Genomics 106 (2015) 221–231

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Dynamics of chloroplast genomes in green plants

Jian-Hong Xu [a,*,1], Qiuxiang Liu [a,1], Wangxiong Hu [a], Tingzhang Wang [a], Qingzhong Xue [a], Joachim Messing [b]

[a] Institute of Crop Science, Zhejiang Key Laboratory of Crop Germplasm, Zhejiang University, Hangzhou 310058, China
[b] Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

ARTICLE INFO

Article history:
Received 9 June 2014
Received in revised form 8 July 2015
Accepted 13 July 2015
Available online 21 July 2015

Keywords:
Genome organization
Gene rearrangement
Inverted repeat
Ka/*Ks*
Chloroplast genome

ABSTRACT

Chloroplasts are essential organelles, in which genes have widely been used in the phylogenetic analysis of green plants. Here, we took advantage of the breadth of plastid genomes (cpDNAs) sequenced species to investigate their dynamic changes. Our study showed that gene rearrangements occurred more frequently in the cpDNAs of green algae than in land plants. Phylogenetic trees were generated using 55 conserved protein-coding genes including 33 genes for photosynthesis, 16 ribosomal protein genes and 6 other genes, which supported the monophyletic evolution of vascular plants, land plants, seed plants, and angiosperms. Moreover, we could show that seed plants were more closely related to bryophytes rather than pteridophytes. Furthermore, the substitution rate for cpDNA genes was calculated to be 3.3×10^{-10}, which was almost 10 times lower than genes of nuclear genomes, probably because of the plastid homologous recombination machinery.

1. Introduction

Chloroplasts are essential organelles in plants that originated from Cyanobacteria by endosymbiosis with the precursor of nucleated ancestral cells more than 1.2 billion years ago [1]. However, chloroplast genome sequences (cpDNA) remained more conserved in gene content than nuclear genomes. The majority of cpDNA have a quadripartite structure that contains two copies of an inverted repeat (IR) separating the small single-copy region (SSC) and the large single-copy region (LSC). Exceptions provide species such as *Chlorella variabilis* that do not have IRs, SSC, and LSC in their cpDNAs [2].

The cpDNAs have been extensively studied in green plants [3–7], and chloroplast genes have widely served as useful markers for systematics and the evolution of species and ecotypes. However, most studies focused on the relationships within angiosperms, seed plants, land plants, and green algae [8–12], whereas less information was known across all embryophyta. The main role of chloroplast is conducting photosynthesis in the presence of sunlight, and it also has functions for synthesis of fatty acids, pigments, starch and amino acids [7,13], its genes could be divided into three types, including genes for photosynthesis, ribosomal protein genes, and other functions [14]. Still, so far it was less clear whether genes for these different categories diverged differently.

Based on systematics, photosynthetic organisms evolved into two main lineages, the Chlorophyta and the Streptophyta. Chlorophyta comprise most of the green algae and the Streptophyta contain the remaining green algae and all embryophytes, commonly known as land plants [15]. Genome sizes of chloroplasts of green plants can range from 120 to 2500 kb, mainly due to duplication of genes and small repeats. However, over time the size of chloroplast genomes tended to decrease [1,16] because of extensive gene losses and/or gene transfer to mitochondrial and/or nucleus genomes [17,18]. The genes that integrated in the nuclear genome could be expressed by the nuclear transcription machinery, and the transit peptides could be imported and function in the organelle, consistent with the endosymbiotic history of organellar and nuclear genomes. Whereas most genes that were retained in the plastid genomes were highly conserved, gene content and inversions of blocks of genes even between closely related species occurred at an astonishing rate [19].

With the advent of the next-generation-sequencing technology, a large amount of cpDNA sequences were completely sequenced and are available in GenBank (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid, July 2013) (369). Here, twenty-four representative species from algae to angiosperms were selected to analyze the genomic organizations and evolution of their cpDNAs. Our results show that gene rearrangements occurred more frequently in the cpDNAs of green algae than in land plants, which include gene gain/loss, sequence inversions, translocations, the expansions, and contractions of IRs. Phylogenetic analysis supports a monophyletic evolution of angiosperms, seed plants, land plants, and vascular plants, whereas seed plants have an evolutionary pattern in ribosomal protein and other genes that is closer to bryophytes rather than pteridophytes. The substitution rate for all 55 conserved protein-coding genes was calculated as 3.3×10^{-10}, which was almost 10 times lower than for nuclear genes.

