 3 protists); *RpoB* was lost in 19 species (1 algae, 14 eudicots, 1 magnoliid, and 3 monocots); *RpoC1* was lost in 21 species (15 eudicots, 1 magnoliid, 5 monocots) and *RpoC2* was lost in 13 species (1 algae, 7 eudicots, 1 magnoliid, and 4 monocots). The loss of *RpoA* occurred across diverse lineages including algae, bryophytes, eudicots, magnoliids, monocots, and protists. Additionally, *RpoB* was lost in algae, eudicots, magnoliids, and monocots; *RpoC1* was lost in eudicots, magnoliids, and monocots and *RpoC2* was lost in eudicots, magnoliids, and monocots (Supplementary File 4).

The majority of chloroplast genomes encode four *Ycf* genes, *Ycf1, Ycf2, Ycf3,* and *Ycf4*. Our analysis indicated a dynamic loss of *Ycf* genes from the chloroplast genome of a variety of taxa (Supplementary File 4). *Ycf1* was lost in 161 species (125 algae, 4 eudicots, 1 magnoliid, 3 monocots, and 28 protists), *Ycf2* was lost in 219 species (185 algae, 1 eudicot and magnoliid each, 2 monocots, and 30 protists). *Ycf3* was lost in 30 species (7 algae, 7 eudicots, 1 magnoliid, 6 monocots, and 9 protists). *Ycf4* was lost in 39 species (6 algae, 24 eudicots, 1 magnoliid, 5 monocots, and 3 protists). Although researchers have yet to elucidate the function of *Ycf* genes, *Ycf3* and *Ycf4* have been reported to be a photosystem I assembly factor. The loss of *Ycf1* and *Ycf2* genes was more prominent in algae and the loss of *Ycf1* and *Ycf2* genes were not found in bryophytes, pteridophytes, and gymnosperms. The loss of *Ycf4* was most prominent in eudicots and the loss of *Ycf3* and *Ycf4* was not observed in bryophytes, pteridophytes, and gymnosperms (Supplementary File 4).

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 lost in algae, eudicots, magnoliids, and monocots (Supplementary Table 1). Only a small number of genes were lost in bryophytes, gymnosperms, protists, and pteridophytes (Supplementary Table 1). 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* were found to be lost in all six lineages; while *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 16, Supplementary Table 1). *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 1, Supplementary Figure 17). The most common 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; while 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 18, Supplementary Table 1). *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.

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. Therefore, a genome-wide analysis of fully sequenced and annotated genomes of 145 plant species was analysed to explore this question. Results indicated the presence of almost all of the chloroplast encoding genes in the nuclear genome. 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 5).

The ratio of nucleotide substitution is highest in Pteridophytes and lowest in Nymphaeales

Determining the rate of nucleotide substitution in the chloroplast genome is 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 2). The ratio of nucleotide substitution in species with deleted IR regions was 2.951 (k1) and 3.42 (k2) (Supplementary Table 2). 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 2). 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 2). 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 2). 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; while high rates of transition/transversion were not observed in bryophytes, gymnosperms, monocots, and dicots (Supplementary Table 2). Notably, G > A transitions were more prominent in chloroplast genomes with deleted IR regions (Supplementary Table 2).

Chloroplast genomes have evolved from multiple common ancestors

A phylogenetic tree was constructed to obtain an evolutionary perspective of chloroplast genomes (Figure 7). All of the 2511 studied species were used to construct a phylogenetic tree (Figure 7). The phylogenetic analysis produced four distinct clusters, indicating that chloroplast genomes evolved independently from multiple common ancestors. 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 in other lineages, they were still distributed sporadically throughout the phylogenetic tree. Time tree analysis indicated that the origin of the cyanobacterial species in this study those used as out-group 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 (Figure 8); 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 (Figure 8). 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 (Figure 8).

Discussion

Chloroplasts are an indispensable part of plant cells which function as a semi-autonomous organelle due to the presence their own genetic material, potential to self-replicate, and capability to modulate cell metabolism [21–24]. 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. The current analysis of however, does not support this premise. The average number of CDS in algae (140.93) was higher than the average number of CDS found in 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 taxa 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 chloroplast genome size. 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; whereas, this was not true in the other plant lineages where the relationship was inversely proportional. Deletion of IR regions has been previously reported in a few species of algae, magnoliids, and other genomes [25–29]. 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 increase in the genome size in algae has largely been attributed to a duplication of chloroplast genome. The evolutionary age of IR-deleted species of algae dates back to ~965-850 Ma. This provides a 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. Zhu et al. (2015) has also suggested a role for duplication in evolution of IR-deleted chloroplast genomes [28]. Characterizing the pattern and frequency of neutral mutations (substitution, insertions, and deletion) is important for deciphering the molecular basis of evolution of genes and genomes. Turmel et al., (2017) 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 [25]. 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 2). 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 analysed. The loss of important genes from the chloroplast genome has been previously reported in some species of green algae, bryophytes, and magnoliids, [30–33]. The loss of the photosynthetic gene, *RBCL*, in the parasitic plant, *Conopholis*, has also been reported. However, the *RBCL* gene was reported to be present in other parasitic plants in the Orobanchacea [34,35]. The results of the present study indicate the loss of *RBCL* gene in at least 17 species among parasitic, myco-parasitic, and saprophytic plant species across

different lineages, including algae, eudicots, magnoliids, monocots, and protists. 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. 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 *RBCL;* and *vice versa*, the loss of *RBCL* is not a pre-requisite of becoming a parasitic plant as well.

Important genes encoded by the chloroplast genome include *AtpA, AtpB, AtpE, AtpF, AtpH, AtpI, CcsA, ChlB, ChlL, ChlN, ClpP, NdhA, NdhB, NdhC, NdhD, NdhE, NdhF, NdhG, NdhH, NdhI, NdhJ, NdhK, Pbf1, PetA, PetB, PetD, PetG, PetL, PetN, PsaA PsaB, PsaC, PsaI, PsaJ, PsaM, Psb30, PsbA, PsbB, PsbC, PsbD, PsbE, PsbF, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM, PsbT, PsbZ, Rpl2, Rpl14, Rpl16, Rpl20, Rpl21, Rpl22, Rpl23, Rpl32, Rpl33, Rpl36, RpoA, RpoB, RpoC1, RpoC2, Rps2, Rps3, Rps4, Rps7, Rps8, Rps11, Rps12, Rps14, Rps15, Rps16, Rps18, Rps19, Ycf1, Ycf2, Ycf3,* and *Ycf4.* Deletions of one or more of these genes have been observed in numerous chloroplast genomes. It is difficult to decipher the 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 analysed 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. 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 plant and photosynthesis [36]. The loss of *Ndh* genes in species across all 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 sub-families [37]. These data suggest that the loss of *Ndh* genes in the parental lineage of orchids led to the loss of *Ndh* genes in the sub-families 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 the gene loss than monocot species. Monocots and dicots chloroplast genome share in 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 make their use as universal markers questionable in studies for DNA barcoding [38–42].

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 [31,43]. The loss of *Psa* and *Psb* was quite prominent in algae, eudicot, magnoliid, monocot, and protist lineages. *Psa* and *Psb* were always found to be present 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 co-divergence events (Table 1). In fact,

co-divergence 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 co-divergent 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 [44–47]. *Chl* genes are involved in photosynthesis in cyanobacteria, algae, pteridophytes, and conifers [48–53]; 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 intact gene in all 2511 of the studied species, suggesting that *Rpl20* is the most evolutionary conserved gene in the chloroplast genome of 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 frequent 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 [13,24,54–56]. 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 other species. One report also indicated that genes transferred to the nuclear genome may not provide a one to one correspondence in function [56]. 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.

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; 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 of one or another species is quite intriguing. Has the chloroplast genome completed 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.

Materials and methods

Sequence retrieval and annotation

All of the sequenced chloroplast genomes available up until December 2018 were downloaded from National Center for Biotechnology Information (NCBI) and used in the current study to analyse 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 1). All of the individual genomes were subjected to OGDRAW to check for the presence and absence of inverted repeats in the genome [57]. Genomes that were found to lack inverted repeats (IR), as determined by OGDRAW, were further searched in NCBI database to cross verify the absence of IR in their genome. The annotated 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 re-checked manually in the NCBI database. Each chloroplast genome was newly annotated using the GeSeq-annotation of organellar genomes pipeline to further extend the study of gene loss in chloroplast genomes [58]. The combined analysis of NCBI and GeSeq-annotation of 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.

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 sequence manipulation suite (http://www.bioinformatics.org/sms2/translate.html). The resulting protein sequences were subjected to a multiple sequence alignment using the Multalin server to identify conserved amino acid motifs [59]. The CDS sequences of *PsaM, psb30, ChlB, ChlL, ChlN* and *RPL21* genes were also subjected to a 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 [60]. 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 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; no. of bootstrap replications, 1000; model/method, general time reversible model; rates among sites, gamma distributed with invariant sites (G+I); no. 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 programme 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 TimeTree [61]. Cyanobacterial species were used as an outgroup to calibrate the time tree for the other species.

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 trees were uploaded in Notung software v. 2.9 followed by uploading of the species tree and subsequent reconciliation of the gene tree with the species tree [62–64]. Once reconciled, deletion and duplication events for the genes were visualized and noted.

Recombination events and time tree construction of the chloroplast genome

The constructed phylogenetic tree of chloroplast genomes was uploaded in IcyTree [65] to analyze the recombination events that occurred in chloroplast genomes. The recombination events in IR-deleted and non-deleted IR species were studied separately.

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 analysing the rate of nucleotide substitutions. The full-length sequences of chloroplast genomes were subjected to multiple sequence alignment to generate a Clustal file. 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 [60]. The converted files were subsequently uploaded in MEGA6 software to analyse the rate of nucleotide substitution. The following statistical parameters were used to analyse 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.

Statistical analysis

Principal component analysis and probability distribution of chloroplast genomes were conducted using Unscrambler software version 7.0 and Venn diagrams were constructed using InteractiVenn (http://www.interactivenn.net/) [66].

References

1. Hedges SB, Blair JE, Venturi ML, Shoe JL. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. BioMed Central; 2004;4:2.

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

3. Sugiura M. The chloroplast genome. In: Schilperoort RA, Dure L, editors. 10 Years Plant Mol Biol. Dordrecht: Springer Netherlands; 1992. p. 149–68.

4. Yu Q-B, Huang C, Yang Z-N. Nuclear-encoded factors associated with the chloroplast transcription machinery of higher plants. Front Plant Sci. Frontiers Media S.A.; 2014;5:316.

5. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011/03/22. Springer Netherlands; 2011;76:273–97.

6. Kolodner R, Tewari KK. Inverted repeats in chloroplast DNA from higher plants. Proc Natl Acad Sci. 1979;76:41 LP – 45.

7. Wolf PG, Der JP, Duffy AM, Davidson JB, Grusz AL, Pryer KM. The evolution of chloroplast genes and genomes in ferns. Plant Mol Biol. 2011;76:251–61.

8. Raubeson L, Jasen R. Chloroplast genomes of plants. In: Henry R, editor. Divers Evol Plants - Genotypic Phenotypic Var High Plants. Wallingford, United Kingdom: CABI Publishing; 2005. p. 45–68.

9. Wu CS, Lai YT, Lin CP, Wang YN, Chaw SM. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. Mol Phylogenet Evol. 2009;52.

10. Daniell H, Lin C-S, Yu M, Chang W-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. Genome Biology; 2016;17:134.

11. Rivas JD Las, Lozano JJ, Ortiz AR. Comparative Analysis of Chloroplast Genomes: Functional Annotation, Genome-Based Phylogeny, and Deduced Evolutionary Patterns. Genome Res. Cold Spring Harbor Laboratory Press; 2002;12:567–83.

12. Stern DB, Goldschmidt-Clermont M, Hanson MR. Chloroplast RNA Metabolism. Annu Rev Plant Biol. Annual Reviews; 2010;61:125–55.

13. Cullis CA, Vorster BJ, Van Der Vyver C, Kunert KJ. Transfer of genetic material between the chloroplast and nucleus: how is it related to stress in plants?

Ann Bot. 2008/09/18. Oxford University Press; 2009;103:625–33.

14. Eckardt NA. Genomic Hopscotch: Gene Transfer from Plastid to Nucleus. Plant Cell. American Society of Plant Biologists; 2006;18:2865–7.

15. 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;23:1324–38.

16. Gao L, Su Y-J, Wang T. Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. J Syst Evol. John Wiley & Sons, Ltd (10.1111); 2010;48:77–93.

17. Downie SR, Palmer JD. Restriction Site Mapping of the Chloroplast DNA Inverted Repeat: A Molecular Phylogeny of the Asteridae. Ann Missouri Bot Gard. Missouri Botanical Garden Press; 1992;79:266–83.

18. Goulding SE, Wolfe KH, Olmstead RG, Morden CW. Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet MGG. 1996;252:195–206.

19. Plunkett GM, Downie SR. Expansion and Contraction of the Chloroplast Inverted Repeat in Apiaceae Subfamily Apioideae. Syst Bot. American Society of Plant Taxonomists; 2000;25:648–67.

20. Guisinger MM, Kuehl J V, Boore JL, Jansen RK. Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage. Mol Biol Evol. 2011;28:583–600.

21. Taylor DL. Chloroplasts as Symbiotic Organelles. In: Bourne GH, Danlelli JF, Jeon KWBT-IR of C, editors. Int Rev Cytol [Internet]. Academic Press; 1970. p. 29–64. Available from: http://www.sciencedirect.com/science/article/pii/S0074769608612450

22. Trench RK. CHLOROPLASTS: PRESUMPTIVE AND DE FACTO ORGANELLES. Ann N Y Acad Sci [Internet]. John Wiley & Sons, Ltd (10.1111); 1981;361:341–55. Available from: https://doi.org/10.1111/j.1749-6632.1981.tb54376.x

23. Stern DS, Higgs DC, Yang J. Transcription and translation in chloroplasts. Trends Plant Sci [Internet]. 1997;2:308–15. Available from: http://www.sciencedirect.com/science/article/pii/S1360138597899530

24. Osteryoung KW, Weber APM. Plastid Biology: Focus on the Defining Organelle of Plants. Plant Physiol [Internet]. 2011;155:1475 LP – 1476. Available from: http://www.plantphysiol.org/content/155/4/1475.abstract

25. Turmel M, Otis C, Lemieux C. Divergent copies of the large inverted repeat in the chloroplast genomes of ulvophycean green algae. Sci Rep [Internet]. Nature Publishing Group UK; 2017;7:994. Available from:

https://www.ncbi.nlm.nih.gov/pubmed/28428552

26. Wolfe KH. The site of deletion of the inverted repeat in pea chloroplast DNA contains duplicated gene fragments. Curr Genet [Internet]. 1988;13:97–9. Available from: https://doi.org/10.1007/BF00365763

27. Strauss SH, Palmer JD, Howe GT, Doerksen AH. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. Proc Natl Acad Sci U S A [Internet]. 1988;85:3898–902. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2836862

28. Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol [Internet]. John Wiley & Sons, Ltd (10.1111); 2016;209:1747–56. Available from: https://doi.org/10.1111/nph.13743

29. 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.

30. Lemieux C, Otis C, Turmel M. Comparative Chloroplast Genome Analyses of Streptophyte Green Algae Uncover Major Structural Alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. Front Plant Sci [Internet]. 2016;7:697. Available from: https://www.frontiersin.org/article/10.3389/fpls.2016.00697

31. 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 Acids Res [Internet]. Oxford University Press; 2003;31:5324–31. Available from: https://www.ncbi.nlm.nih.gov/pubmed/12954768

32. Sinn BT, Sedmak DD, Kelly LM, 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 [Internet]. John Wiley & Sons, Ltd; 2018;105:71–84. Available from: https://doi.org/10.1002/ajb2.1001

33. Alverson AJ, Ruck EC, Theriot EC, Nakov T, Jansen RK. Serial Gene Losses and Foreign DNA Underlie Size and Sequence Variation in the Plastid Genomes of Diatoms. Genome Biol Evol [Internet]. 2014;6:644–54. Available from: https://dx.doi.org/10.1093/gbe/evu039

34. Wolfe AD, dePamphilis CW. The effect of relaxed functional constraints on the photosynthetic gene rbcL in photosynthetic and nonphotosynthetic parasitic plants. Mol Biol Evol [Internet]. 1998;15:1243–58. Available from: https://dx.doi.org/10.1093/oxfordjournals.molbev.a025853