* Corresponding author at: Institute of Crop Science, College of Agriculture & Biotechnology, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China.
E-mail address: jhxu@zju.edu.cn (J.-H. Xu).
[1] These authors contributed equally.

http://dx.doi.org/10.1016/j.ygeno.2015.07.004
0888-7543/

2. Results and discussion

2.1. Gene organization and mobility in chloroplast genomes

Twenty-four representative species with completely sequenced cpDNAs were selected, which include five green algae (*C. variabilis*, *Mesostigma viride*, *Chara vulgaris*, *Staurastrum punctulatum* and *Chaetosphaeridium globosum*), five bryophytes (*Pellia endiviifolia*, *Marchantia polymorpha*, *Anthoceros formosae*, *Syntrichia ruralis* and *Physcomitrella patens*), three pteridophytes (*Equisetum arvense*, *Ophioglossum californicum* and *Psilotum nudum*), three gymnosperms (*Cedrus deodara*, *Pinus thunbergii* and *Ginkgo biloba*), two basal angiosperms (*Nymphaea alba* and *Amborella trichopoda*) two eudicots (*Arabidopsis thaliana* and *Populus alba*), and four monocots (*Oryza sativa*, *Zea mays*, *Hordeum vulgare* and *Dendrocalamus latiflorus*) (Table 1). The size of cpDNAs ranges from 118,360 to 184,933 bp, and the number of protein-coding genes from 70 to 99 in all 24 species, which exclude tRNA, rRNA, and unique ORF genes. There are inverted repeats (IRs) in all investigated cpDNAs except *C. variabilis* and *S. punctulatum*, and the length and the gene content of IRs increase steadily from green algae to monocots except for *C. deodara*, which lacks IRs [20,21] (Table 1).

Because gene content and order are generally conserved, it becomes possible to make syntenic alignments of plastid genomes and reconstruct insertions, deletions, and inversions that have occurred over time. Based on such comparisons, many rearrangements existed especially among green algae with 30 gene-inversion and 33 gene-translocation events, ranging from two to eleven genes (Fig. 1A, Supplementary Fig. 1). Compared to the complex rearrangements of cpDNAs in algae, the cpDNAs of angiosperm exhibit drastically fewer changes. Only one gene inversion of 10 genes between *atpA* and *psbM* and a translocation of three genes (*psbD*, *psbC* and *psbZ*) next to the gene inversion appeared to have occurred (Fig. 1B, Supplementary Fig. 1). One gene-inversion event, which contains 30 genes between *ycf66* and *rps11*, occurred in *P. patens*, two events with 16 genes (*rpoB* to *ycf2*) and three genes (*psbD*, *psbC* and *psbZ*) in *E. arvense*, respectively. Two gene-inversion events are found in *C. deodara* with five (*ycf1* to *rpl32*) and eight (*rps4* to *psbD*) genes (Fig. 1B, Supplementary Fig. 1). All *ndh* genes are deleted in *C. variabilis* and *C. deodara* (Fig. 1, Supplementary Fig. 1) [20,21].

2.2. Dynamic divergence of IR copies

Except for *C. variabilis* and *S. punctulatum* all cpDNAs compared here contain two IR copies. Their lengths differ from 9589 bp (*P. patens*) to 27,659 bp (*P. alba*) with *C. deodara* and *P. thunbergii* having exceptional small IRs of only 424 bp and 495 bp because of significant gene losses [21] (Table 1). To illustrate the nature of sequence mobilities, IR copies were aligned from three algae (*M. viride*, *C. vulgaris* and *C. globosum*), three bryophytes (*A. formosae*, *M. polymorpha* and *P. patens*), two pteridophytes (*P. nudum* and *E. arvense*), three gymnosperms (*G. biloba*, *P. thunbergii* and *C. deodara*), one eudicot (*A. thaliana*), and one monocot (*O. sativa*) (Fig. 2).