35. Wicke S, Müller KF, de Pamphilis CW, Quandt D, Wickett NJ, Zhang Y, et al.

Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. Plant Cell [Internet]. 2013/10/18. American Society of Plant Biologists; 2013;25:3711–25. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24143802

36. Martín M, Sabater B. Plastid ndh genes in plant evolution. Plant Physiol Biochem [Internet]. 2010;48:636–45. Available from: http://www.sciencedirect.com/science/article/pii/S0981942810001099

37. Lin C-S, Chen JJW, Huang Y-T, Chan M-T, Daniell H, Chang W-J, et al. The location and translocation of ndh genes of chloroplast origin in the Orchidaceae family. Sci Rep [Internet]. The Author(s); 2015;5:9040. Available from: https://doi.org/10.1038/srep09040

38. Heckenhauer J, Barfuss MHJ, Samuel R. Universal multiplexable matK primers for DNA barcoding of angiosperms. Appl Plant Sci [Internet]. Botanical Society of America; 2016;4:apps.1500137. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27347449

39. Yu J, Xue JH, Zhou SL. New universal matK primers for DNA barcoding angiosperms. J Syst Evol. 2011;49:176–81.

40. Hollingsworth P, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, et al. A DNA barcode for land plants. Proc Natl Acad Sci [Internet]. 2009;106:12794 LP – 12797. Available from: http://www.pnas.org/content/106/31/12794.abstract

41. Li F-W, Kuo L-Y, Rothfels CJ, Ebihara A, Chiou W-L, Windham MD, et al. rbcL and matK Earn Two Thumbs Up as the Core DNA Barcode for Ferns. PLoS One [Internet]. Public Library of Science; 2011;6:e26597. Available from: https://doi.org/10.1371/journal.pone.0026597

42. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, et al. ycf1, the most promising plastid DNA barcode of land plants. Sci Rep [Internet]. Nature Publishing Group; 2015;5:8348. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25672218

43. Goffinet B, Wickett NJ, Shaw AJ, Cox CJ. Phylogenetic significance of the rpoA loss in the chloroplast genome of mosses. Taxon [Internet]. John Wiley & Sons, Ltd; 2005;54:353–60. Available from: https://doi.org/10.2307/25065363

44. Gabaldón T, Koonin E V. Functional and evolutionary implications of gene orthology. Nat Rev Genet [Internet]. 2013/04/04. 2013;14:360–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23552219

45. Jensen RA. Orthologs and paralogs - we need to get it right. Genome Biol [Internet]. 2001/08/03. BioMed Central; 2001;2:INTERACTIONS1002–INTERACTIONS1002. Available from: https://www.ncbi.nlm.nih.gov/pubmed/11532207

46. Sonnhammer ELL, Koonin E V. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet [Internet]. Elsevier; 2002;18:619–20. Available from: https://doi.org/10.1016/S0168-9525(02)02793-2

47. 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 [Internet]. 2007;104:7705 LP – 7710. Available from: http://www.pnas.org/content/104/18/7705.abstract

48. Suzuki JY, Bauer CE. Light-Independent Chlorophyll Biosynthesis: Involvement of the Chloroplast Gene chlL (frxC). Plant Cell. 2007;4:929.

49. 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;37:313–23.

50. 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;51:289–93.

51. Burke DH, Raubeson LA, Alberti M, Hearst JE, Jordan ET, Kirch SA, et al. 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.

52. 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 Gen Genet. 1996;250:655–64.

53. Karpinska B, Karpinski S, Hällgren JE. The chlB gene encoding a subunit of light-independent protochlorophyllide reductase is edited in chloroplasts of conifers. Curr Genet. 1997;31:343–7.

54. Stegemann S, Hartmann S, Ruf S, Bock R. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci [Internet]. 2003;100:8828 LP – 8833. Available from: http://www.pnas.org/content/100/15/8828.abstract

55. Baldauf SL, Palmer JD. Evolutionary transfer of the chloroplast tufA gene to the nucleus. Nature [Internet]. 1990;344:262–5. Available from: https://doi.org/10.1038/344262a0

56. Martin W, Herrmann RG. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? Plant Physiol [Internet]. 1998;118:9 LP – 17. Available from: http://www.plantphysiol.org/content/118/1/9.abstract

57. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar

genomes. Nucleic Acids Res [Internet]. 2019;doi.org/10.1093/nar/gkz238. Available from: http://biorxiv.org/content/early/2019/02/10/545509.abstract

58. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res [Internet]. 2017/05/09. Oxford University Press; 2017;45:W6–11. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28486635

59. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res [Internet]. 1988;16:10881–90. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2849754

60. Tamura K, Filipski A, Peterson D, Stecher G, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol [Internet]. 2013;30:2725–9. Available from: https://doi.org/10.1093/molbev/mst197

61. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol [Internet]. 2017;34:1812–9. Available from: https://doi.org/10.1093/molbev/msx116

62. 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 [Internet]. 2012/09/03. Oxford University Press; 2012;28:i409–15. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22962460

63. Darby CA, Stolzer M, Ropp PJ, Barker D, Durand D. Xenolog classification. Bioinformatics [Internet]. 2016/12/29. Oxford University Press; 2017;33:640–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27998934

64. Chen K, Durand D, Farach-Colton M. NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. J Comput Biol [Internet]. Mary Ann Liebert, Inc., publishers; 2000;7:429–47. Available from: https://doi.org/10.1089/106652700750050871

65. Vaughan TG. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. Bioinformatics [Internet]. 2017/04/12. Oxford University Press; 2017;33:2392–4. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28407035

66. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics [Internet]. BioMed Central; 2015;16:169. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25994840

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 co-divergence.

Name of the gene	Total No. of Sequences Studied	No. of Duplication	No. of Co-divergence	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

Table 2

Deletion of different genes in the chloroplast genomes. Almost all of the genes have been deleted in the chloropl genome of one or another species. However, Rpl20 was found to be the most intact gene and found in all the spec 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	Ps
12	18	16	17	21	21	20	18	21	13	22	157	23	22	31	21
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				

Figure legends

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 bryophyte fall in one group and eudicots, magnoliids, monocots, and pteridophytes fall in other group; whereas, algae and protists fall distantly.

Figure 3

Principal component analysis of CDS numbers of chloroplast genomes. The CDS number of algae, gymnosperms, and protists fall separately; whereas, bryophytes, eudicots, pteridophytes, and nymphaeales fall together.

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; whereas' 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; whereas, the GC content of the bryophytes and pteridophytes fall distantly.

Figure 6

Phylogenetic tree of 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 of the species containing the inverted repeats (*Floydiella terrestris*, *Carteria cerasiformis*, *B. apyrenoidosa*, *E. grandis*, *O. sativa* and other) to decipher the differences. Deletion of inverted repeats did not have a significant impact on the phylogeny.

Figure 7

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 ancestors. 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.

Figure 8

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.

Supplementary Data

Supplementary File 1

File showing the name and genomic details of the species whose chloroplast genome was used during this study.

Supplementary File 2

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 3

File showing the loss of IR region in the chloroplast genome of different species.

Supplementary File 4

File showing the loss of different chloroplast encoding genes in different species.

Supplementary File 5

List of the chloroplast encoding genes found in the nuclear genomes.

Supplementary Table 1

Loss of chloroplast encoding genes in different species of respective lineages.

Supplementary Table 2

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 respective study are mentioned in the rows. The transition/transversion ratio 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 + non-coding. 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 stand G > A.

Supplementary Figures

Supplementary Figure 1

Conserved amino acid sequences of PsaM proteins. Blue mark indicates conservation of amino acids below 90%.

Supplementary Figure 2.

(A) Phylogenetic tree of *PsaM* genes showing five clusters. (B) Deletion and duplication event of *PsaM* genes. Duplications: 12, co-divergences: 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 3

Conserved amino acid sequences of Psb30 proteins. Red mark indicates conservation of amino acids of 90% or more.

Supplementary Figure 4

(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 5

Conserved amino acid sequences of ChlB proteins. Red mark indicate conservation of 90% or more.

Supplementary Figure 6

(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 7

Conserved amino acid sequences of ChlL proteins. Red mark indicate conservation of 90% or more.

Supplementary Figure 8

(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 9

Conserved amino acid sequences of ChlN proteins. Red mark indicate conservation of 90% or more.

Supplementary Figure 10

(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 11

Conserved amino acid sequences of Rpl21 proteins. Red mark indicate conservation of 90% or more.

Supplementary Figure 12

Phylogenetic tree of *Rpl21* genes. (B) Deletion and duplication event of *Rpl21* genes. Duplications: 3, co-divergences: 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 13

Molecular weight and isoelectric point (pI) of RBCL proteins.

Supplementary Figure 14

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 event in different taxon/groups. The genomic recombination of chloroplast genomes was studied using the IcyTree viewer (https://icytree.org/) server.

Supplementary Figure 15

Recombination event of inverted repeat deleted chloroplast genomes. Each color indicates a locus and their distribution in different cluster indicates they have been undergone vivid recombination.

Supplementary Figure 16

Venn diagram showing group specific loss of chloroplast encoding genes in algae, gymnosperm, bryophyte, monocot, eudicot, and magnoliid.

Supplementary Figure 17

Venn diagram showing group specific loss of chloroplast encoding genes in algae, bryophyte, gymnosperm, angiosperm, pteridophyte and protist.

Supplementary Figure 18

Venn diagram showing group specific loss of chloroplast encoding genes in eudicot, gymnosperm, monocot, and magnoliid.

Figure 1

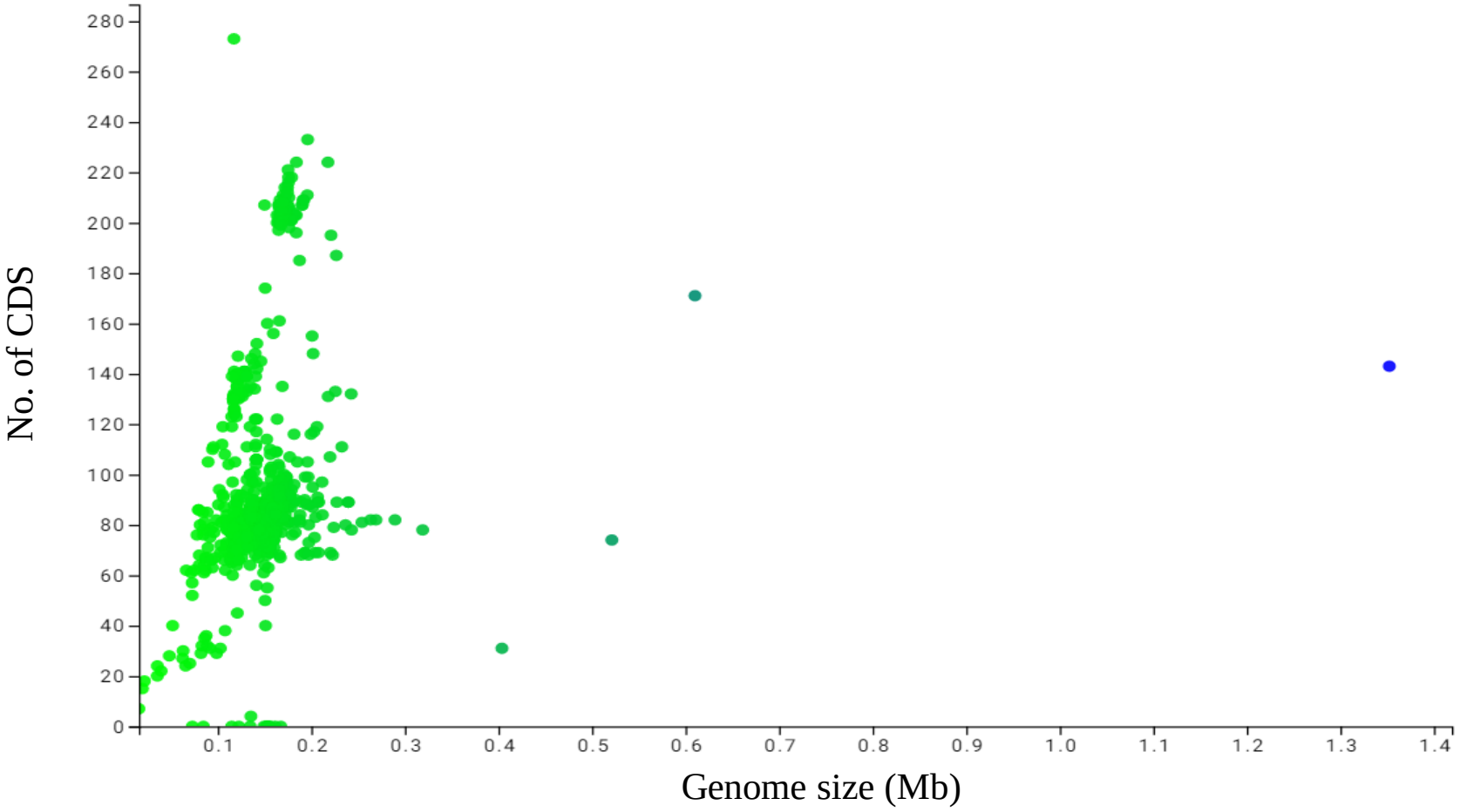

No. of CDS
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
Genome size (Mb)

Figure 2

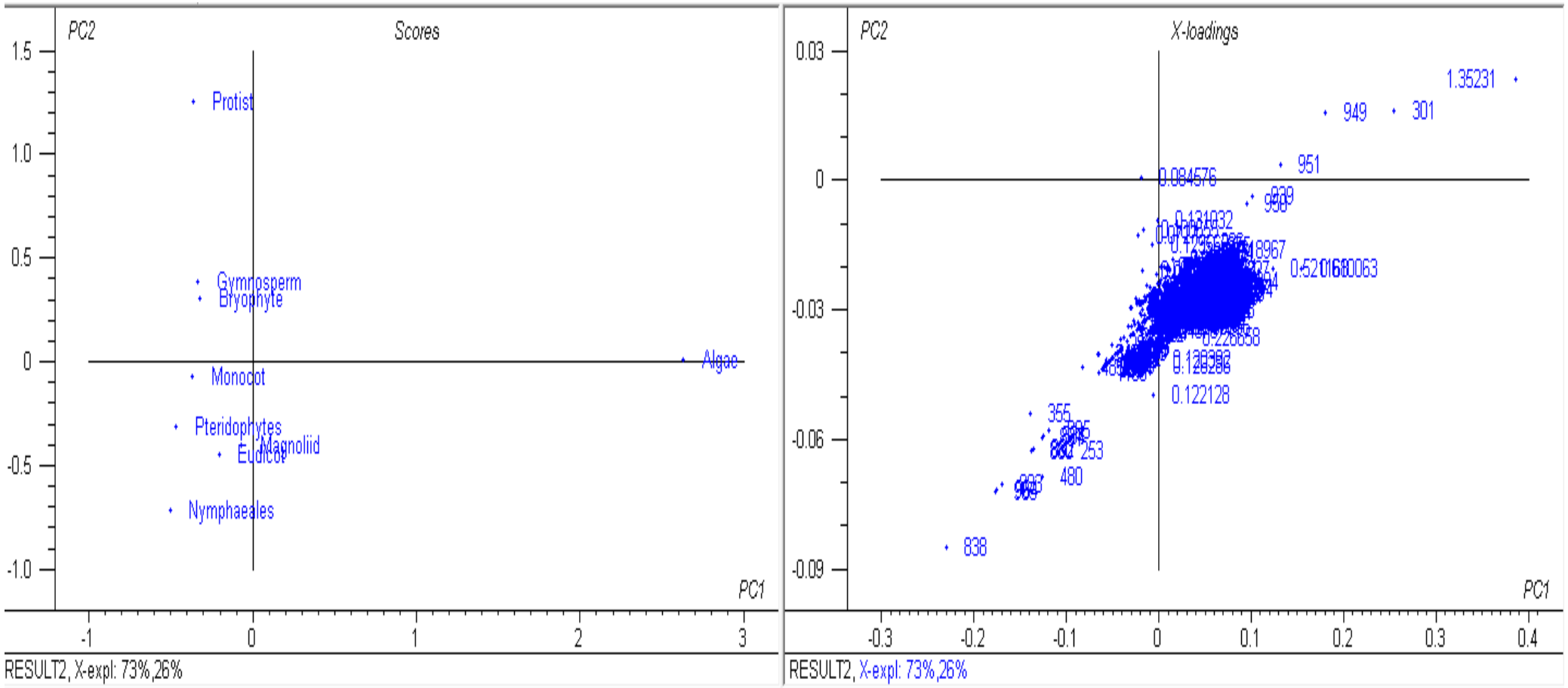

PC2
Scores
Protist
Gymnosperm
Bryophyte
Algae
Monocot
Pteridophytes
Magnoliid
Eudicot
Nymphaeales
PC1
RESULT2, X-expl: 73%,26%
PC2
X-loadings
1.35231
949
301
951
0.084576
0.122128
355
253
480
838
PC1
RESULT2, X-expl: 73%,26%