IRs contain three tRNAs (*trnI*, *trnA* and *trnR*) and three rRNAs (*rrs*, *rr1* and *rrf* in green algae; *rrn16*, *rrn23*, *rrn5* in the land plants) in the same order (*rrs-trnI-trnA-rr1-rrf-trnR* or *rrn16-trnI-trnA-rrn23-rrn5-trnR*). These types of genes spread to LSC and SSC to add *trnV*, *trnN* and 3′-*ndhF* in all species except for *M. viride* (Figs. 2, 3). One rRNA gene, *rrn4.5*, was acquired in IRs of all land plants, *trnI*, *ycf2*, *trnL* and *rps7* in the IRs of seed plants. In angiosperms the IRs have acquired six genes, *rps19*, *rpl2*, *rpl23*, *ndhB*, *rps12*, and *ycf1* (Fig. 3). In green algae, additional gene copies of *trnT*, *ycf20*, and *ChlL*, *ChlN* were inserted in the IRs of *M. viride*, *C. vulgaris*, and *C. globosum*, respectively (Fig. 3). The 3′-*psbA* and *trnH* genes were inserted into the IR only of *C. deodara* and the *trnH*, *rps15*, and 5′-*ndhH* genes in the IRs of the grass species (Fig. 3).

Charophycean green algae and land plants define the green plant lineage of the Streptophyta. It was assumed that IRs be present in the cpDNA of all Charophyceans and land plants [2]. There also was strong evidence that the rRNA genes were present in the ancestral IRs. By inference, IRs also existed with rRNA genes in the original cpDNAs of *S. punctulatum* and *C. deodara*. However, our results showed that one of the IRs was missing in the cpDNAs of *S. punctulatum* and was greatly reduced in *C. deodara* (Figs. 2B, C) [6]. It also has been shown that different IR copies were lost in the cpDNAs of the Pinaceae and Cupressophytes of gymnosperms [21]. IR_B regions were lost in Pinaceae in a two-step model, and only 495 bp were left in *P. thunbergii*, whereas

Table 1
The features of chloroplast genomes in 24 species.

Species	Accession	Size (bp)	LSC (bp)	SSC (bp)	IR (bp)	Protein-coding Gene [a]	Duplicated Gene [b]
C. variabilis	NC_015359	124,579	–	–	–	79	1
M. viride	NC_002186	118,360	83,627	22,619	6057	99	–
C. vulgaris	NC_008097	184,933	135,815	27,280	10,919	94	1
S. punctulatum	NC_008116	157,089	–	–	–	88	–
C. globosum	NC_004115	131,183	88,682	17,639	12,431	94	2
P. endiviifolia	NC_019628	120,546	82,508	19,854	9092	88	–
M. polymorpha	NC_001319	121,024	81,095	19,813	10,058	78	–
A. formosae	NC_004543	161,162	107,503	22,171	15,744	84	2
S. ruralis	NC_012052	122,630	84,078	18,542	10,005	80	–
P. patens	NC_005087	122,890	85,211	18,501	9589	84	–
P. nudum	NC_003386	138,829	84,617	16,304	18,954	80	6
O. californicum	NC_020147	138,270	99,058	19,662	9775	84	–
E. arvense	NC_014699	133,309	93,542	19,469	10,149	85	–
G. biloba	AB684440	156,945	99,221	22,258	17,733	83	2
P. thunbergii	NC_001631	119,707	65,696	53,021	495	70	1
C. deodara	NC_014575	119,299	64,675	53,773	425	76	3
N. alba	NC_006050	159,930	90,015	19,563	25,176	81	4
A. trichopoda	NC_005086	162,686	90,971	18,415	26,650	76	5
A. thaliana	NC_000932	154,478	84,170	17,780	26,264	86	7
P. alba	NC_008235	156,505	84,618	16,567	27,660	85	8
O. sativa	NC_001320	134,525	80,592	12,335	20,799	86	8
Z. mays	NC_001666	140,384	82,352	12,536	22,748	86	8
H. vulgare	NC_008590	136,462	81,671	12,701	21,045	87	9
D. latiflorus	NC_013088	139,394	83,010	12,874	21,755	87	9

[a] Protein coding gene that excludes tRNA, rRNA and unique ORF genes, and genes presented in the IR copies were counted only once.
[b] The number of duplication gene within IR region was counted only once.