Figure 3

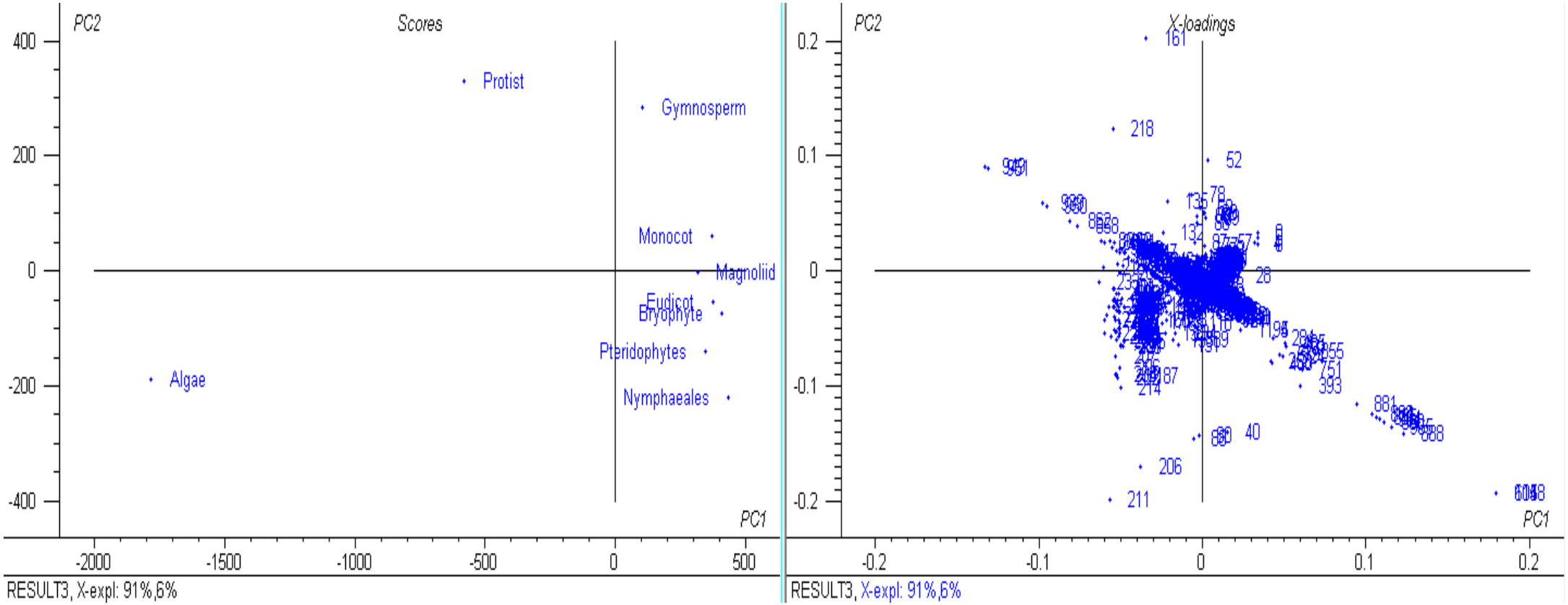

PC2
Scores
Protist
Gymnosperm
Monocot
Magnoliid
Eudicot
Bryophyte
Pteridophytes
Nymphaeales
Algae
PC1
400
200
0
-200
-400
-2000
-1500
-1000
-500
0
500
RESULT3, X-expl: 91%,6%
PC2
X-loadings
161
218
52
28
40
206
211
393
751
881
PC1
0.2
0.1
0
-0.1
-0.2
-0.2
-0.1
0
0.1
0.2
RESULT3, X-expl: 91%,6%

Figure 4

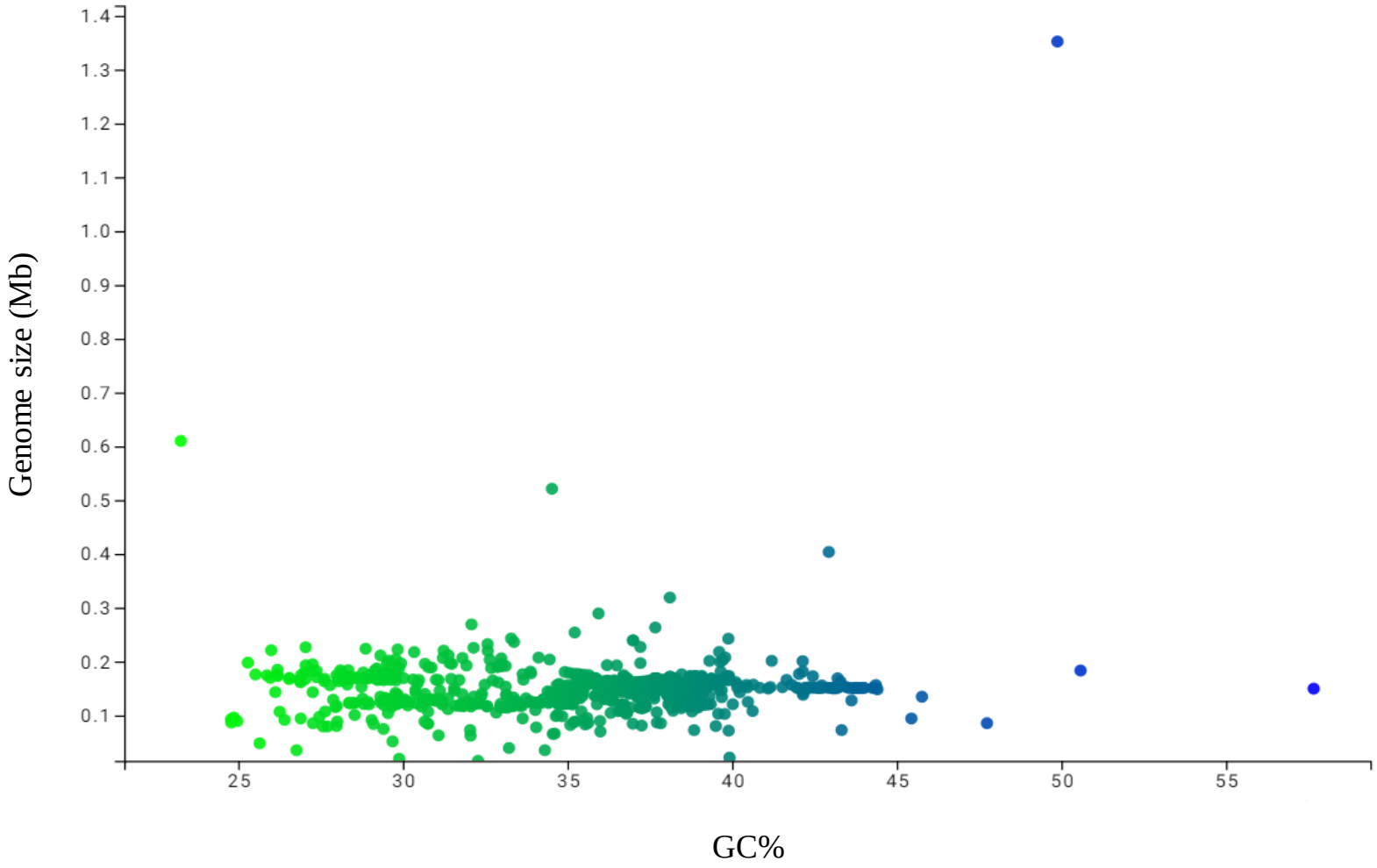
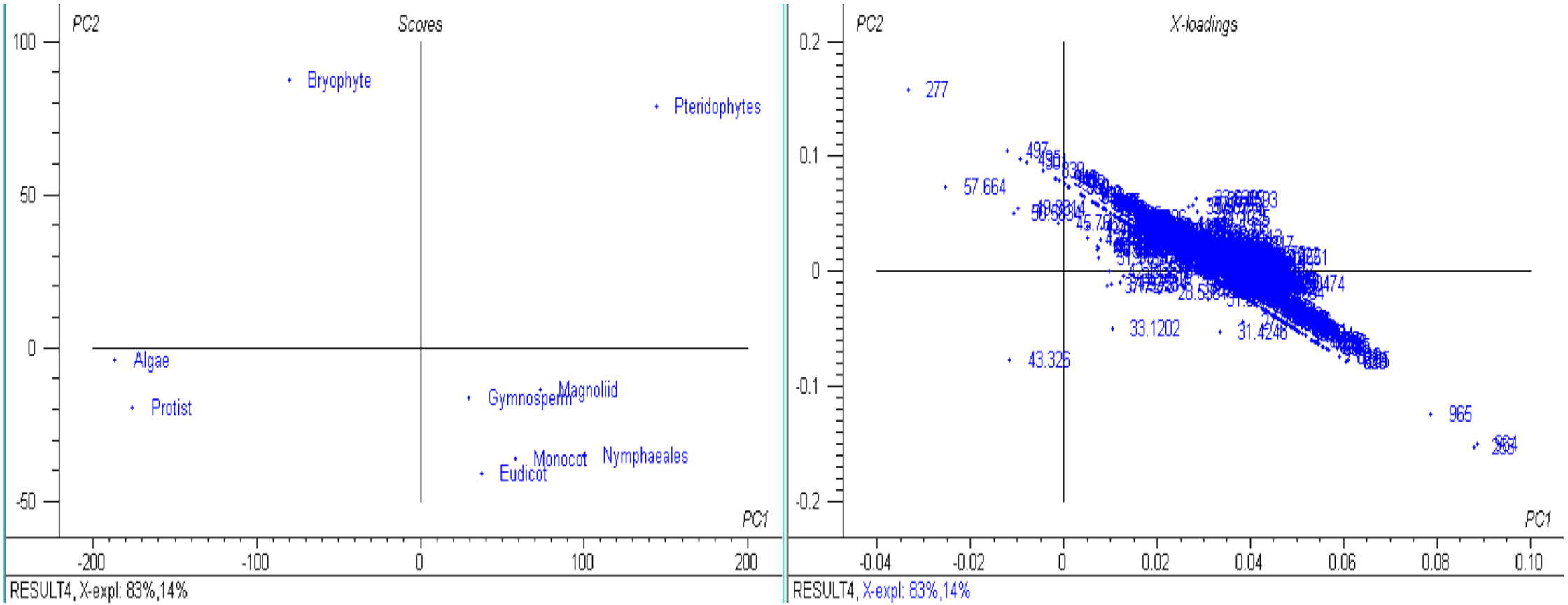
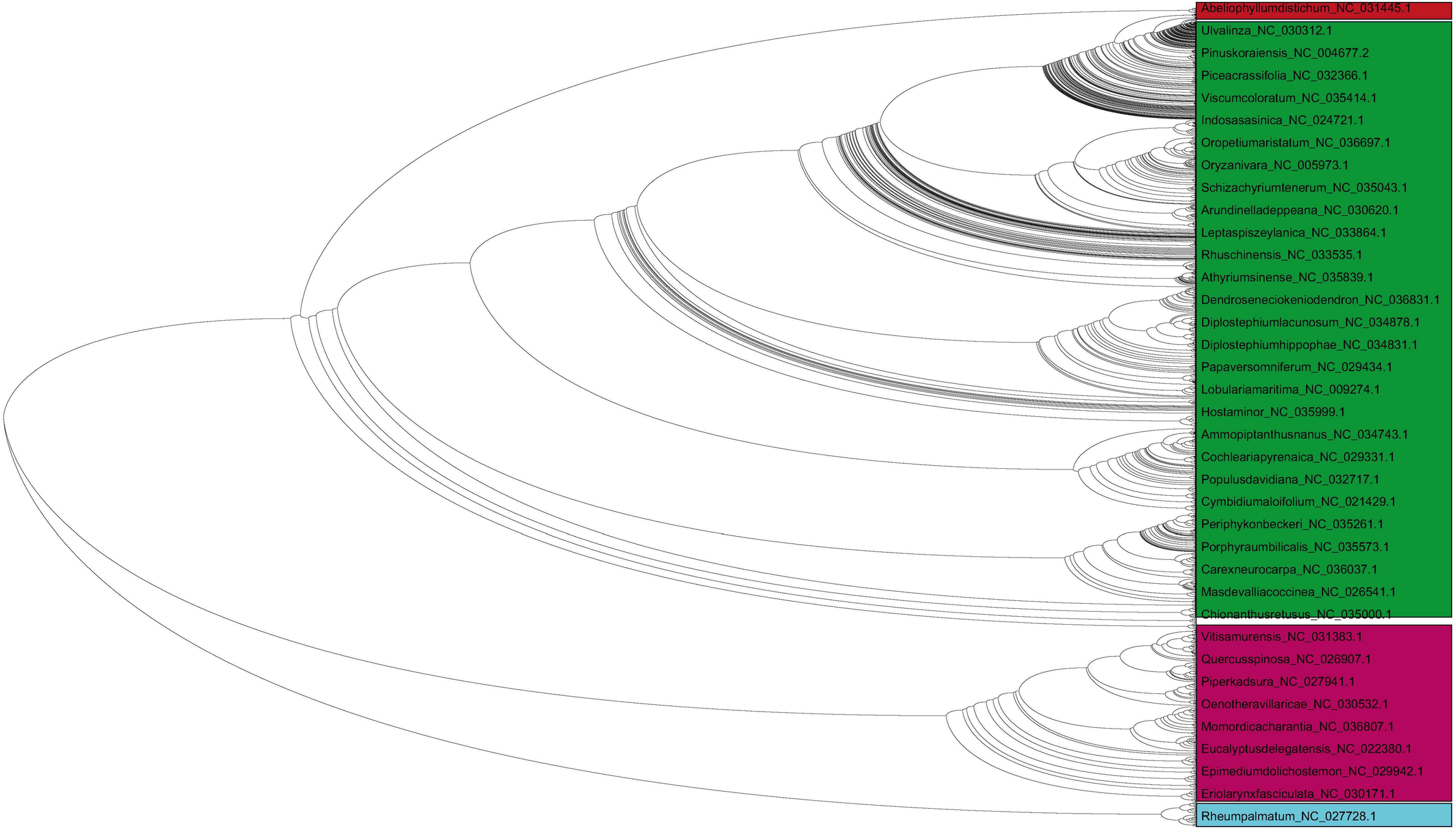
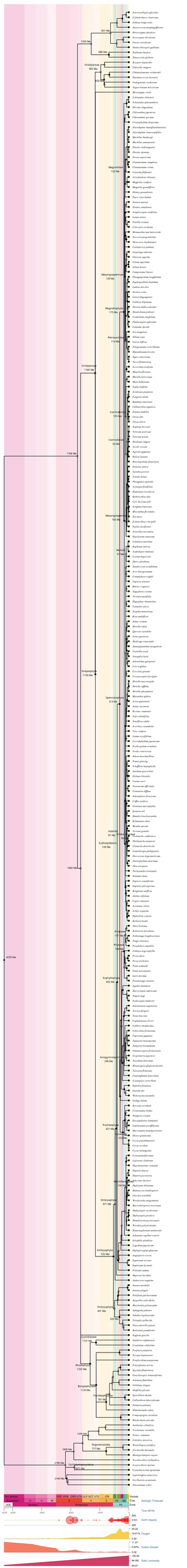

Genome size (Mb)
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
25
30
35
40
45
50
55
GC%

Figure 5


PC2
Scores
Bryophyte
Pteridophytes
Algae
Protist
Gymnosperm
Magnoliid
Monocot
Nymphaeales
Eudicot
PC1
RESULT4, X-expl: 83%,14%
PC2
X-loadings
277
57.664
33.1202
31.4248
43.326
965
PC1
RESULT4, X-expl: 83%,14%

Abieskoreana_Gymnosperm_NC_026892.1
Abiessibirica_Gymnosperm_NC_035067.1
Tsugachinensis_Gymnosperm_NC_030630.1
Picealabies_Gymnosperm_NC_021456.1
Piceasasperata_Gymnosperm_NC_032367.1
Piceacrassifolia_Gymnosperm_NC_032366.1
Piceaglauca_Gymnosperm_KT634228.1
Piceaizoensis_Gymnosperm_NC_029374.1
Pseudotsugasinensis_Gymnosperm_NC_016064.1
Pinustaeda_Gymnosperm_NC_021440.1
Agathisdammara_Gymnosperm_NC_023119.1
Amentotaxusargotaenia_Gymnosperm_NC_027581.1
Amentotaxusformosana_Gymnosperm_NC_024945.1
Torreyafargesii_Gymnosperm_NC_029398.1
Torreyagrandis_Gymnosperm_NC_034806.1
Nageiainagi_Gymnosperm_NC_023120.1
Callitrisrhomboidea_Gymnosperm_NC_034940.1
Calocedrusformosana_Gymnosperm_NC_023121.1
Chamaecyparisformosensis_Gymnosperm_NC_034943.1
Taiwaniacryptomerioides_Gymnosperm_NC_016055.1
Taiwaniaflousiana_Gymnosperm_NC_021441.1
Cunninghamialanceolata_Gymnosperm_NC_021437.1
Sciadopitysverticillata_Gymnosperm_NC_029734.1
Taxodiumdistichum_Gymnosperm_NC_034941.1
Glyptostrobuspensilis_Gymnosperm_NC_031354.1
Metasequoiaglyptostroboides_Gymnosperm_NC_027423.1
Piceamorrisonicola_Gymnosperm_NC_016069.1
Rosachinensis_Eudicot_CM009590.1
Adenocalymmaallamandiiflorum_Eudicot_NC_036494.1
Swertiamussotii_Gymnosperm_NC_031155.1
Nothotsugalongibracteata_Gymnosperm_NC_037407.1
Podocarpusfamberdi_Gymnosperm_NC_023605.1
Trachelium caeruleum_Eudicot_NC_010442.1
Aethionemacordifolium_Eudicot_NC_009265.1
Thlaspiarvense_Eudicot_NC_034362.1
Taraxacumkok-saghyz_Eudicot_NC_032057.1
Taraxacumplatycarpum_Eudicot_NC_031395.1
Utriculariamacrorhiza_Eudicot_NC_025653.1
Veratrumpatulum_Monocot_NC_022715.2
Saccharisgeniculatides_Eudicot_NC_034852.1
Brassaiopsishainla_Eudicot_NC_022811.1
Caraganakorolowii_Eudicot_NC_035228.1
Glycyrrhizalepidota_Eudicot_NC_034229.1
Wisteriafloribunda_Eudicot_NC_027677.1
Wisteriasinensis_Eudicot_NC_029406.1
Caraganamicrophylla_Eudicot_NC_032691.1
Elodeacanadensis_Monocot_NC_018541.1
Albuscarkhii_Monocot_NC_032687.1
Prunushumilis_Eudicot_NC_035880.1
Trapamaximowiczii_Eudicot_NC_037023.1
Amomumkravanh_Monocot_NC_036935.1
Sobralialoucha_Monocot_NC_028205.1
Symplocarpuserenifolius_Monocot_NC_033970.1
Talipariti hamabo_Eudicot_NC_030195.1
Eucalyptusgrandis_Eudicot_NC_014570.1
Villosacerifolia_Eudicot_NC_035878.1
Vitisconglentiae_Eudicot_NC_036371.1
Camelliasinensis_Eudicot_NC_020019.1
Illiciumlanceolatum_Eudicot_NC_035720.1
Vanillaplanifolia_Monocot_NC_026778.1
Aegilopscylindrica_Monocot_NC_023096.1
Aegilopstauschii_Monocot_NC_022133.1
Aegilopsspeltoides_Monocot_NC_022135.1
Aegilopssquarrosa_Monocot_NC_023097.1
Secalecereale_Monocot_NC_021761.1
Ruminanniaoblonga_Monocot_NC_036564.1
Cylindrophycolis_Eudicot_NC_031150.1
Scapiflaliaenifflora_Monocot_NC_027659.1
Monotrophahypopitys_Eudicot_NC_029704.1
Gastrodioelata_Monocot_NC_037409.1
Polysiphoniaserrulatkides_Algae_NC_035270.1
Protothica
Protothecacutis_Algae_NC_037480.1
Prelipancheramosa_Eudicot_NC_023465.1
Prelipanchespurpurea_Eudicot_NC_023132.1
Lepidodinium chlorophorum_Protist_NC_027093.1
Monomorphinaaenigmatica_Protist_NC_020018.1
Prasinodermacoloniale_Algae_NC_024817.1
Palmophyllumcrassum_Algae_NC_033387.1
Monomorphinaparapyrum_Protist_NC_027287.1
Pycnococcuisprovasoli_Algae_NC_012097.1
Codiumsp._Algae_NC_037366.1
Trachelomonasvolvocina_Protist_NC_027288.1
Ulvalinza_Algae_NC_030312.1
Chlorodesmisfastigiata_Algae_NC_037364.1
Ulvaflexuosa_Algae_NC_035623.1
Aureococcusanophagefferens_Protist_NC_012898.1
Euglenavirdis_Protist_NC_020450.2
Euglenaformisproxima_Protist_NC_024154.1
Aureoumbralagunensis_Protist_NC_012903.1
Ulvalasciata_Algae_NC_029040.1
Euglenaarchaeoplastidata_Protist_NC_035156.1
Cryptoglenaskujai_Protist_NC_027286.1
Coleochaetescutata_Algae_NC_030358.1
Phaeocystisglobosa_Protist_NC_021607.1
Cathayaargyrophylla_Gymnosperm_NC_014589.1
Rhipiliapenicilloides_Algae_NC_037365.1
Camegieagigantea_Eudicot_NC_027618.1
Cassythaflliformis_Magnoliids_NC_036001.1
Triticummonococcum_Monocot_NC_021760.1
Triticumurartu_Monocot_NC_021762.1
Derbesiasp._Algae_NC_031367.1
Vaucherialitorea_Protist_NC_016800.1
Nannochloropsisoculata_Protist_NC_022261.1
Nannochloropsisoceanica_Protist_NC_022263.1
Erodiumcarvifolium_Bryophyte_NC_015083.1
Prasinomonas_Gymnosperm_NC_026917.1
Pinuslambertiana_Gymnosperm_NC_011156.4
Pinussibirica_Gymnosperm_NC_028552.2
Pinusnelsonii_Gymnosperm_NC_011159.4
Pinuskesiya_Gymnosperm_NC_011155.4
Pinusmonophylla_Gymnosperm_NC_011158.4
Pinusgerardiana_Gymnosperm_NC_011154.4
Pinuskoraiensis_Gymnosperm_NC_004677.2
Keteleriadavidiana_Gymnosperm_NC_011930.1
Cedrusdeodara_Gymnosperm_NC_014575.1
Piceasitchensis_Gymnosperm_NC_011152.3
Pinuscontorta_Gymnosperm_NC_011153.4
Pinumassoniana_Gymnosperm_NC_021439.1
Pinussylvestris_Gymnosperm_NC_035069.1
Pinustabuliformis_Gymnosperm_NC_028531.1
Pinustaiwanensis_Gymnosperm_NC_027415.1
Pinusthunbergii_Gymnosperm_NC_001631.1
Odontellasinensis_Algae_NC_001713.1
Cephalotaxusoliveri_Gymnosperm_NC_021110.1
Cephalotaxussinensis_Gymnosperm_NC_037245.1
Cauleryanamonensis_Algae_NC_037397.1
Digeneasimplex_Algae_NC_035298.1
Polysiphoniastricta_Algae_NC_035275.1
Spyridiafilamentosa_Algae_NC_035285.1
Thuretiapyercrollia_Algae_NC_035286.1
Polysiphoniabrodiei_Algae_NC_035272.1
Polysiphoniaelongata_Algae_NC_035274.1
Symphyocladiadendroides_Algae_NC_035267.1
Polysiphoniaschneideri_Algae_NC_035266.1
Polysiphoniainfestans_Algae_NC_035277.1
Thaumatellaaounca_Algae_NC_035291.1
Tolypoclaciaglomerulata_Algae_NC_035299.1
Pterphykorbeckeri_Algae_NC_035281.1
Euglenagracilis_Protist_NC_001603.2
Rhodomelaconfervoides_Algae_NC_035271.1
Polysiphoniascopulorum_Algae_NC_035282.1
Leptosiratenerrestris_Algae_NC_009681.1
Taeniomaperpusillum_Algae_NC_035295.1
Bulboplastispayrenoideosa_Algae_NC_034787.1
Gracilariachangii_Algae_NC_038051.1
Gracilariavermicolophylla_Algae_NC_039092.1
Platysiphoniadelicata_Algae_NC_035258.1
Spirogyramaxima_Algae_NC_030355.1
Stauroratrumpunctulatum_Algae_NC_008116.1
Mymecialaraalensis_Algae_NC_025525.1
Schizomeridsolblelnii_Algae_NC_015645.1
Sargassumhomeri_Algae_NC_029856.1
Heterosiphonsakashiwo_Protist_NC_010772.1
Sheathiaarcuata_Algae_NC_035231.1
Grateloupiafilicina_Algae_NC_037841.1
Porphyrapurpurea_Algae_NC_000925.1
Porphyraumbilicalis_Algae_NC_035573.1
Porphyrayezoensis_Algae_NC_007932.1
Pyropiahaitanensis_Algae_NC_021189.1
Cryptomeriajaponica_Gymnosperm_NC_010548.1
Monoraphidiummegalocum_Algae_NW_014013626.1
Sequoiasempervirens_Gymnosperm_NC_030372.1
Lathyruscylmenum_Eudicot_NC_027148.1
Trifoliumglanduliferum_Eudicot_NC_025744.1
Trifoliumstrictum_Eudicot_NC_025745.1
Lathyruspalustris_Eudicot_NC_027078.1
Lensculinaris_Eudicot_NC_027152.1
Trifoliumbeissieri_Eudicot_NC_025743.1
Vicisativa_Eudicot_NC_027155.1
Lathyrusinconspicuus_Eudicot_NC_027149.1
Lathyrusittoralis_Eudicot_NC_027076.1
Lathyrusvenosus_Eudicot_NC_027080.1
Lathyruspubescens_Eudicot_NC_027079.1
Lathyrusochroleucus_Eudicot_NC_027077.1
Lathyrusdavidii_Eudicot_NC_027073.1
Lathyrusgraminifolius_Eudicot_NC_027074.1
Lathyrusjepsonius_Eudicot_NC_027075.1
Lathyrusodoratus_Eudicot_NC_027150.1
Lathyrussativus_Eudicot_NC_014063.1
Lathyrustingitanus_Eudicot_NC_027151.1
Medicagoblakiana_Eudicot_NC_033566.1
Rhodomonassalina_Protist_NC_009573.1
Floydiellaterrestris_Algae_NC_014346.1
Colobanthusapetalus_Eudicot_NC_036424.1
Retrophyllumpiresii_Gymnosperm_NC_044527.1
Wollemianobilis_Gymnosperm_NC_027235.1
Cupressusgigantea_Gymnosperm_NC_028155.1
Juniperusbermudiana_Gymnosperm_NC_024021.1
Juniperusmonosperma_Gymnosperm_NC_024022.1
Juniperusvirginiana_Gymnosperm_NC_024024.1
Juniperuscedrus_Gymnosperm_NC_028190.1
Juniperusmicrosperma_Gymnosperm_NC_037430.1
Juniperusscopulorum_Gymnosperm_NC_024023.1
Adenocalymmadivaricatum_Eudicot_NC_037456.1
Alliumsativum_Monocot_NC_031829.1
Themedaarguens_Monocot_NC_035019.1
Genkwaituberosa_Eudicot_NC_037082.1
Alloeochaetoereogena_Monocot_NC_035518.1
Brachypodiumdistachyon_Monocot_NC_011032.1
Hyparrheniadiplandra_Monocot_NC_035037.1
Urochloadecumbens_Monocot_NC_030066.1
Kerriochloasiamensis_Monocot_NC_035009.1
Oryzasativa_Monocot_NC_031333.1
Amanakuocangshanica_Monocot_NC_034708.1
Tignaradiata_Eudicot_NC_013843.1
Senkiseavolacosa_Eudicot_NC_037083.1
Stachysbyzantina_Eudicot_NC_029825.1
Phyllostegiaevkulina_Eudicot_NC_029820.1
Stenogynebillitis_Eudicot_NC_029818.1
Stenogynekenshonana_Eudicot_NC_029821.1
Senkiseafiliformis_Eudicot_NC_037079.1
Astragalusmongholicus_Eudicot_NC_029826.1
Lesserliatufrutescens_Eudicot_NC_036151.1
Tibetaliangshanensis_Eudicot_NC_036109.1
Scerariefolium_Eudicot_NC_011163.1
Medicagohybrida_Eudicot_NC_027153.1
Medicagopapillosa_Eudicot_NC_027154.1
Trifoliumsubterraneum_Eudicot_NC_011828.1
Trifoliumrepens_Eudicot_NC_036485.1
Taxusmairei_Gymnosperm_NC_020321.1
Chrysochromulinaparva_Protist_NC_036937.1
Ampelopteriseliegans_Pteridophyte_NC_035835.1
Athyriumsinense_Pteridophyte_NC_035839.1
Trebouxiophyceae_Algae_NC_016569.1
Splachnumluteum_Bryophyte_NC_035853.1
Stegnogrammasagittifolia_Pteridophyte_NC_035863.1
Codonopsislanceolata_Eudicot_NC_037949.1
Cycasrevoluta_Gymnosperm_NC_020319.1
Klebsormidiumniteris_Algae_DF236762.1
Monosolenanartoitii_Eudicot_NC_029830.1
Symbiochlorishandae_Algae_NC_025524.1
Asarumsatobaii_Eudicot_NC_037190.1
Utricularierenniformis_Eudicot_NC_029719.2
Salixbabylonica_Eudicot_NC_028350.1
Salixpurpurea_Eudicot_KP010839.1
Salixsuchowensis_Eudicot_NC_026462.1
Chamaecyparishodginsii_Gymnosperm_NC_036996.1
Larixkeiciua_Gymnosperm_NC_016058.1
Larixsibirica_Gymnosperm_NC_036811.1
Monosolenamarginata_Eudicot_NC_029494.1
Pseudotaxiamabilis_Gymnosperm_NC_030631.1
Cyanidioschyzonmerolae_Algae_NC_004799.1
Carlsnaccarosiformis_Algae_NC_028585.1


Abeliophyllumdistichum_NC_031445.1
Ulvalinza_NC_030312.1
Pinuskoraiensis_NC_004677.2
Piceacrassifolia_NC_032366.1
Viscumcoloratum_NC_035414.1
Indosasasinica_NC_024721.1
Oropetiumaristatum_NC_036697.1
Oryzanivara_NC_005973.1
Schizachyriumtenerum_NC_035043.1
Arundinelladeppeana_NC_030620.1
Leptaspiszeylanica_NC_033864.1
Rhuschinensis_NC_033535.1
Athyriumsinense_NC_035839.1
Dendroseneciokeniodendron_NC_036831.1
Diplostephiumlacunosum_NC_034878.1
Diplostephiumhippophae_NC_034831.1
Papaversomniferum_NC_029434.1
Lobulariamaritima_NC_009274.1
Hostaminor_NC_035999.1
Ammopiptanthusnanus_NC_034743.1
Cochleariapyrenaica_NC_029331.1
Populusdavidiana_NC_032717.1
Cymbidiumaloifolium_NC_021429.1
Periphykonbeckeri_NC_035261.1
Porphyraumbilicalis_NC_035573.1
Carexneurocarpa_NC_036037.1
Masdevalliacoccinea_NC_026541.1
Chionanthusretusus_NC_035000.1
Vitisamurensis_NC_031383.1
Quercusspinosa_NC_026907.1
Piperkadsura_NC_027941.1
Oenotheravillaricae_NC_030532.1
Momordicacharantia_NC_036807.1
Eucalyptusdelegatensis_NC_022380.1
Epimediumdolichostemon_NC_029942.1
Eriolarynxfasciculata_NC_030171.1
Rheumpalmatum_NC_027728.1


Arabidopsis thaliana
Cardamine chinensis
Arabis hirsuta
Barbarea vulgaris
Brassica oleracea
Eutrema salsugineum
Capsella bursa-pastoris
Camelina sativa
Brassica rapa
Brassica nigra
Brassica napus
Raphanus sativus
Cleome spinosa
Tarenaya hassleriana
Gynandropsis gynandra
Cleome violacea
Cleome hassleriana
Cleome gynandra
Cleome dolichostyla
Cleome rutidosperma
Cleome droserifolia
Cleome arabica
Cleome amoena
Cleome rosea
Cleome serrata
Cleome polanisioides
Cleome parvipetala
Cleome oligantha
Cleome multinervata
Cleome longipes
Cleome lutea
Cleome lipocarpa
Cleome hirta